DISSERTATION

Non-linear Spatial Filtering for Multi-channel Speech Enhancement and Separation

Kumulative Dissertation zur Erlangung des akademischen Grades
Dr. rer. nat.
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Universität Hamburg

eingereicht von

Kristina Tesch

Hamburg 2023

Kristina Tesch: Non-linear Spatial Filtering for Multi-channel Speech Enhancement and Separation

GUTACHTER:

Prof. Dr.-Ing. Timo Gerkmann

Prof. Dr. Chris Biemann

Prof. Dr.-Ing. Reinhold Häb-Umbach

VORSITZ DER PRÜFUNGSKOMMISSION:

Prof. Dr. Frank Steinicke

TAG DER EINREICHUNG:

17.11.2023

TAG DER DISPUTATION:

05.02.2024

Abstract

A large part of human speech communication takes place in noisy environments and is supported by technical devices. For example, a hearing-impaired person might use a hearing aid to take part in a conversation in a busy restaurant. These devices, but also telecommunication in noisy environments or voiced-controlled assistants, make use of speech enhancement and separation algorithms that improve the quality and intelligibility of speech by separating speakers and suppressing background noise as well as other unwanted effects such as reverberation. If the devices are equipped with more than one microphone, which is very common nowadays, then multi-channel speech enhancement approaches can leverage spatial information in addition to single-channel tempo-spectral information to perform the task.

Traditionally, linear spatial filters, so-called beamformers, have been employed to suppress the signal components from other than the target direction and thereby enhance the desired speech signal. Since the noise reduction is insufficient in acoustically challenging scenarios, a beamformer for spatial filtering is often combined with a single-channel tempo-spectral post-filter. In single-channel speech enhancement and separation, approaches based on deep neural networks (DNNs) have been dominating the research landscape for some time. On the other hand, in multi-channel speech enhancement and separation, a change is currently taking place. Initially, DNNs were only integrated into multi-channel systems for tempo-spectral modeling, e.g., for estimating the beamformer parameters, but the spatial processing continued to be performed with a linear beamformer. Today, however, the number of publications that propose to replace the traditional pipeline with end-to-end trained DNNs is steadily increasing. With such an approach, DNNs can be used to realize a filter that integrates both spatial and temporal-spectral processing into a single non-linear operation. Such joint spatial and tempo-spectral non-linear filters are the subject of this thesis and referred to as *non-linear spatial filters*.

The first part of the thesis aims to clarify the benefits that an analytic non-linear spatial filter can offer compared to the traditional beamformer plus post-filter pipeline from a statistical perspective. A better understanding of the properties of non-linear spatial filters helps to decide if and in which situation a (DNN-based) non-linear spatial filter should replace the traditional approaches. Based on analytical estimators, we show that a non-linear spatial filter outperforms a beamformer plus post-filter approach if the noise distribution is non-Gaussian. Furthermore, by means of experiments, we demonstrate that the non-linear spatial filter enables a more powerful spatial processing that is not bound to the theoretical limits of a linear approach.

The second part focuses on the design and analysis of DNN-based joint spatial and tempo-spectral non-linear filters. We analyze the dependencies between the three available sources of information (spatial, spectral, and temporal) and find that the correlations between the frequency bands are particularly important for achieving a high spatial selectivity. Regarding the network architecture, this implies that spatial and spectral information should be processed together at an early stage. The DNN-based non-linear spatial filter designed according to this principle significantly outperforms an oracle beamformer plus DNN-based post-filter in difficult scenarios with a high number of interfering speakers and a low number of microphones.

In the third part of the thesis, we add a steering mechanism to the DNN-based non-linear spatial filter so that it can be steered in a chosen target direction. We apply the steerable filter to speech separation tasks and find that the explicit focus on the spatial selectivity of the filter during training is not only beneficial for the overall separation performance but also leads to an improved generalization ability compared to a similar network trained based on permutation invariant training (PIT).

As a result, this thesis not only contributes to a better theoretical understanding of non-linear spatial filters and their performance potential, but it also investigates various aspects of a practical implementation using DNNs. The research ultimately culminates in the development of a real-time demonstration of a DNN-based non-linear spatial filter.

Zusammenfassung

Ein großer Teil der menschlichen Sprachkommunikation findet in lauten Umgebungen statt und wird durch technische Hilfsmittel ermöglicht. So kann beispielsweise eine hörgeschädigte Person ein Hörgerät benutzen, um an einem Gespräch in einem belebten Restaurant teilhaben zu können. Diese Geräte, aber auch Telekommunikation in lauten Umgebungen oder sprachgesteuerte Assistenten, nutzen Algorithmen zur Sprachverbesserung und Sprechertrennung. Diese verbessern die Sprachqualität und -verständlichkeit, indem sie die Sprecher separieren und Hintergrundgeräusche sowie andere unerwünschte Effekte wie Nachhall unterdrücken. Wenn die Geräte mit mehr als einem Mikrofon ausgestattet sind, was heutzutage sehr häufig der Fall ist, dann können Ansätze zur mehrkanaligen Sprachverbesserung und Sprechertrennung zusätzlich zu den einkanaligen tempo-spektralen Informationen auch räumliche Informationen nutzen.

Traditionell wurden lineare räumliche Filter, so genannte Beamformer, eingesetzt, um die Signalanteile aus anderen Richtungen als der Zielrichtung zu unterdrücken und so das Sprachsignal zu verbessern. Da die Rauschunterdrückung in akustisch herausfordernden Szenarien meist unzureichend ist, wird ein Beamformer zur räumlichen Filterung oft mit einem einkanaligen tempo-spektralen Post-Filter kombiniert. Im Bereich der einkanaligen Sprachverbesserung und Sprechertrennung dominieren seit einiger Zeit Ansätze auf Basis von tiefen neuronalen Netzen (engl. deep neural networks (DNNs)) die Forschungslandschaft. Im Bereich der mehrkanaligen Sprachverbesserung und Sprechertrennung hingegen findet derzeit ein Umbruch statt. Ursprünglich wurden DNNs nur als tempo-spektrale Modelle in mehrkanalige Systeme integriert, z.B. für die Schätzung der Beamformer-Parameter, aber die räumliche Verarbeitung wurde weiterhin mit einem linearen Beamformer durchgeführt. Heute nimmt jedoch die Zahl der Veröffentlichungen stetig zu, welche die traditionellen Ansätze vollständig durch ein DNN ersetzen. In diesem Fall kann mit dem DNN ein Filter realisiert werden, welches sowohl die räumliche als auch die zeitlich-spektrale Verarbeitung in eine einzige nicht-lineare Operation zusammenfasst. Solche kombiniert räumlich und tempo-spektralen nicht-linearen Filter sind der Forschungsgegenstand dieser Arbeit und werden im Folgenden verkürzend als *nicht-lineare räumliche Filter* bezeichnet.

Der erste Teil der Arbeit untersucht die Vorteile eines analytischen nicht-linearen räumlichen Filters im Vergleich zu einer traditionellen Verkettung aus Beamformer und Post-Filter aus einer statistischen Perspektive. Ein besseres Verständnis der Eigenschaften nicht-linearer räumlicher Filter hilft bei der Entscheidung, ob und in welchen Situationen ein (DNN-basiertes) nicht-lineares räumliches Filter die traditionellen Ansätze ersetzen sollte. Basierend auf analytischen Schätzern zeigen wir, dass ein nicht-lineares räumliches Filter einen Beamformer in Kombination mit einem Post-Filter übertrifft, wenn das Rauschen nicht gaußverteilt ist. Darüber hinaus zeigen wir anhand von Experimenten, dass das nicht-lineare räumliche Filter eine leistungsfähigere räumliche Verarbeitung ermöglicht, die nicht an die theoretischen Grenzen eines linearen Ansatzes gebunden ist.

Der zweite Teil konzentriert sich auf den Entwurf und die Analyse von DNN-basierten kombiniert räumlich und tempo-spektralen nicht-linearen Filtern. Wir analysieren die Abhängigkeiten zwischen den drei verfügbaren Informationsquellen (räumlich, spektral und zeitlich) und stellen fest, dass die Abhängigkeiten zwischen den Frequenzbändern sehr wichtig sind, um eine hohe räumliche Selektivität zu erreichen. Im Hinblick auf die Netzwerkarchitektur bedeutet dies, dass räumliche und spektrale Informationen zu einem frühen Zeitpunkt gemeinsam verarbeitet werden sollten. Unser DNN-basiertes nicht-lineares räumliches Filter, das nach diesem Prinzip entworfen wurde, übertrifft in schwierigen Szenarien mit einer hohen Anzahl von störenden Sprechern und einer geringen Anzahl von Mikrofonen deutlich die Leistung eines Orakel-Beamformers kombiniert mit einem DNN-basierten Post-Filter.

Im dritten Teil der Arbeit fügen wir dem DNN-basierten nicht-linearen räumlichen Filter einen Steuerungsmechanismus hinzu, so dass es in eine bestimmte Zielrichtung ausgerichtet werden kann. Wir verwenden das steuerbare Filter für die Sprechertrennung und stellen fest, dass der explizite Fokus auf die räumliche Selektivität des Filters während des Trainings nicht nur vorteilhaft für die Gesamtleistung ist, sondern auch zu einer verbesserten Generalisierungsfähigkeit im Vergleich zu einem ähnlichen Netzwerk führt, das mit Hilfe einer permutations-invarianten Verlustfunktion trainiert wurde.

Im Ergebnis leistet diese Arbeit damit nicht nur einen Beitrag zu einem besseren theoretischen Verständnis nicht-linearer räumlicher Filter und ihres Leistungspotenzials, sondern untersucht auch verschiedene Aspekte einer praktischen Umsetzung mit DNNs. Die Forschung gipfelt schließlich in der Entwicklung einer Echtzeit-Demonstration eines DNN-basierten nicht-linearen räumlichen Filters.

Acknowledgement

First and foremost, I would like to thank my supervisor Prof. Dr.-Ing. Timo Gerkmann for all the support and advice in the last five years. It was inspiring to see that you never seemed to run out of ideas or suggestions how I could approach a problem or what to try next. This time in your research group was a great learning experience and a lot of fun.

I would like to thank Prof. Dr.-Ing. Reinhold Häb-Umbach and Prof. Dr. Chris Biemann for reviewing my thesis and Prof. Dr. Frank Steinicke for chairing the thesis committee.

Further, I would like to thank my colleagues from the Signal Processing research group. It was great to see how the group grew from just two PhD students in the beginning to ten by now. So many people who have always been very supportive. Thank you for great discussions and memorable conference visits (too few because of the pandemic!).

I thank my family and my partner Frank for their never ending support.

Table of Contents

Introduction 1

1.1 Motivation

Humans are often confronted with the task of understanding speech in noisy environments. The classic example is a conversation at a cocktail party, which serves as the namesake for the task of extracting a target speech signal from a recording that is corrupted by interfering speakers, interfering background noise, and/or reverberation: it is known as the cocktail party problem [1]. However, for many people, it is already a challenge to master communication in less extreme everyday situations. Many people have a problem following a conversation in a busy restaurant or at a dinner table when more than one person is speaking at the same time. Since the ability to understand speech in noisy environments decreases dramatically with age, technical solutions for this problem, for example, in the form of hearing aids, play an increasingly important role in our aging society. The technical solutions are based on algorithms for speech enhancement, speaker extraction, and separation, which have a wide range of applications besides hearing devices, including telecommunication or speech-controlled human-machine interaction based on automatic speech recognition (ASR) systems.

Speech processing algorithms are often classified according to the number of microphone channels that are used to record the noisy signal. While single-channel algorithms use the tempo-spectral properties of the signal to perform a speech processing task, multi-channel algorithms, i.e., those that process noisy recordings obtained from more than one microphone, can additionally leverage spatial information. Spatial information is present since the signal travels along different paths from a source to each microphone of the microphone array. In particular, the length of the direct path between the source and the individual microphones differs depending on the direction of arrival (DOA) of the signal. This is reflected by the time differences of arrival (TDOAs) of the signal in the different microphone channels. The well-known delay-and-sum beamformer [2] is a simple method that utilizes these spatial properties. The idea is to compensate for the TDOAs so that the target signal is time-aligned in all channels. Averaging the time-aligned signals then optimally reduces uncorrelated noise in the individual microphone channels. Another prominent example of a traditional spatial filter is the minimum variance distortionless response (MVDR) beamformer [2], which can be derived by optimizing for minimum noise variance subject to a distortionless constraint. These two spatial filters have a property that they share with all traditional linear spatial filters designed according to the filter-and-sum approach: When a local time-frequency bin is considered, and the filter coefficients are a function of the signals' statistics, the processing model is linear with respect to the noisy observation. In many realistic application scenarios, the noise reduction of the traditional linear spatial filters is insufficient, for example, those with a low number of microphones, many interfering sources, and reverberation. Therefore, a tempo-spectral post-filter is commonly applied to the single-channel output of the traditional linear spatial filters to improve the noise reduction performance.

Traditionally, both the linear spatial filters as well as a plethora of post-filters have been derived to fulfil a chosen statistical optimality criterion. Some publications also investigate the optimality of the two-stage approach combining a beamformer and a post-filter [3]–[5]. For example, a well-known result by Simmer et al. [5] is that the multi-channel Wiener filter can be decomposed in a MVDR beamformer and a single-channel Wiener post-filter. Perhaps under the impression of this result, the linear beamformer plus (non-linear) single-channel post-filter approach is commonly not perceived as a limitation. However, this ignores the fact that the result by Simmer et al. and the other results cited above are linked to the specific assumption that the noise follows a multivariate Gaussian distribution. On the other hand, a derivation of Hendriks et al. [6], who compute the minimum mean square error (MMSE) optimal filter under a Gaussian mixture noise assumption received only very little attention, and the theoretical result was evaluated only much later as part of this work. The resulting filter does not fit the two-stage processing scheme since it jointly performs the spatial and spectral processing in a single non-linear operation. For brevity, we will use the term *non-linear spatial filter* instead of joint spatial and (tempo-)spectral non-linear filter in the connecting text of this cumulative thesis.

Today, such a non-linear spatial filter does not necessarily have to be derived analytically but can also be learned from data with the help of deep learning techniques. Compared to a filter derived in a statistical framework, using a DNN has the advantage that fewer simplifying assumptions are needed to keep the estimation problem tractable and that the performance is often very good. On the other hand, neural networks require a training stage that is costly in terms of energy consumption and time and involves a large portion of unpredictability. It is difficult to foresee if a training will be successful and deliver a spatial filter with good performance that also generalizes to situations unseen during training.

In the deep learning age, the question of the limits of the two-stage approach combining a linear beamformer and a post-filter is of great practical relevance, and it is one of the core questions addressed in this thesis. We investigate the advantages a non-linear spatial filter can offer compared to the much simpler traditional linear spatial filter with a separate post-filter in relevant practical scenarios. A good understanding of the properties and performance potential makes it possible to evaluate if a (DNN-based) non-linear spatial filter should replace the traditional linear spatial filters, in which situations this is advisable, and how such a filter should be designed. In this thesis, we start from a statistical perspective on non-linear spatial filtering and finally arrive at a well-performing implementation based on DNNs. The remainder of the introduction explains the addressed speech processing tasks, describes prior work as well as the development of the research landscape parallel to the work on this thesis, and outlines the research questions that are investigated in detail in the publications included in this cumulative thesis.

1.2 Multi-channel Speech Enhancement, Speaker Extraction and Separation

In this work, we address three speech signal processing tasks that require one or multiple corrupted target speech signals to be recovered from a noisy multi-channel recording. The goal is to improve the speech quality and intelligibility of the target speech signals, which are degraded by interfering noise and reverberation. The task of recovering a single target speech signal from background noise is commonly referred to as *speech enhancement*. The underlying core of all speech enhancement algorithms is the ability to distinguish between the signal components that belong to the target speech signal and those that should be suppressed because they belong to the background noise. It is clear that the task's difficulty increases when the noise is more similar to the target speech signal itself. Accordingly, the so-called cocktail party scenario with many interfering speech signals can be considered particularly difficult. In the literature, the task of recovering a single target speaker from a mixture of speech signals is often referred to as *speaker extraction*, while extracting multiple target speakers is called *speech separation*.

In the single-channel case, only tempo-spectral information is available to distinguish between the target signal and the unwanted interfering signals. Traditional noise reduction techniques, e.g., a Wiener filter [2, Sec. 11.4] or other estimators derived in a statistical framework [7]–[10] along with their associated parameter estimation schemes, e.g., [7], [11]–[13], rely on the assumption that the noise is more stationary than the target speech signal. This limiting assumption is overcome by DNN-based approaches, whose impressive modeling capabilities have been the driving force behind the progress in single-channel speech enhancement in recent years. The neural network revolution resulted in many well-performing single-channel speech enhancement systems, e.g., [14], [15]. Furthermore, in 2016 and 2017, the influential deep clustering [16] and PIT [17] papers have been a major break-through in single-channel speech separation, a problem that seemed hardly solvable before the DNN era. In the years that followed, a variety of systems has been developed that offer impressive single-channel separation performance, e.g., [18]–[22].

In contrast to single-channel approaches, multi-channel algorithms can leverage spatial information in addition to tempo-spectral information. The additional spatial information makes multi-channel approaches potentially much more powerful than single-channel approaches. This is particularly relevant when the scenarios under consideration are challenging due to very low signal-to-noise ratios (SNRs), high reverberation times, diverse and non-stationary noise types, many interfering speakers, and/or tight resource constraints that limit the size of the employed DNNs. Section 1.3 describes the traditional approach to multi-channel speech enhancement with a linear spatial filter, a so-called beamformer, in more detail. However, the fundamental idea is simple: enhance the target speech signal arriving from a specific direction by suppressing signal components from all other directions. This idea provides the basis for decades of research in multi-channel speech enhancement. Numerous publications have addressed the questions of how to design the linear spatial filter for maximum performance, how to estimate the required parameters, which post-filtering schemes are appropriate, and how to increase the robustness under practical constraints [23]–[26].

Not only multi-channel speech enhancement but also multi-channel blind source separation (BSS) has a long-standing research history. More than twenty years of research have led to a variety of approaches. The term blind indicates that no prior knowledge regarding the "recording environment, mixing system, or source locations" [27] is assumed. Many approaches have been proposed to estimate a set of linear de-mixing filters based on modeling assumptions that impose structural constraints on the spatial properties and the spectral structure. For example, techniques based on independent component analysis (ICA) [28]–[30] attempt to optimize a separation criterion that enforces the independence of the individual source signals. ICA was later extended to the independent vector analysis (IVA) method [31], [32] to avoid the frequency-wise permutation problem that arises when separation is performed independently for each frequency bin. The multi-channel non-negative matrix factorization (NMF) technique combines single-channel NMF for low-rank modeling of the sources' power spectra with a model for the spatial covariance matrices. It can also be derived in the more general framework of local Gaussian models [33], [34] and has been combined with IVA, which resulted in the independent low-rank matrix analysis (ILRMA) technique [27], [35]. Another dominant line of research is concerned with clustering-based techniques, which group time-frequency bins with similar spatial and/or tempo-spectral characteristics [34]. For example, [36]–[39] cluster the time-frequency bins according to two spatial features, namely the inter-channel phase differences (IPDs) and inter-channel level differences (ILDs). Clustering-based approaches often involve a statistical model of the observed spatial mixtures with a latent random variable that describes the assignment of each time-frequency bin to a sound source. The parameters of such a statistical model can then be estimated with the expectation-maximization (EM) algorithm [40], [41], and the resulting assignments of time-frequency bins to sources may be used for masking the mixture signal directly or for mask-based beamforming, which is explained in Section 1.4.1.

The success of DNN-based approaches for single-channel enhancement and separation inspired many researchers to explore the applicability of deep learning techniques for multi-channel tasks. The first approaches, which emerged from around 2015, integrate DNNs for tempo-spectral modeling into existing multi-channel processing chains. For example, Heymann et al. [42] and Erdogan et al. [43] proposed to assign time-frequency bins to a speech or a noise mask based on their tempo-spectral characteristics and use these masks for estimating the parameters of a traditional linear spatial filter. Others have explored similar ideas and integrated DNNs as a tempo-spectral model with spatial clustering approaches for multi-channel speech enhancement and separation [44]–[46]. These approaches have been designed to separate the tempo-spectral processing (with a DNN) and spatial processing (with a statistical model). Many newer approaches, just as the DNN-based non-linear spatial filters developed in this thesis, do not separate the spatial and tempo-spectral processing, but they process spatial and tempo-spectral information jointly and end-to-end with a DNN. Since these DNN-based non-linear spatial filters are the research focus of this work, Section 1.4 gives a detailed overview of the related work on DNNs for spatial filtering. Many of the approaches presented there were developed parallel to this work.

For this thesis, we adopt a perspective that views the tasks of multi-channel speaker extraction and separation as a spatial filtering problem in the same way as the task of multi-channel speech enhancement. This is also the perspective underlying separation approaches that work with a collection of fixed beamformers. The beamformers are steered in a large number of directions, which is followed by a "beam selection" step, where the best filter is selected based on the separation quality of the output [47]–[49]. Our underlying assumption is that a good spatial filter should be able to suppress most of the interfering signals regardless of whether they are background noise or other human speakers as long as they have spatial properties that distinguish them from the target signal. Since the only difference between speech enhancement and speaker extraction under this perspective is the type of noise, we do not always differentiate between speech enhancement and speaker extraction in the publications included in this thesis but sometimes use the general term speech enhancement for both. In this perspective, the only difference in handling a speech separation task as opposed to performing speech enhancement or speaker extraction is that multiple spatial filters are required, which are then steered in the direction of the different target speakers. In most cases considered in this thesis, we assume that information about the location of the target speakers is available, either through access to oracle data for parameter estimation, the use of a dataset with a known fixed placement of the target speaker, or oracle DOA information. There is a large body of literature on sound source localization, e.g., see [50] for a review of recent developments. However, the task of sound source localization is mostly out of the scope of this thesis, which focuses on the analysis and design of non-linear spatial filters.

1.2.1 Signal Model and Notation

Next, we give a formal description of the tasks and introduce the notation. The general signal model described here is common to all the publications included in this thesis, but the naming of the variables may differ. We consider a scenario where a C-channel microphone array records a noisy mixture signal composed of overlapping speech signals uttered by P speakers and environmental noise. We denote the time-domain dry speech signal of the p's speaker by $s_p(t)$ with time index t. The speech signal $x_p^\ell(t)$ recorded at the ℓ's microphone can be written as the convolution of the dry speech signal and the acoustic impulse response (AIR) $a_p^\ell(t)$ modeling the propagation path between the p's speaker and the microphone [23], [34], i.e.,

$$x_p^\ell(t) = s_p(t) * a_p^\ell(t). \tag{1.1}$$

In an anechoic scenario, the AIR $a_p^\ell(t)$ describes a time-shift τ_p^ℓ, which corresponds to the propagation time of the signal from the speaker to the microphone along the direct path. In a reverberant environment, the AIR $a_p^\ell(t)$ does not only model the direct path but also the multi-path components caused by reflections from the walls and other obstacles [51]. This way, the AIR reflects the properties of the room in which the signal is transmitted between the source and the recording microphone and, accordingly, the AIR is often referred to as room impulse response (RIR), when enclosed spaces are considered. A typical RIR exhibits distinct peaks related to the direct path and early reflections in the beginning. The reverberation tail models signal components that have been reflected multiple times before reaching the microphone. Since many paths with multiple reflections of similar length exist, their summation in the second part of the RIR has a random structure. The energy level decays exponentially. For a basic description of the acoustic scenario, the reverberation time (RT60) and the direct-to-reverberation ratio (DDR) are usually reported to characterize the properties of the reverberation in a room [51]. The RT60 measures the time it takes for the sound pressure level to decay by 60 decibels (dB), and the DDR is defined as the ratio between the power of the direct path signal (first peak of the RIR) and the power of the signal components reaching the microphone via indirect paths (described by the rest of the RIR).

We transform the noisy signal into the time-frequency domain using a short-time Fourier transform (STFT). Then, by the additive signal model, the STFT-domain noisy signal recorded at microphone ℓ decomposes into a sum of (reverberant) speech signals $X_p^\ell(k,i)$ and additional background noise $V^\ell(k,i)$, i.e.,

$$Y^\ell(k,i) = \sum_{p=1}^{P} X_p^\ell(k,i) + V^\ell(k,i) \tag{1.2}$$

with frequency bin index k and time-frame index i. For a more compact notation, we stack the signals for all microphone channels and denote the multi-channel signal vector with a bold letter, e.g.,

$$\mathbf{Y}(k,i) = [Y^1(k,i), Y^2(k,i), ..., Y^C(k,i)]^T \in \mathbb{C}^C. \tag{1.3}$$

Employing the narrow-band approximation [51], we can rewrite the convolution in (1.1) as a multiplication of the STFT-transformed dry speech signal and the acoustic transfer function (ATF), i.e.,

$$\mathbf{X}_p(k,i) = S_p(k,i) \cdot \mathbf{a}_p(k,i). \tag{1.4}$$

This approximation decouples the individual time-frequency bins but is only equivalent to (1.1) if the AIR is much shorter than a single STFT window [34, Sec. 2.3.2]. Throughout this work, we assume that the acoustic setup, i.e., the positioning of the speakers and the microphone array or the reflection properties of the walls, does not change within a single utterance. Therefore, the ATF can be modeled as time-invariant and denoted without a dependency on the time-frame index i. In a practical application, the ATFs are usually unknown and need to be estimated, which also requires resolving the problem of gain ambiguity. From (1.4), it appears that the magnitude of $X_p^\ell(k,i)$ is influenced by the loudness of the source signal as well as the attenuation due to the length of the propagation path modeled by $a_p^\ell(k,i)$. Estimating relative transfer functions (RTFs) instead of ATFs bypasses these difficulties. The RTFs $\tilde{\mathbf{a}}_p(k,i)$ are obtained by dividing the ATF vector by the ATF of a chosen reference channel, i.e.,

$$\tilde{\mathbf{a}}_p(k) = \left[1, \frac{a_p^2(k)}{a_p^1(k)}, ..., \frac{a_p^C(k)}{a_p^1(k)}\right]^T \tag{1.5}$$

with using the first channel as a reference. The RTF describes the relationship between the target signal received at the reference microphone with the signal received at the other microphones, i.e.,

$$\mathbf{X}_p(k,i) = S_p(k,i) \cdot \mathbf{a}_p(k) = S_p(k,i) \cdot a_p^1(k) \cdot \frac{\mathbf{a}_p(k)}{a_p^1(k)} = X_p^1(k,i) \cdot \tilde{\mathbf{a}}_p(k). \tag{1.6}$$

In an anechoic or free-field scenario and assuming that the distance between the speaker and the microphone array is sufficiently large so that the ILDs can be neglected, the RTF simplifies to the so-called (relative) steering vector [51]

$$\mathbf{d}_p(k) = \begin{bmatrix} 1 \\ e^{-2\pi j \Delta\tau_p^2 f_k} \\ \vdots \\ e^{-2\pi j \Delta\tau_p^C f_k} \end{bmatrix} \tag{1.7}$$

which only depends on the relative TDOAs $\Delta\tau_p^\ell = \tau_p^\ell - \tau_p^1$ with respect to the reference microphone, for which we again picked the first. The variable f_k denotes the continuous frequency associated with the kth frequency bin in Hertz. For an STFT frame length of N samples and with sampling frequency f_s, it is given by $f_k = \frac{kf_s}{N}$.

The noise $\mathbf{V}(k,i)$ recorded at the microphones can be sensor noise, which is independent between microphone channels, or environmental noise, which is likely to be correlated between microphone channels, especially in low frequencies. When adopting a statistical perspective, we model the spectral coefficients as zero-mean random variables and make the common simplifying assumption that they are independent with respect to the frequency bin and the time-frame index. The spatial correlation of the noise signals recorded at the different

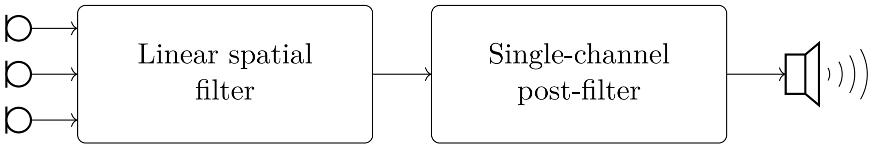


Figure 1.1: Illustration of the traditional two-step processing with a linear spatial filter (beamformer) and single-channel post-filter for tempo-spectral post-processing.

microphone channels is then described by the noise correlation matrix $\mathbf{\Phi}_V(k,i)$:

$$\mathbf{\Phi}_V(k,i) = \mathbb{E}[\mathbf{V}(k,i)\mathbf{V}(k,i)^H] \in \mathbb{C}^{C\times C}. \tag{1.8}$$

Here, we denote with $\mathbb{E}$ the statistical expectation operator and with $(\cdot)^H$ the Hermitian transpose. For a zero-mean signal, the covariance matrix and correlation matrix are identical, so that these terms are often used interchangeably in the context of audio processing.

When the number of speakers P is set to one, then (1.2) describes a speech enhancement task with a single target speech signal potentially corrupted by reverberation and background noise. For speaker extraction and speech separation, we set $P > 1$. During the training of the DNN-based non-linear spatial filters, the goal is to recover the dry speech signal(s) $S_p(k,i)$ or, equivalently, $s_p(t)$, however, delayed by the time it takes the signal to travel from the source to the reference microphone along the direct path. This formulation includes the task of dereverberation if the AIRs model an enclosed room. We choose to include the dereverberation task in the training objective for our neural networks, as this formulation is aligned with the spatial filtering perspective. The objective of the spatial filter is to suppress all signal components arriving from directions other than the target direction, which includes the reflections of the target signal that arrive from directions other than the direct path.

1.3 Traditional Linear Spatial Filtering

Most traditional speech enhancement schemes are designed as a two-step procedure: a linear spatial filter (beamformer) combined with a post-filter. Figure 1.1 illustrates this setup. The multi-channel signal is fed into the spatial filter, which aims to emphasize a signal from a certain direction by suppressing signal components that arrive from other directions. The dimension of the resulting signal is reduced to one. As the noise reduction obtained with a spatial filter is often not sufficient, the output is further processed by a potentially non-linear single-channel post-filter. In this setup, the spatial and tempo-spectral information are processed separately.

Most beamformers are based on the filter-and-sum processing scheme, where the individual microphone signals are filtered and then summed. In the frequency-domain, this can be written as

$$\hat{S}(k,i) = \mathbf{w}(k,i)^H\mathbf{Y}(k,i). \tag{1.9}$$

The speaker index p is omitted in the following sections on spatial filtering to simplify the notation. The vector $\mathbf{w}(k,i) \in \mathbb{C}^C$ contains the filter weights, and $\hat{S}(k,i)$ denotes the estimate of the target speech signal. Like the AIRs, the filter weights may or may not depend on the time-index i. Furthermore, the filter can be data-dependent or data-independent. An example of the latter is the delay-and-sum beamformer [2]. In this case, the filter weights are designed to compensate for the TDOAs at the microphone channels such that the target signal is time-aligned in all channels after applying the respective filter. The summation then leads to a constructive superposition of the target signal and optimally reduces uncorrelated noise, e.g., sensor noise. However, the noise suppression of interfering point sources or diffuse noise, where the microphone signals are highly correlated especially in low frequencies, might be insufficient.

Unlike the delay-and-sum beamformer, the well-known MVDR beamformer requires an estimate of the noise correlation matrix. Its filter weights can be derived by solving the optimization problem [2], [24]

$$\begin{aligned} \mathbf{w}_{\mathrm{MVDR}}(k,i) = \arg\min_{\mathbf{w}\in\mathbb{C}^C} \ & \mathbf{w}^H(k,i)\boldsymbol{\Phi}_N(k,i)\mathbf{w}(k,i) \\ \text{s.t.} \ & \mathbf{w}^H(k,i)\mathbf{d}(k) = 1, \end{aligned} \tag{1.10}$$

where the vector $\mathbf{N}(k,i)$ combines all interfering signal components, i.e., environmental noise and potentially also interfering speakers. The objective function requires that the power of the noise signal at the output of the beamformer be minimized, while the so-called distortionless constraint enforces that the target signal remains unchanged. The target signal is identified by its spatial properties described by the steering vector $\mathbf{d}(k)$. The optimization problem in (1.10) can be solved using the technique of Lagrange multipliers [2], which leads to filter weights

$$\mathbf{w}_{\mathrm{MVDR}}(k,i) = \frac{\boldsymbol{\Phi}_N^{-1}(k,i)\mathbf{d}(k)}{\mathbf{d}^H(k)\boldsymbol{\Phi}_N^{-1}(k,i)\mathbf{d}(k)}. \tag{1.11}$$

The optimization problem in (1.10) can also be posed by replacing the steering vector $\mathbf{d}(k)$ with the RTF vector $\tilde{\mathbf{a}}(k)$ [52], which means that the occurrences of $\mathbf{d}(k)$ in (1.11) are also replaced by $\tilde{\mathbf{a}}(k)$. Besides the steering vector or RTF vector, the solution depends on the second-order statistics of the noise given by $\boldsymbol{\Phi}_N(k,i)$. The noise correlation matrix can be estimated from the noisy observation with strategies explained in Section 1.3.1. In this case, since the statistical properties of the noise are taken into account during filter design, the MVDR using a can be considered to be data-dependent [23]. Alternatively, if a diffuse noise field is assumed, an analytic expression can be used to obtain the corresponding noise correlation matrix, leading to a so-called superdirective beamformer [53], [54]. It is important to note that, also in the data-dependent case, the filter weights only depend on the statistics, i.e., the spatial correlation matrix, of the noise, but not the noisy input signal $\mathbf{Y}(k,i)$ itself. This is the common situation in traditional beamforming, which means filter-and-sum beamforming as in (1.9) is a linear operation with respect to the noisy input signal when considering a local time-frequency bin.

There are many more beamformers besides the delay-and-sum and MVDR beamformer that result from variations in the design criteria. For example:

- While the MVDR minimizes the noise power, the minimum power distortionless response (MPDR) minimizes the overall power of beamformer output subject to a distortionless constraint [51].
- The linearly constraint minimum variance (LCMV) beamformer generalizes the MVDR and introduces a set of linear constraints [24], [55]. For example, these can be set up, so that a distortionless constraint is accompanied by constraints that steer a null in the direction of an interfering source. The maximum number of constraints is bounded by the number of microphones so that the maximum number of undesired sources that can be canceled in an anechoic scenario is $C - 1$.

- The maximum SNR or generalized eigenvalue (GEV) beamformer strives to maximize the SNR at the output of the beamformer [56], [57]. Since this beamformer may introduce speech distortions, Warsitz et al. propose to combine it with a post-filter to compensate for the speech distortions [56].

- The multi-channel Wiener filter (MWF) can be derived by selecting the filter weights of a filter-and-sum beamformer to optimize the MMSE criterion comparing the target speech signal and the estimate in (1.9) [2, Sec. 12.7], i.e.,

  $$\mathbf{w}_{\mathrm{MWF}}(k,i) = \underset{\mathbf{w}\in\mathbb{C}^C}{\arg\min}\ \mathbb{E}\left[\left|S(k,i) - \mathbf{w}(k,i)^H\mathbf{Y}(k,i)\right|^2\right]. \tag{1.12}$$

  The solution involves the inverse of the correlation matrix of the noisy signal $\mathbf{\Phi}_Y$ and the cross-correlation vector of the noisy and clean signal $\mathbf{\Phi}_{YS}$ as follows:

  $$\mathbf{w}_{\mathrm{MWF}}(k,i) = \mathbf{\Phi}_Y^{-1}\mathbf{\Phi}_{YS}. \tag{1.13}$$

  With using the narrow-band approximation for the speech signal and assuming independent speech and noise signals, the multi-channel Wiener filter can be decomposed in a linear spatial filter followed by a single-channel post-filter as shown in Figure 1.1 [5]. The spatial filter then matches the MVDR beamformer, and the single-channel post-filter equals the well-known single-channel Wiener filter [2].

The optimization problem associated with the MWF in (1.12) is based on the filter-and-sum processing model. Solving the optimization problem in (1.12) does not require any assumptions about the probability distributions of the speech and noise signal, which makes the solution very general. However, by optimizing (1.12) with respect to the filter weights $\mathbf{w}(k,i)^H$, the filter is constrained to comply with the filter-and-sum processing model and, therefore, to be a linear filter. Interestingly, the MWF can also be derived in the context of Bayesian estimation, assuming a multivariate Gaussian distribution for the noise signal and univariate Gaussian distribution for the target clean speech signal. The Bayesian MMSE estimate $\mathrm{T}_{\mathrm{MMSE}}(\mathbf{Y}(k,i))$ can be defined as

$$\mathrm{T}_{\mathrm{MMSE}}(\mathbf{Y}(k,i)) = \underset{\hat{S}\in\mathbb{C}}{\arg\min}\ \mathbb{E}\left[\left|S(k,i) - \hat{S}(k,i)\right|^2\right], \tag{1.14}$$

where the target speech estimate $\hat{S}(k,i)$ could be any function of the input $\mathbf{Y}(k,i)$. In particular, no assumption on the linearity of this function is made. Nevertheless, the solution of the minimization task in (1.14) is a filter-and-sum beamformer with filter weights $\mathbf{w}_{\mathrm{MWF}}$ given in (1.13) if a complex Gaussian distribution is assumed for speech and noise spectral coefficients [24, Sec. 6.2.2.2]. In their work, Balan and Rosca [3] generalize these findings to all settings that involve a noise signal that follows a multivariate Gaussian distribution. Using the theory of sufficient statistics, their work leads to the following conclusion: If the noise signal follows a multivariate Gaussian distribution, then the MMSE-optimal spatial filter is an MVDR beamformer followed by a possibly non-linear single-channel post-filter. The probability distribution of the speech signal can be chosen arbitrarily and only affects the type of post-filter. Also the maximum a posteriori (MAP) and maximum likelihood (ML) estimators can be shown to involve an MVDR beamformer for spatial processing. A more detailed review of these relationships is outlined in our publication [P3], which is part of this thesis and includes a new simplified proof for the result of Balan and Rosca. Along similar lines, Schwartz et al. [4] also decompose an MMSE filter derived under a Gaussian assumption

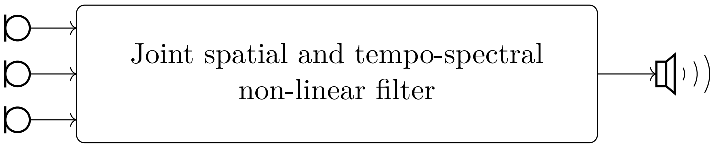


Figure 1.2: Illustration of a non-linear filter, which jointly performs spatial and tempo-spectral processing.

in a multi-speaker case in an LCMV beamformer and a post-filter.

The results reported here may give the impression that a linear beamformer already provides statistically optimal results and, thus, not much can be expected from a non-linear filter. However, it is important to note that these results have a strong underlying assumption, namely that the noise follows a multivariate Gaussian distribution. The Gaussian distribution assumption used to be a common choice also for speech spectral coefficients in statistical single-channel speech enhancement, e.g., [7], but was then dropped in favor of more heavy-tailed distributions [8]–[10], which were found to better resemble the characteristics of speech signals. Clearly, if the multi-channel noise signal involves interfering speakers, which have a non-Gaussian spectral characteristic, different spatial properties, and are sparse in the time-frequency domain [38], a multivariate Gaussian is arguably not the best model. Also in other cases, e.g. real-world noise recordings in different locations, the Gaussian noise assumption may be inaccurate.

Hendriks et al. [6] derived the MMSE filter for a multivariate Gaussian mixture distribution and found it to be a non-linear filter that cannot be separated in a spatial processing and post-filtering stage. A schematic view is given in Figure 1.2. While solving the MMSE estimation problem for a non-Gaussian noise distribution is difficult in most cases, the result by Hendriks et al. is general in the sense that “almost any continuous density can be approximated to arbitrary accuracy” [58, Sec. 2.3.9] with a Gaussian mixture distribution. Therefore, dropping the Gaussian noise assumption will in most cases lead to a joint spatial and tempo-spectral non-linear filter. However, this theoretically quite interesting result derived in [6] has never been evaluated by Hendriks et al. or anyone else prior to our publication [P3]. This thesis investigates the properties, the performance impact, and the implementation (using deep learning techniques) of this type of filter in contrast to the linear beamformers, such as the MVDR beamformer, described earlier in this section. While some recent publications refer to a multi-channel DNN as a “non-linear beamformer”, e.g., [59], in this thesis and all included publications, we use the term beamformer exclusively to refer to filters that are linear with respect to the noisy input signal.

1.3.1 Parameter Estimation

Many beamformers take a relatively simple closed analytic form. However, their use requires the estimation of parameters that are necessary for the computation of the filter weights. For example, the MVDR beamformer requires an estimate of the steering vector or of the RTF vector and an estimate of the noise correlation matrix. The quality of the parameter estimates has a large impact on the performance of the beamformer. In this thesis, when comparing to traditional methods, we estimate the beamformer parameters from oracle data, which means that access to the target speech signal $\mathbf{X}(k,i)$ and the pure noise signal $\mathbf{N}(k,i)$ recorded at the microphones is available. Even though this is not the case in real-world application scenarios, we think that we can learn the most about the potential performance gain of a non-linear spatial filter by comparing it to an oracle linear spatial filter, which reflects the upper bound on the performance that can be obtained with a linear spatial filter. Nevertheless, this section provides an overview of techniques that can be applied in blind settings.

Noise Correlation Matrix Estimation

The definition of the correlation matrix as given in (1.8) uses the statistical expectation operator, i.e., the noise correlation matrix required by the MVDR beamformer is defined as

$$\mathbf{\Phi}_N(k,i) = \mathbb{E}[\mathbf{N}(k,i)\mathbf{N}(k,i)^H] \tag{1.15}$$

for each time-frequency bin (k,i). In practice, this equation comes with challenges:

1. The pure noise signal $\mathbf{N}(k,i)$ is not accessible in real-world application scenarios. Instead, the microphone array records the mixture signal $\mathbf{Y}(k,i)$.

2. The statistical expectation operator must be replaced with an average over a time span with approximately stationary statistical properties. This can be difficult if the statistical properties are changing quickly.

3. Many subsequent processing steps, for example, the computation of MVDR beamformer weights, require that the noise correlation matrix is invertible. Problems can arise when the number of data points that is averaged is too low. Furthermore, in some scenarios, e.g., with a directional noise source, even the ideal matrix is not invertible according to the previously presented signal model. In these cases, a regularization like the diagonal loading technique [24, Sec. 6.6.4] can be applied to ensure that the matrix is invertible.

The easiest way to solve the problem is to assume that the target speech source is inactive for a certain period of time, e.g., at the beginning of each utterance. Then an average along the time axis can be computed to estimate the correlation matrix as

$$\hat{\mathbf{\Phi}}_N(k,i) = \frac{1}{|\mathbb{L}|}\sum_{j\in\mathbb{L}}\mathbf{Y}(k,j)\mathbf{Y}(k,j)^H = \frac{1}{|\mathbb{L}|}\sum_{j\in\mathbb{L}}\mathbf{N}(k,j)\mathbf{N}(k,j)^H \tag{1.16}$$

with $\mathbb{L}$ being the set of time-frame indices in which the target speech signal is not active and, therefore, for which $\mathbf{Y}(k,i) = \mathbf{N}(k,i)$ holds. Numerous voice activity detection (VAD) techniques have been proposed, which estimate if a speaker recorded in background noise is active or not, e.g., [60]–[63]. The formulation in (1.16) relies on a binary decision whether to include a time-frame in the estimate or not. Using a speech presence probability (SPP) estimate, the binary decision can be replaced by a soft decision. The SPP is defined as

$$\rho(k,i) = P(\mathcal{H}_1|\mathbf{Y}(k,i)) \tag{1.17}$$

with $\mathcal{H}_1$ denoting the hypothesis that the target speech source is active. Clearly, the probability of a non-active target speech source $\mathcal{H}_0$ is given by

$$\eta(k,i) = P(\mathcal{H}_0|\mathbf{Y}(k,i)) = 1 - P(\mathcal{H}_1|\mathbf{Y}(k,i)) = 1 - \rho(k,i). \tag{1.18}$$

An estimate of this quantity, $\hat{\eta}(k,i)$, can then be used to control the estimation of the noise correlation matrix as follows:

$$\hat{\mathbf{\Phi}}_N(k,i) = \frac{\sum_j \hat{\eta}(k,j)\mathbf{Y}(k,j)\mathbf{Y}(k,j)^H}{\sum_j \hat{\eta}(k,j)}. \tag{1.19}$$

In many real-world applications, the characteristics of the noise signal can be expected to change over time. In this case, the set of time-frames included in the average can be restricted to a neighborhood of the respective ith time-frame [64] or a recursive averaging strategy can be used, which weights the contribution of close-by time-frames higher than distant ones, i.e.,

$$\hat{\mathbf{\Phi}}_N(k,i) = \alpha'(k,i)\hat{\mathbf{\Phi}}_N(k,i-1) + (1-\alpha'(k,i))\mathbf{Y}(k,i)\mathbf{Y}(k,i)^H \tag{1.20}$$

with a time-varying weighting factor

$$\alpha'(k,i) = \alpha\hat{\eta}(k,i) + (1-\hat{\eta}(k,i)) \tag{1.21}$$

that depends on the estimate of $\hat{\eta}(k,i)$ for the respective time-frequency bin and a forgetting factor $\alpha \in [0,1]$ [51, Eq. 90].

Traditional single-channel SPP estimation schemes, e.g., [13], exploit the tempo-spectral characteristics of the speech signal assuming that the noise signal is more stationary than the speech signal. If interfering speech sources are part of the noise signal that is to be suppressed, this assumption is not valid anymore. However, also in multi-speaker scenarios, it is often assumed that each time-frequency bin is dominated by a single speech or noise source [38]. Clustering-based techniques can then be used to assign each time-frequency bin to a speech source or the noise signal based on the spatial properties [38], [65]–[67]. The resulting time-frequency masks can be used in place of $\hat{\eta}(k,i)$ in (1.19) or (1.20). In this work, a beamforming scheme that relies on SPP estimates or masks for parameter estimation is referred to as mask-based beamforming. Masks can also be estimated with a neural network, which is explained in Section 1.4.1.

Steering Vector Computation

The computation of the relative steering vector in (1.7) requires knowledge of the TDOAs $\Delta\tau^\ell$. These are easy to compute if the array geometry, array position, and the exact position of the target speaker are known. An illustration is shown in Figure 1.3. Based on the coordinates of the microphones and the source, the absolute lengths of the direct paths can be obtained. The TDOAs are then obtained by subtracting the length of the path to the reference microphone m_1 (shown in dark blue) and dividing by the speed of sound c_s. In many practical scenarios, however, the calculation is based on a far-field assumption, which means that it is assumed that the distance between the target speaker and the microphone array is much larger than the distance between the individual microphones in the array. In this case, the sound signal propagation may be modeled by a plane wave, and the attenuation related to differences in the propagation path length may be considered negligible [2], [34].

An illustration of a far-field scenario with a three-channel microphone array can be seen in Figure 1.4. The dashed gray lines represent the propagation paths between the target speaker on the right and the microphone array in the center of the figure. The plane wave is assumed to travel along the gray dashed parallel lines as indicated by the red arrows with a wavefront perpendicular to the propagation direction of the signal, which is shown as a red dashed line in the zoomed-in part of Figure 1.4. In the illustration, the signal first reaches

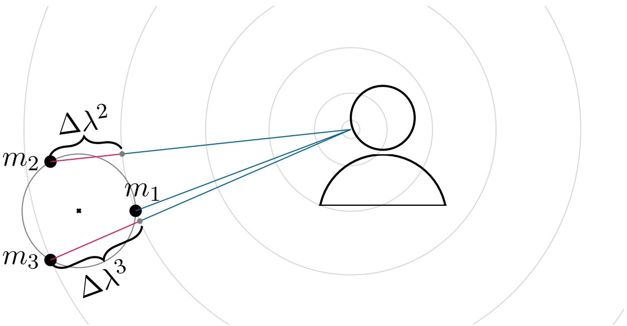


Figure 1.3: Illustration of the relative steering vector computation in a near-field scenario. The relative TDOA $\Delta\tau^\ell$ for the microphone ℓ is computed by dividing the difference in the propagation path length $\Delta\lambda^\ell$ by the speed of sound c_s, i.e., $\Delta\tau^\ell = \frac{\Delta\lambda^\ell}{c_s}$.

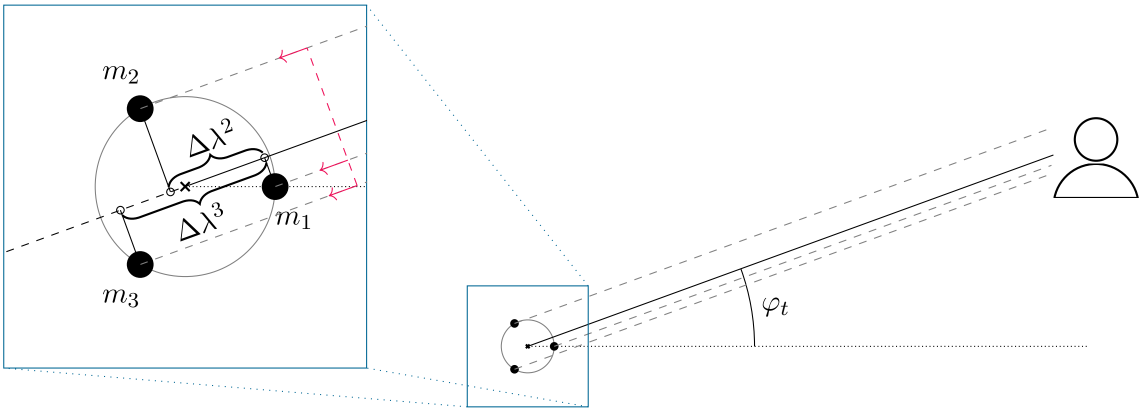


Figure 1.4: Illustration of the relative steering vector computation in a far-field scenario. The sound wave propagation is modeled as a plane wave represented by the dashed red line. The relative TDOA $\Delta\tau^\ell$ for the microphone ℓ is computed by dividing the difference in the propagation path length $\Delta\lambda^\ell$ by the speed of sound c_s, i.e., $\Delta\tau^\ell = \frac{\Delta\lambda^\ell}{c_s}$.

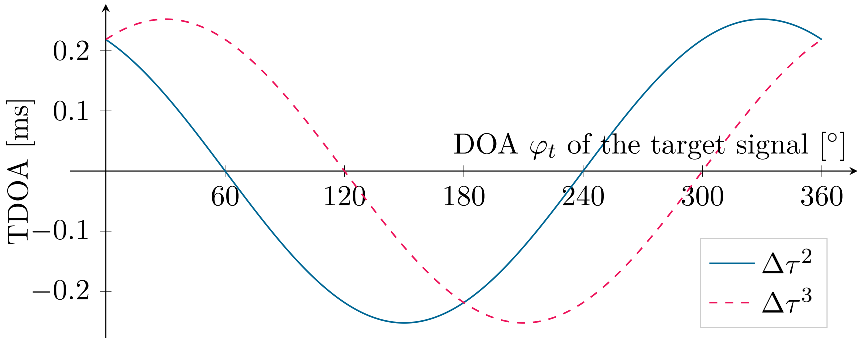


Figure 1.5: Dependency between the target speaker's DOA angle and the TDOAs in milliseconds. The TDOA has been computed for the microphone array shown in Figure 1.4 with three microphones and a circular arrangement. The radius of the microphone array is 5 cm, and the first microphone m_1 is used as a reference.

microphone m_1, then m_2 and finally m_3. The distance that the signal must travel to reach m_2 and m_3 after reaching m_1 has been denoted with $\Delta\lambda^2$ and $\Delta\lambda^3$, respectively. From this, the TDOAs are easily computed by dividing by the speed of sound c_s, i.e., $\Delta\tau^\ell = \frac{\Delta\lambda^\ell}{c_s}$. Under this model, changing the DOA of the target signal φ_t strongly affects the differences in the propagation paths' lengths and, therefore, also the TDOAs, while the distance between the target source and the microphone array is not considered in the computation. The exact relationship between the TDOAs and the target speaker's DOA is depicted in Figure 1.5 for a three-channel circular array with 5 cm radius. Mathematically, the relationships are described by a scaled cosine function that is shifted by 30° and $-30°$ for the TDOA at microphone m_2 and m_3, respectively.

In some applications, the DOA is approximately known. For example, the position of a person in a car seat only varies in a small range. Similarly, hearing aid users will likely turn their heads towards the target speaker during a conversation. In other scenarios, the DOA must be estimated without prior knowledge either from the noisy audio signal alone or with support from other input modalities, e.g., audio-visual sound source localization.

Relative Transfer Function Estimation

In comparison with the steering vector, which is solely based on the direct path between the source and microphone array, the ATF vector in (1.4) is a more flexible model for the propagation of sound since it can include reflections of the signal at walls and obstacles, which occur when the signals are recorded in an enclosed space and not in the free-field. By substituting (1.4) for the recorded speech signal, the speech correlation matrix then computes as

$$\begin{aligned}\mathbf{\Phi}_X(k,i) &= \mathbb{E}\left[\mathbf{X}(k,i)\mathbf{X}(k,i)^H\right] \\ &= \mathbb{E}\left[(S(k,i)\cdot\mathbf{a}(k))(S(k,i)\cdot\mathbf{a}(k))^H\right] \\ &= \sigma_S^2(k,i)\mathbf{a}(k)\mathbf{a}(k)^H\end{aligned} \tag{1.22}$$

with $\sigma_S^2(k,i) = \mathbb{E}[|S(k,i)|^2]$ denoting the power spectral density (PSD) of the speech signal $S(k,i)$. As the speech correlation matrix is given by the outer product of the ATF vector scaled by the speech PSD and, therefore, has rank one, the model in (1.4) is directly linked to the so-called rank-1 assumption. From an estimate of the speech correlation matrix, an estimate of ATF can be obtained by selecting the principal component of the matrix, i.e., the eigenvector corresponding to the largest eigenvalue

$$\hat{\mathbf{a}}(k) = \mathcal{P}\{\hat{\mathbf{\Phi}}_X(k,i)\}. \tag{1.23}$$

However, the scaling of the estimate obtained in (1.23) is ambiguous. This can be resolved by dividing all entries of the ATF estimate by the entry of a selected reference microphone to obtain an estimate of the RTF, i.e.,

$$\hat{\bar{\mathbf{a}}}(k) = \left[1, \frac{\hat{a}^2(k)}{\hat{a}^1(k)}, ..., \frac{\hat{a}^C(k)}{\hat{a}^1(k)}\right]^T. \tag{1.24}$$

Here, the first microphone has been selected as a reference, and $\hat{a}^1(k)$ denotes the first entry of the vector $\hat{\mathbf{a}}(k)$.

Equation (1.23) requires an estimate of the speech correlation matrix. For this, [68], [69] propose to estimate a speech mask which is applied analogously to the noise mask in (1.19). However, for stationary noise, a high SPP value, i.e., $\rho(k, i) \approx 1$, does not mean that only the speech component is observed. As a result, it may be difficult to identify time-frequency bins for which $\mathbf{Y}(k, i) = \mathbf{X}(k, i)$. However, assuming independence of the speech and noise signal, an estimate of the speech correlation matrix can also be obtained via covariance subtraction as [23]

$$\hat{\mathbf{\Phi}}_X(k, i) = \hat{\mathbf{\Phi}}_Y(k, i) - \hat{\mathbf{\Phi}}_N(k, i). \tag{1.25}$$

Care must be taken to ensure that the resulting estimate is a positive semi-definite matrix, which is not necessarily the case if two correlation matrix estimates are subtracted.

An alternative approach to estimating the RTF respectively is to pick the eigenvector $\mathbf{v}(k)$ corresponding to the largest eigenvalue $\lambda(k)$ of the generalized eigenvalue problem

$$\hat{\mathbf{\Phi}}_Y(k, i)\mathbf{v}(k) = \lambda(k)\hat{\mathbf{\Phi}}_N(k, i)\mathbf{v}(k). \tag{1.26}$$

and again use the reference microphone's entry for normalization [52]. In [69], the authors propose to replace the estimate of the noisy correlation matrix $\hat{\mathbf{\Phi}}_Y(k, i)$ by a mask-based estimate of the speech correlation matrix $\hat{\mathbf{\Phi}}_X$.

1.4 Deep Learning Techniques for Spatial Filtering

In the field of single-channel speech enhancement and separation, virtually all high-performing solutions developed today are based on deep learning. In contrast, integrating deep learning methods in multi-channel approaches is a very active field of research to which this work contributes. The research landscape has evolved considerably since the beginning of this research project. While the earlier approaches applied DNNs for estimating the parameters of a linear beamformer and investigated the enrichment of single-channel networks with spatial features, most recent approaches explore end-to-end systems that can be viewed as an instance of a joint spatial and tempo-spectral non-linear filter. This section provides an overview of the current developments in applying DNNs for spatial filtering in multi-channel speech processing tasks.

1.4.1 Integration of DNNs and Traditional Linear Spatial Filters

This section describes approaches that integrate traditional linear spatial filtering with DNNs for multi-channel speech enhancement or separation. This branch of research emerged around 2015 in the wake of the much-noticed publications by Heymann et al. [42], [68] and Erdogan et al. [43], who proposed to estimate the speech and noise masks for a mask-based beamformer using a neural network. Figure 1.6 shows a schematic illustration of their proposed scheme. The second and third block in the diagram illustrate the general mask-based beamforming technique discussed in the previous Section 1.3.1. However, traditional signal processing techniques for SPP estimation are now replaced with a neural network as indicated by the first block. In [42], [43], [68], the core component of the neural network is a long short-term memory (LSTM) layer, which is fed with data from only a single channel, using either the raw STFT magnitudes as input or additionally applying a Mel-filterbank and logarithm. For the choice of training target and loss, multiple variants are proposed:

- The network outputs a mask with elements in the range $[0, 1]$ for speech and noise [42],

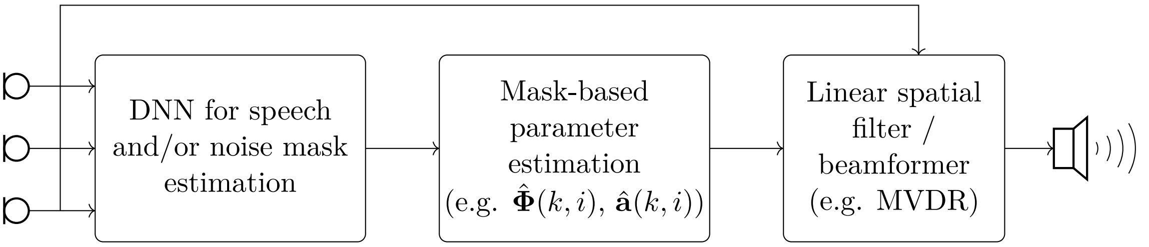

DNN for speech and/or noise mask estimation
Mask-based parameter estimation (e.g. $\hat{\mathbf{\Phi}}(k,i)$, $\hat{\mathbf{a}}(k,i)$)
Linear spatial filter / beamformer (e.g. MVDR)

Figure 1.6: Illustration of mask-based beamforming using a DNN for mask estimation.

[68]. It is trained with a binary cross-entropy loss between the mask output and an ideal binary mask (IBM) computed from clean speech and pure noise data.

- The network outputs a speech mask in the range $[0, 1]$ and is trained according to a mean square error (MSE) loss between the output mask and an ideal ratio mask (IRM) [43]. The noise mask is obtained by subtracting the speech mask from one elementwise.
- The network outputs a speech mask in the range $[0, 1]$, which is applied to one of the noisy channels to obtain a single-channel clean speech estimate. The network is then trained using the MSE loss between the magnitudes of the clean speech signal and the speech estimate [43].

For all approaches, the DNN is applied to each channel separately during inference so that C masks are obtained, which are combined by computing the mean or median. Since the DNN processes only one channel at a time, only tempo-spectral signal characteristics are used for mask estimation, but not spatial characteristics.

From the third configuration listed above, it becomes clear that the mask obtained with the DNN can also directly be applied to the signal. Spectral masking is a very common technique in single-channel enhancement [2], [34]. While pioneering works [70], [71] used DNNs to estimate a real-valued mask in the range $[0, 1]$ similar to the gain of the single-channel Wiener filter [72], recent spectral masking approaches predominantly target a complex ideal ratio mask (CIRM) [73] which is not limited to magnitude enhancement but also alters the phase. Single-channel masking, however, often introduces speech distortions, which are unpleasant for the listener and can heavily degrade ASR performance [74]. Linear spatial filtering, on the other hand, avoids these distortions, which is why it can be advantageous to apply the mask-based beamforming approach as a front-end for ASR. The noise suppression, however, is often not sufficient so that the mask-based beamformer may be combined with a (DNN-based) post-filter for speech enhancement.

The third configuration above furthermore highlights the general fact that the estimated quantity, which is the mask, does not necessarily need to be the quantity that the loss function is defined on, which is the estimated clean speech signal's magnitude in the example above. Rather, it is sufficient that the enhanced signal is the output of a function that involves the mask and through which the gradients can be back-propagated. Clearly, this is the case for the multiplicative application of the mask to the reference channel of the noisy signal. For an ASR system, neither a loss on the estimated mask itself nor a loss on the enhanced signal ensures that the output of the beamformer is a particularly good starting point for ASR. Therefore, it has been proposed to train the mask estimation with an ASR loss [75], [76], which requires back-propagation through the beamformer [77].

It has already been pointed out that no spatial information is used during mask estimation

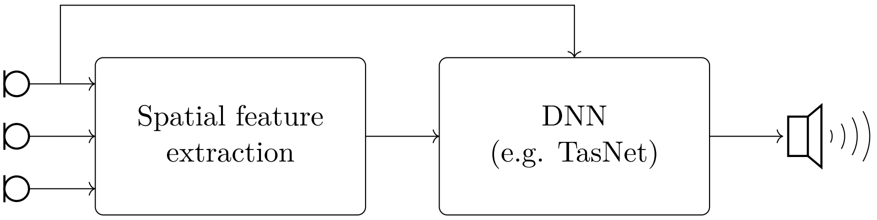


Figure 1.7: Illustration of a spatial feature based approach.

in the initial approaches. Many following works address this limitation [45], [78]–[81]. For example, Masyuama et al. [78] propose to use the multi-channel Itakura-Saito divergence as a loss function, which compares the similarity of the mask-based estimate of the correlation matrices to an oracle estimate of the respective correlation matrix. Nakatani et al. [45] and Zhou et al. [79] suggest combining the mask estimation with spatial clustering with the primary goal of achieving better performance in cases where there is a mismatch between training and testing data. Liu et al. [80] and Yoshioka et al. [81] propose to enrich the input of the mask estimation DNN by appending features that are based on the IPDs.

In order to use the mask-based beamforming approach not only for speech enhancement but also for speech separation, it has been proposed to use a DNN-based speech separation system to create a mask for every speaker or to compute correlation matrices directly from the separation outputs, e.g., [81]–[84]. Furthermore, [85], [86] have addressed scenarios that require a time-varying correlation matrix estimate.

Certainly, the mask-based beamforming approach is the most commonly used method that integrates DNNs with traditional beamforming. However, also other methods exist. One of the early proposals by Xiao et al. [87] was to estimate the complex-valued beamformer weights directly with a DNN, which they provide with generalized cross-correlation (GCC) information as input. Furthermore, there is the idea of using a set of fixed data-independent beamformers steered to a pre-defined selection of look directions for spatial pre-processing, e.g., [47]–[49]. For example, the delay-and-sum beamformer or the MVDR with a noise correlation matrix representing a diffuse noise field [54] can be employed to obtain features that are then further processed with a neural network. The work by Sainath et al. [88], [89] joins both of these ideas: A filter-and-sum structure is implemented using convolutional layers such that a set of linear filters can be learned in a data-driven way. During inference, these linear filters are fixed and provide features that are further processed and finally fed into an acoustic model for ASR.

1.4.2 Spatial Features

Inspired by the success of single-channel DNN-based speech enhancement, another research approach for integrating multi-channel information has been developed from around 2015 in parallel with the DNN-supported mask-based beamforming approach. In this line of research, spatial features are used as additional input to a DNN besides the single-channel noisy input, e.g., [90]–[93]. A schematic view of this approach is shown in Figure 1.7. An early work investigating the benefit of spatial features introduced in this way is by Araki et al. [90]. The authors of this work propose to compute the following features in the feature extraction step depicted in the left block of Figure 1.7: the ILDs, the IPDs and masks obtained by IPD clustering. These spatial features are provided to the network by stacking the single-channel signal feature and the spatial features to obtain the new input feature for the DNN.

For an observed multi-channel mixture $\mathbf{Y}(k,i)$ the ILD and IPD between microphones ℓ and m can be defined as time-varying quantities [34, Sec. 12.1]:

$$\mathrm{ILD}_{\ell m}(k,i) = 10\,\log_{10}\left(\frac{|Y^{\ell}(k,i)|^2}{|Y^{m}(k,i)|^2}\right) \tag{1.27}$$

and

$$\mathrm{IPD}_{\ell m}(k,i) = \angle\frac{Y^{\ell}(k,i)}{Y^{m}(k,i)}. \tag{1.28}$$

This is different from the time-invariant IPD and ILD between each channel ℓ and the reference channel encoded by the RTF [51], i.e.,

$$\mathrm{ILD}_{\ell}(k) = 10\,\log_{10}\left(\left|\tilde{a}^{\ell}(k)\right|^2\right) \tag{1.29}$$

and

$$\mathrm{IPD}_{\ell}(k) = \angle\tilde{a}^{\ell}(k), \tag{1.30}$$

which only depend on the placement of the microphone array and the source in the room. In contrast, the time-varying ILDs and IPDs in (1.27) and (1.28) computed from the noisy observation might be heavily influenced by the interfering signal. However, for time-frequency bins that are dominated by the target source signal, the time-varying ILD and IPD will approximately equal the respective time-invariant ILD and IPD, so that the resulting pattern can be used for mask estimation via clustering and is expected to also be informative to a DNN.

In [92], Wang et al. propose to extend the single-channel deep clustering [16] approach with additional spatial features. Besides IPDs the authors suggest using GCC features computed as

$$\mathrm{GCC}(k,i,\ell,m,\tau) = \cos\left(\angle\frac{Y^{\ell}(k,i)}{Y^{m}(k,i)} - 2\pi f_k\tau\right) \tag{1.31}$$

with τ denoting a candidate value for the TDOA between the microphone pair (ℓ,m). The GCC function is then evaluated for a range of plausible TDOA values, which results in a feature vector for each time-frequency bin, which can be stacked with the single-channel noisy input feature. Another example of a spatial feature is the so-called angle or location-guided feature [47], [93]. For this feature, it is assumed that the DOA of the target speaker and the array geometry are known, so that the steering vector $\mathbf{d}(k)$ can be computed. The angular feature $\mathrm{AF}(k,i)$ then measures the similarity between the relative steering vector and a vector containing the IPDs computed using the first as a reference microphone, i.e.,

$$\mathrm{AF}(k,i) = \Re\left\{\sum_{\ell=1}^{C}\frac{d^{\ell}(k)\exp(\mathrm{IPD}_{\ell 1}(k,i))}{|d^{\ell}(k)\exp(\mathrm{IPD}_{\ell 1}(k,i))|}\right\} \tag{1.32}$$

with $\Re$ denoting the real part of a complex number.

As expected, the additional spatial features are consistently found to lead to a notable performance improvement over the corresponding DNN, which takes only a single-channel noisy signal as input. However, since the features are hand-crafted, it is unclear if all the information that could potentially be exploited by a DNN trained on the raw inputs is still present in the computed features.

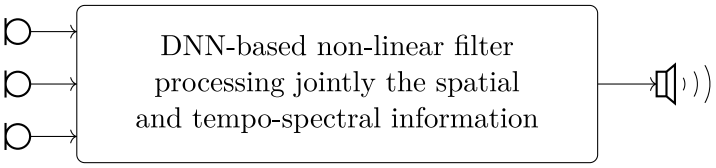

DNN-based non-linear filter processing jointly the spatial and tempo-spectral information

Figure 1.8: Illustration of a DNN implementing a non-linear filter that jointly processes spatial and tempo-spectral information.

1.4.3 DNN-based Joint Spatial and Tempo-spectral Non-linear Filtering

The schematic representation in Figure 1.2 shows a non-linear filter that combines spatial and tempo-spectral processing. As described in Section 1.3, an analytic filter with these properties can be obtained by a statistical derivation under a non-Gaussian noise distribution. On the other hand, this schematic illustration also fits a multitude of deep learning systems for multi-channel speech enhancement and separation, which have recently been proposed. Consequently, the schematic representation of these approaches in Figure 1.8 looks just like the one in Figure 1.2. It can be seen that all channels of the noisy recording on the left are directly used as input for the neural network, and the explicit spatial feature extraction stage is omitted. In most cases, the network expects either time-domain signal inputs, e.g., [94]–[97], or a frequency-domain representation obtained with the STFT, e.g., [59], [98]–[104].

As the DNN depicted in Figure 1.8 is provided with the raw multi-channel inputs, it can, in principle, exploit spatial as well as tempo-spectral information to perform the enhancement or separation task. However, its ability to exploit a source of information depends on the specific choice of network architecture. Some of the first approaches matching the scheme shown in Figure 1.8 are from Chakrabarty et al. [98] and Li and Horaud [100] and have been published in 2018 and 2019, respectively. In both cases, the network expects a frequency-domain STFT input. Since the STFT coefficients are complex-valued, the question arises how to best represent them such that a DNN can exploit the patterns in the data. Chakrabarty et al. choose to stack the magnitude and phase information for each time-frequency bin. A chain of convolutional layers is then applied to each time-frequency bin independently to compute a feature representing cross-channel information. The cross-channel features for all frequencies of a single time step are then stacked and fed into a linear layer to estimate the mask weights for the respective time step. The design focus of the architecture is on the cross-channel spatial as well as spectral information, while temporal information cannot be exploited. In a successive publication [99], the authors propose to resolve this restriction by feeding the stacked feature vectors for each time step into an LSTM layer. In [100], Li and Horaud choose to represent the complex numbers by stacking the real and imaginary parts of all channels. Therefore, the multi-channel data is then stored in a three-dimensional tensor of shape $T \times F \times 2C$, where T and F denote the overall number of time-frames and frequency bins. The proposed architecture consists of two LSTM layers and is fed with slices of shape $T \times 2C$ to produce the mask weights. All frequency bins are processed independently with the same network, which means that the same network weights are used to estimate the mask weights for all frequency bins. Clearly, the focus of this architecture is on spatial and temporal information, while correlations between frequency bins are not considered.

Many model architectures have been proposed in recent years. Some of them are a direct extension of successful architectures for single-channel speech enhancement. For example, Pandey et al. [97] extend the dual-path network for speech separation [19]. Also other proposed architectures reflect the latest trends in network architecture design at the time of publication. For example, Tolooshams et al. [59] suggest an attention mechanism focusing on spatial information, Wang et al. [101] design a network based on an encoder-decoder U-net structure composed of dense blocks [105], Halimeh and Kellermann [103] propose a complex-valued neural network and some of the latest publications use a transformer-based architecture [106], [107]. How to design the network architecture for best performance is an open research question. While the design of the network architecture will continuously evolve to incorporate newly proposed layers, research on the best way to integrate the different sources of information might provide a more general and longer-lasting answer. Along this line of research, Briegleb et al. [108] investigate how a neural network represents spatial information internally, and the publication [P5] included in this thesis studies the interdependency between different sources of information.

Besides the choice of input representation and core network architecture, the output strategy, training target, and loss function are important design decisions. The vast majority of recent publications has converged to using the clean signal as training target in combination with a signal-based loss function, for example, the scale-invariant source to distortion ratio (SI-SDR) [109], an ℓ_1 loss in time and frequency domain [59], or a loss defined on the real and imaginary parts of the signal combined with a magnitude loss [101]. Considering the output strategy of the network, two prevailing solutions can be identified. The first option is that the network estimates a single mask, which is applied to a reference channel of the noisy signal to obtain the clean speech estimate. While older works proposed to use a real-valued mask [98]–[100], recent publications are using a complex-valued mask which enables phase enhancement, e.g., [59]. A similar approach is the so-called complex spectral mapping strategy, which directly estimates the real and imaginary parts of the enhanced speech signal [101]. The second output option is inspired by the filter-and-sum processing model of a traditional beamformer. Here, the network outputs filter coefficients for each channel, either in the time or frequency domain. These are then applied to the noisy signal, and a single-channel enhanced signal is obtained by summing all filtered signals. It is important to note the difference between a traditional linear beamformer and a network with a filter-and-sum output strategy: while the filter coefficients of the traditional beamformer are derived from statistical properties of the signal, e.g., correlation matrices, the filter coefficients estimated by the neural network directly depend on the noisy input signal. As a result, the filter-and-sum operation using these DNN-based filter coefficients cannot be considered linear with respect to the noisy input signal.

Also in Section 1.4.1 on the integration of DNNs and traditional beamformers, we described approaches that estimate the beamformer coefficients using a neuronal network, e.g., the work by Xiao et al. [87] and Sainath et al. [88], [89]. However, these are different from recent DNNs that employ a filter-and-sum output strategy regarding the linearity versus non-linearity of the processing model. In [87], Xiao et al. propose to estimate the filter coefficients from GCC features. Accordingly, the network sees as input not the noisy signal itself but a second-order statistic, and, even more importantly, the estimated filter coefficients are averaged for a whole utterance so that the filter coefficients will depend on the global statistical properties of the noisy signal but not vary based on the noisy observation made for a specific time-frequency bin. Similarly, the filter coefficients in [88], [89] are learned in a data-driven way, but they are fixed during inference and do not change based on the input signal. In contrast, filter coefficients in recent neural network architectures with a filter-and-sum output strategy, for example, the EaBNet [102], or FasNet [94], [95], directly depend on the noisy input signal. Therefore, we classify these as joint spatial and tempo-spectral non-linear filters even though a part of the processing model has a similarity with a traditional linear beamformer. Since these networks can exploit spatial and tempo-spectral information, a fair comparison with a traditional beamformer should always include a post-filtering stage for the beamformer.

Another proposed approach that should be mentioned in this context is the ADL-MVDR proposed by Zhang et al. [110], which seems difficult to categorize at first glance. The authors propose a neural network that consists of two parts. In the first part, a deep filter [111] given spectral information and spatial features is expected to provide estimates of the clean speech and noise signal. These are then used to compute correlation matrix features, which are further processed in a second network. At the output of the network, a vector and a matrix are estimated for each time-frequency bin. These are then inserted into the classic MVDR beamformer equation as RTF vector and inverse noise correlation matrix. According to our definition, this network is also a non-linear spatial filter. The decisive factor for this categorization is that the entire pipeline is trained end-to-end, so that it is by no means guaranteed that the outputs of the first part are clean speech and noise estimates. Accordingly, the correlation matrix features computed from these estimates might not have much in common with the second-order statistics that the traditional MVDR depends on. Furthermore, the formula that is used to estimate the correlation matrices for each time-frequency bin does not perform temporal averaging over neighboring frames. The dependence of the obtained correlation matrix on the specific observation is, therefore, much greater than in estimation schemes that attempt to realize the expected value, for example, with the averaging strategies described in Section 1.3.1.

1.5 Outline of the Thesis

This thesis deals with the topic of non-linear spatial filtering for multi-channel speech enhancement and separation. Three main areas of research can be identified to which this thesis contributes.

1. **Statistical Perspective on Non-linear Spatial Filtering**

   The observation that a non-Gaussian noise distribution leads to a non-linear MMSE filter that jointly processes spatial and spectral information serves as a starting point for the research presented in this thesis. In [6], only the computation result for the MMSE estimator is reported, and its interesting properties (non-linear and non-separable) are noted. In this work, we explore the implications of this result for multi-channel speech enhancement tasks. We aim to answer the question of whether we can expect a relevant performance gain from replacing the traditional linear beamformer plus post-filter with a non-linear spatial filter. Since a non-linear spatial filter has a higher computational demand than a traditional linear beamformer (and if implemented with a neural network, it even needs to go through a costly training stage), we aim to understand if the potential performance gain is worth the effort. Furthermore, we investigate in which application scenarios a non-linear spatial filter could be particularly beneficial and where the performance gain comes from.

   All analyses in this part are based on analytic statistical estimators. To obtain insights into the described research questions, we use the MMSE estimator derived under a Gaussian mixture noise assumption and compare it to an MMSE estimator that we derive under the same assumptions but with the additional constraint that it is composed of a MVDR beamformer and a post-filter. This way, our findings in this first part of the thesis do not depend on architectural choices or hyperparameter tuning for a neural network.

2. **Design and Analysis of Deep Non-linear Spatial Filters**

   Deep neural networks offer a data-driven way to implement a non-linear spatial filter. They can, therefore, be used to circumvent problems that are likely encountered when working with analytical estimators. One of these is that very accurate parameter estimates are required for the MMSE estimator in Gaussian mixture noise, which are difficult to obtain in practice. Another problem is that the estimation problems easily get intractable, for example, if no simplifying assumption like independence of time-frequency bins is made or other noise distributions are assumed.

   The second research focus of this thesis is on the internal functioning of DNN-based non-linear spatial filters. Here, we aim to understand the role of the two properties, which are the non-linearity and the interdependency between spatial and tempo-spectral processing. Is it rather the non-linearity of the processing model or the ability to exploit interdependencies between spatial and tempo-spectral processing that leads to good performance? In Section 1.4.3, we discussed two baseline approaches by Chakrabarty et al. [98], [99] and Li and Horaud [100] to illustrate how the chosen network architecture determines which sources of information (spatial, spectral, temporal) can be exploited by a neural network. We define a set of neural networks with the same underlying architecture and the same number of parameters but access to different sources of information. Based on this, we perform experiments that isolate the impact of different sources of information. Understanding the internal mechanisms of a DNN-based non-linear spatial filter is of great importance since this may provide general guidelines for designing well-performing network architectures for effective spatial filtering.

3. **Steerable Deep Non-linear Spatial Filters for Speech Extraction and Separation**

   The third part of the thesis puts a focus on practical applications, in particular multi-channel speaker extraction and separation. As described in Section 1.2, we adopt a perspective that views these tasks as a spatial filtering problem. Consequently, a possibility to control the steering direction of the deep non-linear spatial filter is required so that a separate DNN-based non-linear spatial filter can be steered in the direction of each target source. For many linear spatial filters, e.g., the delay-and-sum or MVDR beamformer, the look direction is controlled by the steering vector or RTF estimate. However, these may be difficult to obtain and can be inaccurate. For the deep non-linear spatial filters, we are proposing to condition the filter directly on a discretized DOA angle. We then investigate the spatial selectivity of the steered filter and analyze the performance of this spatial filtering focused approach. As a baseline, we are comparing to a DNN that produces an output for every speaker and is trained with a PIT scheme. However, spatial filtering is not directly enforced by the loss function and must implicitly be learned from the training examples. Thus, one of the core research questions addressed in the third part of this thesis is whether speech separation performance and also robustness and generalization ability can be improved by focusing on learning good non-linear spatial filters.

The following Chapter 2 links each of the seven publications related to this thesis to one of the main research areas outlined above and explains the contribution of each co-author. The main part of this cumulative thesis contains three publications, one for every research area, exploring the different aspects of non-linear spatial filters. The thesis concludes with a discussion of the main contributions and an outlook on future research directions in Chapter 6.

2 Overview of the Related Publications

We group the publications related to this thesis with respect to the three main research areas described in Section 1.5. The journal article associated with each thematic area extends the corresponding conference publication(s), which means that these publications have overlapping content. Therefore, we choose to include only the journal publications in the main part of this cumulative thesis. *The papers included in the main part of the cumulative thesis are marked with a gray box in the following list of publications.* All other publications can be found in Appendix A.

1. Statistical Perspective on Non-linear Spatial Filters

[P1] K. Tesch, R. Rehr, and T. Gerkmann, "On Nonlinear Spatial Filtering in Multichannel Speech Enhancement", in *Proceedings of Interspeech*, Graz, Austria, 2019, pp. 91–95.

[P2] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering for multichannel speech enhancement in inhomogeneous noise fields", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 196–200.

[P3] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1795–1805, 2021.

2. Design and Analysis of Deep Non-linear Spatial Filters

[P4] K. Tesch, N.-H. Mohrmann, and T. Gerkmann, "On the Role of Spatial, Spectral, and Temporal Processing for DNN-based Non-linear Multi-channel Speech Enhancement", in *Proceedings of Interspeech*, Seoul, South Korea, 2022, pp. 2908–2912.

[P5] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 563–575, 2023.

3. Steerable Deep Non-linear Spatial Filters for Speech Extraction and Separation

[P6] K. Tesch and T. Gerkmann, "Spatially selective deep non-linear filters for speaker extraction", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhode Island, Greece, 2023.

[P7] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 32, pp. 542–553, 2024.

2.1 Valuation of the Author's Contribution

The author of this thesis, Kristina Tesch, is the first author of all seven previously listed publications. It is the case for all publications that she performed the mathematical derivations, developed and implemented the algorithms, conducted the experiments, performed the evaluations, and wrote the main manuscript texts. Timo Gerkmann contributed to the design of the research projects by providing valuable feedback on intermediate results and ideas, and he reviewed the final manuscripts.

For publication [P1], Robert Rehr gave advice regarding the implementation details of the method and provided feedback on the evaluation strategy.

Niels-Hendrik Mohrmann is the second co-author of the publication [P4]. In his role as a student assistant, he helped to prepare datasets for training the models, and he helped to conduct the evaluations. Furthermore, he reviewed the final manuscript.

Publication 1: Nonlinear Spatial Filtering in Multichannel Speech Enhancement [P3]

3

Abstract

The majority of multichannel speech enhancement algorithms are two-step procedures that first apply a linear spatial filter, a so-called beamformer, and combine it with a single-channel approach for postprocessing. However, the serial concatenation of a linear spatial filter and a postfilter is not generally optimal in the minimum mean square error (MMSE) sense for noise distributions other than a Gaussian distribution. Rather, the MMSE optimal filter is a *joint spatial and spectral nonlinear* function. While estimating the parameters of such a filter with traditional methods is challenging, modern neural networks may provide an efficient way to learn the nonlinear function directly from data. To see if further research in this direction is worthwhile, in this work we examine the potential performance benefit of replacing the common two-step procedure with a joint spatial and spectral nonlinear filter.

We analyze three different forms of non-Gaussianity: First, we evaluate on super-Gaussian noise with a high kurtosis. Second, we evaluate on inhomogeneous noise fields created by five interfering sources using two microphones, and third, we evaluate on real-world recordings from the CHiME3 database. In all scenarios, considerable improvements may be obtained. Most prominently, our analyses show that a nonlinear spatial filter uses the available spatial information more effectively than a linear spatial filter as it is capable of suppressing more than $D-1$ directional interfering sources with a D-dimensional microphone array without spatial adaptation.

Reference

Kristina Tesch and Timo Gerkmann, "Nonlinear Spatial Filtering in Multichannel Speech Enhancement", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1795-1805, 2021. DOI: 10.1109/TASLP.2021.3076372

This publication received the **VDE ITG 2022 award**.

Copyright notice

Nonlinear Spatial Filtering in Multichannel Speech Enhancement

Kristina Tesch, *Student Member, IEEE*, and Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—The majority of multichannel speech enhancement algorithms are two-step procedures that first apply a linear spatial filter, a so-called beamformer, and combine it with a single-channel approach for postprocessing. However, the serial concatenation of a linear spatial filter and a postfilter is not generally optimal in the minimum mean square error (MMSE) sense for noise distributions other than a Gaussian distribution. Rather, the MMSE optimal filter is a *joint spatial and spectral nonlinear* function. While estimating the parameters of such a filter with traditional methods is challenging, modern neural networks may provide an efficient way to learn the nonlinear function directly from data. To see if further research in this direction is worthwhile, in this work we examine the potential performance benefit of replacing the common two-step procedure with a joint spatial and spectral nonlinear filter.**

We analyze three different forms of non-Gaussianity: First, we evaluate on super-Gaussian noise with a high kurtosis. Second, we evaluate on inhomogeneous noise fields created by five interfering sources using two microphones, and third, we evaluate on real-world recordings from the CHiME3 database. In all scenarios, considerable improvements may be obtained. Most prominently, our analyses show that a nonlinear spatial filter uses the available spatial information more effectively than a linear spatial filter as it is capable of suppressing more than $D-1$ directional interfering sources with a D-dimensional microphone array without spatial adaptation.

***Index Terms*—Multichannel, speech enhancement, nonlinear spatial filtering, neural networks**

I. Introduction

In our everyday life, we are surrounded by background noise for example traffic noise or competing speakers. Hence, speech signals that are recorded in real environments are often corrupted by noise. Speech enhancement algorithms are employed to recover the target signal from a noisy recording. This is done by suppressing the background noise or reducing other unwanted effects such as reverberation. This way, speech enhancement algorithms aim to improve speech quality and intelligibility. Their fields of application are manifold and range from assisted listening devices to telecommunication all the way to automatic speech recognition (ASR) front-ends [1], [2].

If the noisy speech signal is captured by a microphone array instead of just a single microphone, then not only tempo-spectral properties can be used to extract the target signal but also spatial information. Spatial filtering aims at suppressing signal components from other than the target direction. The filter-and-sum beamforming approach [3, Sec. 12.4.2] achieves this by filtering the individual microphone signals and adding them. In the frequency domain, this means to compute the scalar product between a complex weight vector and the vector of spectral representations of the multichannel noisy signal. Hence, the beamforming operation is linear with respect to the noisy input.

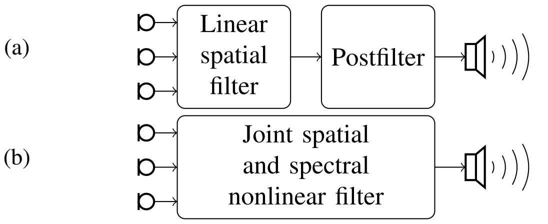


Fig. 1: (a) Illustration of the commonly employed two-step processing using a linear spatial filter (beamformer) followed by a single-channel postfilter. (b) Illustration of the nonlinear spatial filter investigated in this paper, which joins the spatial and spectral processing into a non-separable nonlinear operation.

The beamforming weights are chosen to optimize some performance measure. For example, minimizing the noise variance subject to a distortionless constraint leads to the well-known minimum variance distortionless response (MVDR) beamformer [4, Sec. 3.6]. The noise suppression capability of such a spatial filter alone is often not sufficient and a single-channel filter is applied to the output of the spatial filter to improve the speech enhancement performance. The second processing stage in this two-step processing scheme is often referred to as the postfiltering step.

Single-channel speech enhancement has a long research history that has led to a variety of solutions like the classic single-channel Wiener filter [3, Sec 11.4] or other estimators derived in a statistical framework [5]–[7]. Many recent advances in single-channel speech enhancement are driven by the modeling capabilities of deep neural networks (DNNs) [8]–[11].

It seems convenient to independently develop a spatial filter and a postfilter and combine them into a two-step procedure afterward as shown in Figure 1a. If the noise follows a Gaussian distribution, this approach can even be regarded as optimal in the MMSE sense as Balan and Rosca [12] have shown that the MMSE solution can always be separated into the linear MVDR beamformer and a postfilter. However, this separability into a linear spatial filter and a postfilter only holds under the restrictive assumption that the noise is Gaussian distributed. The work of Hendriks et al. [13] points out that the MMSE optimal solution for non-Gaussian noise joins the spatial and spectral processing into a single nonlinear operation. Throughout this

The authors are with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany (e-mail: kristina.tesch@uni-hamburg.de; timo.gerkmann@uni-hamburg.de).

work, we call such an approach a *nonlinear spatial filter* for brevity even though spectral processing steps are also included. An illustration is given in Figure 1b.

The result of Hendriks et al. reveals that the common two-step multichannel processing scheme cannot be considered optimal for more general noise distributions than a Gaussian distribution. This leads to the question if we should invest in the development of nonlinear spatial filters for example using DNNs. Today, single-channel approaches often use the possibilities of DNNs to learn complex nonlinear estimators directly from data. In contrast, the field of multichannel speech enhancement is dominated by approaches that use DNNs only for parameter estimation of a beamformer [14], [15] or restrict the network architecture in a way that a linear spatial processing model is preserved [16]. Only a few approaches with and without DNNs [17]–[19] have been proposed that extend the spatial processing model to be nonlinear. Still, the questions of how much we can possibly gain by doing this, in which situations, and also where the benefit of using a nonlinear spatial filter comes from have not been addressed adequately. These are the questions that we aim to investigate in this paper.

This work is based on a previous conference publication [20]. In [21] we have studied related aspects of these questions. Here, we extend our previous work by more detailed derivations and new analyses that provide some insight into the functioning of the nonlinear spatial filter. In Section III, we provide a detailed overview of the theoretical results from a statistical perspective. We include the previously outlined results and also provide a new simplified proof for the finding of Balan and Rosca in [12]. We then evaluate the performance benefit of a nonlinear spatial filter for heavy-tailed noise in Section IV-A, for an inhomogeneous noise field created by five interfering human speakers in Section IV-B, and real-world noise recordings in Section IV-C. In Section V, we investigate the improved exploitation of spatial information by the nonlinear spatial filter and discuss practical issues of the used analytic nonlinear spatial filter. Even though nonlinear spatial filters would most likely be implemented using DNNs in the future, in our analyses we rely on statistical MMSE estimators to provide more general insights than by using DNN-based nonlinear spatial filters which would be highly dependent on the network architecture and training data.

II. Assumptions and notation

We assume that the signals recorded by a D-dimensional microphone array decompose into a target speech and a noise component. For each microphone-channel $\ell \in \{1, ..., D\}$, we segment the time-domain signal into overlapping windows and transform the signal to the frequency domain using the discrete Fourier transform (DFT) to obtain the DFT coefficients $Y_\ell(k,i)$ with frequency-bin index k and time-frame index i. Throughout this work, we use segments of length 32 ms with 16 ms shift and apply the square-root Hann function for spectral analysis and synthesis. By the additive signal model, the noisy DFT coefficient can be written as the sum of the clean speech and the noise DFT coefficients $S_\ell(k,i) \in \mathbb{C}$ and $N_\ell(k,i) \in \mathbb{C}$, i.e.,

$$Y_\ell(k,i) = S_\ell(k,i) + N_\ell(k,i). \tag{1}$$

As we model the DFT coefficients to be random variables and assume independence with respect to the frequency-bin and time-frame index, we drop the indices (k,i) to simplify the notation. We indicate random variables with uppercase letters and use lowercase letters for their respective realization. Furthermore, we assume all DFT coefficients to be zero-mean and speech and noise to be uncorrelated.

We stack the noisy and noise DFT coefficients into vectors $\mathbf{Y} = [Y_1, Y_2, ..., Y_\ell]^T \in \mathbb{C}^D$ and $\mathbf{N} = [N_1, N_2, ..., N_\ell]^T \in \mathbb{C}^D$and obtain the vector of speech DFT coefficients $\mathbf{S} \in \mathbb{C}^D$ by multiplying the clean speech signal coefficient $S \in \mathbb{C}$ with the so-called steering vector $\mathbf{d} \in \mathbb{C}^D$, which accounts for the propagation path between the target speaker and the microphones. We can then rewrite the signal model as

$$\mathbf{Y} = \mathbf{d}S + \mathbf{N}. \tag{2}$$

The noise correlation matrix is denoted by $\mathbf{\Phi}_n = \mathbb{E}[\mathbf{N}\mathbf{N}^H] \in \mathbb{C}^{D\times D}$ with the statistical expectation operator $\mathbb{E}$ and $(\cdot)^H$ denoting the Hermitian transpose. The spectral power of the target speech signal is given by $\sigma_s^2 = \mathbb{E}[|S|^2] \in \mathbb{R}^+$. When appropriate, we use the polar representation for complex-valued quantities, e.g., $s = |s|e^{j\varphi_s} \in \mathbb{C}$, and then let φ denote the phase of the complex number.

III. Linearity of the optimal spatial filter

In this section, we aim to provide a more complete picture of the nature of the optimal spatial filter by aggregating existing results and presenting more straightforward derivations for some of these. We identify the noise distribution as the key to linearity versus non-linearity of the spatial filter and also to the separability of spatial and spectral processing. Accordingly, in our considerations we distinguish the two cases of Gaussian distributed noise and non-Gaussian distributed noise or, more precisely, noise that follows a Gaussian mixture distribution.

A. Gaussian Noise

We start with revisiting the results from Balan and Rosca [12] and then provide a simplified proof that may be easier to follow. We assume that the vector of noise DFT coefficients $\mathbf{N}$ follows a multivariate complex Gaussian distribution with zero mean and covariance matrix $\mathbf{\Phi}_n$, i.e., $\mathbf{N} \sim \mathcal{CN}(0, \mathbf{\Phi}_n)$. As we employ an additive signal model, the conditional distribution of the noisy DFT coefficient vector $\mathbf{Y}$ given information on the reference clean speech DFT is a multivariate complex Gaussian distribution centered around the vector of clean speech DFT coefficients $\mathbf{d}s$ with the same covariance matrix $\mathbf{\Phi}_n$. The corresponding conditional probability density function (PDF) is given by [22, Thm. 15.1]

$$p_{\mathbf{Y}}(\mathbf{y}|s) = \frac{1}{\pi^D|\mathbf{\Phi}_n|}\exp\left\{-(\mathbf{y}-\mathbf{d}s)^H\mathbf{\Phi}_n^{-1}(\mathbf{y}-\mathbf{d}s)\right\}. \tag{3}$$

Our goal is to show that the linear MVDR beamformer defined as

$$T_{\text{MVDR}}(y) = \frac{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{y}}{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d}} \tag{4}$$

is the optimal spatial filter with respect to the maximum a posteriori (MAP), MMSE and maximum likelihood (ML) optimization criterion if the noise follows a Gaussian distribution.

Balan and Rosca [12] rely on the concept of sufficient statistics to prove the property in question for the MMSE optimization criterion. In our context, the MVDR beamformer T_{MVDR} is a *sufficient statistic in the Bayesian sense* if

$$p_S(s|\mathbf{y}) = p_S(s|T_{\mathrm{MVDR}}(\mathbf{y})) \tag{5}$$

holds for every observation $\mathbf{y}$ and every prior distribution of S [23, Thm. 2.4]. We infer from (5) that all information about S contained in the noisy observation is retained in the output of the MVDR beamformer despite the fact that the MVDR beamformer reduces the dimension of the multidimensional input to one dimension. Note that the variable of interest S in the above definition is a random variable. In contrast, T_{MVDR} is a *sufficient statistic in the classical sense* for the true clean speech DFT coefficient s, which is not assumed to be a random variable, if the conditional distribution of the noisy observation $\mathbf{Y}$ given $T_{\mathrm{MVDR}}(\mathbf{Y})$ does not depend on s [24, Def. IV.C.1].

As a first step, Balan and Rosca deduce that the MVDR beamformer is a sufficient statistic in the classical sense from the Fisher-Neyman factorization theorem [24, Prop. IV.C.1] [25, Cor. 2.6.1], which is applicable since the conditional PDF of the observation $\mathbf{Y}$ given S in (3) can be rewritten as

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}|s) &= \underbrace{\frac{1}{\pi^D|\mathbf{\Phi}_n|}\exp\{-\mathbf{y}^H\mathbf{\Phi}_n^{-1}\mathbf{y}\}}_{h(\mathbf{y})} \\ &\quad \underbrace{\times\exp\left\{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d}\left(2\,\mathrm{Re}\left\{s^*T_{\mathrm{MVDR}}(\mathbf{y})\right\}-|s|^2\right)\right\}}_{g(s,T_{\mathrm{MVDR}}(\mathbf{y}))} \\ &= h(\mathbf{y})g(s,T_{\mathrm{MVDR}}(\mathbf{y})) \\ &= h(\mathbf{y})g(s,z). \end{aligned} \tag{6}$$

under the Gaussian noise assumption. In the last line of equation (6), we replaced the random variable $T_{\mathrm{MVDR}}(\mathbf{Y})$ with Z, i.e.,

$$Z = T_{\mathrm{MVDR}}(\mathbf{Y}), \tag{7}$$

and will now continue to use this substitute when it improves the readability. In a second step, Balan and Rosca conclude that the MVDR beamformer is a sufficient statistic in the Bayesian sense because any statistic that is sufficient in the classical sense is also sufficient in the Bayesian sense [23, Thm. 2.14.2].

We now provide a proof of the T_{MVDR} being a sufficient statistic of S in the Bayesian sense, which does not require a reference to advanced stochastic theorems. For this, we compute a factorization of the likelihood PDF of Z $p_Z(z|s)$ with Z defined in (7) as the output of the MVDR beamformer for the noisy input $\mathbf{Y}$. From the properties of the multivariate complex Gaussian distribution undergoing a linear transformation [22, Appx. 15B], we infer that Z given S is distributed according to a one-dimensional complex Gaussian distribution with mean s and variance $(\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d})^{-1}$, i.e.,

$$p_Z(z|s) = \mathcal{CN}\left(s,(\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d})^{-1}\right). \tag{8}$$

The corresponding PDF at the output of the beamformer can be factorized as

$$\begin{aligned} p_Z(z|s) &= \underbrace{\frac{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d}}{\pi}\exp\{-\mathbf{y}^H\mathbf{\Phi}_n^{-1}\mathbf{y}\,|z|\}}_{f(\mathbf{y})} \\ &\quad \underbrace{\times\exp\left\{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d}\left(2\,\mathrm{Re}\left\{s^*z\right\}-|s|^2\right)\right\}}_{g(s,z)} \\ &= f(\mathbf{y})g(s,z). \end{aligned} \tag{9}$$

Using (6) we rewrite the posterior distribution as

$$\begin{aligned} p_S(s|\mathbf{y}) &= \frac{p(\mathbf{y}|s)p(s)}{\int_{\mathbb{C}}p(\mathbf{y}|s)p(s)ds} \\ &= \frac{h(\mathbf{y})g(s,z)p(s)}{\int_{\mathbb{C}}h(\mathbf{y})g(s,z)p(s)ds}. \end{aligned} \tag{10}$$

Since the term $h(\mathbf{y})$ in the denominator does not depend on the integration variable s, this term cancels with the corresponding term in the numerator. Next, we extend the fraction with the term $f(\mathbf{y})$ from (9) to obtain

$$\begin{aligned} p_S(s|\mathbf{y}) &= \frac{f(\mathbf{y})g(s,z)p(s)}{\int_{\mathbb{C}}f(\mathbf{y})g(s,z)p(s)ds} \\ &= p_S(s|z) \\ &= p_S(s|T_{\mathrm{MVDR}}(\mathbf{y})), \end{aligned} \tag{11}$$

which is the identity we wanted to prove (cf. (5)). Consequently, as the posterior given the noisy observation $\mathbf{y}$ equals the posterior given the output of the MVDR beamformer $T_{\mathrm{MVDR}}(\mathbf{y})$, we find that

$$T_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_{s\in\mathbb{C}}\ p_S(s|T_{\mathrm{MVDR}}(\mathbf{y})) \tag{12}$$

holds. The MVDR beamformer reduces its multidimensional input to a single-channel output and, therefore, the right-hand side of (12) can be seen as a single-channel postfilter working on the output of the MVDR beamformer. Since the MMSE estimator complies with the mean of the posterior, a similar decomposition in a linear spatial filter and a spectral postfilter is given by

$$\begin{aligned} T_{\mathrm{MMSE}}(\mathbf{y}) &= \mathbb{E}[S|\mathbf{y}] \\ &= \mathbb{E}[S|T_{\mathrm{MVDR}}(\mathbf{y})]. \end{aligned} \tag{13}$$

Because the relationship (5) holds for all prior distributions of S, a decomposition of the MAP and MMSE estimators into a linear spatial filter followed by a postfilter exist independently from any further assumptions regarding the prior distribution of the clean speech DFT coefficient.

Finally, we consider the ML estimator. Starting from (6) and exploiting the monotony of the logarithm and Euler's formula, we find the representation

$$\begin{aligned} T_{\mathrm{ML}}(\mathbf{y}) &= \arg\max_{s\in\mathbb{C}}\ p_{\mathbf{Y}}(\mathbf{y}|s) \\ &= \arg\max_{s\in\mathbb{C}}\ 2\,\mathrm{Re}\{s^*\underbrace{T_{\mathrm{MVDR}}(\mathbf{y})}_{=z}\}-|s|^2 \\ &= \arg\max_{s\in\mathbb{C}}\ 2\cdot|s|\cdot|z|\cdot\cos(\varphi_z-\varphi_s)-|s|^2. \end{aligned} \tag{14}$$

Clearly, this function is maximized when the phase of s matches the phase of $T_{\text{MVDR}}(\mathbf{y})$, as then the cosine function is maximized. Equating the derivative with respect to $|s|$ to zero and solving for $|s|$ reveals that the magnitude of the MVDR beamformer maximizes the likelihood. Thus, $T_{\text{ML}}(\mathbf{y}) = T_{\text{MVDR}}(\mathbf{y})$, i.e. the MVDR beamformer is the maximum likelihood estimator of the clean speech DFT coefficient as also stated in [26, Sec. 6.2.1.2].

B. Non-Gaussian Noise

As we have seen, if the noise DFT coefficients follow a Gaussian distribution, then a linear spatial filter can be considered optimal. However, Hendriks et al. [13] have shown that this does not need to be the case for non-Gaussian distributed noise. In their work, they model the noise distribution with a multivariate complex Gaussian mixture distribution. The M Gaussian mixture components with respective covariance matrix $\mathbf{\Phi}_m$, $m \in \{1, ..., M\}$, are assumed to be zero-mean such that the conditional PDF given the clean speech is given by

$$p_{\mathbf{Y}}(\mathbf{y}|s) = \sum_{m=1}^{M} c_m \mathcal{CN}(\mathbf{d}s, \mathbf{\Phi}_m). \tag{15}$$

with mixture weights c_m that sum to one. Hendriks et al. assume that the amplitude A_S and phase φ_S of the clean speech DFT coefficient are independent. They model the phase to be uniformly distributed over the interval $[0, 2\pi)$ and assume the amplitude to be generalized-Gamma distributed ([7, Eq. 1], with $\gamma = 2$ and $\beta = \nu/\sigma_s^2$). The corresponding PDF

$$p_{A_S}(a) = 2\frac{\left(\frac{\nu}{\sigma_s^2}\right)^{\nu}}{\Gamma(\nu)} a^{2\nu-1} \exp\left\{-\frac{\nu}{\sigma_s^2}a^2\right\} \text{ with } \nu > 0,\ a \geq 0 \tag{16}$$

depends on the speech shape parameter ν, and $\Gamma(\cdot)$ is the Gamma function. Under these assumptions, Hendriks et al. derive the MMSE estimator

$$T_{\text{MMSE}}(\mathbf{y}) = \nu \frac{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_m|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right]} \frac{\sigma_s^2 T_{\text{MVDR}}^{(m)}(\mathbf{y}) \mathcal{M}(\nu+1, 2, P_m)}{\nu(\mathbf{d}^H \mathbf{\Phi}_m^{-1}\mathbf{d})^{-1} + \sigma_s^2}}{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_m|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right]} \mathcal{M}(\nu, 1, P_m)} \tag{17}$$

with

$$T_{\text{MVDR}}^{(m)}(\mathbf{y}) = \frac{\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{y}}{\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d}}, \quad Q_m = (\nu + \mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d} \sigma_s^2)^{-\nu},$$

$$\text{and} \quad P_m = \frac{\sigma_s^2 \mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d} \left|T_{\text{MVDR}}^{(m)}(\mathbf{y})\right|^2}{\nu(\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d})^{-1} + \sigma_s^2}$$

with $\mathcal{M}(\cdot, \cdot, \cdot)$ being the confluent hypergeometric function [27, Sec. 9.21]. From (17) it is apparent that the MMSE estimator *cannot* be decomposed in a linear spatial filter and a spectral postfilter. This is because the linear term $T_{\text{MVDR}}^{(m)}$ as well as the quadratic term $\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}$ depend on the summation index m. The spatial nonlinearity is particularly evident from the aforementioned quadratic term.

Throughout this work, we compare the results of the optimal spatially nonlinear MMSE estimator with a classical setup comprised of a linear spatial filter and (nonlinear) spectral postfilter. Figure 1 provides an illustration of the compared estimators: part (b) represents the nonlinear spatial filter T_{MMSE} given in (17) and part (a) corresponds to a combination of the MVDR beamformer with an MMSE-optimal postfilter. We now derive the postfilter under the same distributional assumptions as T_{MMSE}.

Since the MVDR beamformer is linear, we can infer the distribution of the beamformer output and observe that it follows a one-dimensional complex Gaussian mixture distribution with PDF

$$p(T_{\text{MVDR}}(\mathbf{y})|s) = \sum_{m=1}^{M} c_m \mathcal{N}_{\mathbb{C}}\Bigg(s, \underbrace{\frac{\mathbf{d}^H \mathbf{\Phi}_n^{-1} \mathbf{\Phi}_m \mathbf{\Phi}_n^{-1} \mathbf{d}}{(\mathbf{d}^H \mathbf{\Phi}_n^{-1} \mathbf{d})^2})}_{\sigma_m^2}\Bigg) \tag{18}$$

for an input $\mathbf{Y}$ that is distributed according to a multivariate complex Gaussian mixture distribution. The Gaussian mixture components have the mean s and variance σ_m^2, $m \in \{1, ..., M\}$. Based on this observation, we compute the MMSE-optimal spectral postfilter using [27, Eq. 3.339, Eq. 6.643.2, Eq. 9.220.2] and [28, Eq. 10.32.3] and obtain the estimator

$$T_{\text{MVDR-MMSE}}(\mathbf{y}) = \nu \frac{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_{\text{MVDR}}(\mathbf{y})|^2}{\sigma_m^2}\right]} \frac{\sigma_s^2 T_{\text{MVDR}}(\mathbf{y}) \mathcal{M}(\nu+1, 2, P_m)}{\nu\sigma_m^2 + \sigma_s^2}}{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_{\text{MVDR}}(\mathbf{y})|^2}{\sigma_m^2}\right]} \mathcal{M}(\nu, 1, P_m)} \tag{19}$$

with

$$\mathbf{\Phi}_n = \sum_{m=1}^{M} c_m \mathbf{\Phi}_m, \quad \sigma_m^2 = \frac{\mathbf{d}^H \mathbf{\Phi}_n^{-1} \mathbf{\Phi}_m \mathbf{\Phi}_n^{-1} \mathbf{d}}{(\mathbf{d}^H \mathbf{\Phi}_n^{-1} \mathbf{d})^2},$$

$$Q_m = (\frac{1}{\sigma_m^2} + \frac{\nu}{\sigma_s^2})^{-\nu} \quad \text{and} \quad P_m = \frac{\sigma_s^2 \sigma_m^{-2} |T_{\text{MVDR}}(\mathbf{y})|^2}{\nu\sigma_m^2 + \sigma_s^2}.$$

that sequentially combines linear spatial processing with MMSE-optimal spectral postprocessing as depicted in Figure 1a.

IV. Evaluation of the benefit of a nonlinear spatial filter in non-Gaussian noise

Section III points out that using a nonlinear spatial filter is MMSE-optimal and, thus, may be beneficial if the noise does not follow a Gaussian distribution. It is well known that the DFT coefficients of speech are often better modeled by a more heavy-tailed distribution than a Gaussian if originating from short-time Fourier transform (STFT) segments with short duration [6], [29]. Consequently, one may argue that this as well applies to noise DFT coefficients if the background noise stems from human speakers. In any case, Martin [29] observed that heavy-tailed distributions also provide a good fit for DFT coefficients of some types of noise in the one-dimensional case.

In this section, we investigate the potential of the optimal nonlinear spatial filter versus the classical separated setup with a linear spatial filter and a spectral postfilter for noise with a non-Gaussian distribution. Section IV-A presents our findings for noise that departs from Gaussianity by means of heavier tails but with a rather simple spatial structure. We published parts of this analysis and of the analysis in Section IV-C in [20]. However, here we also include the multichannel Wiener filter for comparison, compute more detailed performance metrics, and have made changes to the speech power parameter estimation scheme. In Section IV-B we provide results for noise that is modeling a spatially more diverse noise field created by five interfering human speakers and in Section IV-C we evaluate the nonlinear spatial filtering approach based on real-world noise recordings from the CHiME3 database. Please find audio examples for all experiments on our website[1].

A. Heavy-tailed noise distribution

In our first experiment, we investigate the performance of the nonlinear spatial filter T_{MMSE} by mixing the target speech signal at the microphones with multichannel noise that is sampled from a heavy-tailed Gaussian mixture distribution.

1) Noise distribution model: We construct a Gaussian mixture distribution with an adjustable heavy-tailedness by combining Gaussian components with scaled versions of the same covariance matrix. Therefore, we set the mth mixture component's covariance matrix $\mathbf{\Phi}_m$ to be

$$\mathbf{\Phi}_m = \frac{b^{m-1}}{r}\mathbf{\Phi}_n \quad \text{with} \quad r = \sum_{m=1}^{M} c_m b^{m-1} \tag{20}$$

and scaling factor $b \in \mathbb{R}^+$. The constant r ensures correct normalization such that the overall covariance matrix of our scaled Gaussian mixture distribution remains $\mathbf{\Phi}_n$.

We rely on the kurtosis to quantify the heavy-tailedness of the scaled Gaussian mixture distributions. It is a statistical measure that accounts for the likelihood of the occurrence of outliers [30] and it has been extended for real-valued multivariate distributions by Mardia [31]. We extend it to complex-valued random vectors $\mathbf{X} \in \mathbb{C}^n$ with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}_x$ by defining its kurtosis to be

$$\kappa_{\mathbb{C}}(\mathbf{X}) = \mathbb{E}\left[(2(\mathbf{X}-\boldsymbol{\mu})^H\mathbf{C}_x^{-1}(\mathbf{X}-\boldsymbol{\mu}))^2\right]. \tag{21}$$

A complex-valued n-dimensional Gaussian distribution can equivalently be formulated as a real-valued $2n$-dimensional Gaussian distribution [22, Thm. 15.1]. The additional factor of two in (21) ensures that the same kurtosis value results for both formulations of the same distribution. Using [32, Sec. 8.2.4], we compute the kurtosis of a vector $\mathbf{N}$ distributed according to a scaled Gaussian mixture distribution to obtain

$$\kappa_{\mathbb{C}}(\mathbf{N}) = 2D(2+2D)\underbrace{\sum_{m=1}^{M} c_m \frac{b^{2(m-1)}}{r^2}}_{q} \tag{22}$$

and observe that $\kappa_{\mathbb{C}}(\mathbf{N})$ is given by the kurtosis of a D-dimensional complex Gaussian distribution multiplied by a factor q that depends on the scaling factor b and the number of mixture components M.

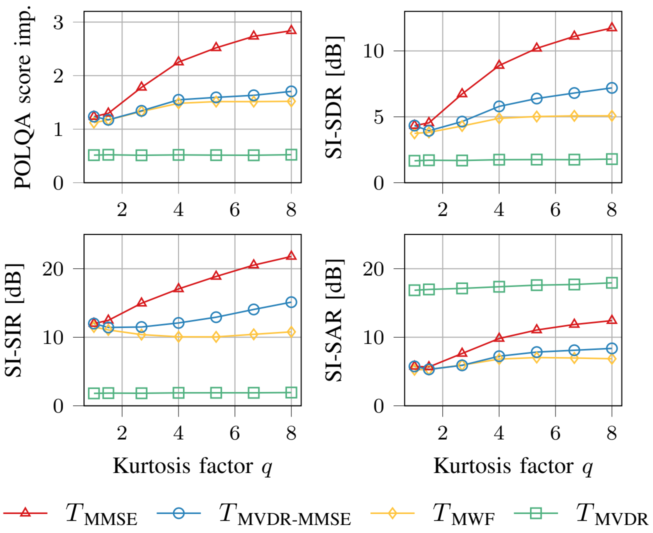


Fig. 2: POLQA, SI-SDR, SI-SIR and SI-SAR results for scaled Gaussian mixture noise distributions with increased heavy-tailedness in diffuse noise.

2) Experimental setup: In our test scenario, we use five microphones arranged in a linear array with 5 cm spacing and broadside orientation towards the target signal source and model the propagation path between the target speaker and the microphones based on time delays only. We perform the evaluation using 48 clean speech signals taken from the WSJ0 dataset [33] that are balanced between female and male speakers. The noise DFT coefficients are samples from a scaled Gaussian mixture distribution with scale factor $b = 2$ and a variable number of mixture components with equal weight $c_m = \frac{1}{M}$, $m \in \{1, ..., M\}$. The noise covariance matrix $\mathbf{\Phi}_n$ models a diffuse noise field with a small portion (factor of 0.05) of additional spatially and spectrally white noise as in [34, Eq. 27]. The noise and speech are combined such that a signal-to-noise ratio (SNR) of 0 dB is obtained.

3) Performance evaluation: Figure 2 provides a performance comparison of the jointly spatial and spectral nonlinear T_{MMSE} and the spatially linear $T_{\text{MVDR-MMSE}}$ with a nonlinear postfilter. The speech shape parameter is set to $\nu = 0.25$ for both estimators. Furthermore, we display results obtained with the well-known linear spatial filter T_{MVDR} without a postfilter and the multichannel Wiener filter T_{MWF}, which is the MMSE-optimal solution if noise *and* speech follow a Gaussian distribution, i.e., $T_{\text{MVDR-MMSE}}$ with $\nu = 1$ and $M = 1$. The performance results are displayed with respect to the kurtosis factor q on the x-axis indicating an increased heavy-tailedness of the noise distribution from left to right.

The plot in the upper left corner shows the performance with respect to the improvement of the POLQA measure [35], which is the successor of perceptual evaluation of speech quality (PESQ) [36] and returns the expected mean opinion score (MOS), which takes values from one (bad) to five (excellent). In any plot of Figure 2, we are particularly interested in the

[1] https://uhh.de/inf-sp-nonlinear-spatial-filter-tasl2021

performance difference of the T_{MMSE} (red) and $T_{\text{MVDR-MMSE}}$ (blue) as this gap characterizes the potential performance gain of a nonlinear spatial filter. For POLQA, we observe an increase of the performance difference up to 1.1 POLQA score improvement as the noise distribution shifts towards a more heavy-tailed distribution.

The estimators including a postfilter, T_{MMSE}, $T_{\text{MVDR-MMSE}}$, and T_{MWF}, require an estimate of the speech power spectral density (PSD) σ_s^2. In contrast to our previous paper [20], here we do not rely on oracle knowledge of the clean speech signal to estimate this parameter but obtain an estimate from the noisy signal based on the cepstral smoothing technique [37]. This results in an increased performance gap between T_{MMSE} and $T_{\text{MVDR-MMSE}}$. From this finding, we conclude that a nonlinear spatial filtering approach is even more beneficial if the performance of the postfilter decreases due to estimation errors of the spectral power of the target speech signal.

The next three plots (upper right and second row) display the performance results for the SI-SDR, SI-SIR, and SI-SAR measures as defined in [38]. We compute the SI-SDR, SI-SIR, and SI-SAR for segments of length 10 ms without overlap and include only segments with target speech activity similar to the computation of the segmental SNR in [39]. The performance results based on the SI-SDR measure show a similar structure to the ones obtained with POLQA. For high kurtosis values, we observe a performance gap of 4.5 dB for T_{MMSE} and $T_{\text{MVDR-MMSE}}$. Furthermore, the difference between $T_{\text{MVDR-MMSE}}$ and T_{MWF} for high kurtosis values, which results exclusively from the different postfilter, is more obvious. The observed performance gaps, in particular the performance advantage of the nonlinear spatial filter, coincide with our own listening experience[1].

For the computation of the SI-SIR and SI-SAR measure displayed in the second row, the residual noise is split into interference noise and artifacts. It is striking to see that the red graph of T_{MMSE} runs above the blue graph of $T_{\text{MVDR-MMSE}}$ in both plots meaning that the nonlinear spatial filter achieves better noise reduction and fewer speech distortions at the same time. The better performance with respect to the SI-SAR measure is quite notable as we can see that the linear MVDR beamformer introduces very few speech distortions but its combination with different postfilters ($T_{\text{MVDR-MMSE}}$ and T_{MWF}) still performs worse than the joint spatial and spectral nonlinear processing by T_{MMSE}.

B. Inhomogeneous noise field (interfering speech)

Instead of sampling a Gaussian mixture distribution as in the previous section or in [21], we now use a setup with five interfering point sources arranged as illustrated in Figure 3 and, this way, move closer towards realistic noise scenarios. The estimators T_{MMSE} and $T_{\text{MVDR-MMSE}}$ have been derived under a Gaussian mixture noise assumption. To be consistent with this modeling assumption, we require the five interfering point sources to not be Gaussian distributed or not be simultaneously active per time-frequency bin, because otherwise the overall noise resulting from the different interfering sources would also be Gaussian distributed. Choosing human speakers as interfering sources, this assumption is commonly assumed to hold and referred to as w-disjoint orthogonality [40].

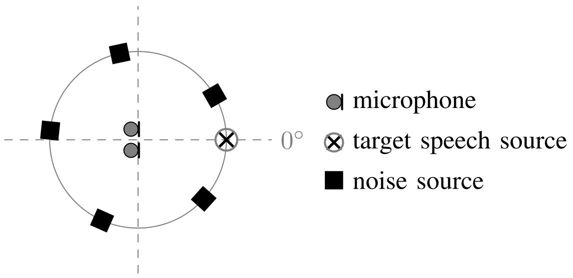


Fig. 3: Illustration of the experiment setup with a two-dimensional linear microphone array, a target speech source in broadside direction and five interfering point sources (human speakers (Section IV-B) or Gaussian bursts (Section V)).

	Interfering speech	Gaussian bursts
Δ POLQA	0.84 ± 0.04	2.64 ± 0.08
Δ SI-SDR	4.63 ± 0.15	9.92 ± 0.30
Δ SI-SAR	3.91 ± 0.16	8.39 ± 0.26
Δ SI-SIR	6.44 ± 0.22	14.95 ± 0.46
ESTOI (noisy)	0.49 ± 0.01	0.57 ± 0.02
ESTOI (T_{MMSE})	0.85 ± 0.01	0.94 ± 0.00
ESTOI ($T_{\text{MVDR-MMSE}}$)	0.72 ± 0.01	0.67 ± 0.02

TABLE I: Performance results (mean and the 95% confidence interval) of the T_{MMSE} and $T_{\text{MVDR-MMSE}}$ estimators for an inhomogeneous noise field with interfering speech and Gaussian sources as described in Section IV-B and Section V respectively.

1) Experimental setup: As can be seen in Figure 3, the target speech source is placed in the broadside direction of the two-dimensional linear microphone array with 6 cm microphone spacing. We sample the target speech signal and the interfering signals from distinct subsets of the WSJ0 dataset. The two-dimensional noise signal is then obtained by multiplying the interfering speech signals with the steering vectors $\mathbf{d}_i$, $i \in \{0, ..., 4\}$, and adding the individual interfering sources' signals. The steering vector $\mathbf{d}_i$ of the ith interfering source positioned at $\theta_i = \frac{\pi}{6} + \frac{2\pi}{5} i$ radians models the relative time difference of arrival at the microphones. The target speech signal and noise signal are rescaled to correspond to an SNR of 0 dB.

We now compare the performance of the T_{MMSE} and $T_{\text{MVDR-MMSE}}$ estimators for the inhomogeneous noise field. For this, we require estimates of the Gaussian mixture components' covariance matrices $\mathbf{\Phi}_m$ and the mixture weights c_m. We choose the number of components equal to the number of interfering sources, i.e., $M = 5$, and estimate the Gaussian mixture parameters using the expectation maximization (EM) algorithm [41] applied to overlapping signal segments of length 250 ms and with an overlap of 50% from the pure noise signal. As before, we estimate the spectral power of speech using the cepstral smoothing technique and use a speech shape parameter $\nu = 0.25$.

2) Performance evaluation: The first column of Table I displays the performance results for the described simulation. For the performance measures in the first four rows, preceded with a Δ symbol, we report the performance difference between

$T_{\text{MVDR-MMSE}}$ and T_{MMSE} averaged over 48 samples. We observe that the nonlinear spatial filter delivers a considerable performance gain that amounts to 4.63 dB SI-SDR and a POLQA score of 0.84. The bottom part of Table I presents ESTOI [42] scores for the noisy signal and the enhancement results obtained with T_{MMSE} and $T_{\text{MVDR-MMSE}}$. The ESTOI scores provide a measure of speech intelligibility. As for the other performance measures, we find that the nonlinear spatial filter outperforms the combination of a linear spatial filter and a postfilter as the T_{MMSE} estimator yields an ESTOI score of 0.85 as opposed to the result of 0.72 achieved by $T_{\text{MVDR-MMSE}}$.

C. Real-world CHiME3 noise

Furthermore, we investigate the performance of the nonlinear spatial filter for real-world noise from the CHiME3 database [43] that has been recorded in four different environments: a cafeteria, a moving bus, next to a street, and in a pedestrian area.

1) Experimental setup: The CHiME3 data has been recorded using six microphones that are attached to a tablet computer. For this experiment, we use the simulated training subset of the official dataset, which has been created by mixing the recording of real-world background noise with a spatialized version of WSJ0 utterances. A detailed description of the data generation process can be found in [43]. We evaluate on 48 randomly selected samples that are balanced between male and female speakers.

As before, we require an estimate of the time-varying Gaussian mixture distribution parameters and estimate them using oracle knowledge of the noise signal. For this, we apply the EM algorithm to overlapping signal segments of length 750 ms. For both, T_{MMSE} and $T_{\text{MVDR-MMSE}}$, we use $\nu = 0.25$ and estimate the speech power σ_s^2 using the cepstral smoothing technique. In addition, we need to estimate the steering vector for the target speaker. For this, we employ oracle knowledge of the clean speech signal and extract the steering vector estimates as principal eigenvectors of the time-varying covariance matrix estimates obtained by recursive smoothing.

2) Performance evaluation: Again, we assess the performance gap between the nonlinear spatial filter T_{MMSE} and the separated setup with a linear spatial filter and a postfilter $T_{\text{MVDR-MMSE}}$. Figure 4 displays the SI-SDR results for these estimators and also the MVDR beamformer T_{MVDR} with respect to the number of mixture components that have been fitted to a cafeteria background noise using the EM algorithm. While the performance of the T_{MMSE} estimator (red) improves with increased modeling capabilities of the mixture distribution, neither the T_{MVDR} nor the $T_{\text{MVDR-MMSE}}$ estimator benefits from using more mixture components. As a result, we observe a performance gap of 3.17 dB SI-SDR between the best result obtained with the nonlinear spatial filter and the best result obtained with the separated setup. This value coincides with the first entry of Table II. For the POLQA measure displayed in the second row, we find a difference of 0.59 POLQA score. The table shows bigger performance differences for the cafeteria (CAF) and pedestrian area (PED) noise than the bus (BUS) and street (STR) noise. We suppose that this reflects that the cafeteria and pedestrian area noise is less stationary as we hear the most significant differences for impulse like background noise. The ESTOI scores displayed at the bottom of Table II indicate that the nonlinear spatial filter is not only beneficial to the speech quality but also the speech intelligibility. Overall, we conclude that the Gaussian noise assumption does not seem to be valid for the examined real-world noise as the nonlinear spatial filter provides a notable benefit also for these recordings.

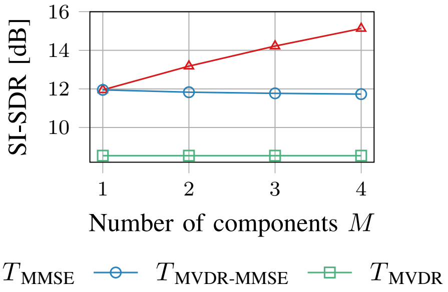


Fig. 4: SI-SDR results for cafeteria noise with respect to the number of mixture components used to fit the noise distribution.

	CAF	BUS	PED	STR
Δ SI-SDR	3.17±0.19	2.48±0.26	3.31±0.24	2.07±0.28
Δ POLQA	0.59±0.06	0.38±0.07	0.56±0.05	0.28±0.04
ESTOI (noisy)	0.60±0.03	0.71±0.02	0.56±0.03	0.69±0.03
ESTOI (T_{MMSE})	0.94±0.01	0.97±0.01	0.93±0.01	0.96±0.01
ESTOI ($T_{\text{MVDR-MMSE}}$)	0.89±0.02	0.95±0.01	0.88±0.02	0.94±0.01

TABLE II: Performance results (mean and the 95% confidence interval) of the nonlinear spatial T_{MMSE} and linear spatial filter combined with a postfilter $T_{\text{MVDR-MMSE}}$ for noise from the CHiME3 databse, which has been fitted with a Gaussian mixture distribution with four mixture components as described in Section IV-C.

V. Interpretation: a nonlinear spatial filter enables superior spatial selectivity

We assume that the performance benefit of the nonlinear spatial filter reported in Section IV-B and IV-C is due to the more efficient use of spatial information by the T_{MMSE} estimator. Here we support this conjecture by an experiment that provides an insight into the functioning of the nonlinear spatial filter.

1) Experimental setup: We use the same geometric setup as described previously in Section IV-B (Figure 3) but replace the interfering speech sources with sources that emit spectrally white Gaussian signals. To match the long-term non-Gaussianity assumption, only one interfering source emits a signal at a specific time instance. We implement this by using short (336 ms) non-overlapping Gaussian bursts for the interfering sources. The so created noise signal can be viewed as stationary regarding its spectral characteristics except at the segment boundaries. By applying the EM algorithm to the full-length noise signal we also model the spatial characteristics as long-

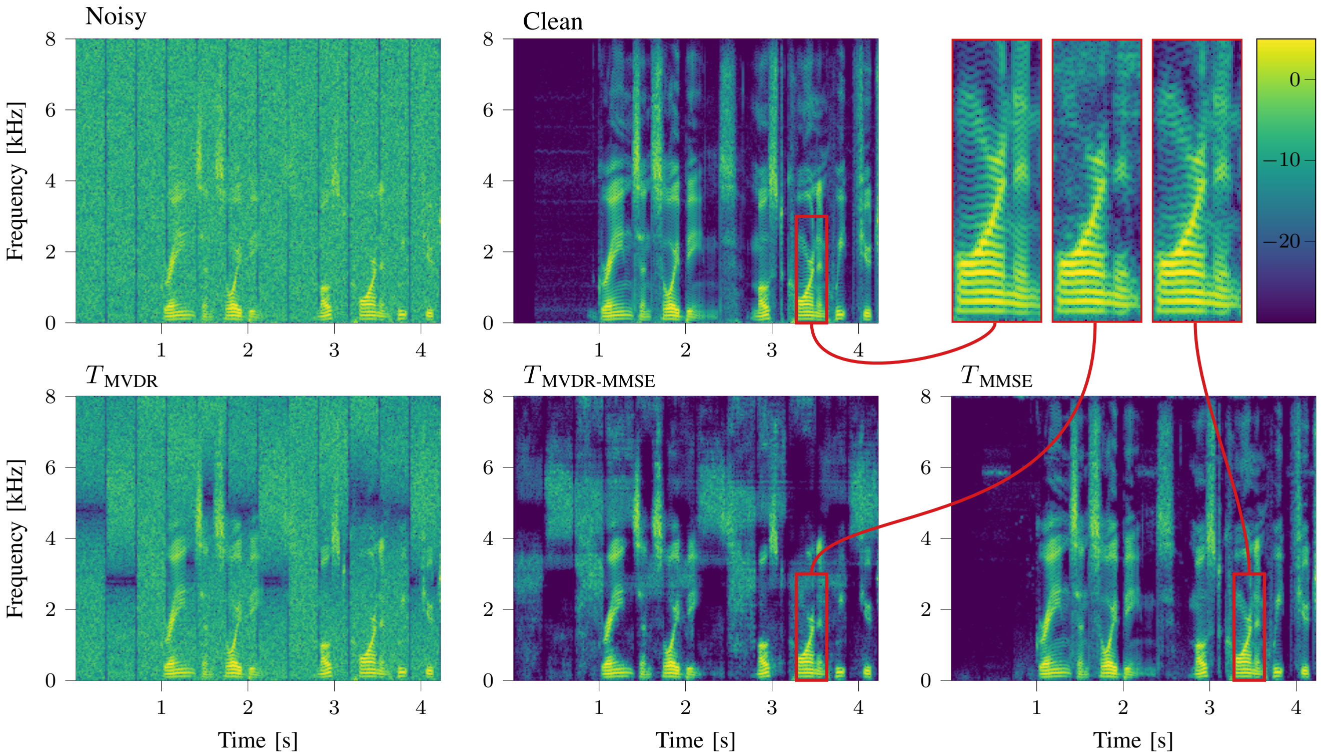


Fig. 5: Spectrograms of an example in an inhomogeneous noise field with five interfering sources emitting Gaussian noise bursts. The second row visualizes the processing results obtained with T_{MVDR}, $T_{\text{MVDR-MMSE}}$ and T_{MMSE} and the top row shows the clean and noisy spectrograms as well as close-ups of the fine-structure of a voiced speech segment.

term stationary. All other experiment settings remain unchanged as described before in Section IV-B.

2) Performance evaluation: The performance results are displayed in the second column of Table I. For this artificial type of noise, we observe an even greater performance difference of 9.9 dB SI-SDR and 2.6 POLQA score. In fact, the T_{MMSE} estimator seems to be able to recover the original signal almost perfectly except from minor residual high-frequency noise while $T_{\text{MVDR-MMSE}}$ suffers from clearly audible speech degradation and residual noise. Audio examples can be found online[1].

Figure 5 depicts the spectrograms of the clean and noisy signals in the top row and the spectrograms of the enhancement results obtained by the T_{MVDR}, $T_{\text{MVDR-MMSE}}$, and T_{MMSE} estimators in the bottom row. The uniform green coloration of the vertical stripes in the noisy spectrogram reflects the spectral stationarity. The vertical dark blue lines separate segments with different spatial properties. While the spatial diversity cannot be seen from the spectrogram, it becomes visible from the result of the MVDR beamformer (first in bottom row). Here, the MVDR beamformer suppresses different frequencies for signal segments with different spatial properties as can be seen from the displaced horizontal dark blue lines. The described differences between the $T_{\text{MVDR-MMSE}}$ (middle) and T_{MMSE} (right) estimators' results are also found in the spectrograms. A close look reveals that the nonlinear spatial filter preserves much more of the target signal's fine structure. Furthermore, a comparison with the spectrogram of the clean speech signal highlights that it suppresses background noise much better than the $T_{\text{MVDR-MMSE}}$ estimator. Residual noise is visible in the spectrogram only in some segments at a frequency of about 6 kHz.

3) Discussion: To explain these observations, we examine the covariance matrices $\mathbf{\Phi}_m$, $m \in \{1, ..., 5\}$, of the Gaussian mixture noise distribution estimated with the EM algorithm. In Figure 6 we visualize their spatial structure based on the directivity pattern [3, Sec. 12.5.2] that they produce when used as noise correlation matrix in the MVDR beamformer, which is denoted with $T^{(m)}_{\text{MVDR}}$ in (17). Furthermore, we visualize the directivity pattern of the MVDR beamformer. The correlation matrix $\mathbf{\Phi}_n$ required to compute T_{MVDR} is related to the mixture component covariance matrices via

$$\mathbf{\Phi}_n = \sum_{m=1}^{M} c_m \mathbf{\Phi}_m. \tag{23}$$

The directivity pattern produced by T_{MVDR} is displayed at the top left followed by visualizations of the five mixture component covariance matrices. For each of these, a pronounced spatial characteristic can be observed by means of the horizontal dark lines. On the right side of the directivity patterns, we indicate the incidence angles θ_i of the noise sources. We notice that each component's covariance matrix models one of the noise sources as apparent from the zero placed in the respective direction by the MVDR beamformer. The second horizontal line

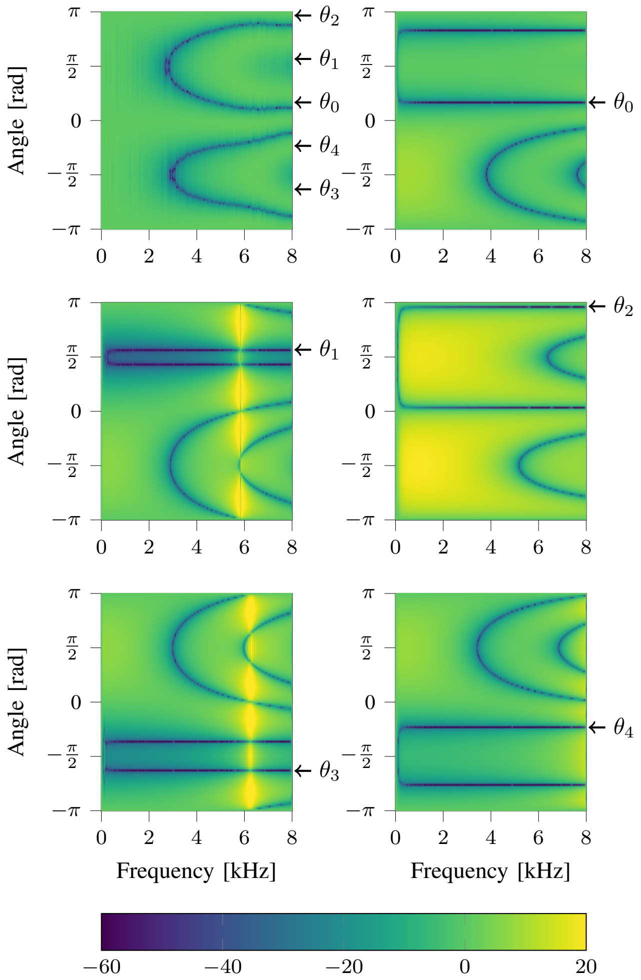


Fig. 6: Directivity patterns of T_{MVDR} (top left) and $T^{(m)}_{\text{MVDR}}$ (MVDR beamformer with noise correlation matrix $\mathbf{\Phi}_m$), $m \in \{1, ..., 5\}$. The arrows on the right side indicate the incidence angle θ_i, $i \in \{0, ..., 4\}$ of the ith interfering point source.

originates from the symmetry requirements of the directivity pattern, which are determined by the array geometry.

In comparison, the directivity pattern of the T_{MVDR} (top left) does not show zeros placed into the directions of the interfering point sources. Instead, the weighted combination in (23) seems to eliminate some of the spatial information, which corresponds to the well-known fact that a two-microphone MVDR beamformer can suppress only one directional interfering source but not five of them. As a result, only some frequencies are suppressed for each interfering source as we observed in Figure 5 before.

Figure 6 reveals how much more spatial information can be utilized by T_{MMSE} in comparison with $T_{\text{MVDR-MMSE}}$, whose spatial processing relies on an estimate of $\mathbf{\Phi}_n$ as visualized in the top left plot. The initial spatial filtering step using the MVDR beamformer is not capable of suppressing the directional sources and the remaining noise has to be filtered by the postfilter, which then leads to some improvements. On the other hand, the nonlinear spatial T_{MMSE} estimator can utilize the spatial information provided by the estimates of covariance matrices $\mathbf{\Phi}_m$.

One could argue that a time-varying MVDR beamformer $T^{(m)}_{\text{MVDR}}$ with correctly chosen m would suffice to solve the problem on the short signal segments with noise from a single point source and a complicated nonlinear approach is not required. However, we must point out that the step of choosing the 'right' covariance matrix is not required for the nonlinear spatial filter. Instead, we provide the Gaussian mixture parameters reflecting the spatial properties of the full utterance and, nevertheless, the T_{MMSE} estimator is capable of suppressing five directional noise sources with only two microphones without the need for spatial adaptation. Note that this is an exciting finding as traditional linear spatial filters can only suppress $D-1$ interfering point sources with D microphones without spatial adaptation [26, Sec. 6.3].

Despite the impressive performance results achieved in this experiment, the analytic nonlinear spatial filter has some weaknesses: it requires a very accurate estimation of the spatial and spectral characteristics of the noise signal and is also computationally quite demanding. In addition to the presented experiments, we carried out simulations using measured impulse responses between the microphones and speakers and observed a much lower benefit from using a nonlinear spatial filter for the experiment with five interfering Gaussian sources even with access to oracle noise data. This is because the spatial and spectral diversity of the noise signal increase and many more mixture components would be required to model the noise accurately which then results in a data problem. Similarly, estimating the parameters of the Gaussian mixture from a noisy signal is difficult. We approached this using masks to identify time-frequency bins that are dominated by noise but did not obtain reliable estimates this way.

Therefore, we conclude that the analytical estimators allow us to study the potential of nonlinear spatial filters in principle, but because of high sensitivity to parameter estimation errors and high computational costs, practical nonlinear spatial filters may be better implemented using modern machine learning tools like DNNs.

VI. Conclusions

In a detailed theory overview, we have revisited the fact that the multichannel MMSE-optimal estimator of the clean speech signal is in general a jointly spatial-spectral nonlinear filter. Therefore, the state-of-the-art concatenation of a linear spatial filter and a postfilter is MMSE-optimal only in the special case that the noise follows a Gaussian distribution. The experimental section of this paper studied the performance advantage that can be gained by replacing the generally suboptimal sequential setup with a nonlinear spatial filter in three different non-Gaussian noise scenarios.

First, we have shown that considerable performance improvements result if the noise distribution deviates from a Gaussian distribution by an increased heavy-tailedness as the nonlinear spatial filter enables a higher noise reduction and lower speech distortions at the same time. Second, we report a performance benefit of 4.6 dB SI-SDR and of 0.8 POLQA score for an inhomogeneous noise field created by five interfering speech sources and, furthermore, we have observed a benefit of about 3.2 dB SI-SDR and 0.6 POLQA score for the real-world cafeteria noise recordings from the CHiME3 database. In addition, we have performed experiments that revealed that the nonlinear spatial filter has some notably increased spatial processing capabilities allowing for an almost perfect elimination of five Gaussian interfering point sources with only two microphones.

The presented findings on the performance potential of a nonlinear spatial filter motivate further research on the implementation of nonlinear spatial filters, e.g., using DNNs to learn the nonlinear spatial filter directly from data and overcome the parameter estimation issues and other limitations of the analytic nonlinear spatial filter that we have used for this analysis.

Acknowledgment

We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

References

[1] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.

[3] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Chichester, England Hoboken, NJ: John Wiley, 2006.

[4] J. Benesty, *Microphone Array Signal Processing*. Berlin: Springer, 2008.

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

[6] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, May 2005.

[7] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1741–1752, 2007.

[8] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[9] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5069–5073.

[10] S. R. Park and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," in *Proc. Interspeech 2017*, Stockholm, Sweden, 2017, pp. 1993–1997.

[11] A. Pandey and D. Wang, "TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 6875–6879.

[12] R. Balan and J. P. Rosca, "Microphone Array Speech Enhancement by Bayesian Estimation of Spectral Amplitude and Phase," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, Rosslyn, Virginia, Aug. 2002, pp. 209–213.

[13] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On Optimal Multichannel Mean-Squared Error Estimators for Speech Enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, 2009.

[14] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 196–200.

[15] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5745–5749.

[16] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.

[17] H. Lee, H. Y. Kim, W. H. Kang, J. Kim, and N. S. Kim, "End-to-End Multi-Channel Speech Enhancement Using Inter-Channel Time-Restricted Attention on Raw Waveform," in *Proc. Interspeech 2019*. Graz, Austria: ISCA, Sep. 2019, pp. 4285–4289.

[18] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-Attention Dense U-Net for Multichannel Speech Enhancement," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 836–840.

[19] G. Itzhak, J. Benesty, and I. Cohen, "Nonlinear Kronecker product filtering for multichannel noise reduction," *Speech Communication*, vol. 114, pp. 49–59, Nov. 2019.

[20] K. Tesch, R. Rehr, and T. Gerkmann, "On Nonlinear Spatial Filtering in Multichannel Speech Enhancement," in *Proc. Interspeech 2019*, Graz, Austria, 2019, pp. 91–95.

[21] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering for multichannel speech enhancement in inhomogeneous noise fields," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, May 2020, pp. 196–200.

[22] S. M. Kay, *Fundamentals Of Statistical Signal Processing*. Pearson, 2009.

[23] M. Schervish, *Theory of Statistics*. New York, NY: Springer New York, 1995.

[24] H. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY: Springer New York, 1994.

[25] E. L. Lehmann, *Testing Statistical Hypotheses*. New York: Springer, 2005.

[26] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Apr. 2004.

[27] I. S. Gradshteyn and J. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego: Academic Press, 2000.

[28] "NIST Digital Library of Mathematical Functions," f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[29] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[30] P. H. Westfall, "Kurtosis as Peakedness, 1905–2014. R.I.P." *The American Statistician*, vol. 68, no. 3, pp. 191–195, 2014.

[31] K. V. Mardia, "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, vol. 57, no. 3, pp. 519–530, 1970.

[32] K. B. Petersen and M. S. Pedersen, "The Matrix Cookbook," Nov. 2012.

[33] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," Linguistic Data Consortium, Philadelphia, May 2007.

[34] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, Jan. 2014.

[35] "P.863: Perceptual objective listening quality prediction," International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[36] "P.862.3: Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," International Telecommunication Union, Nov. 2007, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.3-200711-I/en

[37] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, Mar. 2008, pp. 4897–4900.

[38] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 626–630.

[39] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[40] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[41] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[42] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[43] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, USA, Dec. 2015, pp. 504–511.

Publication 2: Insights into Deep Non-linear Filters for Improved Multi-channel Speech Enhancement [P5]

4

Abstract

The key advantage of using multiple microphones for speech enhancement is that spatial filtering can be used to complement the tempo-spectral processing. In a traditional setting, linear spatial filtering (beamforming) and single-channel post-filtering are commonly performed separately. In contrast, there is a trend towards employing deep neural networks (DNNs) to learn a joint spatial and tempo-spectral non-linear filter, which means that the restriction of a linear processing model and that of a separate processing of spatial and tempo-spectral information can potentially be overcome. However, the internal mechanisms that lead to good performance of such data-driven filters for multi-channel speech enhancement are not well understood.

Therefore, in this work, we analyse the properties of a non-linear spatial filter realized by a DNN as well as its interdependency with temporal and spectral processing by carefully controlling the information sources (spatial, spectral, and temporal) available to the network. We confirm the superiority of a non-linear spatial processing model, which outperforms an oracle linear spatial filter in a challenging speaker extraction scenario for a low number of microphones by 0.24 POLQA score. Our analyses reveal that in particular spectral information should be processed jointly with spatial information as this increases the spatial selectivity of the filter. Our systematic evaluation then leads to a simple network architecture, that outperforms state-of-the-art network architectures on a speaker extraction task by 0.22 POLQA score and by 0.32 POLQA score on the CHiME3 data.

Reference

Kristina Tesch and Timo Gerkmann, "Insights into Deep Non-linear Filters for Improved Multi-channel Speech Enhancement", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563-575, 2023. DOI: 10.1109/TASLP.2022.3221046

Copyright notice

A Note on the Term Non-linear Spatial Filter

Please note that in this publication, the term *non-linear spatial filter* takes a different meaning than in the other chapters of this cumulative thesis. The publication differentiates between the following DNN-based filters:

- A joint spatial and spectral non-linear filter (F-JNF)
- A joint spatial and temporal non-linear filter (T-JNF)
- A joint spatial and tempo-spectral non-linear filter (FT-JNF)
- A filter focused on spatial information, which is referred to as a non-linear spatial filter (NSF), with variants F-NSF, T-NSF, and FT-NSF, where the prefix indicates which type of global information is available.

Insights into Deep Non-linear Filters for Improved Multi-channel Speech Enhancement

Kristina Tesch, *Student Member, IEEE*, and Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—The key advantage of using multiple microphones for speech enhancement is that spatial filtering can be used to complement the tempo-spectral processing. In a traditional setting, linear spatial filtering (beamforming) and single-channel post-filtering are commonly performed separately. In contrast, there is a trend towards employing deep neural networks (DNNs) to learn a joint spatial and tempo-spectral non-linear filter, which means that the restriction of a linear processing model and that of a separate processing of spatial and tempo-spectral information can potentially be overcome. However, the internal mechanisms that lead to good performance of such data-driven filters for multi-channel speech enhancement are not well understood.**

Therefore, in this work, we analyse the properties of a non-linear spatial filter realized by a DNN as well as its interdependency with temporal and spectral processing by carefully controlling the information sources (spatial, spectral, and temporal) available to the network. We confirm the superiority of a non-linear spatial processing model, which outperforms an oracle linear spatial filter in a challenging speaker extraction scenario for a low number of microphones by 0.24 POLQA score. Our analyses reveal that in particular spectral information should be processed jointly with spatial information as this increases the spatial selectivity of the filter. Our systematic evaluation then leads to a simple network architecture, that outperforms state-of-the-art network architectures on a speaker extraction task by 0.22 POLQA score and by 0.32 POLQA score on the CHiME3 data.

***Index Terms*—Multi-channel, speech enhancement, joint non-linear spatial and tempo-spectral filtering**

I. Introduction

In our everyday life, speech understanding often takes place in noisy environments. This can be, for example, a conversation in a crowded restaurant, a phone call in a busy train station or the use of a voice control system in a driving car. To enable devices such as hearing aids or voice-controlled assistants to function in these challenging acoustic environments, speech enhancement algorithms are employed to improve the speech quality and intelligibility of the target speech signal.

Traditionally, many algorithms utilized a short-time Fourier transform (STFT) signal representation and derived an analytical clean speech estimator from a statistical model, e.g., [1]–[4]. While this has led to many interpretable and computationally lightweight algorithms, the derivations often require restricting and simplifying assumptions, e.g., independent time-frequency bins, to keep the problem tractable. This is in contrast to DNN-based algorithms, which do not need an explicit model, but learn to recognize complex dependencies directly from training data. In the domain of single-channel speech enhancement, these DNN-based algorithms, have been dominating the state of the art for a couple of years now, e.g., [5]–[8].

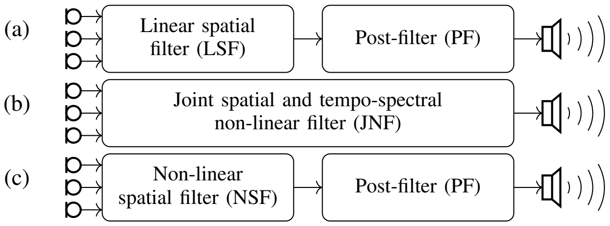

(a)
Linear spatial filter (LSF)
Post-filter (PF)
(b)
Joint spatial and tempo-spectral non-linear filter (JNF)
(c)
Non-linear spatial filter (NSF)
Post-filter (PF)

Fig. 1: (a) The traditional two-step processing using a linear spatial filter (beamformer) followed by a single-channel postfilter. (b) A joint spatial and tempo-spectral non-linear processing scheme that we implement using DNNs in this work. (c) Two-step processing scheme, however, not only the postfilter performs non-linear filtering but also the spatial filter.

While single-channel speech enhancement approaches exploit tempo-spectral signal characteristics to perform the enhancement, multi-channel approaches can additionally leverage spatial information by using multiple microphones. Commonly, this is done by employing a linear spatial filter, a so-called beamformer. Figure 1a illustrates a traditional multi-channel processing pipeline, which first applies the linear spatial filter and then adds a single-channel post-filter in a second step. The post-filter can be either linear or non-linear. In our prior work [9], [10], we have demonstrated that separation into a linear spatial filter and a post-filter is generally not optimal in the minimum mean square error (MMSE) sense unless we restrict the noise distribution to be Gaussian. However, if a non-Gaussian distribution is assumed, the resulting analytical solution is overall non-linear and joins the spatial and spectral processing as illustrated in Figure 1b. Our experimental evaluation in [10] has shown great potential for a joint non-linear spatial-spectral filter, but has also led to the conclusion that the estimation of required higher-order parameters limits the practical applicability of the analytic estimator. However, DNNs provide a data-driven way to implement practical joint spatial and tempo-spectral non-linear filters (JNF).

A very influential paper on multi-channel speech enhancement using DNNs has been the paper by Heymann et al. [11], who propose to use a DNN for estimating the parameters of a linear spatial filter. Also others have proposed approaches along this line of research, e.g., [12]–[14]. However, using a DNN for parameter estimation does not allow for a more general non-linear processing model nor does it permit the exploitation of interdependencies between spatial and tempo-spectral information during processing. In contrast, a variety of

The authors are with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany (e-mail: kristina.tesch@uni-hamburg.de; timo.gerkmann@uni-hamburg.de).

data-driven multi-channel filters have been proposed recently [15]–[20]. These implicitly drop the linearity assumption and integrate spatial and tempo-spectral processing steps such that this class of joint non-linear approaches is fundamentally different and potentially more powerful than DNN-driven linear spatial filters, aka neural beamformers. Good performance as been reported for these deep non-linear filters, but the internal mechanisms that lead to good performance are not well understood. This, however, is essential for a deliberate design of a neural network architecture that fully unlocks the potential of neural networks for multi-channel speech enhancement.

In this work, we investigate the internal functioning of DNN-based (joint) non-linear filters for multi-channel speech enhancement. To learn about the role of a non-linear spatial filter and the interdependency between spatial and tempo-spectral information, we consider a second separated approach, which combines a non-linear spatial filter with an independent post-filter. An illustration of this setup is given in Figure 1c. A systematic comparison of the three approaches outlined in Figure 1 then allows us to assess what makes for good spatial filtering performance: Is non-linear as opposed to linear spatial filtering the main factor for good performance? Or is it rather the interdependency between spatial and tempo-spectral processing? And do temporal and spectral information have the same impact on spatial filtering performance?

This paper is based on our recent conference paper [21], but the experimental evaluation here goes far beyond the results presented previously. Specifically, we propose new experimental designs to investigate the spatial filtering performance of a DNN-based joint filter. This then allows for a discussion of the spatial selectivity of the different approaches. We include a comparison with state-of-the-art approaches, showing that the joint non-linear filter obtained by our systematic evaluation outperforms them and, furthermore, we extend our evaluations from the speaker extraction task to the CHiME3 dataset. The latter then enables us to assess the role of the dataset characteristics with respect to the previously mentioned research questions.

In a recent study, also Tan et al. [22] compare the performance of a joint spatial and tempo-spectral non-linear filter with a DNN-driven linear spatial filter plus additional post-filter (i.e. Figure 1a versus Figure 1b). While they report comparable performance for these two approaches, in this paper, in line with our theoretical findings in [10], we demonstrate the conceptual superiority of a joint non-linear spatial and tempo-spectral filter by outperforming an *oracle* linear spatial filter plus post-filter. Furthermore, our work adds additional value beyond a general performance comparison of the two approaches by presenting experiments that allow for insights into the internal mechanisms underlying a well-performing joint non-linear filter.

The remainder of this paper is structured as follows. Section II introduces the signal model and provides a detailed overview of traditional and DNN-based spatial filtering. In Section III, we introduce a set of DNN-based filter variants, which will be analyzed thoroughly to provide insights into the separability of spatial processing and post-filtering (Section IV-B) and the interdependency between spatial and tempo-spectral processing (Section IV-C). In Section IV, we provide a comparison with recent state-of-the-art methods and, in Section VI, we report results for the CHiME3 dataset.

II. Background and related work

A. Signal model

We consider the task of extracting a single target speaker from a recording obtained in a noisy and reverberant environment. The noise signals may be environmental noise or concurrent speakers. The noisy mixture signals are captured by a microphone array with C microphone channels. In the time-domain, the speech signal uttered by the target speaker and recorded by the ℓth microphone can be written as the convolution of the non-reverberant speech signal $s(t)$ and the room impulse response (RIR) $h_\ell(t)$ describing the propagation path between the speaker and the ℓth microphone [23]:

$$x_\ell(t) = s(t) * h_\ell(t). \tag{1}$$

Note that, besides the room characteristics, $h_\ell(t)$ also allows to model the characteristics of the loudspeaker and the microphones.

We transform the time-domain signal $x_\ell(t)$ into the frequency domain using a short-time Fourier transform (STFT) to obtain complex spectral coefficients $X_\ell(k,i) \in \mathbb{C}$ with frequency-bin index k and time-frame index i. Based on an additive signal model, the mixing process in the frequency domain is given by

$$Y_\ell(k,i) = X_\ell(k,i) + V_\ell(k,i). \tag{2}$$

with $V_\ell(k,\, i)$ denoting the noise signal recorded at the ℓth microphone. We use bold face symbols to refer to the vector stacking the STFT coefficients for all channels, e.g., $\mathbf{Y}(k,i) = [Y_1(k,i), ..., Y_C(k,i)]^T \in \mathbb{C}^C$ and drop the time-frequency indices (k,i) to denote the tensor with shape $(C \times F \times T)$ comprising the time-frequency points for all C microphones and with F and T being the number of frequency-bins and time-indices respectively.

B. Traditional spatial filtering

Most traditional multi-channel speech enhancement schemes involve a spatial filter that is usually implemented following a filter-and-sum beamforming approach [24, Sec. 12.4.2]. Such a filter-and-sum beamformer aims to suppress signal components not originating from the target direction by filtering the individual microphone signals and adding them. Using vector notation, the processing model of a filter-and-sum beamformer in the frequency domain can be formulated as

$$\hat{S}(k,i) = \mathbf{h}(k,i)^H \mathbf{Y}(k,i) \tag{3}$$

with $\hat{S}(k,i) \in \mathbb{C}$ being an estimate of the target signal, a filter $\mathbf{h}(k,i) \in \mathbb{C}^C$ that may or may not be depending on the time index i (time-variant vs. time-invariant filter) and $(\cdot)^H$ denoting the Hermitian transpose.

The simplest form is a delay-and-sum beamformer [24, Sec. 12.4.1] that applies a filter to compensate for different time delays at the microphones caused by the differing lengths of propagation paths for the signal to reach each microphone. This approach implicitly assumes the noise signals recorded at the different microphones to be uncorrelated [24, Sec.12.6.1], which is a reasonable assumption for sensor noise, but not for environmental noise or interfering point sources.

Another commonly used spatial filter is the minimum variance distortionless response (MVDR) beamformer [24, Sec. 12.6.1] that takes into account the correlation between microphone channels. The filter weights $\mathbf{h}_{\text{MVDR}}(k,\, i)$ are obtained by solving the optimization problem

$$\mathbf{h}_{\text{MVDR}}(k,i) = \arg\min_{\mathbf{h}\in\mathbb{C}^C} \mathbf{h}^H(k,i)\mathbf{\Phi}_V(k,i)\mathbf{h}(k,i) \quad \text{s.t.} \quad \mathbf{h}(k,i)^H\mathbf{d}(k,i) = 1, \tag{4}$$

with the so-called steering vector $\mathbf{d}(k,i)$ modelling the direct path of the target signal $S(k,\, i)$ to the microphones and noise correlation matrix $\mathbf{\Phi}_V(k,i) = \mathbb{E}[\mathbf{V}(k,i)\mathbf{V}(k,i)^H]$ with $\mathbb{E}$ denoting the statistical expectation operator. Thus, the MVDR beamformer tries to minimize the noise variance at the output of the beamformer while leaving the target signal unchanged. The latter condition is referred to as the distortionless constraint of the MVDR. The solution of the optimization problem posed in (4) is given by [24, Sec. 12.6.1]

$$\mathbf{h}_{\text{MVDR}}(k,i) = \frac{\mathbf{\Phi}_V^{-1}(k,i)\mathbf{d}(k,i)}{\mathbf{d}^H(k,i)\mathbf{\Phi}_V^{-1}(k,i)\mathbf{d}(k,i)}. \tag{5}$$

Adhering to the filter-and-sum processing model, and using filter weights that do not depend on the value of the noisy signal $\mathbf{Y}(k,i)$ itself as in (5), traditional spatial filtering clearly is a linear operation with respect to the noisy input.

It has been shown that the MVDR beamformer is the optimal spatial filter under a Gaussian noise assumption [10], [25]. That is, any filter jointly performing spatial filtering and postfiltering can (in theory) be decomposed into an MVDR beamformer for spatial processing followed by a single-channel postfilter. A prominent example is the multi-channel Wiener filter, which can be decomposed in an MVDR plus single-channel Wiener filter [26]. The work by Hendriks et al. [27] and our prior work [10] reveal that this is not the case for more general noise distributions. The analytic filter derived in [10], [27] joins the spatial and spectral filtering into a non-separable non-linear operation which is in contrast to the simple and linear processing model of a beamformer. Our own previous work [10] demonstrates that such a joint spatial-spectral nonlinear processing may overcome the limitations of a linear beamformer, which is restricted to suppressing $M-1$ directional interfering point sources (maximum number of sources in a reverberation-free setting). However, oracle knowledge of the target and noise signals are required for accurate parameter estimation to obtain good results with the analytic joint spatial-spectral nonlinear filter.

C. DNN-based spatial filtering

While state-of-the-art single-channel speech enhancement nowadays completely relies on DNN-based approaches, DNN-based multi-channel approaches have become a vivid research topic recently. An important step towards using the capabilities of neural networks for multi-channel speech enhancement was taken by Heymann et al. [11], who design a DNN-based parameter estimation scheme for computing estimates of the steering vector and noise correlation matrix to be used in a traditional MVDR beamformer. This method has gained a reputation for its ease of use as well as good and robust results. Similarly, Togami [12] proposes to extract speaker masks for facilitating covariance matrix and speech power estimation to be used in a multi-channel speech separation scheme. Liu et al. [13] extend the masking-based beamforming approach of [11] by processing multi-channel instead of a collection of single-channel inputs and providing cross-channel features. Xiao et al. [14] train a network to directly estimate the time-invariant filter weights $\mathbf{h}(k)$ of a filter-and-sum beamformer from cross correlation features. The main drawback of a method that uses the impressive modeling capabilities of neural networks only for parameter estimation to be used in classical linear processing scheme is that the limitations of the linear model itself cannot be overcome.

In another line of research, spatial features are used as additional input to a neural network to increase speech separation or enhancement performance, e.g., [28]–[31]. The most common spatial features are inter-channel time or phase differences (ITD/IPD), inter-channel level differences (ILD), cross-correlation based features, as well as features computed with fixed beamformers. Most of these works, show notable performance improvements over single-channel approaches proving that spatial information is very valuable for speech separation and enhancement tasks. As for all approaches using hand-crafted features, a major concern is the question whether the chosen feature design is optimal for the task at hand. For example, in [32] and [33] the authors propose to estimate beamforming weights from speech and noise second-order statistics (covariance matrix estimates) using a DNN, while our analysis in [10] suggests that higher-order statistics are a valuable source of information, which can not be exploited this way.

An increasing number of recent works skips the spatial feature design part and trains a DNN-based filter to perform speech enhancement or separation based on raw multi-channel signals, either providing the time-domain signals [34]–[38] or frequency-domain signals [15]–[20] as input to the network. In many of these works, the authors claim that the network architecture has been designed with the goal in mind to implicitly learn a spatially selective filter from data. Nonetheless, the architectures proposed in these papers differ notably from each other. While some authors propose to learn a mask that is applied to a reference channel of the noisy signal, e.g., [15], [17], others propose a network that outputs the real and imaginary part of the target clean speech signal [18] or to learn a set of coefficients $\mathbf{h}$ and apply them to the signal adhering to the filter-and-sum processing model (cf. (3)) [19], [35].

For the last mentioned approach, it is clear that the authors have derived their architecture design from traditional linear filter-and-sum beamforming, but also others claim their architecture to be inspired be the traditional spatial filters, e.g., [17]. The authors of EaBNet [19] even propose to append the DNN-based spatial filter with a (DNN-based) post-filter following the traditional two-step procedure. However, it is important to be aware that their "spatial filter" as well as all other DNN-based approaches referenced in the last paragraph are in principle not only capable of performing non-linear spatial filtering but will likely perform spatial filtering jointly with tempo-spectral postfiltering. As a consequence, a direct comparison with a traditional linear beamformer, for example the MVDR beamformer, without a post-filter can therefore not be considered a fair comparison.

Overall, we conclude that many interesting architectures for implementing a DNN-based filter for multi-channel speech enhancement have been proposed, but also a lot of open questions remain. Most approaches haven been evaluated with respect to their overall speech enhancement performance. However, this is not very informative with regard to the internal mechanisms of the network. For example, it is unclear whether a network architecture inspired from traditional beamforming performs particularly well in spatial filtering as hypothesized by many authors, since performance improvements could also be achieved by better exploitation of tempo-spectral information. Thus, a more systematic evaluation is required to provide insights in the internal mechanisms of these DNN-based filters.

III. Proposed approach

In this work, we aim to investigate the contribution of different sources of information, that is spatial, spectral and temporal information, to a DNN-based filter for a speech enhancement or speech extraction problem. We are particularly interested in understanding the nature of a non-linear spatial filter and its interdependencies with temporal and spectral information. To provide insights into the "black box" of a DNN-powered filter, we use a simple network structure that allows us to easily control the integration of different sources of information and a dataset that makes it easy to assess the quality and properties of the spatial filter. This section describes the network design used in our experiments.

A. Base network architecture (F-JNF, T-JNF)

For our experiments, we adapt the architecture proposed by Li and Horaud [15], [39], that performs speech enhancement using a mask estimated from narrow-band multi-channel inputs. The distinctive feature of their approach is that the network processes all frequency bands separately. The network weights, however, are shared between frequencies. In the following, we propose a number of alternative network architectures to enable a detailed analysis. Figure 2 depicts the base network architecture. As can be seen in the bottom part, the network consists of only three layers, two (bi-directional) long short-term memory (LSTM) layers followed by a feed-forward (FF) layer. An LSTM layer [40] is commonly used for sequence modeling. In our setup, the feature dimension (vertical) mostly corresponds to channel information (real and imaginary parts stacked) while the sequence dimension (horizontal) is chosen according to the second source of information which could be time (narrow-band) or spectral (wide-band) information. As spatial information is processed jointly with a second source of information, we denote this network as joint non-linear filter (JNF) prefixed with T or F in the narrow-band and wide-band case respectively. Thus, the narrow-band version (T-JNF) as proposed in [15] has access to fine-grain spatial and temporal information but only global spectral statistics, while our proposed variant F-JNF can leverage fine-grained spectral information in addition to spatial information.

B. Combining temporal and spectral information (FT-JNF)

The basic architecture described in Section III-A combines spatial information with spectral *or* temporal information. Next, we propose a variant that can exploit all three sources of information combining spatial with tempo-spectral processing. In order to ensure comparability of the results, we do not change the basic architecture or the number of parameters. Instead, we manipulate the data arrangement at the position marked with a circled two. The filter denoted by FT-JNF then feeds wide-band data into the first LSTM layer. The obtained features are then switched to a narrow-band arrangement before input to the second LSTM layer. This way, the FT-JNF can potentially exploit all three sources of information.

C. Non-linear spatial filtering (T-NSF, F-NSF, FT-NSF)

To study the properties of a non-linear spatial filter (NSF) separately from the tempo-spectral processing, we define three additional variants of the the network architecture: T-NSF, F-NSF, and FT-NSF. The underlying idea is to prevent the network from employing fine-grained temporal and spectral information by randomly permuting the data along the sequence dimension before feeding it into the LSTM at position ①. The inverse permutation operation is then applied before the FF layer at position ③. Accordingly, only global statistics with respect to the frequency or time dimension are available but correlations between neighboring frequencies or time steps cannot be exploited. Preliminary experiments have shown that the spatial processing using a wide-band data arrangement (F-NSF) performs poorly if the frequency-bin index is unknown to the network. This is likely because the spatial characteristics of the data depend strongly on the frequency-bin index. To ensure that this information is still available after shuffling along the sequence dimension, we append the frequency-bin index to the feature dimension. We do this also for a narrow-band data arrangement (T-JNF) but in this case the effect on the performance is minor. Analogous to the procedure described in Section III-B, we define a non-linear spatial filter that incorporates both global spectral and global temporal information. This is achieved by again switching from wide-band to narrow-band data arrangement at position ② and requires both LSTM layers to be wrapped in permutation and inverse permutation operations with respect to the respective sequence dimension.

D. DNN-based post-filtering (PF)

Finally, we introduce a single-channel post-filter that jointly processes temporal and spectral information. For consistency, we stick to the simple base architecture shown in Figure 2. Here, the real and imaginary parts of the single-channel input data are stacked along the frequency dimension to form the feature dimension. The time axis is then used as sequence dimension.

E. Lossfunction and training details

We train the networks based on a complex ideal ratio mask (cIRM) [41] in favor of a magnitude ideal ratio mask (IRM) as in [15] to facilitate phase enhancement. For this reason, the FF layer is followed by a *tanh* activation function, which

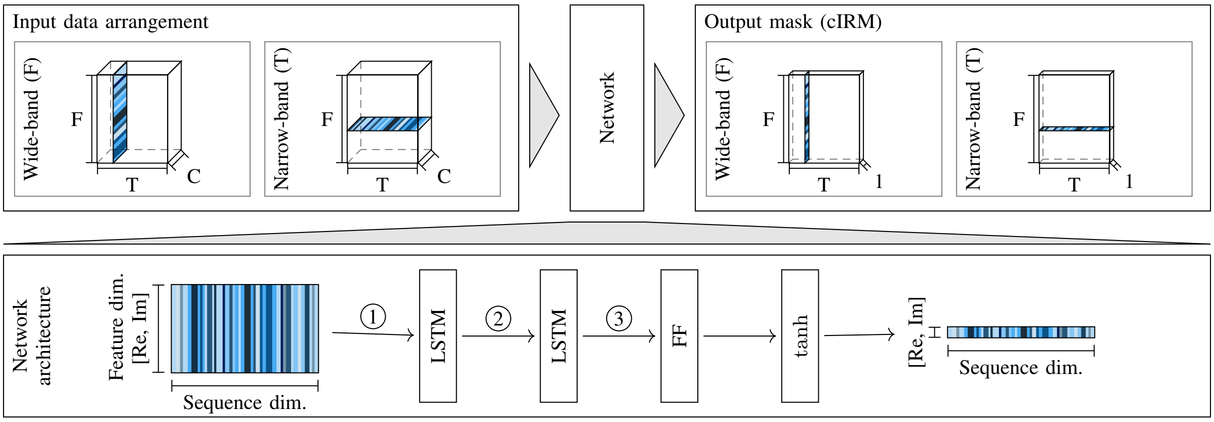


Fig. 2: Illustration of the base system architecture. The input data is arranged according to a wide-band or narrow-band input and fed into a network with two LSTM layers, an FF layer and tanh activation to obtain an estimate of a cIRM.

outputs a compressed mask estimate. We use compression parameters $K=C=1$ as defined in [41]. The enhanced signal is then obtained by multiplication of the uncompressed target speech cIRM $\mathcal{M}_{\mathrm{S}}(k,i)\in\mathbb{C}$ with the noisy recording $Y_0(k,i)$ using the first channel as reference, i.e.,

$$\hat{S}(k,i)=\mathcal{M}_{\mathrm{S}}(k,i)\cdot Y_0(k,i). \tag{6}$$

The real and imaginary parts of the noise cIRM $\mathcal{M}_{\mathrm{V}}$ can be obtained from the real and imaginary part of the target speech cIRM using [17]:

$$\mathrm{Re}(\mathcal{M}_{\mathrm{V}})=1-\mathrm{Re}(\mathcal{M}_{\mathrm{S}}), \tag{7}$$

$$\mathrm{Im}(\mathcal{M}_{\mathrm{V}})=-\mathrm{Im}(\mathcal{M}_{\mathrm{S}}). \tag{8}$$

The noise cIRM estimate can be used to obtain an estimate of the pure noise component contained in the signal, i.e,

$$\hat{V}(k,i)=\mathcal{M}_{\mathrm{V}}(k,i)\cdot Y_0(k,i). \tag{9}$$

We use the loss function proposed by Tolooshams et al. [17], which is composed of time and frequency domain ℓ_1 loss terms:

$$L(s,\hat{s})=\sum_{u\in\{s,v\}}\alpha\|u-\hat{u}\|_1+\left\||U|-|\hat{U}|\right\|_1. \tag{10}$$

Here, the frequency-domain terms $\hat{S}$ and $\hat{V}$ are estimated as given in (6) and (9) and time-domain quantities are obtained by an inverse STFT. We set $\alpha=10$ to equalize the contribution of either domain in the loss term.

As can be seen from the loss function, our training scheme uses the noisy observations $\mathbf{y}(t)$, which serves as network input, as well as the ground truth noise signals $\mathbf{v}(t)$ recorded at the microphones and the non-reverberant signal $s(t)$, which has been aligned with the noisy observation to include the propagation delay. If the ground truth for the noise signal is unknown, we only use the clean speech related parts of the loss function. During training, we randomly extract three seconds of audio from an utterance and compute the STFT using a 32 ms long window with 50% overlap. The $\sqrt{\text{Hann}}$ window is applied for analysis and synthesis. We train the networks with batch size six until convergence with maximum 250 epochs and select the best model with respect to the validation loss. The number of LSTM units is set to 256 and 128 for all networks, except PF, for which 256 units are used in both layers. The Adam optimizer [42] with learning rate 0.001 is used.

TABLE I: For each sample, the room characteristics are obtained by sampling uniformly from the value ranges given in this table.

Width	Length	Height	T60
2.5−5 m	3−9 m	2.2−3.5 m	0.2−0.5 s

IV. Analysis of the Interplay of Spatial with Tempo-Spectral Information

In this section, we evaluate the previously described networks in a speaker extraction scenario with a single speaker that is to be extracted and five additional interfering speech sources. Such a scenario seems particularly suitable to study the spatial filtering capabilities of a processing method since a spatially selective filter, as opposed to a filter that mainly exploits tempo-spectral information, is expected to be the key to good performance on this task. This is because the target signal has very similar tempo-spectral properties as the interfering signal (five speakers) but the signals differ decisively in their spatial properties.

A. Dataset

We generate a simulated dataset using `pyroomacoustics` [43], which provides an implementation of the source-image model [44]. The setup is illustrated in Figure 3. For each sample, the room dimensions and the reverberation time are uniformly sampled from the value ranges given in Table I. We use a circular microphone array with a diameter of 10 cm and between two and five channels. The microphone array is placed at a random position in the xy-plane but at least 1 m away from the walls, and it is located at a height of 1.5 m. As depicted in Figure 3, a rotation φ is applied to the microphone array sampled from the interval $[0,2\pi)$. In our setup, the target speaker has to be identified by its spatial location. Accordingly, we place the target speaker in a fixed position relative to the microphone orientation

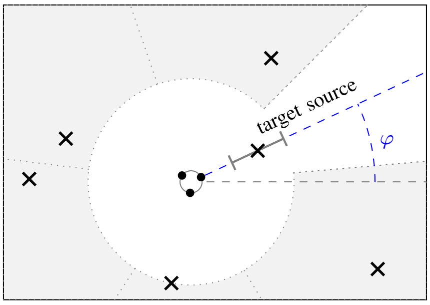


Fig. 3: Illustration of the simulation setup. The target source is located in a fixed orientation with respect to microphone array. The five interfering sources are placed in the gray area (one per segment). Room properties are sampled from the given ranges.

on the blue dotted line in Figure 3. Its distance to the microphone array ranges between 0.3 m and 1 m. The five interfering sources are placed in the gray area with a minimum distance of 1 m to the microphone array location. As indicated by the white area, a room spanned by the 20° angle to either side of the target source is also kept free of interferers. To ensure an even distribution of sources in the room, we place one interfering source per segment as indicated by the dashed gray lines. The height of the interfering speech sources is sampled from a normal distribution with mean 1.6 m and standard deviation 0.08.

We generate 6000, 1000, and 600 samples with a sampling frequency of 16 kHz for training, validation and testing respectively using clean speech signals from the WSJ0 dataset [45]. Signals between the different sets do not overlap. The signal-to-noise ratio (SNR) is not explicitly controlled but obtained from the the simulation setup with varying distances of the sources to the microphone array. The average SNR is -4 dB and 95% of the data samples distribute between -9 dB and 2 dB SNR.

B. Separability of spatial processing and post-filtering

Figure 1a illustrates the traditional two-step approach with a spatial filter that is applied first and a single-channel post-filter for tempo-spectral processing that is applied in a second processing step. Such a modular design is desirable as it offers flexibility and interpretability, however, the analytical MMSE solution in a non-Gaussian noise scenario is non-linear and non-separable [10]. The MMSE-optimal solution thus corresponds to the joint spatial and (tempo-)spectral filter depicted in Figure 1b. However, it is unclear if the third option of using a non-linear spatial filter as depicted in Figure 1c is a meaningful concept or if non-linear spatial processing is only useful if tempo-spectral information and spatial information are processed jointly as in Figure 1b. For this reason, in this section, we investigate if a DNN-based non-linear filter can be separated into spatial processing and single-channel tempo-spectral post-filtering by comparing the performance of all three configurations shown in Figure 1.

The left plot in Figure 4 shows the mean perceptual objective listening quality analysis (POLQA) score [46] and the 95% confidence interval for a varying number of microphones. The POLQA algorithm is the successor to the perceptual evaluation of speech quality (PESQ) measure [47]. It measures speech quality based on mean opinion score (MOS) scale ranging from one (bad) to five (excellent). The dashed lines correspond to spatial-only filters. That is the traditional MVDR beamformer (green) and the FT-NSF described in Section III-C (blue). The parameters of the MVDR beamformer are estimated from oracle data. We compute the time-varying noise covariance estimate by recursive averaging of the pure noise data and estimate the acoustic transfer function (ATF) by multiplying the principal eigenvector of the generalized eigenvalue problem for speech and noise covariance matrices with the speech covariance matrix as described in [48].

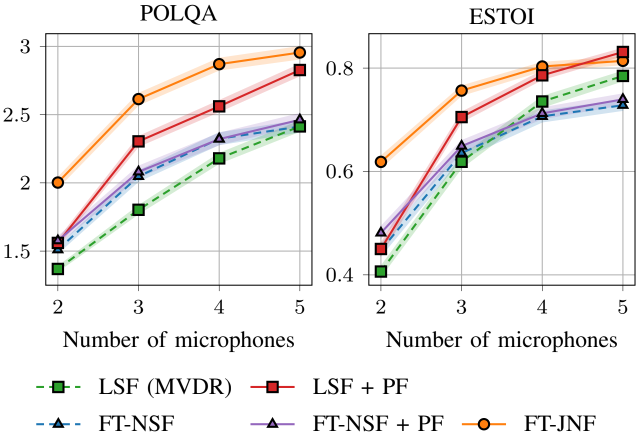


Fig. 4: We report the mean POLQA and ESTOI scores along with the 95% confidence interval for a set of multi-channel filters. This figure shows that joint spatial and tempo-spectral filtering (FT-JNF) outperforms a nonlinear spatial filter plus a postfilter (FT-NSF+PF).

Even though the MVDR parameters were accurately estimated from oracle data, which means that the MVDR should be considered as an upper bound on the spatial filtering performance achievable with a linear processing model, the non-linear spatial filter excluding a tempo-spectral post-filtering yields higher POLQA scores, in particular for a small number of microphones. A spectrogram visualization for three microphones is shown in Figure 5. The results obtained with a linear spatial filter (LSF) and a non-linear spatial filter (FT-NSF) are depicted in the middle row. Differences in the behavior are clearly visible: While the MVDR is distortionless by design at the cost of little noise suppression in this difficult noise scenario, the non-linear spatial filter aggressively reduces noise, but introduces quite some speech distortions. Please find audio examples on our website[1]. Next, we combine each spatial filter with an independent single-channel post-filter. For this, the DNN described in Section III-D is trained using the output of the MVDR and FT-NSF evaluated on the training set as network input. The results for these two-step approaches are shown in Figure 4 using the same marker as the corresponding spatial filter but with a solid line. We find that the post-filter added to the non-linear spatial filter (FT-NSF+PF) does not result in a notable performance

[1] https://uhh.de/inf-sp-deep-non-linear-filter

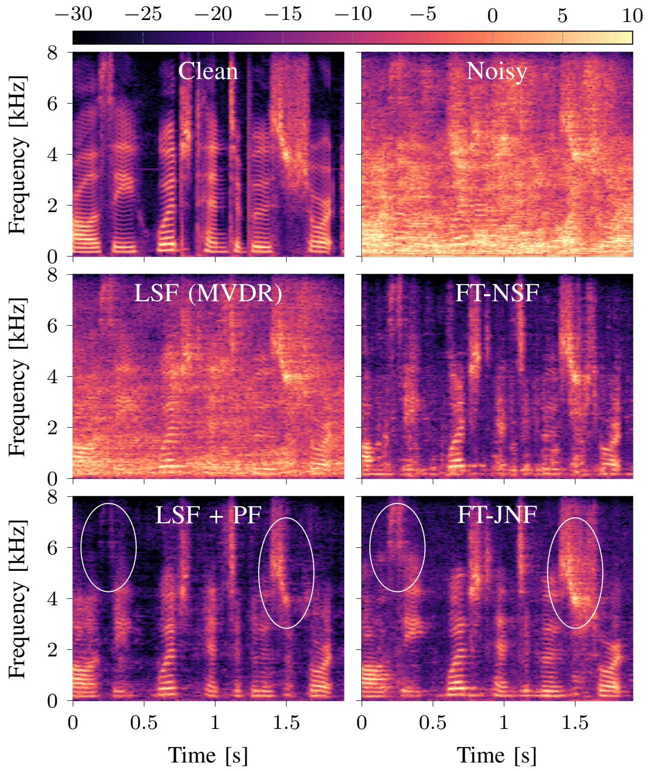


Fig. 5: Spectrogram visualization of an example utterance. The target signal and the noisy observation are displayed in the top row. The middle row shows two spatial filters, a linear MVDR on the left and a DNN-based non-linear on the right. The bottom depicts the MVDR with an independent post-filter and the joint spatial and tempo-spectral filter.

improvement. In effect, the purple line runs almost exactly on top of the blue dashed line. This can be explained by the fact that speech information, which was lost already during spatial processing, cannot be recovered by multiplication with the post-filter mask. In contrast, the MVDR beamformer does not distort the clean speech signal, and adding a single-channel post-filter, represented by the red solid line, is very effective. Here, we observe a performance boost between 0.18 POLQA score (two microphones) and 0.5 POLQA score (three microphones) in comparison with the linear spatial filter only.

Finally, we compare with the joint non-linear spatial and tempo-spectral filter FT-JNF. As visible in Figure 1b, the separation of spatial and tempo-spectral processing has been removed, which allows the network described in Section III-B to exploit the interdependencies between spatial and tempo-spectral information. This joint approach, depicted by the solid orange line, clearly outperforms the separated linear spatial filter plus post-filter approach for a low number of microphones. For two microphones the difference even amounts to 0.44 POLQA score. With an increased number of microphones, the gap between the orange line (FT-JNF) and the red line (LSF+PF) decreases or inverts even for ESTOI [49] depicted in the right plot. This is not very surprising as the number of anechoic point sources that can be canceled by the oracle MVDR increases by one for every added microphone. Accordingly, the performance of the spatial filter improves considerably with every microphone added, and when combined it with a strong post-filter, it becomes increasingly difficult to outperform the *oracle* MVDR plus post-filter with a data-driven filter.

TABLE II: Impact of different sources of information (spectral (F) and temporal (T)) used besides spatial information. We report mean improvements and the 95% confidence interval.

	Δ POLQA	ESTOI
F-NSF	0.78 ± 0.03	0.62 ± 0.012
T-NSF	0.46 ± 0.03	0.54 ± 0.013
FT-NSF	0.87 ± 0.03	0.64 ± 0.011
F-JNF	1.15 ± 0.04	0.70 ± 0.011
T-JNF [15]	0.74 ± 0.03	0.63 ± 0.012
FT-JNF (proposed)	**1.43** ± **0.04**	**0.76** ± **0.009**

Overall, two conclusions emerge from these results: First, the joint non-linear spatial and tempo-spectral filter (orange) drastically outperforms the non-linear spatial filter with an independent post-filter (purple) in terms of speech quality and intelligibility. This means that the dependencies between spatial and tempo-spectral information are successfully exploited by the neural network. And second, the DNN-based joint non-linear filter (FT-JNF) significantly outperforms the oracle MVDR with an added single-channel post-filter for a small number of microphones.

C. Interdependency of spatial processing with spectral and temporal information

The experiment in the previous section demonstrated that spatial processing should not be separated from tempo-spectral processing, as these two seem to mutually enrich each other. In this section, we will further investigate the interdependencies between spatial processing and temporal and spectral processing.

In the top three rows of Table II, we report the results obtained with a non-linear spatial filter that has access to global spectral, temporal or tempo-spectral information using three microphones. The corresponding neural network architectures have been explained in Section III-C. As expected, we observe that the highest performance is obtained with a non-linear spatial filter that incorporates both, global temporal and spectral information, denoted by FT-NSF. However, the comparison of F-NSF and T-NSF reveals that spatial processing here benefits much more from global spectral than global temporal information. The difference even amounts to 0.32 POLQA score and is also reflected in the ESTOI measurements. A similar pattern is also observed for the joint non-linear filter that can not only exploit global statistics but also fine-grained information including correlations between neighboring frequency bins and/or time steps. The performance differences between F-JNF and T-JNF amount to 0.41 POLQA score and 0.07 ESTOI score.

The impact of different sources of information on the spatial selectivity of the filter is visualized in Figure 6 in more detail. For this, we present the trained networks with a clean speech signal originating from varying directions with 1 m distance from the microphone array. For this experiment, we use a simulated anechoic room as we want to measure the

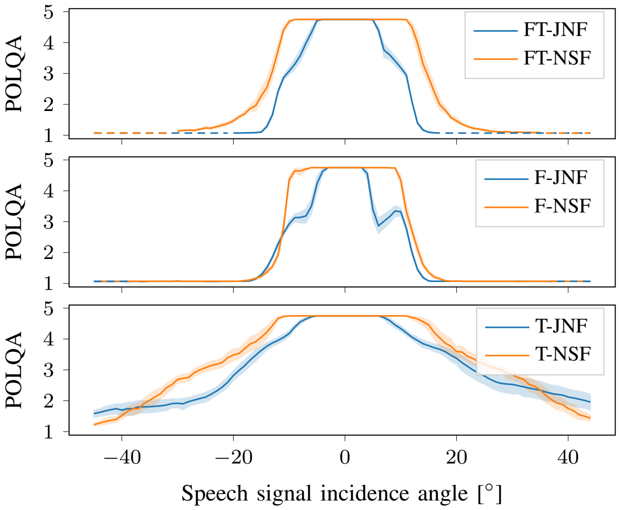


Fig. 6: Visualization of the spatial selectivity of the learned filters. The plots show the the mean POLQA score and 95% confidence interval for a clean and anechoic signal arriving from a given incidence angle. A low POLQA score here corresponds high suppression of the signal, while a very high POLQA score (around 0°) means that the signal has passed through the filter unaltered. Signals for which no POLQA score can be computed are marked with a dashed line.

filter's response to a signal from a specific direction. The plots in Figure 6 show the POLQA score for the filtered signals averaged over 15 examples. A high POLQA score, which is attained by all filters near 0°, corresponds to a signal that has passed through the filter unaltered, while a low POLQA score indicates high suppression of the signal. The POLQA algorithm does not provide a result if the signal is not speech-like anymore and has very low energy. For these processed signals, which retain less then 0.1% of their original energy, we indicate high suppression with a dashed line at the minimum POLQA score.

Comparing the two bottom plots of Figure 6, it is clearly visible that exploiting frequency information as opposed to time information increases the spatial selectivity, which can serve as an explanation of the performance differences observed before. While all plots show a "distortionless" response for signals with an incidence angle between $-4°$ and $4°$, signals arriving from a larger angles are much less suppressed (resulting in a higher POLQA score) for the network using temporal information. In particular, even signals that arrive from the interference region are not fully suppressed. Furthermore, considering the upper two plots, it is interesting to observe that adding fine-grained spectral information in FT-JNF and F-JNF narrows down the spatial selectivity even beyond the $-20°$ and $20°$ angle that can be expected from the dataset configuration. Yet, a narrower selectivity pattern might be helpful to resolve the spatial characteristics in a noisy scenario.

V. Comparison to state-of-the-art methods

In Table II, the overall best performance is obtained with the joint non-linear filter FT-JNF that exploits tempo-spectral in addition to spatial information. Comparing to T-JNF which has originally been proposed by Li and Horaud [15], we find that our systematic evaluation of the interplay between spatial and temporal as well as spectral information leads to a drastic performance improvement of 0.69 POLQA score in a speaker extraction scenario.

A. Baselines

In this section, we compare the proposed FT-JNF with four additional baseline network architectures besides T-JNF. This ensures that the study we conducted with a rather simple network provides meaningful results also in comparison with recent and more elaborate state-of-the-art network architectures and it furthermore allows us to assess the question whether a network design inspired by a traditional filter-and-sum beamformer, e.g., [19], [20], [35], is likely to exhibit enhanced spatial filtering capabilities.

As our primary focus in this work is to better understand the consequences of architectural choices for implementing multichannel DNN-based filters, we train all baseline architectures following the same procedure outlined in Section III-E and using the loss function defined in (10). For most baselines, we use the code provided by the authors with the default parameter settings and focus our parameter search mostly on the learning rate. The selected values are given in Table III. It is likely that an extensive hyper-parameter tuning might lead to better results, but we nevertheless consider the results representative of their respective network architecture on the used dataset. Deviations from the training procedure or the settings described in the respective paper will be noted in the following. These are the baselines that we compare the proposed FT-JNF to:

- T-JNF: We consider the architecture T-JNF as an instance of the network proposed by Li and Horaud [15]. However, in order to facilitate phase processing, we have changed the network output from IRM to cIRM and also replaced the final output layer with a tanh layer accordingly.
- CRNN: We reimplement a variant of the convolutional recurrent neural network (CRNN) for mask estimation proposed by Chakrabarty and Habets [16]. The authors propose a convolutional neural network (CNN) for spatial feature extraction. For this small convolution kernels are used on the channel dimension such that a series of convolutional layers reduces the channel dimension to one. These spatial features are then processed with a bi-directional LSTM and fed into a FF layer to produce a mask. We use real and imaginary parts as input and estimate a cIRM.
- FaSNet+TAC: FasNet [50] is a time-domain approach mimicking a traditional filter-and-sum beamformer. The authors proposed an extension, denoted FaSNet+TAC [35], which enables variable microphone array configurations. As the authors report improved performance also for fixed array geometries, we choose to evaluate FaSNet+TAC on the speaker extraction dataset. We use the implementation provided by the authors.
- EaBNet: Li et al. [19] propose the Embedding and Beamforming Network (EaBNet). It uses a U-Net structure to estimate an embedding that incorporates

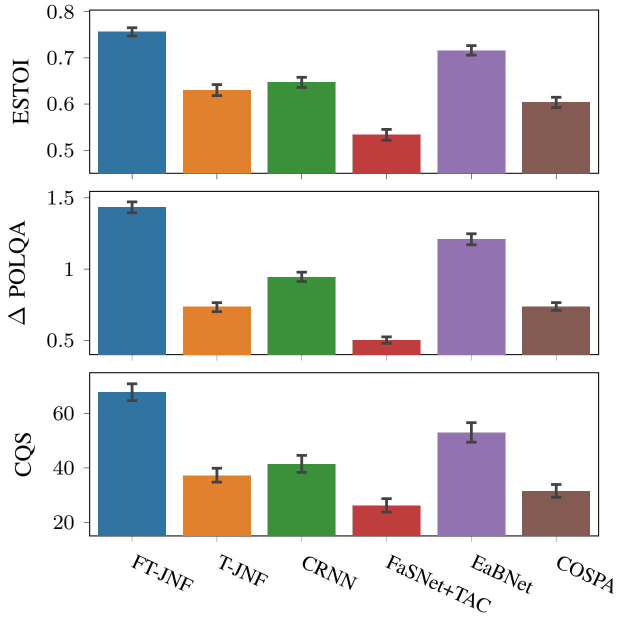


Fig. 7: Performance comparison of the proposed architecture FT-JNF and five baselines. The two upper plots show the mean ESTOI and POLQA performance on the speaker extraction dataset and the 95% confidence interval. The bottom plot shows the CQS results obtained by a MUSHRA listening experiment on twelve randomly selected examples.

- spatial and tempo-spectral information and then employs a "beamformer" network to obtain weights that are applied in a filter-and-sum beamforming manner. We use the implementation provided by the authors using the LSTM branch. We do not apply a single-channel DNN (post-filter network) to the output of EaBNet and use uncompressed network inputs and targets. This baseline uses shorter STFT windows of length 20 ms and 50% overlap.
- COSPA: The Complex-valued Spatial Autoencoder (COSPA) has been proposed by Halimeh and Kellermann [20]. Similar to EaBNet it adopts a filter-and-sum approach with frequency-domain complex-valued coefficients estimated by the network. The network architecture is composed of an encoder, a compandor and a decoder part. All of these are complex-valued networks. We use the implementation provided by the authors, which uses 64 ms long STFT windows and an overlap of 50%. We train using the clean speech terms in the loss function only.

B. *Performance analysis*

We train and evaluate all networks on the speaker extraction dataset. The results with respect to the POLQA improvements and ESTOI scores are displayed in the two upper plots of Figure 7. Here, we observe that the proposed FT-JNF consistently outperforms all other baselines by at least 0.22 POLQA score and 0.04 ESTOI score. In addition to using objective performance measures, we also conducted a MUSHRA [51] listening experiment with eleven participants using the webMUSHRA framework [52]. The participants have rated the overall quality of the algorithms based on twelve randomly sampled examples. The results are reported on a continuous quality scale (CQS) and presented in the bottom plot. The test results align very well with the objective measures and we find that FT-JNF performs best with a score of 67.9 outperforming EaBNet in second place with a score of 53.1. This is despite the fact that our proposed FT-JNF has the least number of learnable parameters. The number of parameters for each network architecture are given in Table III. It is apparent that the number of parameters is not the decisive factor for good performance here. Since all networks were trained in the same way (data, loss, optimizer etc.), we attribute the performance differences to the architectural choices of how to integrate different sources of information in the processing.

TABLE III: Baseline configurations.

	LR	STFT [ms]	#Param. [M]	Implementation / Github repository
FT-JNF	0.001	32	1.2	own (sp-uhh/deep-non-linear-filter)
T-JNF	0.001	32	1.2	own
CRNN	0.0001	32	17.4	own
FaSTAC	0.0001	–	4.1	ylou42/TAC
EaBNet	0.001	20	2.8	Andong-Li-speech/EaBNet
COSPA	0.0001	64	2.1	ModarHalimeh/COSPA

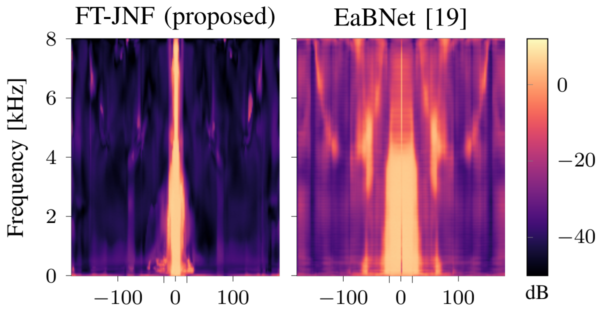


Fig. 8: Visualization of the spatial selectivity of the learned filters. The patterns are created by presenting white noise signals to the networks and averaging the resulting STFT signal along the time dimension for each incidence angle and converting to decibel.

While the architectures described in Section III as well as the CRNN adopt a mask-based approach, the baselines FasNet+TAC, EaBNet and COSPA resort to the filter-and-sum technique from traditional beamforming, where the filter weights are learned by the respective network. As the speaker extraction dataset is very challenging with low SNR and many interfering speech sources that have a similar tempo-spectral structure as the target signal, we can interpret the results in Figure 7 to reflect to a large extend the spatial selectivity of the DNN-based filters. Contrary to the common belief that a network design guided by the traditional beamforming paradigm is beneficial to spatial filtering capabilities, the best performance is obtained by FT-JNF that employs a mask-based approach, while the beamformer-inspired EaBNet only performs second best with an audible performance difference.

In order to investigate further the spatial selectivity of the different approaches, we perform an experiment similar to the one presented in Figure 6. Here, we present the trained networks

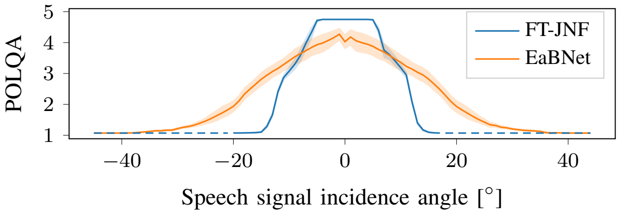


Fig. 9: Visualization of the spatial selectivity of the learned filters. The plots show the the mean POLQA score and 95% confidence interval for a clean and anechoic signal arriving from a given incidence angle.

FT-JNF and EaBNet with spectrally white noise signals originating from variable directions in an anechoic room. Clearly, these signals are out-of-distribution data for a network trained on speech mixtures. However, the spatial properties are still consistent with the ones the network has seen during training. Figure 8 displays the filters' response to these spatial cues. The incidence angle of the white noise signals is plotted on the x-axis. For each direction the STFT of eight filtered signals are averaged along the time axis. These white-noise response patterns seem to resemble the traditional directivity patterns [24, Sec. 12.5.2]. However, it must be noted that these white-noise response patterns do not allow for the same interpretation as a traditional beampattern. The reason for this is the non-linearity of the DNN-based filters. While a traditional beamformer, due to its linear nature, can in principle process all directional components of a signal separately and compose the final result after processing, this is not possible for a non-linear approach.

Bearing this in mind, the plots in Figure 8 nevertheless provide interesting insights into the spatial processing performed by the two networks. The FT-JNF shows a very clear spatial selectivity oriented towards the known position of the target source at zero degree. The width of the beam here coincides quite well with the two additional ticks at -20° and 20°, which mark the noise-free spatial section. On the other hand, the beam produced by EaBNet is much wider and suppression in the non-target direction does not work as well in particularly for high frequencies. What is is also noticeable is that the pattern suggest that signals near zero degree are slightly low-pass filtered, while the signal originating from an exact zero degree angle is high-pass filtered to some extend.

This loss in overall signal quality is also visible in Figure 9 for EaBNet. Here, we repeat the previously described experiment, where we present clean speech signals from different directions as input to the network (Figure 6). Comparing the orange line representing EaBNet with the blue line for FT-JNF, we find that EaBNet reduces the quality of the clean speech signal even if it is presented from the target direction. Considering this and also the width of the beam in both figures, we conclude that the performance differences that we have found in Figure 7 are well explained by the spatial properties of the filters.

VI. Implications of training DNN-based multi-channel filters on CHiME3

Finally, we evaluate on the CHiME3 data [53], which has been recorded in four real-world noisy environments: a cafeteria, a bus, a pedestrian area and next to a busy street. This dataset is frequently used to train and evaluate DNN-based multi-channel algorithms. The recordings have been conducted with a six-channel microphone array attached to a tablet that is held by the recorded speaker.

A. *Dataset*

The T-JNF network proposed by Li and Horaud [15] has originally been trained on the CHiME3 data. The authors propose in [15] to create a simulated dataset, which combines the pure noise recordings provided in the CHiME3 dataset with clean booth recordings instead of artificially spatialized target signals. We use their data generation script to obtain 2400, 476, and 3251 utterances for training, validation and test respectively. The signals in the test set are mixed with a SNR in $\{-4{,}0{,}4{,}8\}$ dB and we use the last four channels for our experiments.

B. *Performance analysis*

First, we assess the interaction between spatial and spectral as well as temporal information also on the CHiME3 dataset. Therefore, in Table IV, we report the POLQA improvement scores for FT-JNF, F-JNF and T-JNF. As before and as expected, we find that the best performance is obtained by FT-JNF in the top row that can exploit all available sources of information, that is spatial, spectral and temporal information. However, a comparison with the bottom part of Table II showing results for the speaker extraction dataset reveals that the performance benefit of including spectral versus temporal information is reversed here. While a spatial-spectral filter performs better on the speaker extraction dataset, a spatial-temporal filter prevails on the CHiME3 dataset even though with a smaller performance gap. This behavior can be explained by considering the differences in the signal characteristics of the two datasets. While the speaker extraction dataset requires high spatial selectivity for good performance, which means that multi-channel processing is required, a single-channel filter performing tempo-spectral enhancement is expected to obtain solid results on the CHiME3 dataset. This is because the noise signals in the CHiME3 dataset have a tempo-spectral structure that is quite different from that of the target speech signal and are, in most cases, much more stationary.

Consistent with the results of Section IV-C, in Figure 10 we show that a spatio-temporal filter (T-JNF) has a substantially lower spatial selectivity than a spatio-spectral filter: The plots have been obtained by providing the network trained on the CHiME3 data with a clean speech input from a variable direction. For this, we simulate the CHiME3 microphone array in a room with a clean speech source in a variable position with 40 cm distance to the microphone array. Signal suppression (blue) or signal pass-through (yellow) are measured by POLQA scores. The centered yellow blob for F-JNF (right plot) corresponds to the position of target speech sources in

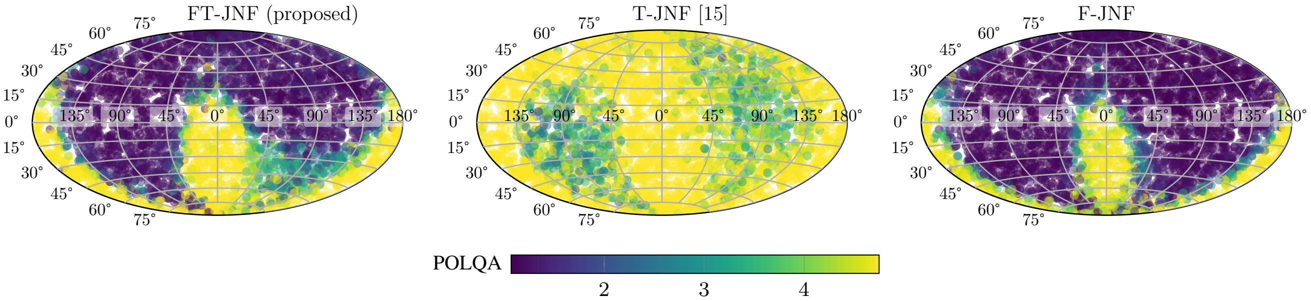


Fig. 10: Spatial selectivity maps for filters trained on the CHiME3 data. The scatter plots show the POLQA scores for a clean and anechoic signal arriving from a given incidence angle. Examples for which POLQA could not be computed because there is so little energy retained in the signal were assigned the minimum POLQA score. The data points, which lie on a sphere of radius 40 cm, are projected into the plane using the Hammer projection.

TABLE IV: POLQA improvement scores (mean and 95% confidence interval) for the proposed network architectures and baselines evaluated on the CHiME3 data.

	BUS	CAF	PED	STR
F-JNF	1.16±0.05	1.17±0.05	1.08±0.04	1.35±0.03
T-JNF	1.30±0.03	1.23±0.03	1.11±0.03	1.45±0.03
FT-JNF	**1.53±0.04**	**1.56±0.04**	**1.45±0.04**	**1.76±0.03**
CRNN	0.89±0.04	0.90±0.04	0.83±0.04	1.02±0.03
FaSNet+TAC	0.61±0.03	0.53±0.03	0.51±0.02	0.61±0.02
EaBNet	1.19±0.04	1.18±0.04	1.08±0.04	1.31±0.03
COSPA	0.60±0.03	0.61±0.03	0.56±0.03	0.65±0.03

the dataset. A speech source positioned at the origin represents a speaker that holds the recording tablet frontally at face level. Most speakers in the dataset however tilt the tablet to look at it a bit from above corresponding to a negative latitude value. The yellow blob at the left and right bottom edge shows that the filter cannot differentiate between signals impinging on the microphones attached to the tablet from front-side or back-side, which is expected for a planar microphone array. As the T-JNF has only little spatial selectivity but nevertheless obtains better performance than F-JNF, we conclude that temporal information, which is not reflected in this plot, plays an important role. However, based on the first spatial selectivity plot for FT-JNF, we find that this information can be incorporated without sacrificing a lot of the spatial selectivity, which gives a great performance boost of 0.23 POLQA score.

In addition, we draw two more general conclusions from the above analysis: First, the plots show clearly that the CHiME3 dataset resembles a scenario with a fixed (only slightly variable) target speaker position relative to the microphone array orientation. This is easily forgotten as the target speaker positions in the CHiME3 dataset are unknown. And second, we have seen that performance improvements observed for a joint multi-channel filter evaluated on the CHiME3 dataset can not directly be attributed to an improved spatial filtering, but that a much more detailed analysis is necessary to understand the internal functioning of such a filter.

Finally, we compare our proposed algorithm with the four additional baselines described in Section V-A. The results are presented in the bottom part of Table IV. The results are consistent with the performances reported on the speaker extraction dataset. Only T-JNF [15] improves in comparison with the other baselines and now slightly outperforms EaBNet [19]. Overall, we find that our proposed architecture FT-JNF, which has been designed to use all three sources of information, outperforms all other baselines regardless of the noise type.

VII. Conclusion

In this work, we have presented a detailed analysis of the internal mechanisms of a DNN-based filter for multi-channel speech enhancement. While traditional approaches combine a linear spatial filter with a separate tempo-spectral post-filter, DNN-based filters can potentially overcome the linear processing model and exploit interdependencies between spatial and tempo-spectral information. Here, we have shown that a non-linear spatial filter indeed outperforms an oracle MVDR on a challenging speaker extraction task with a low number of microphones. Furthermore, our analyses reveal that the interdependencies between spatial and spectral information can successfully be exploited by a DNN-based filter showing that additional spectral information increases the spatial selectivity of the filter. Our systematic review of this interplay of spatial and and tempo-spectral information leads to a simple network architecture with only two LSTM layers and a single feed-forward layer, that outperforms state-of-the-art network architectures for multi-channel speech enhancement by at least 0.22 POLQA score on the speaker extraction task and 0.32 POLQA score on the CHiME3 noise data.

Acknowledgment

We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

References

[1] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Audio, Speech, Language Proc.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] T. Lotter and P. Vary, “Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model,” *EURASIP J. Adv. Signal Proc.*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[3] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, pp. 1741–1752, 2007.

[4] M. Krawczyk-Becker and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 12, pp. 2251–2262, 2016.

[5] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *ISCA Interspeech*, Hyderabad, India, 2018, pp. 3229–3233.

[6] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for speech enhancement," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2019, pp. 249–253.

[7] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "DF-conformer: Integrated architecture of Conv-Tasnet and Conformer using linear complexity self-attention for speech enhancement," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2021, pp. 161–165.

[8] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 6633–6637.

[9] K. Tesch, R. Rehr, and T. Gerkmann, "On nonlinear spatial filtering in multichannel speech enhancement," in *ISCA Interspeech*, Graz, Austria, 2019, pp. 91–95.

[10] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1795–1805, 2021.

[11] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *IEEE Workshop Autom. Speech Recog. and Underst. (ASRU)*, Scottsdale, USA, 2015, pp. 444–451.

[12] M. Togami, "Multi-channel Itakura Saito distance minimization with deep neural network," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brighton, UK, 2019, pp. 536–540.

[13] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 6717–6721.

[14] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, 2016, pp. 5745–5749.

[15] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2019, pp. 298–302.

[16] S. Chakrabarty and E. A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[17] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense U-net for multichannel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 836–840.

[18] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1778–1787, 2020.

[19] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, 2022, pp. 6487–6491.

[20] M. M. Halimeh and W. Kellermann, "Complex-valued spatial autoencoders for multichannel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, 2022, pp. 261–265.

[21] K. Tesch, N.-H. Mohrmann, and T. Gerkmann, "On the role of spatial, spectral, and temporal processing for DNN-based non-linear multi-channel speech enhancement," in *ISCA Interspeech*, 2022. [Online]. Available: https://arxiv.org/abs/2206.11181

[22] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 30, pp. 605–621, 2022.

[23] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Proc. Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[24] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Chichester, England Hoboken, NJ: John Wiley, 2006.

[25] R. Balan and J. P. Rosca, "Microphone Array Speech Enhancement by Bayesian Estimation of Spectral Amplitude and Phase," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, Rosslyn, Virginia, 2002, pp. 209–213.

[26] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 39–60.

[27] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On optimal multichannel mean-squared error estimators for speech enhancement," *IEEE Signal Proc. Letters*, vol. 16, pp. 885–888, Oct. 2009.

[28] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brisbane, Australia, 2015, pp. 116–120.

[29] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brisbane, Australia, 2018, pp. 1–5.

[30] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *ISCA Interspeech*, 2019, pp. 4290–4294.

[31] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 2, pp. 457–468, 2019.

[32] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning mvdr beamformer for target speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 6089–6093.

[33] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized Spatio-Temporal RNN Beamformer for Target Speech Separation," in *ISCA Interspeech*, Brno, Czech Republic, 2021, pp. 3076–3080.

[34] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1888–1900, 2020.

[35] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 6394–6398.

[36] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," in *ISCA Interspeech*, Graz, Austria, 2019, pp. 86–90.

[37] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Sigapore, May 2022.

[38] D. Lee, S. Kim, and J.-W. Choi, "Inter-channel Conv-Tasnet for multichannel speech enhancement," *arXiv preprint:2111.04312*, 2021. [Online]. Available: https://arxiv.org/abs/2111.04312

[39] X. Li and R. Horaud, "Narrow-band Deep Filtering for Multichannel Speech Enhancement," *arXiv preprint arXiv:1911.10791*, 2019. [Online]. Available: http://arxiv.org/abs/1911.10791

[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[41] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 3, pp. 483–492, 2016.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015.

[43] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 351–355.

[44] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[45] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," Linguistic Data Consortium, Philadelphia, May 2007.

[46] "P.863: Perceptual objective listening quality prediction," International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[47] "P.862.3: Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," International Telecommunication Union, Nov. 2007, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.3-200711-I/en

[48] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, USA, 2017, pp. 681–685.

[49] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[50] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *IEEE Workshop Autom. Speech Recog. and Underst. (ASRU)*, Sentosa, Singapore, 2019, pp. 260–267.

[51] "BS.1534 : Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Oct. 2015, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534

[52] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA – A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[53] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop Autom. Speech Recog. and Underst. (ASRU)*, Scottsdale, USA, 2015, pp. 504–511.

5 Publication 3: Multi-channel Speech Separation Using Spatially Selective Deep Non-linear Filters [P7]

Abstract

In a multi-channel separation task with multiple speakers, we aim to recover all individual speech signals from the mixture. In contrast to single-channel approaches, which rely on the different spectro-temporal characteristics of the speech signals, multi-channel approaches should additionally utilize the different spatial locations of the sources for a more powerful separation especially when the number of sources increases. To enhance the spatial processing in a multi-channel source separation scenario, in this work, we propose a deep neural network (DNN) based spatially selective filter (SSF) that can be spatially steered to extract the speaker of interest by initializing a recurrent neural network layer with the target direction. We compare the proposed SSF with a common end-to-end direct separation (DS) approach trained using utterance-wise permutation invariant training (PIT), which only implicitly learns to perform spatial filtering. We show that the SSF has a clear advantage over a DS approach with the same underlying network architecture when there are more than two speakers in the mixture, which can be attributed to a better use of the spatial information. Furthermore, we find that the SSF generalizes much better to additional noise sources that were not seen during training and to scenarios with speakers positioned at a similar angle.

Reference

Kristina Tesch and Timo Gerkmann, “Multi-channel Speech Separation Using Spatially Selective Deep Non-linear Filters”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 542-553, 2024. DOI: 10.1109/TASLP.2023.3334101.

Copyright notice

Multi-channel Speech Separation Using Spatially Selective Deep Non-linear Filters

Kristina Tesch, *Student Member, IEEE*, and Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—In a multi-channel separation task with multiple speakers, we aim to recover all individual speech signals from the mixture. In contrast to single-channel approaches, which rely on the different spectro-temporal characteristics of the speech signals, multi-channel approaches should additionally utilize the different spatial locations of the sources for a more powerful separation especially when the number of sources increases. To enhance the spatial processing in a multi-channel source separation scenario, in this work, we propose a deep neural network (DNN) based spatially selective filter (SSF) that can be spatially steered to extract the speaker of interest by initializing a recurrent neural network layer with the target direction. We compare the proposed SSF with a common end-to-end direct separation (DS) approach trained using utterance-wise permutation invariant training (PIT), which only implicitly learns to perform spatial filtering. We show that the SSF has a clear advantage over a DS approach with the same underlying network architecture when there are more than two speakers in the mixture, which can be attributed to a better use of the spatial information. Furthermore, we find that the SSF generalizes much better to additional noise sources that were not seen during training and to scenarios with speakers positioned at a similar angle.**

***Index Terms*—Multi-channel, speech separation, DNN-based, spatially selective filter (SSF)**

I. Introduction

Speech separation algorithms target the so-called cocktail party problem, where several (two or more) human speakers are speaking at the same time. The goal is to recover the original speech signals from a mixture recording that may also contain additional background noise and reverberation. This task is particularly challenging because all target speech signals have similar tempo-spectral characteristics. But nevertheless, normal-hearing people are very good at focusing their attention on a single target speaker, so that they can even enjoy a conversation at a cocktail party. This ability is mainly due to the fact that humans have two ears, which enables them to perceive and process spatial information. Similarly, also speech processing algorithms can leverage spatial information by using multiple microphones to record the mixture signals.

The most traditional form of spatial processing is to employ a linear spatial filter, a so-called beamformer, which is designed to enhance the signal arriving from a target direction by suppressing signal components that arrive from a direction other than the target direction. Two prominent examples are the Delay-and-Sum beamformer and the minimum variance distortionless response (MVDR) beamformer [1]. Both of these employ a linear processing model: first, the individual microphone signals are filtered, and then added. The underlying idea for the Delay-and-Sum beamformer is to compensate for the relative time differences of arrival (TdoAs) of the signal at the microphone channels in the filtering step. Therefore, accurate TdoA or related direction of arrival (DoA) estimates are required. The MVDR beamformer additionally takes the second-order statistics of the interfering signal into account so that it can form a superdirective beamformer or steer nulls in the direction of interfering point sources. However, the number of point-sources that can be eliminated is bounded by the number of microphone channels minus one [2, Sec. 6.3]. Consequently, the MVDR beamformer deteriorates in a reverberant setting as reflections of the interfering sources arrive from all directions. As the performance of these linear spatial filters is limited, a single-channel post-filter is commonly applied to the output of the linear spatial filter.

Linear spatial filters have been, and still are, a popular choice for source separation problems because of their ability to focus on a single target source based on its spatial characteristics [3]–[15]. The challenging part here is accurate parameter estimation: a data-dependent implementation of the MVDR requires the localization of speakers or a direct estimation of the relative transfer function (RTF) as well as an estimation of the interfering signal's covariance matrix. While older works employed statistical modeling, e.g., [5], [7], recent ones rely on neural networks for this purpose. In the so-called masked-based beamforming approach, a neural network is used to estimate time-frequency masks for each speaker and use these to obtain the target speech and interfering noise signal's covariance matrix, e.g., [10], [11], [13], [16]. Other researchers suggest to sample the space with a fixed beamformer and use a neural network for beam selection and post-filtering [12], [14], [15].

However, with the rise of the neural network era, there is also an increasing number of multi-channel speech separation approaches that do not perform explicit spatial filtering. Instead, the neural networks are presented with multi-channel inputs and/or directional features and are trained to estimate the speech sources directly from the mixture, e.g., [17]–[20]. Throughout this work, we will refer to these end-to-end regression-based systems as direct separation (DS) approaches. Typically, these systems output as many speech signals as there are speech sources in the mixture, which gives rise to a permutation problem. For this reason, most DS approaches are trained with an utterance-wise permutation invariant training (PIT) loss. Unlike in the case of using a beamformer for spatial filtering, the spatial processing takes place only implicitly in the DS networks. It is clear, however, that maximum separation performance can only be reached if such a network learns to

The authors are with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany (e-mail: kristina.tesch@uni-hamburg.de; timo.gerkmann@uni-hamburg.de).

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 508337379.

perform powerful spatial processing directly from training examples.

While the traditional spatial filters are constraint by a linear processing model, the nature of neural networks enables non-linear spatial processing and, furthermore, an integration of the spatial and tempo-spectral processing steps, which are separated in a traditional beamformer plus tempo-spectral post-filter setup. In a previous analysis based on statistical minimum mean square error (MMSE) estimators [21], we have shown that this indeed leads to more powerful spatial processing in non-Gaussian interferences, which is arguably always the case in speech separation. As a consequence, the upper bound for the number of sources that can be canceled by a linear spatial filter does not hold anymore if the filter is non-linear and jointly performs spatial and tempo-spectral processing. In further experiments with DNN-based joint non-linear spatial and tempo-spectral filters, we have confirmed that neural networks can implement spatial filters that drastically outperform an oracle MVDR plus additional DNN-driven post-filter [22]. Accordingly, DNN-based DS approaches can offer a potentially better spatial filtering than a traditional linear spatial filter. On the other hand, these networks have to learn the spatial processing implicitly from data. How well this is accomplished can only be determined indirectly. Since the DS approaches are less modular than the spatial filtering approach that separates parameter estimation, e.g., DoA estimation, and spatial filtering, they are also less flexible with respect to, for example, a variable number of sources. In this paper, we investigate the separation performance of DNN-based non-linear joint spatial and tempo-spectral filters, which have been trained according to two different strategies: (1) using PIT, which means that the spatial filtering must be learned implicitly from the provided examples and (2) with an explicit focus on the spatial filtering by steering a filter towards a target speaker with a given DoA. A filter obtained by using the second strategy is referred to as a spatially selective filter (SSF) in this work.

Many researchers have proposed to enhance multi-channel speech separation with so-called directional features [17], [23] or use location-information to guide speaker extraction tasks [18], [24], [25]. For example, Gu et al. and [24] and Chen et al. [25] proposed an angle feature indicating which time-frequency bins are dominated by a signal from a particular DoA. Other common features are related to the inter-channel phase differences (IPDs) of the microphone pairs, cross-correlation features or features computed with fixed beamformers, e.g., a Delay-and-Sum beamformer steered to a set of candidate locations [12], [25]. In contrast, in our proposed approach, we do not rely on hand-crafted features, but use a neural network to learn the spatial processing from raw multi-channel data.

In our recent ICASSP 2023 paper [26], we have introduced a conditioning mechanism to flexibly steer a DNN-based non-linear spatial filter in a desired target direction. Given the noisy mixture and the target look-direction of the filter, the SSF then extracts the speech signal corresponding to the speaker located in that direction, similar to traditional linear spatial filters. This ability to flexibly steer the filter in a desired direction is a major improvement over a filter with a fixed look-direction [22], [27], or with fixed spatial target regions [28], [29]. The conditioning mechanism we proposed in [26] does not need a steering vector like a classic linear beamformer or the related work by Jenrungrot et al. [30] but is conditioned on the one-hot encoded angle. This avoids an implicit far-field assumption and leads to better performance, as we showed in [26]. Similarly, Kindt et al. [31] have shown that a learned encoding based on a one-hot encoded angle used as a feature to improve separation of closely spaced speakers is more valuable than a hand-crafted feature based on expected phase differences.

In this paper, we extend our previous work and investigate the use of SSFs for speech separation. We aim to understand if the explicit spatial filtering in SSFs is advantageous over the implicit spatial filtering learned by the widely adopted DS approach in terms of overall performance, but also in terms of generalization ability to conditions unseen during training. Furthermore, we investigate the robustness of the SSF to errors in the DoA input as well as pertubations in the microphone array geometry.

The rest of the paper is structured as follows: In the next section, we give a formal problem description and introduce the notation. In Section III-A, we describe two neural network architectures for joint spatial and tempo-spectral non-linear filtering, which we use to compare a DS and SSF approach using the same underlying network architecture. In addition, we explain the steering mechanism of the SSFs in Section III-B. Section IV describes the dataset generation, and in Section V, we compare the speech separation performance of the two approaches. Detailed investigations on the robustness and generalization ability are presented in Section VI and Section VII.

II. Problem Definition

In this work, we consider a multi-channel reverberant speech separation scenario. The goal is to recover the speech signals uttered by P concurrently speaking persons in a reverberant room. The mixture signal is recorded by an omni-directional microphone array with C channels. We denote the dry speech signal of the p's speaker by $s_p(t)$ with time-index t. The recording of $s_p(t)$ at the ℓ's microphone includes not only a time-shift due to the propagation delay but also reflections on the walls and is denoted by $x_p^\ell(t)$. Given the room impulse response (RIR) $h_p^\ell(t)$ describing the propagation path of the signal uttered by the p speaker to the ℓ's microphone, the dry and recorded signal are related via a convolution operation, i.e.,

$$x_p^\ell(t) = s_p(t) * h_p^\ell(t). \tag{1}$$

Using the short-time Fourier transform (STFT), we transform the time-domain signals $x_p^\ell(t)$ into their complex-valued frequency-domain representations $X_p^\ell(k,i)$ with frequency-bin index k and time-frame index i. The letters F and T denote the total number of frequency bins and time frames respectively. Following the additive signal model, the observed mixture signal is then given by

$$Y^\ell(k,i) = \sum_{p=1}^{P} X_p^\ell(k,i) + V^\ell(k,i), \tag{2}$$

where $V^\ell(k,i)$ denotes the sensor and environmental noise possibly recorded at the ℓ's microphone in addition to the speech signals. Given the mixture signal, the task now is to recover the original speech signals $S_p(k,i)$ or, equivalently $s_p(t)$, except the propagation delay caused by the length of the direct path between source and microphone array. Accordingly, we use the direct-path dry speech signals as training target and to compute metrics that require a reference signal.

III. Spatially Selective Non-Linear Filter (SSF)

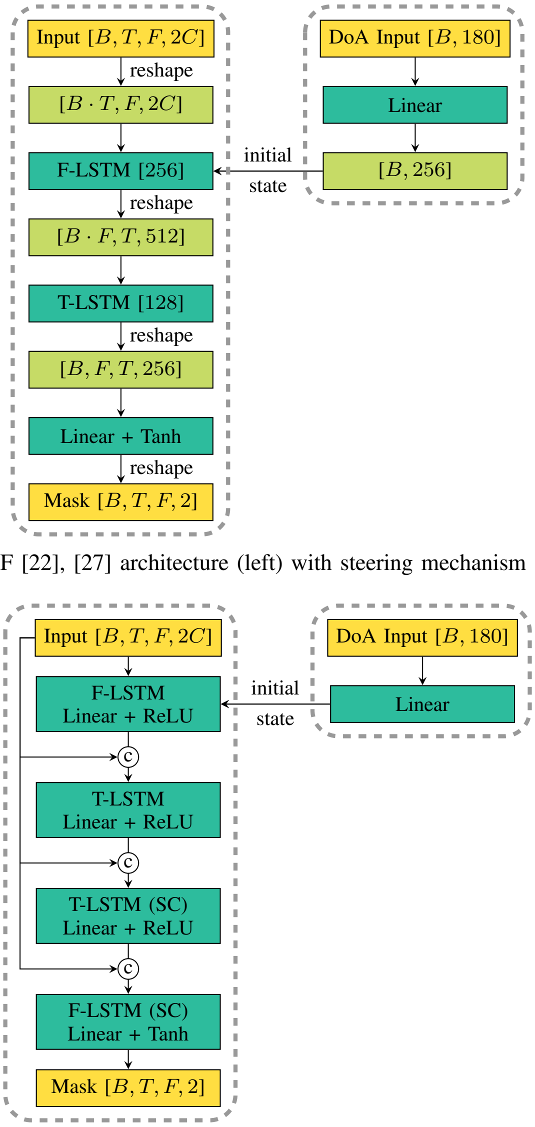
In this work, we investigate the use of a spatially selective deep non-linear filter (SSF) for multi-channel speech separation. Our proposed method is in line with the common approach to separate the localization task from the actual speaker extraction, which is then performed in a second step using a spatial filter steered to the speaker locations. In this section, we describe two network architectures for joint spatial and tempo-spectral non-linear filtering (JNF [22], [27] and McNet [20]) and explain the proposed mechanism for flexibly steering the filter in the desired target direction.

A. Network architectures for joint spatial and tempo-spectral non-linear filtering

In our prior work [22], [27], we have proposed the FT-JNF architecture, which we refer to as JNF in this work. The network architecture is depicted on the left side of Figure 1a. The JNF network expects a three-dimensional frequency-domain input. The yellow box on the top left of Figure 1a visualizes the input of the filter including the batch dimension denoted by B. The last dimension expects the real and imaginary part of all C channels stacked into a vector of length $2C$. The multi-channel input provides three sources of information, which should be exploited by the network: spatial, spectral and temporal information. For this, we have previously proposed the depicted architecture with two LSTM layers at its core. The F-LSTM has been designed to extract features related to spatial and spectral information as well as their relationship, while excluding temporal correlations. This is achieved by a rearrangement of the data, which moves the time dimension into the batch dimension such that all time-frames are processed independently with network weights shared across all time-frames. The reshaping operations on the data are represented by the light green boxes in Figure 1a. The correlations along the time axis are then processed by the T-LSTM layer, which performs independent processing of all frequency-bins. The design has been inspired by the work of Li and Horaud [32], who propose a network that stacks two T-LSTM layers. However, the replacement of the first T-LSTM by the F-LSTM significantly enhances the spatial selectivity of the resulting filter such that the JNF outperformed other state-of-the-art methods on a speaker extraction task in [22].

The JNF outputs a compressed complex-valued mask in the range $[-1,1]$, which is expanded into an uncompressed mask following the description in [33] with the steepness parameter set to one to obtain the single-channel time-frequency mask $\mathcal{M}_p(k,i)$ for the pth speaker located in the direction the filter was steered toward. The corresponding estimate $\hat{S}_p$ is given by multiplication of the mask with the reference channel of the noisy observation, i.e.,

$$\hat{S}_p(k,i) = Y^0(k,i) \cdot \mathcal{M}_p(k,i). \qquad (3)$$



(a) JNF [22], [27] architecture (left) with steering mechanism (right)

(b) McNet [34] architecture (left) with steering mechanism (right)

Fig. 1: Schematic view of a spatially selective filter (SSF) based on the JNF (top) and McNet (bottom) network architecture. The proposed conditioning on the target DoA is depicted on the right side.

The successful combination of different sources of information in the JNF architecture [22], has inspired Yang et al. [34] to improve it further by appending two more LSTM layers that are focused on the single-channel (SC) spectral correlations in time and frequency dimension. A schematic view of the resulting network architecture, named McNet, is shown in Figure 1b. Besides two additional LSTM layers, the authors have introduced skip connections and add additional feed-forward layers after every LSTM layer. The first skip connection concatenates the noisy multi-channel signal to the input of the T-LSTM and the second and third skip connection concatenates the noisy magnitude of the reference channel to the input of the two single-channel LSTM layers. Please refer to [34] for a more detailed illustration, which also includes the reshaping steps for McNet. For all experiments, we use the default configuration of McNet. Since the steering mechanism, proposed in the next section, targets the first F-LSTM layer, which is the same in both networks, we can steer both DNN-based filters in the same way and perform experiments with a spatially selective filter based on the JNF architecture (JNF-SSF) and based on the McNet architecture (McNet-SSF).

B. Proposed steering of the non-linear spatial filter (conditioning on target direction)

In addition to the multichannel signal input, the proposed SSF requires the steering direction as a second input, as shown on the right side of Figure 1a and Figure 1b. The direction information is presented to the network as a one-hot encoded vector, whose dimension depends on the chosen angular resolution. Figure 1 illustrates a 2° angle resolution, which leads to 180 possible input vectors. The one-hot vector is then fed into a linear layer, which provides an encoding of the direction information that matches with the number of units in the F-LSTM layer, which we set to 256 in the JNF architecture. The encoded DoA information is then used to initialize the forward and backward initial states of the bi-directional F-LSTM layer. A similar conditioning mechanism has also been used by Vinyals et al. [35] to initialize a network for image caption generation with information about the image.

In contrast to our previous paper [26], we only initialize the first F-LSTM layer with the direction information and omit this step for the second T-LSTM layer. Preliminary experiments have shown that conditioning the first LSTM layer leads to much better performance than conditioning the second LSTM. Furthermore, we observed that conditioning both layers does not provide a benefit but slightly increases the computational demands. These findings are in line with our previous observations in [22] that the spatial selectivity is mainly controlled by the F-LSTM layer.

In [26], we compared the proposed conditioning mechanism based on a one-hot angle encoding to the method suggested by Jenrungrot et al. [30]: using knowledge of the microphone array geometry and based on a far-field assumption, the individual input channels are shifted such that signals arriving from the target DoA are time-aligned. These time-aligned signals are then used as input signal of the network. We found that the target speaker is reliably extracted with this competing method, but the proposed conditioning on the encoded DoA angle consistently performs better and does not rely on a far-field assumption.

IV. Dataset Generation

Using `pyroomacoustics` [36], we generate a simulated dataset for training and evaluation based on the image-source method [37]. An illustration of the geometric setup is given in Figure 2. All rooms have a rectangular shape with their dimensions and reverberation characteristics, described by the T_{60} reverberation time, uniformly sampled from the ranges given in Table I. We use a circular microphone array with three omni-directional microphones. The diameter of the microphone array is 10 cm. With respect to the x and y axis, we position the microphone array randomly in the room, however with a minimum distance of 1.2 m to the walls. The height of the array is fixed at 1.5 m. As illustrated in Figure 2, the microphone array rotation is denoted by $\varphi_m \in [0, 2\pi)$.

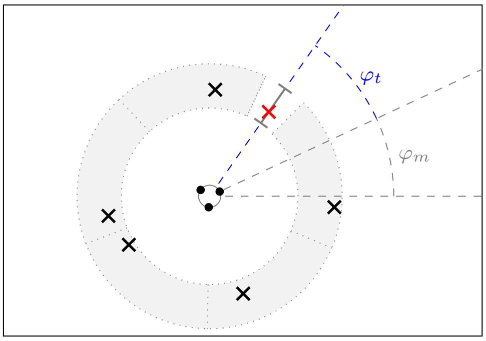


Fig. 2: Illustration of the dataset generation. The target source is marked with a red cross and its DoA angle φ_t is computed relative to the microphone orientation in the room given by φ_m. Interfering sources are placed in the gray area.

TABLE I: The room characteristics are sampled uniformly from the displayed ranges.

Width	Length	Height	T_{60}
$2.5-5$ m	$3-9$ m	$2.2-3.5$ m	$0.2-0.5$ s

During training, the spatially selective filter learns to extract a single target speaker from the mixture given its angle φ_t. The target speaker is represented by a red cross in Figure 2. Its corresponding DoA angle φ_t is measured with respect to the microphone orientation as indicated by the blue dashed line. Interfering speakers are placed in the gray area at a minimum distance of 0.8 m and a maximum distance of 1.2 m just like the target speaker itself. We leave a 10° space with no interfering speakers around the target speaker as indicated by the white space in Figure 2. The area of the gray annulus is divided into equally spaced segments as indicated by the gray dotted lines and one interfering speaker is randomly placed per segment, also with a minimum angular distance of 10° to speakers in a neighboring segment. In Figure 2, the interfering speakers are marked by a black cross. The height of the speakers is sampled from a normal distribution with mean 1.6 m and standard deviation 0.08 m.

For training the spatially selective filter, we use a setup with five interfering speakers as shown in the Figure 2. We discretize the target speaker location φ_t using a 2° resolution, which results in 180 target speaker directions and provide 300 examples per direction. This results in a total of 54, 000 training examples.

For training or testing on a speech separation task, we do not change any simulation parameter, but may vary the number of “interfering speakers”, which are then also considered as additional target speakers. For validation and testing, we generate $2,700$ and $1,800$ examples respectively. The dry clean speech utterances are taken from the WSJ0 dataset [38], with no overlap between training, validation and test datasets. The sampling rate is 16 kHz. The average direct-to-reverberation ratio (DRR) for each individual speaker's signal is -0.8 dB and 95 % of the samples lie in the interval $[-5.9,\ 4.8]$ dB. For a separation scenario, we can characterize the distribution of the signal-to-noise ratio (SNR) with respect to all included speaker extraction tasks considering one speaker as target signal and the mixture of interfering speakers as noise. In our setup, the SNR is mainly influenced by the number of interfering speakers and the distance of the speakers to the microphone array. For two speakers, the SNR of the extraction tasks range between $[-9.4,\ 9.4]$ dB for 95% of the data. For three and five speakers, the separation problem gets more difficult as the SNR ranges shift to $[-11.8,\ 4.9]$ dB and $[-14.5,\ 0.5]$ dB respectively.

V. Evaluation of the Separation Performance

A well-performing multi-channel speech separation system can be expected to benefit from knowledge of the speakers' locations. This assumption has led researchers to propose a variety of direction-based features, which are used as additional inputs to enhance DNN-based speech separation [17], [18], [23]. While a localization might happen implicitly in a regression-based direct separation (DS) approach, the spatially selective filter (SSF) separates the localization from the speaker extraction task and puts a strong focus on the spatial properties as the networks learns to focus on a single speaker using its direction as cue. In this section, we aim to investigate the impact of the chosen method, DS or SSF, on the separation performance.

To ensure a fair comparison, we use the same underlying network structure, JNF and McNet as described in Section III-A. For the DS approach, we omit the conditioning mechanism shown on the right side of Figure 1a and Figure 1b and only provide the multi-channel mixture STFT as input. The output dimension of the last layer is changed so that not only one mask is predicted but as many masks as there are speakers in the mixture. We assume that the number of speakers is known. The network then produces an estimate for every speaker, and we use the same ℓ_1 loss in time and frequency domain as for the SSF, but we apply it in a PIT [39] scheme. The loss function and other DNN training settings are described in Appendix A.

In Table II, we report the separation performance for mixtures of two, three and five speaker mixtures measured using the perceptual objective listening quality analysis (POLQA) score [40] improvement, scale-invariant source-to-distortion ratio (SI-SDR) [41] improvement and DNSMOS [42] score. The POLQA and DNSMOS score predict values on a mean opinion score (MOS) score ranging from one (bad) to five (excellent). In contrast to POLQA and SI-SDR, DNSMOS does not require a reference signal, but is a neural network trained on user ratings according to the P.835 standard. We report the overall quality rating.

Row number 1 displays the results for the DS networks based on the JNF architecture and trained with PIT. Separate networks have been trained for the different numbers of speakers. In contrast, all results in row 2 have been obtained with the same network implementing the SSF approach based on the JNF architecture and evaluated given oracle DoA information for the individual speakers. As the speakers are likely not positioned on the 2° grid that has been used during training, we map the oracle target speaker direction onto the closest point in the grid before computing the one-hot encoding to condition the network as described in Section III-B. A comparison of the first two rows of Table II reveals that the JNF-SSF outperforms the JNF-DS approach in all metrics and for all numbers of speakers in the mixture. Furthermore, we observe that the performance difference is larger for a higher number of speakers in the mixture. For example, the POLQA performance difference between JNF-DS and JNF-SSF increases from 0.21 for two speakers to 0.43 for three and five speakers.

Both networks JNF-DS and JNF-SSF used for the comparison in this table have approximately the same number of parameters. However, since the SSF is evaluated multiple times with each speaker as the target, the DS approach has a smaller number of learnable parameters per speaker. To investigate the influence of this effect, we also train JNF-DS with an increased number of parameters. For this we scale both LSTM layers by the same factor. The F-LSTM then has 364, 448 and 576 units for two, three and five speakers. The results are shown in Table III. Comparing with row 1 in Table II, we find that increasing the network size improves the performance of JNF-DS. However, as can be seen from the second block of Table III, the JNF-SSF still outperforms JNF-DS. We therefore conclude that the superiority of the SSF approach can not be solely explained by an increased amount of parameters per speaker.

As expected, the extension of the JNF architecture to McNet leads to a significant performance improvement in all metrics and for both the DS and SSF configuration. Comparing row 4 and 6, we find that the performance of McNet-SSF for two speaker mixtures is only slightly better than McNet-DS considering the POLQA score improvement and DNSMOS. It is even worse by 0.3 dB with respect to the SI-SDR measure. However, for more speakers the previously observed trend that the SSF output performs the DS network persists. For example, the performance difference amounts to 0.36 and 0.56 POLQA improvement score for three and five speakers, which is in line with a strong preference for the SSF result in a listening experiment as shown in Figure 3. In this listening experiment, which evaluates the listener's preference for the SSF or DS result. Ten test subjects have been asked to rate a statement regarding their preference for one of two separation results obtained with the SSF or DS method. We use eight examples for each of the four comparisons. Of course, the test has been conducted blindly without a labeling of the methods and a random order of the compared items. The statement to be rated has the form: "Example 1 is preferable over Example 2." We then aggregated the results to comply with the statement displayed in Figure 3. While the metric scores in Table II for two speaker mixtures are quite similar for McNet-DS and McNet-SFF, we observe a clear preference of the test subjects towards the SSF result. In total, more than 55%

TABLE II: Speech separation performance for reverberant mixtures of two, three and five speakers. We compare an approach based on a spatially selective filter (SSF) with a direct separation (DS) approach using the same network architecture: JNF [22], [27] or McNet [34].

No.		DoA	2 speakers			3 speakers			5 speakers		
			ΔPOLQA	ΔSI-SDR	DNSMOS	ΔPOLQA	ΔSI-SDR	DNSMOS	ΔPOLQA	ΔSI-SDR	DNSMOS
1	JNF-DS	–	1.20	11.7	2.80	0.87	11.5	2.46	0.53	10.7	2.11
2	JNF-SSF	oracle	1.41	12.7	2.94	1.30	14.2	2.79	0.96	15.1	2.52
3	JNF-SSF	search	1.40	12.6	2.94	1.29	13.9	2.78	0.93	14.4	2.51
4	McNet-DS	–	1.82	15.0	3.03	1.40	15.4	2.79	0.87	14.2	2.39
5	McNet-iDS	oracle	1.82	15.7	3.07	1.61	15.9	2.85	0.96	15.0	2.43
6	McNet-SSF	oracle	1.85	14.7	3.13	1.76	16.3	3.04	1.43	17.3	2.84
7	McNet-SSF	search	1.91	15.0	3.15	1.80	16.3	3.06	1.43	16.6	2.85
8	McNet-SSF	DNN	1.85	14.7	3.13	1.76	16.2	3.04	1.42	16.9	2.84
9	MVDR + PF	oracle	0.42	3.8	2.47	0.23	2.8	2.20	0.14	3.1	1.90
10	McNet-SSF (HCF)	oracle	1.49	11.6	2.90	1.38	12.6	2.78	1.03	12.4	2.53

TABLE III: Speech separation performance of JNF-DS with the number of network parameters scaled according to the number of speakers to extract and the performance of JNF-SSF with multiple evaluations of the same network.

	# Speakers	Param. [M]	ΔPOLQA	ΔSI-SDR	DNSMOS
JNF-DS	2	2.4	0.83	10.5	2.82
	3	3.6	1.07	13.2	2.62
	5	6.0	0.70	12.7	2.28
JNF-SSF	2	1.2×2	1.41	12.7	2.94
	3	1.2×3	1.30	14.2	2.79
	5	1.2×5	0.96	15.1	2.52

of the SSF test examples have been rated to be preferable over the corresponding DS result, which is favored only for 10% of the examples. Please find audio examples on our website[1].

The SSF results discussed so far, have been obtained using oracle knowledge about the speaker DoA angles. For classic beamforming, e.g. with an MVDR beamformer, it is well-known that errors in the speaker DoA estimates used to construct a steering vector are likely to cause significant performance degradation [43]. To get an idea of the performance that can be expected in a blind setting, we also report results for two different DoA estimation strategies in Table II. The sensitivity to errors in the DoA estimates is then investigated in more detail in Section VI-A. The first strategy is search-based and evaluates the SSF for a set of potential target directions. The results for the individual speakers are then selected based on the energy of the filtered signal. The search-based strategy is illustrated in Figure 4. The top row shows the energy of the filtered signals evaluated on a grid with 4° resolution. We compute the energy on 10 ms long non-overlapping segments and plot the average energy for all segments, in which speech is active. A -45 dB threshold with respect to the maximum energy in the mixture signal is used for detection of speech activity as in [44]. We observe clear peaks at the true speaker locations, which are marked with a dashed gray line. The green crosses visualize the DoA angle estimates, which have been obtained with a peak-finding heuristic applied to the energy curve of which the details are outlined in Appendix B.

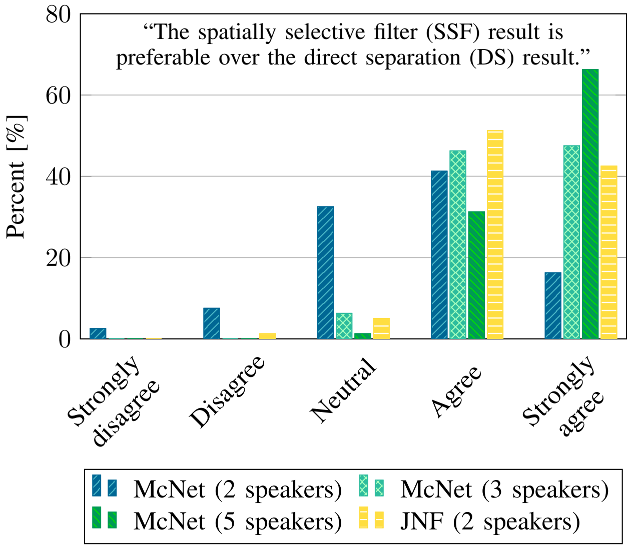


Fig. 3: Results for a listening experiment assessing the participants' preference for separation results obtained with a spatial filter (SF) or a direct separation (DS) result. Speaker locations are assumed to be known for the spatial filter. The test is conducted blindly without test subjects knowing which example corresponds to which algorithm. The results have then been aggregated to match with the displayed statement.

Comparing the vertical dashed lines and the position of the green crosses in Figure 4, we observe that the search-based DoA estimates are quite accurate for the selected examples. The evaluation of the localization accuracy shown in Table IV asserts that these examples can be considered representative of the dataset, since also the mean angular error is low. As the search-based DoA estimates depend on the DNN-based filter output, we display separate results for the JNF-SSF and the McNet-SSF in the first and second row. Somewhat unexpectedly, the mean angular error of the search-based DoA estimates is smaller for JNF-SFF than McNet-SSF. However, in both cases the estimates are accurate enough to replace the oracle DoA information without a loss in separation

[1] https://uhh.de/inf-sp-ssf

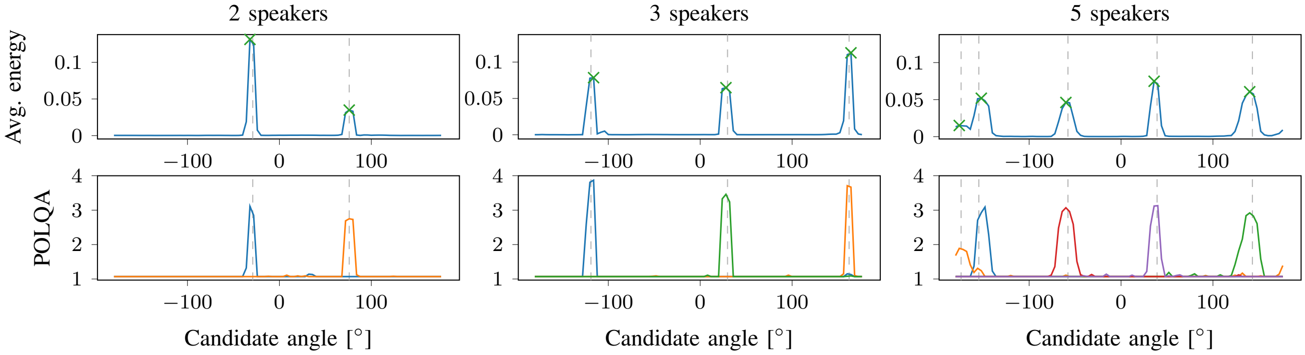


Fig. 4: Examples for blind speaker separation and localization by peak-searching for a mixture of two, three and five speakers using non-linear filters steered in all candidate directions. The vertical dashed gray lines indicate the true positions of the speakers and the green cross marks the speaker location estimated based on the energy peaks in the filter output.

TABLE IV: The speaker localization accuracy for mixtures of two, three and five speakers in a reverberant room. We report the mean angular error in degree and the 95% confidence interval.

DoA estimation	2 speakers	3 speakers	5 speakers
search (JNF-SSF)	1.57 ± 0.12	2.06 ± 0.19	3.54 ± 0.25
search (McNet-SSF)	2.07 ± 0.07	2.53 ± 0.15	3.99 ± 0.23
DNN	1.06 ± 0.03	1.24 ± 0.09	2.13 ± 0.19

performance as can be seen in rows 3 or 7 of Table II. For the McNet architecture the results obtained with the search-based strategy are even slightly better than those obtained with oracle DoA information. While this observation might raise questions at the first sight, it can be explained by the second row of Figure 4, which shows POLQA scores. We compute the POLQA measure for each candidate location's McNet-SSF output with respect to one speaker's reference signal to obtain one of the colored curves. The plots clearly show that the spatial filter has a high spatial selectivity for its steering direction, which is also why the search-based DoA estimation works well. Consider now the first peak in the left-most plot. The DoA estimate denoted by a green cross is slightly off to the left. However, the POLQA score at this estimated position is higher than the POLQA score at the true position. As a result, slight deviations in the search-based DoA estimation are not harmful to the overall performance, but can even be helpful as they are correlated with the filter's behavior. As we will show in Section VI-A, an uncorrelated DoA error of 2° causes a performance degradation.

Even though it provides interesting insights in the spatial selectivity of the proposed SSF, the search-based approach is too computationally demanding for most realistic applications. Therefore, we also evaluate the SSF with DoA estimates provided by a DNN-based classifier, which is trained to detect for every 2° bin if there is a speaker or not. More details are provided in Appendix C. Table IV shows that the DNN-based classifier is not only much more efficient, but also outperforms the search-based strategy in terms of localization accuracy by up to 1.86° mean angular error for five speakers. The separation results for McNet-SSF are given in row number 8. Also for these DNN-based DoA estimates, we do not see major deviations from the oracle performance, which demonstrates that the SSF approach is well applicable also to blind separation tasks.

The two bottom rows provide results for two baseline systems. The first one is a traditional MVDR beamformer with a DNN-based post-filter (PF). The parameters of the MVDR are estimated from oracle data. The time-varying noise covariance matrices are estimated by recursive averaging of the pure noise data and the time-invariant RTF estimate is obtained by multiplying the principal eigenvector of the generalized eigenvalue problem for speech and noise covariance matrices with the speech covariance matrix as described in [45]. The post-filter is a single-channel DNN with two LSTM layers trained on MVDR outputs as described in [22]. The comparison with JNF-SSF and McNet-SSF highlights that drastic improvements can be achieved by replacing the linear spatial filter with a non-linear one. While the former does not achieve a substantial performance improvement over the mixture recording in a setting with three microphones and five speakers, the DNN-based SSFs provide respectable performance also in this difficult scenario.

The second baseline uses hand-crafted features (HCF). For a fair comparison, we use the same McNet architecture as before but exchange the input signal. Previously, we provided the network with raw frequency domain data, i.e., the real and imaginary parts stacked. Following the approach proposed in [24], we replace this input by a compilation of features: the real and imaginary part of the reference channel as spectral feature, the IPDs between all pairs of microphones and the location-guided angle feature [24], [25], which provides information about which speaker to extract. Comparing the results in row 10 with those in row 6, we find that our proposed approach without using hand-crafted features and with explicit conditioning on the one-hot encoded angle is beneficial. It outperforms the baseline with hand-crafted features by about 0.4 POLQA improvement score for two, three and five speakers. The performance benefit is particularly audible for three and five speaker mixtures.

A question that might come to mind is whether the measured performance difference can be attributed to the difference in the approach (explicit spatial filtering in SSFs versus implicit spatial filtering in DS) or whether the explicit provisioning of DoA information introduces a bias since speaker DoA information is generally known to be helpful for a multi-channel speaker separation problem. To investigate this, we train the DS networks again, but now also provide additional DoA information for all speakers in the mixture. For this, we use the conditioning mechanism shown in Figure 1 using a multi-hot encoding.

We report the DoA-informed direct separation (iDS) results using the McNet architecture in the fifth row of Table II. For two speakers, the performance obtained with the DoA-informed network is quite similar to the performance of McNet-DS in row 4 of Table II. For three and five speakers, the additional DoA information leads to a performance improvement over the McNet-DS. However, the performance is still worse than that of McNet-SSF for three speakers and five speakers by 0.11 and 0.47 POLQA score respectively. Thus, providing the DoA information does not close the performance gap between McNet-DS and McNet-SSF, especially not for a large number of speakers.

VI. Robustness Experiments

For an approach to be applicable in practice, it is not only the maximum performance that is of interest, but also the robustness of the system, which is its capability to tolerate perturbations, for example in the geometric setup of the microphone array. In the following subsections, we evaluate the robustness of the SSF approach with respect to errors in the DoA estimates and variances in the microphone array setup. All experiments in this and the next section have been conducted using the McNet architecture. Data generation parameters that are not explicitly mentioned in the experiment description are kept constant at the value or range of values specified previously in Section IV.

A. *DoA Estimation Errors*

Figure 5 displays the results of an experiment that investigates the sensitivity of the SSF to errors in the DoA estimates, which are used to steer the filter. For this, we add an error term to the oracle DoA angle of all speakers in the mixture. The magnitude of the error is displayed on the x-axis. The y-axis represents the POLQA improvement score obtained with the McNet-SSF approach. We report average results and the 95% confidence interval for 100 randomly selected mixture signals. The left plot shows the results for McNet-SSF, which has been trained using the exact DoA angle of the target speaker. As expected, the performance decreases as the DoA error, which is added only during evaluation, increases. On the positive side, though, the performance drop is not very drastic for a small error of 2°. In this case the performance loss is about 5%, 7% and 3% for two, three and five speaker mixtures respectively. Furthermore, we note that the more difficult problem with five speakers in the mixture is less affected by an DoA estimation error. This observation can be explained by looking at the peak plots in Figure 4. Here we can see in the bottom row that the peaks for five speakers have become wider and, therefore, an erroneous DoA estimate to steer the filter has less consequences.

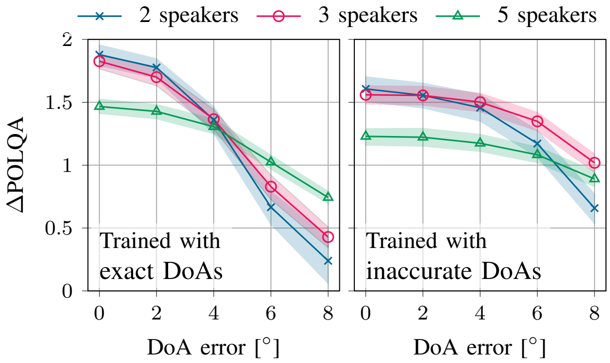


Fig. 5: Separation results for the McNet-SSF conditioned on a target angle that is subject to a localization error of varying magnitude. During evaluation, the respective error is added to all speakers' DoA angles. The results shown in the left plot are obtained with with a McNet-SSF that has been conditioned on the exact DoA location during training, while the results displayed on the left side are obtained with a McNet-SSF that has been trained with inaccurate DoAs that include an error of up to 4°.

The right side of Figure 5 displays results for a McNet-SSF trained with inaccurate DoA information. During training we add a DoA error of up to 4° to the true DoA. As a consequence, we observe that the performance stays approximately the same for a DoA error of up to 4° degrees. The performance drop for larger errors is also significantly reduced for two and three speakers. However, this increased robustness comes at the cost of a reduced performance if the DoA estimates are accurate, which is expected to some extend as there will always be a trade-off between sensitivity to DoA estimation errors and the spatial selectivity of the filter. In line with this, we observe that the peaks reflecting the spatial selectivity as in Figure 4 have widened for the filter trained with inaccurate DoAs.

B. *Perturbations in the Microphone Placement*

In a multi-channel scenario, the spatial information is very closely related to the geometric configuration of the microphone array, since the main source of information is the relative TdoAs of a signal at the microphones. So far, we have used a fixed and exact placement of the microphones in the array for generation of the simulated training and testing data. Now we add some noise to the positioning of the microphones to evaluate the sensitivity of the SSF and DS networks to these kinds of perturbations. We sample the noise that we add to the x-, y- and z-coordinate of the microphones in the circular array from a zero-mean normal distribution with a standard deviation between 0.1 cm and 1 cm. The standard deviation, i.e. the amount of perturbation in the microphone array geometry, is shown on the x-axis of Figure 6 against the POLQA improvement score.

The plots from left to right in Figure 6 show the results for two, three and five speaker mixtures. Clearly, the SSF

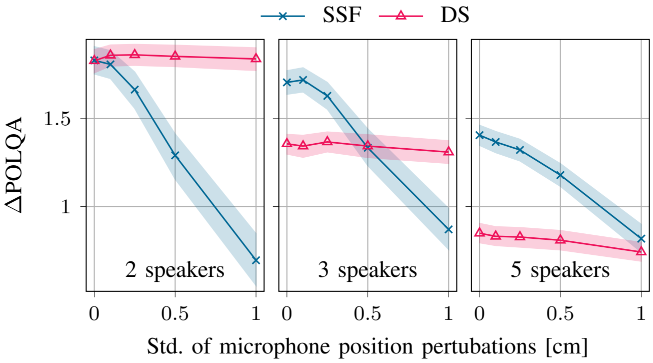


Fig. 6: Separation results for test examples with pertubated microphone positioning. The noise added to the microphone placement is sampled from a zero-mean normal distribution with a standard deviation shown on the x-axis.

approach is highly sensitive to perturbations in the microphone array geometry. While small perturbations with a standard deviation of 1 mm are tolerated without a significant loss in performance, large perturbations render the method useless. In contrast, the DS performance only slightly decreases for five speakers and is approximately constant in the other cases. From a perspective of robustness this can be considered a favorable behavior. However, the DS performance for the three and five speaker mixtures is far from the maximum performance reached by the SSF method even with some perturbations in the microphone array geometry. Therefore, we would not consider the DS approach robust to variations in the spatial characteristics related to the microphone array geometry but view the results of this experiment as a strong indication that the DS approach performs worse than the SSF approach for a higher number of speakers because it does not fully exploit the spatial information in the multi-channel data in the first place.

VII. Generalization Experiments

As neural networks learn to extract patterns from data, for example the general spectro-temporal structure of speech or the spatial characteristics of a signal arriving from a particular direction, it is insightful to evaluate the performance of a DNN-based approach also for inputs whose characteristics vary from those present in the training set. Here we perform an experiment that varies the distance between speech source and microphone array, an experiment investigating the performance for speech sources with similar DoA and one that adds an additional noise source.

A. *Far-field vs Near-field Scenario*

The relative TdoAs of the direct-path signal are not only related to the microphone configuration in the array, but also influenced by the distance of the source to the microphone array. If the distance between the source and the microphone array is large compared to the distances between the microphones, we commonly make a far-field assumption and model the propagation of sound as a plane wave as opposed to spherical wave in the near-field [1]. We have trained the SSF with mixtures of sources that were placed at a distance of $0.8 - 1.2$ m to the microphone array, which itself has a diameter of 10 cm. Here we investigate the performance of the SSF for input signals that deviate from the data seen during training with respect to their distance. For this experiment, we present the SSF network with a single reverberant speech signal originating from a varying distance and report the POLQA improvement score in Figure 7. The blue plot with square markers represents a network trained with data generated according to the configuration given in Section IV. The range that is covered by the training data is shown be the blue shaded rectangle. We can see that the network reaches the maximum performance for examples that fall in this range. The performance gradually decreases as the source moves further away, which will also decrease the DRR. If the source moves closer toward to microphone array on the left side of the blue area, the DRR increases. Nevertheless, the performance drops even to negative improvement scores as the source moves very close to the microphone array. For our previous work on speaker extraction [22], [27], we have trained the SSF with a target source that is placed closer to the microphone array with a distance of $0.3 - 1.0$ m. The results for this filter are shown in red. Here we can see that including near-field examples in the training data drastically improves the performance for close-by sources. However, as the network spends more parameters modeling spatial characteristics in the near-field, the performance for far-distant sources is decreased.

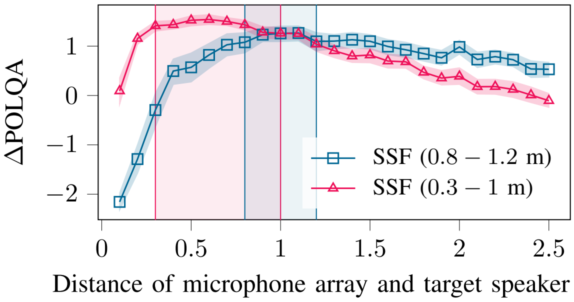


Fig. 7: POLQA improvement scores for a single speaker placed at a varying distance to the microphone array. Results for MCNet-SSF trained with speakers at a distance of $0.8 - 1.2$ m are shown by the blue plot with square markers and the results of a MCNet-SSF trained with a target speaker positioned at a distance of $0.3 - 1.0$ m (as in [26]) are represented by red triangles. The range covered by the training data has been marked by the respective color. The diameter of the microphone array is 0.1 m in both cases.

B. *Sources with Similar DoA*

All speech sources in the examples generated according to the dataset description in Section IV have a minimum 10° angle difference between them and every other source. While such such a constraint may be realistic for example in a meeting scenario with the microphone array positioned on the table, in other scenarios the speakers may stand closer together. We therefore investigate the generalization ability of the DS and

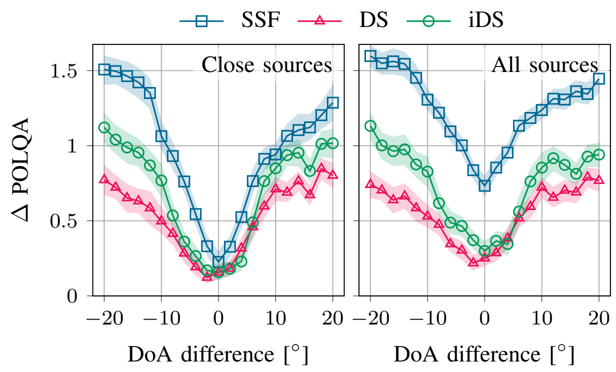


Fig. 8: Separation results for test examples that include three speakers, of which one is placed at a 60° angle, one at a 0° angle and one speaker with at a variable angle between −20° and 20°. The left plot shows the average POLQA improvement for the two close sources and the right plot shows the average improvement for all three sources.

SSF approach for close sources. For this, we generate test examples with mixtures of three speakers which are positioned at a 60° angle, a 0° angle and a variable angle between −20° and 20°. We evaluate on 60 examples for every angle difference. The average POLQA improvement for the two close sources is shown in the left plot of Figure 8. Clearly, neither the SSF nor the DS approach can provide good separation results for sources arriving from the same direction as can be seen by the performance dip at 0° angle difference. However, listening to the results it becomes clear that they handle the task in different ways. For sources positioned in the same direction, the SSF returns a mixture of only the two close sources excluding the third speaker for both speakers, while the DS approach returns different results for every speaker, which are of very low quality.

The plot on the right side of Figure 8 includes the separation result for the third speaker positioned at a 60° angle when computing the average POLQA improvement score. Here we can see that for the SSF, shown in blue, the average POLQA improvement score increases by about 0.5. In contrast, the average performance of the DS (red curve) remains about the same as in the left plot. This means that while the SSF is struggling to separate the two close sources, it has no problem to accurately extract the third source. On the other hand, the DS approach is not able to provide a reasonable result for the third speaker if two speakers are close. From these findings, we conclude that the decoupling of the separation results for the individual speakers in the SSF leads to a better generalization in a scenario that contains sources with similar DoA. The green curve shows the performance for iDS, for which we provide the DoA information as an additional input. Here we can see that such a desirable decoupling is not achievable by simply providing DoA information to a network that is trained with PIT.

TABLE V: Separation results for two speaker mixtures with an additional music noise source at a random position. The number of sources refers to the number of sources predicted by each system. We then select the outputs corresponding to the two speech sources for evaluation.

	DoA	#Sources	ΔPOLQA	ΔSI-SDR	DNSMOS
McNet-SSF	oracle	2	1.66	15.2	2.98
McNet-DS		2	0.65	6.5	2.28
McNet-iDS	oracle	2	0.84	10.0	2.35
McNet-DS		3	1.29	14.0	2.70
McNet-iDS	oracle	3	1.51	14.8	2.80

C. Noise (Unseen During Training)

It is a major advantage of the SSF in comparison with the DS approach that the neural network does not need to be re-trained for evaluations on mixtures of different numbers of speakers as it focuses on extraction of speech signal from a particular direction. In our last experiment, we now add a music noise source to a mixture of two speakers and compare the ability of the SSF and DS approach to generalize to this new scenario. For the music noise source, we sample a random position between 0.8 m and 1.2 m away from the microphone array (same distance as speakers) and add the music signal to the mixture signal with an SNR of 5 dB. The music signals are taken from the *jamendo* subset of the MUSAN dataset [46].

The performance results are shown in Table V. The first row shows the performance of the SSF given oracle knowledge of the two speakers' directions. Compared to the results in the fifth row of Table II, the POLQA score improvement and DNSMOS score are reduced by about 10% and 5% respectively. However, the system still reliably separates the two speakers without the music signal leaking into one of the estimates. In contrast, McNet-DS trained on two speaker mixtures cannot handle the additional noise source, which is reflected by the low performance scores in the second row of Table V. Results obtained with the DoA-informed McNet-iDs are given in the third row. Interestingly, the performance notably improves if oracle knowledge about the speaker DoAs is provided, which was not the case for the noise-free scenario. However, the performance is still 0.82 POLQA and 5.2 dB SI-SDR lower than that of McNet-SSF.

Since the music noise source could be seen as a third source, we also test the DS networks trained on three sources. For computing the performance results, we then only consider the outputs that correspond to the two speech sources. We can see in the two bottom rows of Table V that this indeed improves the performance by a large margin. The best performance is obtained with McNet-iDS trained on three speech sources. However, there remains a performance gap of 0.15 POLQA score, 0.4 dB SI-SDR and 0.18 DNSMOS to the McNet-SSF, which, unlike McNet-iDS in the last row, does not require information about the noise source's DoA. Since detecting the number of noise point sources as well as their DoA will be difficult in most real-world scenarios, the SSF, which only needs DoA estimates for the speech sources, is not only performing better but is also much more practical compared to all tested DS approaches.

VIII. Conclusion

Based on our conference paper [26], we proposed a steerable DNN-based spatially selective filter (SSF). Beyond [26], here we have investigated the separation performance of a DNN-based spatially selective filter (SSF) steered in the direction of each speaker in the mixture in comparison with a classic end-to-end direct separation (DS) approach trained with PIT. We find that the SSF, which has been trained for high spatial selectivity in the given DoA, outperforms a DS approach by a large margin if there are more than two speakers in the mixture. Experiments on the robustness of either system provides evidence that this is because the SSF better exploits spatial information. Furthermore, we have shown that the SSF generalizes much better to unseen noise conditions than the DS approach.

Appendix

A. Network Training Details

All neural networks (JNF-DS, JNF-SSF, McNet-DS, McNet-iDS and McNet-SSF) have been trained on an ℓ_1 loss in time and frequency domain [47]:

$$L(s_p, \hat{s}_p) = \alpha \|s_p - \hat{s}_p\|_1 + \left\| |S_p| - |\hat{S}_p| \right\|_1. \tag{4}$$

The frequency-domain term $\hat{S}_p$ is estimated as given in (3) by multiplication of the noisy signal's reference channel with the network-estimated mask and $\hat{s}$ is its inverse STFT. The parameter α is set to $\alpha = 10$ to approximately equalize the contribution of either domain. We use the Adam optimizer [48] with an initial learning rate of 0.001 and reduce the learning rate by a factor $\gamma = 0.75$ every 50 epochs. We train for a maximum of 500 epochs and select the best weights based on the validation loss. We use 32 ms windows for computing the STFT with a 50% overlap and use the square-root Hann window for both analysis and synthesis.

B. Peak-finding Heuristic

As a first step, we normalize the highest peak in the energy curve to one and then run `scipy.signal.find_peaks` with a prominence of 0.009, a height of 0.05 and a width of one. If fewer peaks than the expected number of speakers are detected, we re-execute the function with relaxed parameters settings (no width requirement, decreased prominence of 0.001 and height of 0.025) and merge peaks that likely to represent the same speaker as they are close together and have a similar height. If more peaks than expected speakers are found, we pick the highest ones.

C. DNN-based DoA Classifier

We train a DNN-based classifier for DoA estimation, which is composed of an F-LSTM layer as described in Section III-A and two feed-forward layers with 256 and 180 hidden units. We use an exponential linear unit activation for the first feed-forward layer and a sigmoid activation for the second. The network is trained to detect for every 2° bin if there is a speaker or not based on the full utterance. We train for 100 epochs with an average binary cross-entropy loss on the dataset of two speaker mixtures. Even though trained on two speaker mixtures, the classifier performs sufficiently well in detecting other numbers of speakers. We use the same peak-finding heuristic as for the search-based strategy, which is necessary as the classifier provides an output between zero and one for every angle bin.

Acknowledgment

We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

References

[1] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Chichester, England Hoboken, NJ: John Wiley, 2006.

[2] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Apr. 2004.

[3] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 7, pp. 1900–1912, 2011.

[4] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[5] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 4, pp. 692–730, 2017.

[7] M. Taseska and E. A. Habets, "Doa-informed source extraction in the presence of competing talkers and background noise," *EURASIP J. Adv. Signal Proc.*, vol. 2017, no. 1, p. 60, 2017.

[8] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for BSS with deep clustering embeddings," in *Proceedings of Interspeech*, 2017, pp. 2650–2654.

[9] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2018, pp. 5739–5743.

[10] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, D. Raj, S. Watanabe, Z. Chen, and J. R. Hershey, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 905–911.

[11] S. Sivasankaran, E. Vincent, and D. Fohr, "Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, 2020, pp. 346–350.

[12] R. Liu, Y. Zhou, H. Liu, X. Xu, J. Jia, and B. Chen, "A new neural beamformer for multi-channel speech separation," *Journal of Signal Processing Systems*, vol. 94, no. 10, pp. 977–987, 2022.

[13] A. S. Subramanian, C. Weng, M. Yu, S.-X. Zhang, Y. Xu, S. Watanabe, and D. Yu, "Far-field location guided target speech extraction using end-to-end speech recognition objectives," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2020, pp. 7299–7303.

[14] Z. Chen, T. Yoshioka, X. Xiao, L. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2018, pp. 5384–5388.

[15] A. Aroudi and S. Braun, "Dbnet: Doa-driven beamforming network for end-to-end reverberant sound source separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2021, pp. 211–215.

[16] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, “BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge,” in *IEEE Workshop Autom. Speech Recog. and Underst. (ASRU)*, Scottsdale, USA, 2015, pp. 444–451.

[17] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2018, pp. 1–5.

[18] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, “Enhancing end-to-end multi-channel speech separation via spatial feature learning,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2020, pp. 7319–7323.

[19] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “End-to-end microphone permutation and number invariant multi-channel speech separation,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2020, pp. 6394–6398.

[20] C. Quan and X. Li, “Multichannel Speech Separation with Narrow-band Conformer,” in *Proceedings of Interspeech*, 2022, pp. 5378–5382.

[21] K. Tesch and T. Gerkmann, “Nonlinear spatial filtering in multichannel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1795–1805, 2021.

[22] ——, “Insights into deep non-linear filters for improved multi-channel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 563–575, 2023.

[23] Z.-Q. Wang and D. Wang, “Combining spectral and spatial features for deep learning based blind speaker separation,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 2, pp. 457–468, 2019.

[24] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, “Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information,” in *Proceedings of Interspeech*, 2019, pp. 4290–4294.

[25] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, “Multi-channel overlapped speech recognition with location guided speech extraction network,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 558–565.

[26] K. Tesch and T. Gerkmann, “Spatially selective deep non-linear filters for speaker extraction,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2023.

[27] K. Tesch, N.-H. Mohrmann, and T. Gerkmann, “On the role of spatial, spectral, and temporal processing for dnn-based non-linear multi-channel speech enhancement,” in *Proceedings of Interspeech*, Sep. 2022, pp. 2908–2912.

[28] D. Markovic, A. Defossez, and A. Richard, “Implicit Neural Spatial Filtering for Multichannel Source Separation in the Waveform Domain,” in *Proceedings of Interspeech*, Sep. 2022, pp. 1806–1810.

[29] J. Wechsler, S. R. Chetupalli, W. Mack, and E. A. Habets, “Multi-microphone speaker separation by spatial regions,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2023.

[30] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, “The cone of silence: Speech separation by localization,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 20 925–20 938.

[31] S. Kindt, A. Bohlender, and N. Madhu, “Improved separation of closely-spaced speakers by exploiting auxiliary direction of arrival information within a u-net architecture,” in *18th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, 2022, pp. 1–8.

[32] X. Li and R. Horaud, “Multichannel speech enhancement based on time-frequency masking using subband long short-term memory,” in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2019, pp. 298–302.

[33] D. S. Williamson, Y. Wang, and D. Wang, “Complex Ratio Masking for Monaural Speech Separation,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, no. 3, pp. 483–492, 2016.

[34] Y. Yang, C. Quan, and X. Li, “Mcnet: Fuse multiple cues for multichannel speech enhancement,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2023.

[35] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. of the IEEE Conf. on Comp. Vision and Pattern Recog. (CVPR)*, Jun. 2015.

[36] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 351–355.

[37] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[38] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) Complete LDC93S6A,” Linguistic Data Consortium, Philadelphia, May 2007.

[39] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[40] “P.863: Perceptual objective listening quality prediction,” International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[41] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2019, pp. 626–630.

[42] C. K. A. Reddy, V. Gopal, and R. Cutler, “DNSMOS p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2022, pp. 886–890.

[43] J. Li, P. Stoica, and Z. Wang, “Doubly constrained robust capon beamformer,” *IEEE Trans. Audio, Speech, Language Proc.*, vol. 52, no. 9, pp. 2407–2423, 2004.

[44] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[45] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, “Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, USA, 2017, pp. 681–685.

[46] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.

[47] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, “Channel-attention dense U-net for multichannel speech enhancement,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2020, pp. 836–840.

[48] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Int. Conf. Learning Repr. (ICLR)*, May 2015.

Discussion and Conclusions 6

6.1 Main Contributions of this Thesis

In this section, we summarize the main contributions presented in detail in the three included publications. We structure this along the three main areas of research as described in Section 1.5. In particular, we aim to highlight the cross-connections between a highly theoretical perspective on the topic of non-linear spatial filtering in the first part and the practical development of DNN-based non-linear spatial filters in parts two and three. Our research culminates in the development of a real-time demo, which we presented at the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2023. A video of this demo can be found on our website[1]. Finally, we conclude this thesis by pointing out directions for future research.

6.1.1 Statistical Perspective on Non-linear Spatial Filtering

A central motivation of our investigations, starting from a statistical perspective, was to answer the question of whether investing in the development of neural networks to replace classical beamformers could be worthwhile. Here, it should be borne in mind that the current rapid development of many well-performing DNN-based non-linear spatial filters described in Section 1.4.3 was only about to begin, and the answer to this question was by no means clear at the start of the research project that led to this thesis. After all, a traditional beamformer has many obvious advantages over a data-driven neural network: lower computational complexity, no training data, and no training period required, which comes with good generalization ability to unseen types of data and good interpretability. As of today, our own research [P4], [P5] and a plethora of works from others (see Section 1.4.3) in recent years has demonstrated the enormous performance potential of neural networks to implement non-linear spatial filters. The comparison of a DNN-based non-linear spatial filter in [P5, Fig. 4] clearly shows that the DNN-based filter outperforms the oracle beamformer plus DNN-based post-filter by a large margin for a low number of channels in the microphone array. Accordingly, the question of whether research in neural networks for spatial filtering is worthwhile can clearly be answered positively. However, our investigations in the first part of this thesis (with related publications [P1]–[P3]) are, to our knowledge, the only publications to date that offer a statistical perspective on what makes the DNN-based non-linear spatial filters as successful as they are.

The key insight of the first part of this thesis is the finding that a non-linear spatial filter enables a more powerful processing model than a linear spatial filter. The corresponding experiment is described in [P3, Sec. V]: A single target speech source and five directional

[1] `https://uhh.de/inf-sp-jnf-demo`

interfering sources are arranged on a circle around a two-channel microphone array. In one of the investigated settings, the interfering sources alternately play short white Gaussian burst signals. Since the statistical estimators were derived under a multi-variate Gaussian mixture noise assumption, an estimate of the covariance matrices of the mixture components is required, which we obtain from oracle noise data using the EM algorithm. A weighted sum of the mixture components' covariance matrices results in an estimate of the overall covariance matrix of the noise signal. A graphical representation of the covariance matrices (overall and for the individual mixture components) based on the MVDR beampattern created by them is shown in [P3, Fig. 6]. It is clearly visible that the mixture components' covariance matrices reflect the spatial properties of the individual interfering sources. To understand the value of this experiment, it is important to realize that the estimation of the spatial properties, i.e., the covariance matrices, is not adapted over time. Only one set of matrices is estimated for a whole utterance. The covariance matrix that goes into the calculation of the MVDR beamformer, therefore, represents the spatial properties of all interfering sources equally. As a result, the MVDR beamformer, as a linear spatial filter, cannot cancel out all five interfering sources with only two available microphone signals. In contrast, the non-linear spatial filter can draw on the more detailed spatial information in the various covariance matrices of the individual mixture components. Just as for the MVDR, no temporal adaptation takes place, and yet the non-linear spatial filter can almost perfectly suppress the interfering sources. From this, we conclude that the known limit of $C-1$ cancelable anechoic noise sources for C microphones does not apply to the non-linear spatial filter.

The analytic non-linear spatial filter also has a clear performance advantage when speech signals are emitted by the interfering sources instead of Gaussian bursts. In contrast to the experiment explained above, the individual time-frequency bins are dominated by one of the interfering sources with a pattern that is difficult to predict. From this more realistic scenario, it becomes clear why a temporal adaptation of the MVDR in the previous experiment, which is expected to result in a cancelation of the interfering sources, is not as usable in a practical scenario as a non-linear spatial filter, which does not require adaptation. For many practical scenarios like the speaker extraction task, a fast enough adaptation is nearly impossible to achieve. Listening to the audio examples for this speaker extraction task, it is noticeable that the analytic non-linear spatial filter clearly delivers better performance than the MVDR plus post-filter. However, there is much more residual noise than in the experiment with the Gaussian bursts and also more than in the results obtained with a DNN-based non-linear spatial filter in [P5]. Likely, this is due to the fact that for the derivation of the analytical estimators, it is assumed that the time-frequency bins are independent of each other, which is a heavily simplifying assumption for speech signals.

A second interesting property of non-linear spatial filters can be deduced from the compiled theory overview that outlines the concept of sufficient statistics. Under a Gaussian noise assumption, it can be shown, and in [P3], we even present a new simplified proof, that the MVDR beamformer output is a sufficient statistic for the clean speech signal in the Bayesian sense. This means that the beamformer output $T_{\text{MVDR}}(\mathbf{Y}(k,i))$ contains all information about the searched quantity $S(k,i)$ that was available in the original noisy observation $\mathbf{Y}(k,i)$ despite the dimension reduction. An alternative formulation to the one given in [P3] is based on the information-theoretic concept of mutual information. A sufficient statistic requires the mutual information between the statistic and the clean speech signal, which is to be estimated, to be equal to the mutual information between the original noisy and the clean speech signal, i.e.,

$$\mathcal{I}\left(T_{\mathrm{MVDR}}(\mathbf{Y}(k,i)), S(k,i)\right) = \mathcal{I}(\mathbf{Y}(k,i), S(k,i)), \tag{6.1}$$

where $\mathcal{I}$ denotes the information theoretic concept of mutual information [112], [113]. When the noise follows a Gaussian distribution, this means that the two-stage approach shown in Figure 1.1 does not stand in the way of statistical optimality if the MVDR is used for spatial filtering since no information about the speech signal is lost in the first processing stage. However, the non-linear spatial filter derived under a Gaussian mixture noise assumption cannot be separated in a MVDR plus post-filter scheme. Consequently, this means that information required to estimate the speech signal $S(k,i)$ is lost if a MVDR beamformer is applied as the first processing step. The effect of this information loss on the performance is what we report in the experimental sections. It might feel counter-intuitive that information about the speech signal is lost despite the MVDR obeying a distortionless constraint. The explanation here is that it is not information about speech that is lost but information about the noise, which is used by the non-linear spatial filter to distinguish between speech and noise. A consequence of this insight is that it might not be advisable to restrict oneself to spatial features primarily focused on the properties of the speech signal as one could easily miss out on the relevant information contained in the noise part of the multi-channel signal.

6.1.2 Design and Analysis of Deep Non-linear Spatial Filters

In the second part of this thesis, we turn to the development and analysis of data-driven non-linear spatial filters implemented using a neural network. The main contribution of our research are the insights into the internal mechanisms of the DNN-based non-linear spatial filter. As explained in Section 1.4.3, three sources of information, spatial, spectral, and temporal, can, in principle, be exploited by the network to perform spatial filtering. When using the analytic MMSE filter derived under a Gaussian mixture noise assumption as inspiration, it becomes clear that the spatial processing model should be non-linear, which is naturally the case for a neural network, and that tempo-spectral information should be processed jointly with the spatial information. Our experiments shed light on the importance of either of these design objectives for high overall performance. By twisting the input data of the network, we are isolating the effect of a non-linear spatial processing model and can also investigate the interdependencies between spatial and spectral versus spatial and temporal information. The experiment shows that the speech quality of the enhanced signal greatly improves when interdependencies are considered. In line with our observations based on the analytic estimators in the first part, we find that such a DNN-based joint non-linear spatial filter outperforms an MVDR plus DNN-based post-filter approach by a large margin when the number of microphones is small.

We perform the experiments in [P5] for two scenarios. One is a reverberant speaker extraction task, where a single target speech signal is to be extracted from the noisy mixture. The target speaker is identified by its position relative to the microphone array. The training and test data are simulated based on the image-source method [114]. The second task is speech enhancement in environmental noise. The background noise is taken from the real-world recordings in the CHiME3 [115] dataset. Our goal is to investigate the influence of interdependencies between spatial and spectral as well as spatial and temporal information on the overall performance. We consider a speaker extraction task as particularly suitable for assessing the quality of a spatial filter. The reason for this is that all interfering signals have similar tempo-spectral properties as the target signal, and, therefore, we assume that a good extraction of the target speech signal is linked to a good use of spatial information. From our results, it is very clear that correlations along the frequency axis in an STFT time frame have a very large influence on the spatial selectivity of the filter. This also makes sense intuitively since the phase shifts of a directional signal depend on the frequency. A consequence of this insight is that when designing a network architecture for spatial filtering, much attention should be paid to the joint processing of spatial and spectral information. Temporal correlations can also be exploited by the neural network and improve the performance. However, they are not as important for the spatial selectivity of the resulting filter, as can be seen in Figure 6 of [P5]. A comparison with state-of-the-art methods in [P5] shows that our proposed filter, which processes spatial and spectral information jointly in the first layer and is given access to temporal information only in the second layer, outperforms all baselines. This is a remarkable result given that the network structure is very simple, with only three layers, and that the network also has the lowest number of learnable parameters.

The evaluation based on the CHiME3 data shows that, for the enhancement task, temporal information plays a significantly greater role than spectral information. Looking at Table IV in [P5], which displays the overall performance, one could conclude that the spatial filtering here improves when temporal information instead of spectral information is incorporated. However, Figure 10 shows that this conclusion would be premature. An examination of the spatial selectivity of the filters shows the same patterns as we observed for the speaker extraction task: spectral information increases the spatial selectivity of the filter. However, this is only partially reflected in the performance results, which is probably due to the fact that spatial filtering plays a less important role in solving the enhancement task. In fact, for the CHiME3 data, the characteristics of noise and speech are quite different (especially with respect to their temporal structure) so that the gain of an improved (more selective) spatial filter by taking spectral information into account does not compensate for the loss of information that one has when temporal information is not taken into account. Accordingly, and as expected, the best results are obtained when both spectral and temporal information are considered in addition to spatial information. A noteworthy insight here is that a higher spatial selectivity of a filter, i.e., a better spatial filtering performance, is not necessarily linked to a better overall enhancement performance. Therefore, when the spatial selectivity of a filter is to be assessed, a performance comparison on a dataset like CHiME3 could be misleading.

6.1.3 Steerable Deep Non-linear Spatial Filters for Speech Extraction and Separation

All filters discussed in the second part with related publications [P4], [P5] have a fixed look direction. For the speaker extraction task, this is quite obvious from the experiment setup since the target speaker to be extracted is always positioned at a zero-degree angle relative to the microphone array orientation. In a practical application, this would require a constrained setup in which the target speaker is always in the same position. In some applications, this might be a reasonable assumption, e.g., a person in the driver seat of a car can only move in a limited range. However, most applications require a more flexible solution. For our experiments using the well-known CHiME3 dataset in [P5], it is less obvious that the learned filter has a fixed look direction since the speaker and microphone array position are unknown for the examples in the dataset. From the description of the dataset, it appears that the speakers read sentences from a tablet to which also the microphone array is attached. They are encouraged to take different positions during the recording session [115]. Our investigation of the spatial selectivity of the learned filter in [P5] clearly shows that the filter trained on the CHiME3 data might be limited to a small range of geometric setups (position of the microphone array relative to the target speaker position) depending on how well the chosen architecture supports the learning of a spatially selective filter. This makes sense as the participants of the recording session can only change their speaking position relative to the tablet so much that they are still able to read the words on the screen. Therefore, the geometric variation in the dataset is rather limited, and a good performance cannot be expected in test conditions that include spatial locations of the target speaker not matching the training set. In comparison, a classic beamformer is much more flexible, as the look direction can be controlled with the help of the steering vector or RTF vector.

In the third part of this thesis, we aim to achieve such flexibility for the DNN-based non-linear spatial filter and, therefore, develop a conditioning mechanism to control the look direction of the filter. For this, only an estimate of the target DOA is needed, which is fed into the neural network as a one-hot encoded vector. An accurate estimate of the DOA would also be needed to construct the steering vector for an MVDR under a far-field assumption, and the accurate estimation of the RTF can be considered even more difficult. In [P6], the information about the desired look direction is then fed into the LSTM layers as an initial state. Follow-up experiments revealed that it is not necessary to condition both layers. Instead, it is sufficient to only condition the first spatial-spectral LSTM layer, which is consistent with our finding in [P5] that joint processing of spatial and spectral information leads to a high spatial selectivity of the filter. Since the first layer has no dependency on the time index, the look direction can almost instantaneously be adjusted for every time-frame. Thus, in the real-time demo that we presented at WASPAA 2023, we can adjust the look direction with a control wheel without a noticeable time lag every 8 ms.

In our publication [P6], we show that this flexible filter achieves virtually the same performance as a fixed filter explicitly trained on the corresponding look direction. This is a rather astonishing result since the flexible filter has to learn not only one spatial filter but many at the same time. We use a resolution of two degrees in the azimuth direction and train the filter with sources of variable height. This means that the steerable filter then has to learn filters for 180 potential look directions. We provide much less training data per direction, 300 (flexible) versus 6000 (fixed) examples because the training would take too long otherwise. But still, the performance of the learned flexible filter matches with that of the fixed filter trained explicitly for the respective directions, which means that we can increase the flexibility of the filter at virtually no performance cost.

In the journal publication [P7] included in the main part of this thesis, we investigate how the performance is affected by treating the separation problem as a spatial filtering problem. For this purpose, we use the flexibly steerable spatially selective filter (SSF) from [P6] and steer one instance of the filter towards each speaker in the mixture to obtain the separated speech signals. This requires that the position of all speakers of interest is known or can be estimated from the noisy mixture. The predominantly chosen alternative is to train an end-to-end direct separation (DS) network that generates a separate output for each speaker. Since the output ordering of the speakers is subject to a permutation problem, such a separation network is often trained with a loss applied in a PIT scheme. Alternatively, Tan et al. [116] have proposed to enforce an ordering of the outputs according to the DOA and Wechsler et al. [107] found that networks may even resort to a spatially consistent output order if trained with PIT. In any case, and in contrast to using an SSF, the number of speakers must be known in advance. Changing the number of outputs requires a re-training of the network.

The advantage, on the other hand, is that the network does not require any target DOAs and estimates all separated signals at the same time.

For a DS approach, the spatial filtering is expected to be implicitly learned during the training stage of the DNN. In contrast, the SSF approach is more similar to a traditional beamformer as it is steered towards a specific direction and learns to extract signals from this target direction. On the other hand, a difference compared to a traditional linear beamformer is that it performs this task by joint spatial and tempo-spectral non-linear filtering. During training, by performing speaker extraction conditioned on a DOA, the focus is put on the ability to identify signal components from the target direction. Our publication [P7] addresses the research question which influence the employed training strategy has on the separation performance but also the robustness of the learned filter. The investigated strategies are (1) direct separation (DS) based on PIT or similar techniques, which means that the spatial filtering must be learned implicitly by the network from the provided examples, and (2) spatially selective filtering (SSF) with an explicit focus on the spatial properties of the target signal.

From the results in Table II of [P7], it can be seen that the SSF outperforms a DS approach by a large margin for challenging scenarios with more than two speakers. The superiority becomes especially clear when a small network is used. Here, the participants of a blind listening experiment overwhelmingly agree that the SSF separation result is preferable to the DS network output.

We use the same core network architecture for both the SSF and the DS network. The only difference is in the output layer (one speaker vs the number of speakers in the mixture) and the conditioning mechanism, which is not needed for DS. However, the SSF is evaluated as many times as there are speakers in the mixture. This means that the number of learnable parameters per speaker is smaller for the DS network. Table III in [P7], therefore, presents the results of the DS network, which has been scaled up to keep the number of parameters per speaker constant. Still, we observe the same trend: the SSF outperforms the DS approach also in this case. We perform a second experiment in which we provide additional DOA information for all speakers in the mixture to the DS network. For this, we employ the same conditioning mechanism as for the SSF but with a multi-hot encoding. Interestingly, this additional knowledge of the speakers' DOAs still does not boost the performance so much that it matches that of the SSF. These experiments make it very clear that not only the architecture of the neural networks plays an important role, but that attention must also be paid to the choice of the training strategy. Here, our results imply that it is quite beneficial to focus on the network's ability to extract a speech signal with specific spatial properties during training.

In [P7], we not only investigate the overall performance but also perform experiments regarding the robustness and generalization ability of the SSF versus DS approach. Interestingly, we find that the performance of DS is much less affected by perturbations in the microphone placement (randomly added noise to the microphone positions) than the SSF. While this is generally desirable, this behavior, in combination with lower overall performance, can be seen as a clear indication that the spatial properties like the IPDs, which change drastically with changed microphone positions, are not fully exploited by the DS network. Hence, we conclude that the performance benefit of the SSF in comparison with the DS approach results from the fact that spatial information is better taken into account by the SSF, which is also the focus of the training strategy.

However, focusing on the spatial properties of the target signal has another advantageous effect: the learned filter much better generalizes to scenarios that have not been seen during training. In a two-speaker separation experiment with an additional interfering music source, the SSF delivers a very good performance despite being trained on speech mixtures only. In contrast, the performance of a DS network with two outputs is drastically decreased, and good results can only be obtained when the music source is treated as an additional source of interest and the DOA of all sources, including the noise source, is provided to the network. A second experiment clarifies the value of decoupled separation outputs. In a setting with two sources located at a similar angle and a third source further away, we find that neither the network trained according to the DS nor the SSF approach can separate the close sources. On the other hand, the SSF can extract the third source without a problem, while the quality of the output of DS network for the third speaker is heavily reduced by its inability to separate the two close speakers.

6.2 Directions for Future Research

In this thesis, we have investigated non-linear spatial filters for multi-channel speech enhancement, speaker extraction, and separation. The focus is on filters that join the spatial and tempo-spectral filtering into a single non-linear operation. We found that these filters are more powerful than traditional linear spatial filters from a statistical perspective and also confirmed this observation by implementing DNN-based non-linear spatial filters that outperform an oracle MVDR beamformer followed by a post-filter. However, the MVDR beamformer still has a highly valued advantage over a non-linear filter: at least in theory, i.e., with accurately estimated parameters, it does not introduce any speech distortions. This is ensured by the distortionless constraint in (1.10), which is the equation that corresponds to the constrained optimization problem that leads to the MVDR. Note that the formulation of the distortionless constraint involving the steering vector in (1.10) requires the linearity of the filter to be meaningful.

In contrast, consider the following optimization problem, where the time-frequency indices have been dropped to improve the readability:

$$\min_{T} \quad \mathcal{I}(\boldsymbol{Y}, T) \quad \text{subject to} \quad \mathcal{I}(T, S) = \mathcal{I}_c, \tag{6.2}$$

where $\mathcal{I}(\cdot, \cdot)$ is the mutual information as before in (6.1). We are searching for a scalar speech estimate T, which should be a function of the noisy multi-channel observation $\mathbf{Y}$. Since the dimension is reduced to one, it is reasonable to assume that some kind of spatial filtering is performed. If $\mathcal{I}_c = \mathcal{I}(\mathbf{Y}, S)$, then T is a sufficient statistic for the clean speech signal S [113]. In this case, the processed signal T carries the same information about the target signal S as the noisy observation Y. This means that irreversible distortions introduced by the processing will be minimal. It could be an interesting direction for future research to relate the above optimization problem, which can be recognized as an instance of the so-called information bottleneck [117], with the MVDR. It would be interesting to see if one can derive the MVDR as a special case, assuming Gaussian noise, in this general framework allowing for non-linear solutions. With choosing $\mathcal{I}_c < \mathcal{I}(\boldsymbol{Y}, S)$ one could trade noise reduction by minimization of $\mathcal{I}(\boldsymbol{Y}, T)$ against allowed speech distortions reflected by the constraint that bounds the information loss $\mathcal{I}(T, S) = \mathcal{I}_c$.

A framework that includes the well-known linear MVDR beamformer as well as a non-linear spatial filter might be helpful to learn about the characteristics of such a filter. Unfortunately, it is likely that the optimization problem in (6.2) is intractable for distributions other than the Gaussian distribution, for which a solution has been presented in [118]. Therefore, in most works on the information bottleneck, the problem in (6.2) is not directly optimized but instead a Lagrangian relaxation,

$$\mathcal{L}(T) := \mathcal{I}(\boldsymbol{Y}, T) - \beta \mathcal{I}(T, S), \tag{6.3}$$

is minimized [117]. The mutual information is generally difficult to estimate, but some works integrate it as a training loss into a DNN for learning representations, e.g., [119], [120]. Research questions that could be investigated here would be how a distortionless constraint and non-linear spatial filtering relate to each other. For example, can there be a (nearly) distortionless filter that also has the powerful capabilities of a non-linear spatial filter observed in the first part of this thesis for the analytic filter? Or does a distortionless constraint inherently lead to a filter that resembles the properties of a linear filter?

Furthermore, there are a number of more practically oriented questions worth investigating. In [P7], we observed (not to our surprise) a strong dependency on the exact microphone arrangement in the array. While very small deviations still lead to good results, a large deviation leads to a drastic drop in performance. In preliminary experiments, we found that it is not sufficient to disturb the microphone positions in the training set. Instead, this led to a significantly worse performance overall. Nevertheless, working on this problem could provide further insights into the functioning of a non-linear spatial filter. While the concrete task to be solved and the relevant parameters, such as the phase differences, change drastically, the spatial filtering task itself retains its general structure. From a meta-learning perspective, i.e., using a learning-to-learn paradigm, there should still be commonalities between the tasks that can be exploited and affect the core of spatial filtering. These commonalities could be the key to an even deeper understanding of non-linear spatial filtering.

We demonstrate the practical applicability of the DNN-based non-linear spatial filters developed in parts two and three with a real-time multi-channel speech enhancement and speaker extraction system. The look direction of the filter is interactively controlled by the user. Since the filter has a very high spatial selectivity, setting the look direction correctly in a live scenario is not always easy, even if DOA estimates are provided as visual guidance. Furthermore, already slight movements of the target speaker will shift him or her out of the pass-through direction of the filter. Therefore, we added a heuristic tracking algorithm that picks the most dominant source detected by the source location algorithm in the region selected by the user. Currently, the source localization is performed by a second neural network that is run in parallel with the actual deep non-linear spatial filter. Since the computational complexity of the algorithm plays a major role in many practical applications (on which hardware can the program be executed in real-time and how much power is consumed in the process), all synergies between the related tasks of localization, tracking, and filtering should be exploited as far as possible.

Furthermore, the comparison in [P5] shows that we achieve very good state-of-the-art performance with a comparatively small network in terms of the number of learnable parameters. However, the bi-LSTM along the frequency axis makes up a large sequential component, so that a hardware architecture oriented towards parallelism cannot be optimally utilized. Therefore, research into even more efficient network architectures is required to make the non-linear spatial filters developed in this thesis applicable to a wider range of applications.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears", *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment.* Chichester, England Hoboken, NJ: John Wiley, 2006.

[3] R. Balan and J. P. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase", in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, Rosslyn, Virginia, 2002, pp. 209–213.

[4] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multispeaker LCMV beamformer and postfilter for source separation and noise reduction", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 5, pp. 940–951, 2017.

[5] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques", in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 39–60.

[6] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On optimal multichannel mean-squared error estimators for speech enhancement", *IEEE Signal Proc. Letters*, vol. 16, pp. 885–888, 2009.

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

[8] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model", *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[9] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors", *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, pp. 1741–1752, 2007.

[10] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors", *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 5, pp. 845–856, 2005.

[11] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Las Vegas, USA, 2008, pp. 4897–4900.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 5, pp. 504–512, 2001.

[13] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay”, *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[14] X. Hao, X. Su, R. Horaud, and X. Li, “Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 6633–6637.

[15] H. Schröter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, “Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, Singapore, 2022, pp. 7407–7411.

[16] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, 2016, pp. 31–35.

[17] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, USA, 2017, pp. 241–245.

[18] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation”, *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[19] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 46–50.

[20] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 21–25.

[21] N. Zeghidour and D. Grangier, “Wavesplit: End-to-end speech separation by speaker clustering”, *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 2840–2849, 2021.

[22] S. Zhao and B. Ma, “Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023.

[23] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones”, *IEEE Signal Proc. Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[24] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory.* John Wiley & Sons, 2004.

[25] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing.* Springer Science & Business Media, 2008, vol. 1.

[26] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications.* Springer Science & Business Media, 2001.

[27] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[28] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis.* New York: Wiley, 2001.

[29] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain", *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[30] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources", *IEEE Trans. on Speech and Audio Proc.*, vol. 8, no. 3, pp. 320–327, 2000.

[31] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions", in *Independent Component Analysis and Blind Signal Separation: 6th International Conference*, Charleston, USA, 2006, pp. 601–608.

[32] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies", *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, no. 1, pp. 70–79, 2007.

[33] S. Makino, *Audio source separation.* Springer, 2018.

[34] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement.* John Wiley & Sons, 2018.

[35] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ilrma originating from ica and nmf", *APSIPA Transactions on Signal and Information Processing*, vol. 8, e12, 2019.

[36] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, Istanbul, Turkey, 2000, pp. 2985–2988.

[37] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization", *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[38] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", *IEEE Trans. on Signal Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[39] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization", *IEEE Trans. Audio, Speech, Language Proc.*, vol. 18, no. 2, pp. 382–394, 2010.

[40] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Dallas, USA, 2010, pp. 241–244.

[41] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment", *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 3, pp. 516–527, 2011.

[42] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge", in *IEEE Automatic Speech*

Recognition and Understanding Workshop (ASRU), Scottsdale, USA, 2015, pp. 444–451.

[43] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks", in *Proceedings of Interspeech*, San Francisco, USA, 2016, pp. 1981–1985.

[44] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[45] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, USA, 2017, pp. 286–290.

[46] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation", *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815–826, 2019.

[47] Z. Chen, J. Li, X. Xiao, *et al.*, "Cracking the cocktail party problem by multi-beam deep attractor network", in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, 2017, pp. 437–444.

[48] Z. Chen, T. Yoshioka, X. Xiao, L. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 5384–5388.

[49] R. Liu, Y. Zhou, H. Liu, X. Xu, J. Jia, and B. Chen, "A new neural beamformer for multi-channel speech separation", *Journal of Signal Processing Systems*, vol. 94, no. 10, pp. 977–987, 2022.

[50] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods", *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[51] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 4, pp. 692–730, 2017.

[52] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals", *IEEE Trans. Audio, Speech, Language Proc.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[53] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays", in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 19–38.

[54] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 22, no. 1, pp. 67–79, 2013.

[55] E. A. P. Habets, J. Benesty, S. Gannot, P. A. Naylor, and I. Cohen, “On the application of the LCMV beamformer to speech enhancement”, in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, New Paltz, USA, 2009, pp. 141–144.

[56] E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition”, *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[57] S. Araki, H. Sawada, and S. Makino, “Blind speech separation in a meeting situation with maximum SNR beamformers”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 1, Honolulu, USA, 2007, pp. I-41–I–44.

[58] C. Bishop, *Pattern recognition and machine learning.* New York: Springer, 2006.

[59] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, “Channel-attention dense U-net for multichannel speech enhancement”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 836–840.

[60] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection”, *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[61] F. Beritelli, S. Casale, and G. Ruggeri, “Performance evaluation and comparison of ITU-T/ETSI voice activity detectors”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 3, 2001, pp. 1425–1428.

[62] J.-H. Chang, N. S. Kim, and S. Mitra, “Voice activity detection based on multiple statistical models”, *IEEE Trans. on Signal Proc.*, vol. 54, no. 6, pp. 1965–1976, 2006.

[63] T. Hughes and K. Mierle, “Recurrent neural networks for voice activity detection”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Vancouver, Canada, 2013, pp. 7378–7382.

[64] C. Blandin, A. Ozerov, and E. Vincent, “Multi-source tdoa estimation in reverberant audio using angular spectra and clustering”, *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[65] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, “Relaxed disjointness based clustering for joint blind source separation and dereverberation”, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, 2014, pp. 268–272.

[66] N. Ito, S. Araki, and T. Nakatani, “Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing”, in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, IEEE, Budapest, Hungary, 2016, pp. 1153–1157.

[67] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, “Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, 2016, pp. 5210–5214.

[68] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, 2016, pp. 196–200.

[69] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, “Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant

environments", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2017, pp. 681–685.

[70] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation", *IEEE Trans. Audio, Speech, Language Proc.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[71] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[72] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications.* The MIT press, 1949.

[73] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation", *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 3, pp. 483–492, 2016.

[74] Y. Xu, M. Yu, S.-X. Zhang, *et al.*, "Neural Spatio-Temporal Beamformer for Target Speech Separation", in *Proceedings of Interspeech*, Shanghai, China, 2020, pp. 56–60.

[75] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, USA, 2017, pp. 3246–3250.

[76] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, USA, 2017, pp. 5325–5329.

[77] C. Boeddeker, P. Hanebrink, L. Drude, J. Heymann, and R. Haeb-Umbach, "Optimizing neural-network supported acoustic beamforming by algorithmic differentiation", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, New Orleans, USA, 2017, pp. 171–175.

[78] Y. Masuyama, M. Togami, and T. Komatsu, "Multichannel loss function for supervised speech source separation by mask-based beamforming", in *Proceedings of Interspeech*, Graz, Austria, 2019, pp. 2708–2712.

[79] Y. Zhou and Y. Qian, "Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 536–540.

[80] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 6717–6721.

[81] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 5739–5743.

[82] T. Higuchi, K. Kinoshita, M. Delcroix, K. Žmolíková, and T. Nakatani, "Deep Clustering-Based Beamforming for Separation with Unknown Number of Sources", in *Proceedings of Interspeech*, Stockholm, Sweden, 2017, pp. 1183–1187.

[83] L. Yin, Z. Wang, R. Xia, J. Li, and Y. Yan, “Multi-talker Speech Separation Based on Permutation Invariant Training and Beamforming”, in *Proceedings of Interspeech*, Hyderabad, India, 2018, pp. 851–855.

[84] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, “Beamtasnet: Time-domain audio separation network meets frequency-domain beamformer”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 6384–6388.

[85] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, “Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brighton, United Kingdom, 2019, pp. 6855–6859.

[86] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, “Mask-based neural beamforming for moving speakers with self-attention-based tracking”, *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 835–848, 2023.

[87] X. Xiao, S. Watanabe, H. Erdogan, *et al.*, “Deep beamforming networks for multi-channel speech recognition”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, 2016, pp. 5745–5749.

[88] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, “Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms”, in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, USA, 2015, pp. 30–36.

[89] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, “Factored spatial and spectral multichannel raw waveform CLDNNs”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, 2016, pp. 5075–5079.

[90] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, “Exploring multi-channel features for denoising-autoencoder-based speech enhancement”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brisbane, Australia, 2015, pp. 116–120.

[91] X. Zhang and D. Wang, “Deep learning based binaural speech separation in reverberant environments”, *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 5, pp. 1075–1084, 2017.

[92] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Calgary, Canada, 2018, pp. 1–5.

[93] R. Gu, L. Chen, S.-X. Zhang, *et al.*, “Neural spatial filter: Target speaker speech separation assisted with directional information”, in *Proceedings of Interspeech*, Graz, Austria, 2019, pp. 4290–4294.

[94] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, “FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing”, in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Sentosa, Singapore, 2019, pp. 260–267.

[95] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “End-to-end microphone permutation and number invariant multi-channel speech separation”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 6394–6398.

[96] N. Tawara, T. Kobayashi, and T. Ogawa, “Multi-channel speech enhancement using time-domain convolutional denoising autoencoder”, in *Proceedings of Interspeech*, Graz, Austria, 2019, pp. 86–90.

[97] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, “TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Sigapore, 2022.

[98] S. Chakrabarty, D. Wang, and E. A. P. Habets, “Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks”, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, 2018, pp. 476–480.

[99] S. Chakrabarty and E. A. P. Habets, “Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[100] X. Li and R. Horaud, “Multichannel speech enhancement based on time-frequency masking using subband long short-term memory”, in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, New Paltz, USA, 2019, pp. 298–302.

[101] Z.-Q. Wang, P. Wang, and D. Wang, “Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR”, *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1778–1787, 2020.

[102] A. Li, W. Liu, C. Zheng, and X. Li, “Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, 2022, pp. 6487–6491.

[103] M. M. Halimeh and W. Kellermann, “Complex-valued spatial autoencoders for multichannel speech enhancement”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Singapore, 2022, pp. 261–265.

[104] Y. Yang, C. Quan, and X. Li, “McNet: Fuse multiple cues for multichannel speech enhancement”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023.

[105] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, 2017.

[106] C. Quan and X. Li, “Multichannel speech separation with narrow-band conformer”, in *Proceedings of Interspeech*, Incheon, Korea, 2022, pp. 5378–5382.

[107] J. Wechsler, S. R. Chetupalli, W. Mack, and E. A. Habets, “Multi-microphone speaker separation by spatial regions”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhodes Island, Greece, 2023.

[108] A. Briegleb, T. Haubner, V. Belagiannis, and W. Kellermann, “Localizing spatial information in neural spatiospectral filters”, Helsinki, Finland, 2023.

[109] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Brighton, United Kingdom, 2019, pp. 626–630.

[110] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, “ADL-MVDR: All deep learning MVDR beamformer for target speech separation”, in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Toronto, Canada, 2021, pp. 6089–6093.

[111] W. Mack and E. A. P. Habets, “Deep filtering: Signal extraction and reconstruction using complex time-frequency filters”, *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2020.

[112] T. Cover and J. A. Thomas, *Elements of information theory.* Chichester, England Hoboken, NJ: Wiley-Interscience, 2006.

[113] H. Hafez-Kolahi and S. Kasaei, “Information bottleneck and its applications in deep learning”, *J. Information Systems and Telecommunication (JIST)*, vol. 6, no. 3, 2018.

[114] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics”, *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[115] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines”, in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, USA, 2015, pp. 504–511.

[116] K. Tan, Z.-Q. Wang, and D. Wang, “Neural spectrospatial filtering”, *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 30, pp. 605–621, 2022.

[117] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method”, in *Proc. of the 37-Th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[118] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, “Information bottleneck for Gaussian variables”, *J. Machine Learning Research*, vol. 6, no. 6, pp. 165–188, 2005.

[119] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, *et al.*, “Learning deep representations by mutual information estimation and maximization”, in *Int. Conf. Learning Repr. (ICLR)*, New Orleans, USA, 2019.

[120] M. Ravanelli and Y. Bengio, “Learning speaker representations with mutual information”, in *Proceedings of Interspeech*, Graz, Austria, 2019, pp. 1153–1157.

[121] I. S. Gradshteyn, *Table of integrals, series, and products.* San Diego: Academic Press, 2000.

[122] *NIST Digital Library of Mathematical Functions*, Release 1.0.20 of 2018-09-15, F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[123] S. M. Kay, *Fundamentals Of Statistical Signal Processing.* Pearson, 2009.

List of Acronyms

AIR	acoustic impulse response
ASR	automatic speech recognition
ATF	acoustic transfer function
BSS	blind source separation
CIRM	complex ideal ratio mask
dB	decibels
DDR	direct-to-reverberation ratio
DFT	discrete Fourier transform
DNN	deep neural network
DOA	direction of arrival
DS	direct separation
EM	expectation-maximization
GCC	generalized cross-correlation
GEV	generalized eigenvalue
IBM	ideal binary mask
ICA	independent component analysis
ILD	inter-channel level difference
IPD	inter-channel phase difference
IRM	ideal ratio mask
IVA	independent vector analysis
LCMV	linearly constraint minimum variance
LSTM	long short-term memory
MAP	maximum a posteriori
ML	maximum likelihood
MMSE	minimum mean square error
MPDR	minimum power distortionless response
MSE	mean square error
MVDR	minimum variance distortionless response
MWF	multi-channel Wiener filter
NMF	non-negative matrix factorization
PDF	probability density function

PIT	permutation invariant training
PSD	power spectral density
RIR	room impulse response
RTF	relative transfer function
SI-SDR	scale-invariant source to distortion ratio
SNR	signal-to-noise ratio
SPP	speech presence probability
SSF	spatially selective filter
STFT	short-time Fourier transform
TDOA	time difference of arrival
VAD	voice activity detection

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, 17.11.2023

Kristina Tesch

Kristina Tesch

Appendices

A Related Peer-Reviewed Conference Publications

This appendix contains four additional conference publications. These have not been included in the main body of the thesis since the journal publications cover the topic of each part in more depth. The journal publications extend the previous conference publications, which is explicitly allowed by the author guidelines of the IEEE/ACM Transactions of Audio, Speech, and Language Processing, so that there is a substantial overlap between the respective papers. Nevertheless, some experiments presented in the conference publications have been omitted in the corresponding journal publication so that they are included here for completeness.

A.1 On Nonlinear Spatial Filtering in Multichannel Speech Enhancement [P1]

Abstract

Using multiple microphones for speech enhancement allows for exploiting spatial information for improved performance. In most cases, the spatial filter is selected to be a linear function of the input as, for example, the minimum variance distortionless (MVDR) beamformer. For non-Gaussian distributed noise, however, the minimum mean square error (MMSE) optimal spatial filter may be nonlinear.

Potentially, such nonlinear functional relationships could be learned by deep neural networks. However, the performance would depend on many parameters and the architecture of the neural network. Therefore, in this paper, we more generally analyze the potential benefit of nonlinear spatial filters as a function of the multivariate kurtosis of the noise distribution.

The results imply that using a nonlinear spatial filter is only worth the effort if the noise data follows a distribution with a multivariate kurtosis that is considerably higher than for a Gaussian. In this case, we report a performance difference of up to 2.6 dB segmental signal-to-noise ratio (SNR) improvement for artificial stationary noise. We observe an advantage of 1.2 dB for the nonlinear spatial filter over the linear one even for real-world noise data from the CHiME-3 dataset given oracle data for parameter estimation.

Reference

Kristina Tesch, Robert Rehr and Timo Gerkmann, “On Nonlinear Spatial Filtering in Multichannel Speech Enhancement”, in *Proceedings of Interspeech*, Graz, Austria, 2019, pp. 91-95. DOI: 10.21437/Interspeech.2019-2751

On Nonlinear Spatial Filtering in Multichannel Speech Enhancement

Kristina Tesch, Robert Rehr, and Timo Gerkmann

Signal Processing, Universität Hamburg, Germany

kristina.tesch@uni-hamburg.de, robert.rehr@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

Using multiple microphones for speech enhancement allows for exploiting spatial information for improved performance. In most cases, the spatial filter is selected to be a linear function of the input as, for example, the minimum variance distortionless response (MVDR) beamformer. For non-Gaussian distributed noise, however, the minimum mean square error (MMSE) optimal spatial filter may be nonlinear.

Potentially, such nonlinear functional relationships could be learned by deep neural networks. However, the performance would depend on many parameters and the architecture of the neural network. Therefore, in this paper, we more generally analyze the potential benefit of nonlinear spatial filters as a function of the multivariate kurtosis of the noise distribution.

The results imply that using a nonlinear spatial filter is only worth the effort if the noise data follows a distribution with a multivariate kurtosis that is considerably higher than for a Gaussian. In this case, we report a performance difference of up to 2.6 dB segmental signal-to-noise ratio (SNR) improvement for artificial stationary noise. We observe an advantage of 1.2 dB for the nonlinear spatial filter over the linear one even for real-world noise data from the CHiME-3 dataset given oracle data for parameter estimation.

Index Terms: Multichannel, speech enhancement, nonlinear filtering, acoustic beamforming, neural networks

1. Introduction

Many speech signals recorded in everyday environments, for example in a restaurant or next to a busy street, are corrupted by additional background noise. Therefore, speech enhancement algorithms that improve the perceived quality or intelligibility of a recorded speech signal by reducing noise or other disturbing effects such as reverberation are of great importance in a wide range of communication applications.

Noise reduction methods such as the Wiener filter [1, Sec. 11.3.1] and nonlinear optimal estimators of the clean speech Fourier coefficient [1, Sec. 11.4] or its magnitude [2] effectively reduce noise in single-channel microphone recordings. However, multichannel approaches often outperform single-channel methods as they incorporate not only tempo-spectral properties of the signals but can also include spatial information in the processing.

In most cases, the spatial filtering is based on a linear processing model, called beamforming, that weights the DFT coefficient of the different microphone channels with complex-valued coefficients before summation to suppress signal components from others than the target direction [3, Sec. 3.1]. The MVDR beamformer is a prominent example of a linear spatial filter that exploits the time delay of signal arrival determined by the spatial arrangement and further takes the correlation of the noise signals between the microphones into account.

It seems natural to include well-developed single-channel methods into multichannel speech enhancement by applying a single-channel algorithm, called a postfilter, to the output of a spatial filter. For Gaussian distributed noise, it has been shown that the sequential coupling of the spatially linear MVDR filter and a postfilter yields optimal results with respect to the MMSE, maximum a posteriori (MAP) and maximum likelihood (ML) criterion [4, 5]. In contrast, Hendriks et al. show that the optimal spatial filter is nonlinear and cannot be separated from spectral processing if the noise is not Gaussian distributed [6]. However, it remains open how large the potential benefit of using nonlinear spatial filters really is. This question gained importance in the context of the rise of neural networks in recent years: while it is demanding to derive optimal nonlinear spatial filters in a statistical framework, neural networks can learn to approximate nonlinear functions directly from data [7].

Neural networks have successfully been incorporated into single-channel speech enhancement [8, 9, 10, 11] often in the context of automatic speech recognition (ASR) [12] and they have also been very successful in estimating the parameters of linear spatial beamformers [13]. Sainath et al. propose a multichannel neural network approach to ASR that includes a spatial filtering layer [14, 15, 16]. Interestingly, the structure of their proposed time-convolutional layer imposes a linearity constraint on the spatial filter even though fixing a linear spatial filter might not lead to an optimal solution.

The goal of our research is to answer the question if investing in the development of neural networks that learn optimal *nonlinear* spatial filters is worth the effort. As a first step towards answering this question, in this paper, we analyze the potential benefit of nonlinear spatial filtering as compared to a standard linear spatial filter like the MVDR under ideal conditions.

In order to gain a better understanding of the role and potential of nonlinear spatial filters, we proceed as follows: Section 3 reviews the most relevant theoretical results on the optimality of linear versus nonlinear spatial filters. In section 4, we analyze the potential performance gain of an optimal nonlinear spatial filter in contrast to a linear spatial filter for noise with a known super-Gaussian distribution. Section 5 assesses the improvement potential of nonlinear spatial filters for real noise recordings from the CHiME-3 dataset [17].

2. Notation and Assumptions

We assume that a microphone array composed of D microphones records the target speech signal along with interfering noise. The time domain signals are windowed and transformed into the frequency domain using the discrete Fourier transform (DFT), which leads to the noisy DFT coefficients $Y_\ell(k,\ i)$ with microphone-channel index $\ell \in \{1,...,D\}$, frequency-bin index k and time-frame index i. We assume an additive noise signal model so that the noisy DFT coefficient $Y_\ell(k,i)$ can be represented as a sum of the clean speech DFT coefficient $S_\ell(k,i)$ and noise DFT coefficient $N_\ell(k,i)$ received at the ℓth microphone, i.e.,

$$Y_\ell(k,i) = S_\ell(k,i) + N_\ell(k,i). \qquad (1)$$

The DFT coefficients of the speech and noise signals are modeled as random variables. We denote random variables by uppercase letters, while lowercase letters are used for their respective realizations. The speech and noise coefficients are assumed to be uncorrelated and all DFT coefficients to be zero-mean and independent with respect to time and frequency. As a consequence, we can drop the indices (k,i) from the notation. Let $\mathbf{Y}=[Y_1,...,Y_D]\in\mathbb{C}^D$ be the vector containing the noisy DFT coefficients for all D channels and let $\mathbf{S}\in\mathbb{C}^D$ and $\mathbf{N}\in\mathbb{C}^D$ denote the vectors of speech and noise DFT coefficients, respectively. We work in a single source scenario, which means that there is only one target speaker, and model the signal propagation from the speaker to the microphones as a plane wave. Thus, the vector of clean speech DFT coefficients $\mathbf{S}$ can be obtained by multiplying the reference clean speech DFT coefficient S with a frequency-dependent vector $\mathbf{d}\in\mathbb{C}^D$, i.e., $\mathbf{S}=\mathbf{d}S$. We denote the noise correlation matrix by $\mathbf{\Phi}_n=\mathbb{E}[\mathbf{N}\mathbf{N}^H]$, while $\sigma_s^2=\mathbb{E}[|S|^2]$ denotes the spectral power of S.

3. Linearity of the Optimal Spatial Filter

In this section, we review optimal multichannel estimators of the clean speech DFT coefficient to address the question under which conditions an optimal solution decomposes into a linear spatial filter and a spectral postfilter. First, we consider the case of multivariate complex Gaussian distributed noise DFT coefficients with zero mean and covariance matrix $\mathbf{\Phi}_n$. Since we assume that the noise is additive, the distribution of $\mathbf{Y}$ given the reference speech DFT coefficient s is Gaussian distributed with mean $\mathbf{d}s$ and covariance matrix $\mathbf{\Phi}_n$, i.e., $\mathbf{Y}\sim\mathcal{N}_{\mathbb{C}}(\mathbf{d}s,\mathbf{\Phi}_n)$. The corresponding conditional probability density function (PDF) is given by [18, Thm. 15.1]

$$p(\mathbf{y}|s)=\frac{1}{\pi^D|\mathbf{\Phi}_n|}\exp\left\{-(\mathbf{y}-\mathbf{d}s)^H\mathbf{\Phi}_n^{-1}(\mathbf{y}-\mathbf{d}s)\right\}. \quad (2)$$

Balan and Rosca [4] use the concept of sufficient statistics to show that the MMSE estimator of the clean speech DFT coefficient S

$$T_{\text{MMSE}}(\mathbf{y})=\arg\min_{\hat{s}\in\mathbb{C}}\mathbb{E}\left[|S-\hat{s}|^2\,|\mathbf{y}\right] \quad (3)$$

separates into the well-known MVDR beamformer and a spectral postfilter under a Gaussian noise assumption. The MVDR beamformer T_{MVDR} is a sufficient statistic *in the classical sense* for the true clean speech DFT coefficient s if the conditional distribution of the noisy observation $\mathbf{Y}$ given $T_{\text{MVDR}}(\mathbf{y})$ does not depend on s [19, Def. IV.C.1], i.e. $T_{\text{MVDR}}(\mathbf{Y})$ contains all the information in $\mathbf{Y}$ that is useful for estimating s. Furthermore, T_{MVDR} is said to be a sufficient statistic of S *in the Bayesian sense* if

$$p(s|\mathbf{y})=p(s|T_{\text{MVDR}}(\mathbf{y})) \quad (4)$$

holds for all observations $\mathbf{y}$ regardless of the prior distribution of S [20, Def. 2.4]. If the MVDR beamformer is a sufficient statistic, then no information about S is lost during spatial filtering even though the dimension of the output is reduced to one dimension. As a result, spatial processing and spectral processing can be performed separately in sequence. Since a statistic that is sufficient in the classical sense is also sufficient in the Bayesian sense [20, Thm. 2.14.2], we can infer that (4) holds by showing that the MVDR beamformer is a sufficient statistic in the classical sense. Resorting to the Fisher-Neyman factorization theorem [19, Prop. IV.C.1][21, Cor. 2.6.1], we deduce this property of the MVDR beamformer from the finding that the conditional PDF $p(\mathbf{y}|s)$ can be factorized as

$$\begin{aligned}p(\mathbf{y}|s)&=\underbrace{\frac{1}{\pi^D|\mathbf{\Phi}_n|}\exp\{-\mathbf{y}^H\mathbf{\Phi}_n^{-1}\mathbf{y}\}}_{h(\mathbf{y})}\\&\quad\times\underbrace{\exp\left\{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d}\left(2\text{Re}\{s^*T_{\text{MVDR}}(\mathbf{y})\}-|s|^2\right)\right\}}_{g(s,T_{\text{MVDR}}(\mathbf{y}))}\\&=h(\mathbf{y})g(s,T_{\text{MVDR}}(\mathbf{y}))\end{aligned} \quad (5)$$

with

$$T_{\text{MVDR}}(\mathbf{y})=\frac{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{y}}{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d}}. \quad (6)$$

Using the fact that the MMSE estimator complies with the mean of the posterior [19, IV.B.1], we infer from (4) that

$$\begin{aligned}T_{\text{MMSE}}(\mathbf{y})&=\mathbb{E}[S|\mathbf{y}] && (7)\\&=\mathbb{E}[S|T_{\text{MVDR}}(\mathbf{y})] && (8)\end{aligned}$$

holds. The quantity $\mathbb{E}\left[S|T_{\text{MVDR}}(\mathbf{y})\right]$ can be seen as a single-channel filter working on the output of the MVDR beamformer. Because the relationship (4) holds for any prior distribution of S, a decomposition of the MMSE estimator into an MVDR beamformer and single-channel postfilter results independent of any further assumptions about the prior distribution of the reference speech DFT coefficient. The decomposition of the MMSE estimator is also described by Hendriks et al. [6] but derived without the concept of sufficient statistics.

From (4) we conclude that the MAP estimator also separates into a linear spatial filter and a single-channel postfilter. Furthermore, the MVDR beamformer can be identified as the ML estimator of the clean speech DFT coefficient S [5, Sec. 6.2.1.2].

However, the work of Hendriks et al. [6] reveals that the Gaussian noise assumption is fundamental to both the decomposability of the optimal estimator into a spatial and a spectral processing step and the linearity of the spatial filter. They derive an MMSE estimator for noise that follows a multivariate Gaussian mixture distribution. The M Gaussian mixture components are modeled as zero-mean with covariance matrices $\mathbf{\Phi}_m$, $m=1,...,M$, and combined by positive weighting factors c_m that fulfill the constraint $\sum_{m=1}^{M}c_m=1$. The resulting conditional PDF of $\mathbf{Y}$ is given by [22, Sec. 9.2]

$$p(\mathbf{y}|s)=\sum_{m=1}^{M}\frac{c_m}{\pi^D|\mathbf{\Phi}_m|}\exp\left\{-(\mathbf{y}-\mathbf{d}s)^H\mathbf{\Phi}_m^{-1}(\mathbf{y}-\mathbf{d}s)\right\}. \quad (9)$$

Hendriks et al. assume the clean speech amplitude A to be distributed according to the PDF

$$p(a)=\frac{2\left(\frac{\nu}{\sigma_s^2}\right)^{\nu}}{\Gamma(\nu)}a^{2\nu-1}\exp\left\{-\frac{\nu}{\sigma_s^2}a^2\right\}\text{ with }\nu>0,\,a\leq 0 \quad (10)$$

and the phase $\Psi\in[0,2\pi)$ to be uniformly distributed and independent of the speech amplitude. Then the MMSE estimator is given by

$$\widetilde{T}_{\text{MMSE}}(\mathbf{y})=\nu\frac{\sum\limits_{m=1}^{M}\frac{c_mQ_m}{|\mathbf{\Phi}_m|}e^{\left[-\mathbf{y}^H\mathbf{\Phi}_m^{-1}\mathbf{y}\right]}\frac{\sigma_s^2T_{\text{MVDR}}^{(m)}(\mathbf{y})\mathcal{M}(\nu+1,2,P_m)}{\nu(\mathbf{d}^H\mathbf{\Phi}_m^{-1}\mathbf{d})^{-1}+\sigma_s^2}}{\sum\limits_{m=1}^{M}\frac{c_mQ_m}{|\mathbf{\Phi}_m|}e^{\left[-\mathbf{y}^H\mathbf{\Phi}_m^{-1}\mathbf{y}\right]}\mathcal{M}(\nu,1,P_m)} \quad (11)$$

with

$$T_{\text{MVDR}}^{(m)}(\mathbf{y})=\frac{\mathbf{d}^H\mathbf{\Phi}_m^{-1}\mathbf{y}}{\mathbf{d}^H\mathbf{\Phi}_m^{-1}\mathbf{d}},\quad Q_m=(\nu+\mathbf{d}^H\mathbf{\Phi}_m^{-1}\mathbf{d}\sigma_s^2)^{-\nu},$$

$$\text{and}\quad P_m=\frac{\sigma_s^2\mathbf{d}^H\mathbf{\Phi}_m^{-1}\mathbf{d}\left|T_{\text{MVDR}}^{(m)}(\mathbf{y})\right|^2}{\nu(\mathbf{d}^H\mathbf{\Phi}_m^{-1}\mathbf{d})^{-1}+\sigma_s^2}$$

with $\mathcal{M}(\cdot,\cdot,\cdot)$ being the confluent hypergeometric function [23, Sec. 9.21]. Interestingly, the result shows that the MMSE estimator for the considered non-Gaussian model cannot be separated into an MVDR beamformer and a single-channel postfilter. Furthermore, the optimal spatial filter is not even linear [6].

4. Potential of Nonlinear Spatial Filters

In this section, we investigate the improvement potential of using the optimal spatially nonlinear MMSE estimator for Gaussian mixture distributed noise as opposed to a setup that combines a linear spatial filter and a spectral postfilter. To our knowledge, the MMSE estimator for non-Gaussian noise derived by Hendriks et al. has not been evaluated before.

We use a segment length of 32 ms and a square-root Hann window with 50% overlap for spectral analysis and synthesis. The clean speech signals have been taken from the WSJ0 dataset [24] and are balanced between male and female speakers (30 utterances each).

The noise DFT coefficients are generated by sampling a zero-mean Gaussian mixture distribution. The covariance matrix $\mathbf{\Phi}_n$ of the distribution is chosen to represent one of three scenarios [25]: spatially white noise, diffuse noise, and a directional noise source positioned at a 45 degree angle to the target source. In the latter cases, we add a small portion of spatially white noise ($\alpha_{\text{wn}}=0.05$) to ensure that the noise correlation matrix is invertible. We obtain noise distributions that depart from normality by means of heavier tails by combining mixture components with scaled versions of the same covariance matrix. Thus, we set the mth mixture component's covariance matrix $\mathbf{\Phi}_m$ to be

$$\mathbf{\Phi}_m=\frac{b^{m-1}}{r}\mathbf{\Phi}_n\quad\text{with}\quad r=\sum_{m=1}^{M}c_m b^{m-1}\tag{12}$$

and scaling factor $b\in\mathbb{R}^+$. The constant r takes care of normalization so that the covariance matrix $\mathbf{\Phi}_n$ of the mixture distribution remains unchanged.

The kurtosis is a statistical measure that accounts for the shape of a distribution, specifically its heavy-tailedness [26, 27]. We extend Mardia's multivariate kurtosis definition [28] to complex-valued random vectors $\mathbf{X}\in\mathbb{C}^D$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}_x$ to obtain

$$\kappa_{\mathbb{C}}(\mathbf{X})=\mathbb{E}\Big[(2(\mathbf{X}-\boldsymbol{\mu})^H\mathbf{C}_x^{-1}(\mathbf{X}-\boldsymbol{\mu}))^2\Big].\tag{13}$$

Using [29, Sec. 8.2.4], we find the multivariate complex kurtosis of the random vector $\mathbf{N}$ following a scaled Gaussian mixture distribution to be

$$\kappa_{\mathbb{C}}(\mathbf{N})=2D(2+2D)\underbrace{\sum_{m=1}^{M}c_m\frac{b^{2(m-1)}}{r^2}}_{q}.\tag{14}$$

The factor $2D(2+2D)$ corresponds to the kurtosis of the D-dimensional complex Gaussian distribution. Thus, the kurtosis of the scaled Gaussian mixture distribution equals the kurtosis of a Gaussian distribution multiplied by a factor that we name q. We see that the multivariate kurtosis depends on the dimensionality of the distribution and that the scaling factor b and the number components allow us to adjust the degree of heavy-tailedness of the noise distribution.

We use the MVDR beamformer as a linear spatial filter for the comparison setup because it is optimal with respect to the maximum likelihood criterion if the noise follows a scaled Gaussian mixture distribution as given in (12). This property can be deduced from the fact that the MVDR beamformer is the ML estimator for each Gaussian mixture component and that the MVDR beamformer is invariant against scaling of the noise correlation matrix. We then combine the MVDR beamformer with an MMSE optimal single-channel postfilter.

Since the input vector given the reference speech DFT coefficient s follows a multivariate complex Gaussian mixture distribution, i.e., $\mathbf{Y}\sim\sum_{m=1}^{M}c_m\mathcal{N}_{\mathbb{C}}(\mathbf{d}s,\mathbf{\Phi}_m)$, the output of the MVDR beamformer is distributed according to a one-dimensional complex Gaussian mixture distribution. More precisely, it is

$$p(T_{\text{MVDR}}(\mathbf{y})|s)=\sum_{m=1}c_m\mathcal{N}_{\mathbb{C}}\Bigg(s,\underbrace{\frac{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{\Phi}_m\mathbf{\Phi}_n^{-1}\mathbf{d}}{(\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d})^2})}_{\sigma_m^2}\Bigg).\tag{15}$$

We adhere to the assumptions regarding speech phase and amplitude that Hendriks et al. introduced in [6] to compute the spatially nonlinear MMSE estimator and derive the postfilter using [23, Eq. 3.339, Eq. 6.643.2, Eq. 9.220.2] and [30, Eq. 10.32.3] in an analog way. We find the estimator $T_{\text{MVDR-MMSE}}$ that combines the MVDR beamformer with the MMSE postfilter to be

$$T_{\text{MVDR-MMSE}}(\mathbf{y})=\\ \nu\frac{\sum_{m=1}^{M}\frac{c_mQ_m}{\sigma_m^2}e^{\left[-\frac{|T_{\text{MVDR}}(\mathbf{y})|^2}{\sigma_m^2}\right]}\frac{\sigma_s^2T_{\text{MVDR}}(\mathbf{y})\mathcal{M}(\nu+1,2,P_m)}{\nu\sigma_m^2+\sigma_s^2}}{\sum_{m=1}^{M}\frac{c_mQ_m}{\sigma_m^2}e^{\left[-\frac{|T_{\text{MVDR}}(\mathbf{y})|^2}{\sigma_m^2}\right]}\mathcal{M}(\nu,1,P_m)}\tag{16}$$

with

$$\mathbf{\Phi}_n=\sum_{m=1}^{M}c_m\mathbf{\Phi}_m,\quad\sigma_m^2=\frac{\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{\Phi}_m\mathbf{\Phi}_n^{-1}\mathbf{d}}{(\mathbf{d}^H\mathbf{\Phi}_n^{-1}\mathbf{d})^2},$$

$$Q_m=(\frac{1}{\sigma_m^2}+\frac{\nu}{\sigma_s^2})^{-\nu}\quad\text{and}\quad P_m=\frac{\sigma_s^2\sigma_m^{-2}|T_{\text{MVDR}}(\mathbf{y})|^2}{\nu\sigma_m^2+\sigma_s^2}.$$

Both the spatially nonlinear MMSE estimator and the MMSE postfilter require an estimate of the spectral power of the speech signal σ_s^2. We estimate the parameter for a given time frame by time-averaging over five successive segments of the clean speech data. The speech parameter ν in (10) is set to 0.25 for the nonlinear MMSE estimator and to 0.5 for the postfilter of the $T_{\text{MVDR-MMSE}}$ estimator because this gives the best results for scaled Gaussian mixture noise distributions with higher kurtosis values.

We model the microphone array as a linear array with five microphones at a distance of 5 cm and generate the vector of speech DFT coefficients $\mathbf{S}$ for a source that is located in endfire position. The noise and speech DFT coefficients are combined to give an SNR of 0 dB.

The left column of Figure 1 shows the segmental SNR improvement of the MVDR beamformer T_{MVDR}, the spatially nonlinear MMSE estimator $\widetilde{T}_{\text{MMSE}}$ derived by Hendriks et al. [6], and the MVDR beamformer combined with an MMSE postfilter $T_{\text{MVDR-MMSE}}$ with respect to the kurtosis factor q defined in (14). We compute the segmental SNR using a segment length of 10 ms as described in [31]. To measure the improvement of the segmental SNR, we compare the mean segmental SNR of

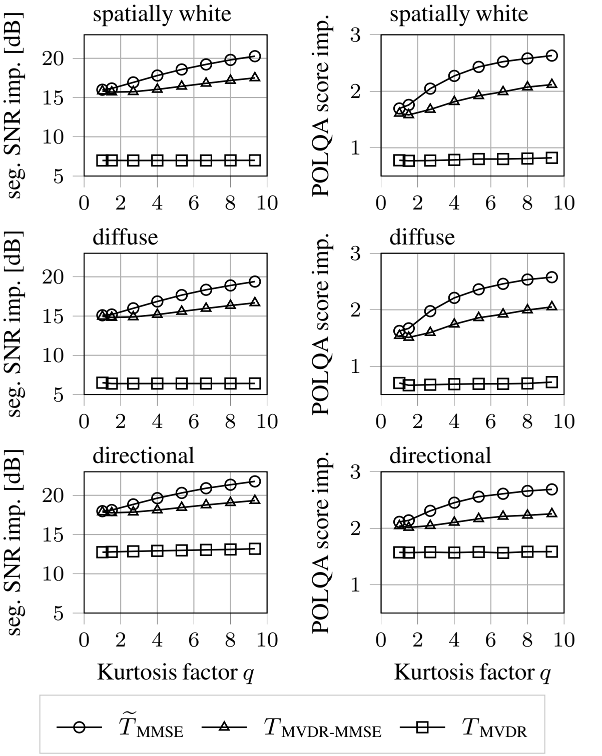


Figure 1: *Segmental SNR and POLQA improvement for noise distributions with increasing kurtosis in three noise scenarios (spatially white, diffuse and directional).*

the noisy microphone recordings to the segmental SNR of the enhanced speech signal. The gap between the top curves (circle and triangle) quantifies the advantage of the nonlinear spatial filter over the linear spatial filter. The difference amounts to values in the order of 2.6 dB for noise that obeys a significantly more heavy-tailed distribution than a Gaussian.

The right column of Figure 1 depicts the perceptual objective listening quality analysis (POLQA) score [32] improvement achieved by the three processing methods. POLQA is the successor of the perceptual evaluation of speech quality (PESQ) measure [33] and returns the expected mean opinion score (MOS) [34] that ranges from one (bad) to five (excellent). As for the segmental SNR improvement, there is a measurable performance difference ($\sim$ 0.5 POLQA score improvement) between the spatially linear and nonlinear estimator. We conclude that the use of a nonlinear spatial filter could be worthwhile if real noise follows a distribution that is considerably more heavy-tailed than a Gaussian distribution.

5. Evaluation on Real-World Noise Data

Using the same estimators as in the previous section, we aim to assess if performance improvements obtained by a nonlinear spatial filter also hold for real-world noise recordings, as provided by the CHiME-3 dataset [17]. We use the five recordings that correspond to the front-facing microphones placed in a frame around a tablet computer that has been used to record noise in four different locations: a bus, a cafeteria, a pedestrian area, and a busy street. We place the target source in the same plane as the tablet, perpendicular to the upper edge, and combine the speech noise signals to obtain an SNR of 0 dB.

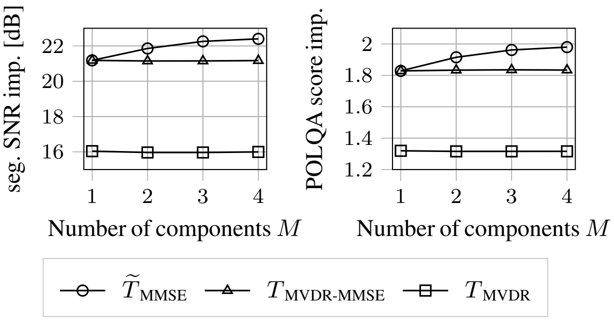


Figure 2: *Segmental SNR and POLQA improvement for CHiME-3 noise recordings from four locations (bus, cafeteria, pedestrian area, street) with respect to the number of mixture components used to fit the noise distribution.*

The estimators $\widetilde{T}_{\mathrm{MMSE}}$ and $T_{\mathrm{MVDR\text{-}MMSE}}$ require the parameters of a zero-mean Gaussian mixture distribution to be estimated from data. We obtain time-variant estimates of the component covariance matrices with the expectation maximization (EM) algorithm [22] applied to signal segments of length 750 ms that overlap by 50% and set the speech parameter $\nu = 0.25$ for both estimators as this gave the best results for the CHiME-3 data.

The left side of Figure 2 depicts the segmental SNR improvement results with respect to the number of components M in the mixture distributions that have been fitted to the data. We find that the use of a postfilter significantly increases the segmental SNR improvement (the difference of 5 dB between the results of T_{MVDR} and $T_{\mathrm{MVDR\text{-}MMSE}}$), but the postfilter following the linear spatial filter in $T_{\mathrm{MVDR\text{-}MMSE}}$ delivers a very similar performance regardless of the number of components of the distribution model. In contrast, we observe that the $\widetilde{T}_{\mathrm{MMSE}}$ estimator with a nonlinear spatial filter achieves better results when we model the distribution through a Gaussian mixture with more components. The performance difference between $\widetilde{T}_{\mathrm{MMSE}}$ and $T_{\mathrm{MVDR\text{-}MMSE}}$ that we attribute to the usage of a nonlinear spatial filter amounts to 1.2 dB averaged over all locations. We make similar observations for the individual locations.

The right plot of Figure 2 shows the improvement with respect to the POLQA measure. The results obtained with the perceptively motivated POLQA measure exhibit the same structure as the results obtained with the segmental SNR and, thus, we find that using a nonlinear spatial filter instead of a linear spatial filter increases the speech quality predicted by POLQA for real-world noise data.

6. Conclusions

In this paper, we showed that using the MMSE optimal nonlinear spatial filter instead of a classical concatenation of a linear spatial filter and a postfilter may yield a performance gain of up to 2.6 dB segmental SNR improvement if the noise follows a distribution with considerably higher multivariate kurtosis than a Gaussian distribution. Also for the real-world noise recordings from the CHiME-3 dataset still moderate improvements of 1.2 dB are achieved when the parameters are estimated on oracle speech and noise data. Future work will analyze the achievable benefit when the filter parameters are estimated blindly from noisy data.

7. Acknowledgment

We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

8. References

[1] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Chichester, England Hoboken, NJ: John Wiley, 2006.

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

[3] J. Benesty, *Microphone array signal processing*. Berlin: Springer, 2008.

[4] R. Balan and J. P. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, Rosslyn, Virginia, Aug. 2002, pp. 209–213.

[5] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2004.

[6] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On optimal multichannel mean-squared error estimators for speech enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, Oct. 2009.

[7] Y. B. Ian Goodfellow and A. Courville, *Deep Learning*. MIT Press, 2016.

[8] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[9] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, Lyon, France, Aug. 2013, pp. 436–440.

[10] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[11] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Interspeech*, San Francisco, USA, Sep. 2016, pp. 3768–3772.

[12] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 196–200.

[14] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and Andrew, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, USA, Dec. 2015, pp. 30–36.

[15] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5075–5079.

[16] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, USA, Dec. 2015, pp. 504–511.

[18] S. M. Kay, *Fundamentals Of Statistical Signal Processing*. Pearson, 2009.

[19] H. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY: Springer New York, 1994.

[20] M. Schervish, *Theory of Statistics*. New York, NY: Springer New York, 1995.

[21] E. L. Lehmann, *Testing statistical hypotheses*. New York: Springer, 2005.

[22] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[23] I. S. Gradshteyn, *Table of integrals, series, and products*. San Diego: Academic Press, 2000.

[24] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," Linguistic Data Consortium, Philadelphia, May 2007.

[25] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, Jan. 2014.

[26] P. H. Westfall, "Kurtosis as peakedness, 1905–2014. R.I.P." *The American Statistician*, vol. 68, no. 3, pp. 191–195, Aug. 2014.

[27] L. T. DeCarlo, "On the meaning and use of kurtosis," *Psychological Methods*, vol. 2, pp. 292–307, Sep. 1997.

[28] K. V. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, no. 3, pp. 519–530, Dec. 1970.

[29] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," https://www.math.uwaterloo.ca/ hwolkowi/matrixcookbook.pdf, November 2012.

[30] "NIST Digital Library of Mathematical Functions," Release 1.0.20 of 2018-09-15, f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds. [Online]. Available: http://dlmf.nist.gov/

[31] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[32] "P.863: Perceptual objective listening quality prediction," International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[33] "P.862.3: Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," International Telecommunication Union, Nov. 2007, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.3-200711-I/en

[34] "P.800: Methods for subjective determination of transmission quality," International Telecommunication Union, Aug. 1996, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I/en

A.2 Nonliner Spatial Filtering for Multichannel Speech Enhancement in Inhomogeneous Noise Fields [P2]

Abstract

A common processing pipeline for multichannel speech enhancement is to combine a linear spatial filter with a single-channel postfilter. In fact, it can be shown that such a combination is optimal in the minimum mean square error (MMSE) sense if the noise follows a multivariate Gaussian distribution. However, for non-Gaussian noise, this serial concatenation is generally suboptimal and may thus also lead to suboptimal results. For instance, in our previous work, we showed that a joint spatial-spectral nonlinear estimator achieves a performance gain of 2.6 dB segmental signal-to-noise ratio (SNR) improvement for heavy-tailed large-kurtosis multivariate noise compared to the traditional combination of a linear spatial beamformer and a postfilter.

In this paper, we show that a joint spatial-spectral nonlinear filter is not only advantageous for noise distributions that are significantly more heavy-tailed than a Gaussian but also for distributions that model inhomogeneous noise fields while having rather low kurtosis. In experiments with artificially created noise we measure a gain of 1 dB for inhomogenous noise with low kurtosis and up to 2 dB for inhomogeneous noise fields with moderate kurtosis.

Reference

Kristina Tesch and Timo Gerkmann, "Nonlinear Spatial Filtering for Multichannel Speech Enhancement in Inhomogeneous Noise Fields", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Barcelona, Spain, 2020, pp. 196-200. DOI 10.1109/ICASSP40776.2020.9053210

Copyright notice

NONLINEAR SPATIAL FILTERING FOR MULTICHANNEL SPEECH ENHANCEMENT IN INHOMOGENEOUS NOISE FIELDS

Kristina Tesch and Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany
kristina.tesch@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

ABSTRACT

A common processing pipeline for multichannel speech enhancement is to combine a linear spatial filter with a single-channel postfilter. In fact, it can be shown that such a combination is optimal in the minimum mean square error (MMSE) sense if the noise follows a multivariate Gaussian distribution. However, for non-Gaussian noise, this serial concatenation is generally suboptimal and may thus also lead to suboptimal results. For instance, in our previous work, we showed that a joint spatial-spectral nonlinear estimator achieves a performance gain of 2.6 dB segmental signal-to-noise ratio (SNR) improvement for heavy-tailed large-kurtosis multivariate noise compared to the traditional combination of a linear spatial beamformer and a postfilter.

In this paper, we show that a joint spatial-spectral nonlinear filter is not only advantageous for noise distributions that are significantly more heavy-tailed than a Gaussian but also for distributions that model inhomogeneous noise fields while having rather low kurtosis. In experiments with artificially created noise we measure a gain of 1 dB for inhomogenous noise with low kurtosis and up to 2 dB for inhomogeneous noise fields with moderate kurtosis.

***Index Terms*—** Multichannel, speech enhancement, nonlinear filtering, acoustic beamforming

1. INTRODUCTION

Speech enhancement algorithms are used to recover a target speech signal from microphone recordings that are corrupted by background noise. These techniques are fundamental to many communication applications such as telephony, hearing aids, and the emerging field of human-machine interaction with an automatic speech recognition (ASR) system.

If several recordings of the target signal from multiple microphones are available, multichannel speech enhancement methods can be used. The advantage of these methods over single-channel approaches, e.g., [1, 2, 3], is that not only tempo-spectral but also spatial information can be included in the processing [4]. In many cases, the spatial filtering is carried out using a so-called *beamformer* that emphasizes a signal from a certain direction and suppresses the signal components originating from other directions. Beamforming is a linear operation: the discrete Fourier transform (DFT) coefficients of all channels are multiplied by complex weights and summed [5].

Commonly, a single-channel method is applied to the output of such a linear spatial filter to further exploit spectral characteristics for suppressing the remaining noise. It is often referred to as a *postfilter*. However, this common processing pipeline, despite its prevalence, is in general suboptimal if the noise does not follow a multivariate complex Gaussian distribution.

Balan and Rosca [6] have shown that the clean speech MMSE estimator for multivariate complex Gaussian noise can be separated into an minimum variance distortionless response (MVDR) beamformer and a single-channel postfilter. In contrast, the work of Hendriks et al. [7] revealed that the MMSE solution for noise that follows a multivariate complex Gaussian mixture distribution inseparably joins the spatial and spectral processing and is even nonlinear in the spatial filter. From these results, it becomes clear that the noise distribution plays an important role in determining whether joint spatial-spectral nonlinear processing could lead to an improved performance. In the sequel, we may refer to an estimator that joins the spatial and spectral processing into a single nonlinear operation a *nonlinear spatial filter*.

It is important to note that characterizing the noise scenarios in which a nonlinear filtering is advantageous gains particular relevance in the context of the neural network revolution. Evermore often, neural networks are trained to solve single-channel and multichannel speech enhancement tasks, e.g., [8, 9, 10, 11]. While neural networks could potentially be used to elegantly approximate nonlinear joint spatial-spectral filters, most neural network approaches for multichannel speech enhancement restrict the spatial filter to be linear [10, 12] or use neural networks just for estimating the beamformer parameters [13]. In contrast, using neural networks for modeling a nonlinear spatial filter is far less common, e.g., [11]. This is also because the potential benefit of using nonlinear spatial filters is not fully understood. Tackling this problem experimentally by trying out different network architectures does not seem to be a satisfying approach to fundamentally understand the potential gain of nonlinear spatial filtering. For instance, network architectures that are more complex than necessary are generally undesirable as they require more data and training time. Therefore, it is important to understand for which noise scenarios learning a nonlinear spatial filter is worthwhile and for which not. For this, we compare the performance obtained by statistical MMSE-optimal estimators to be able to gain more general insights without depending e.g. on specific neural network architectures.

Already in our recent previous work [14], we evaluated the benefit of the optimal MMSE solution of Hendriks et al. with joint spatial-spectral nonlinear filtering by comparing it to the best matching estimator composed of an MVDR beamformer and an MMSE single-channel postfilter. However, in this analysis we obtained Gaussian mixtures by combining Gaussian components with the *same spatial structure* but different scaling. We observed for noise distributions with a high kurtosis, which measures the heavy-tailedness of a distribution [15, 16], a gain of 2.6 dB segmental SNR and 0.5 POLQA score improvement. Furthermore, in [14] we observed a gain of 1.2 dB segmental SNR improvement for noisy mixtures with real-world noise recordings taken from the CHiME-3 data set [17] when fitting a zero-mean multivariate complex Gaussian mixture with four components to the data. Since the nonlinear spatial filter delivers better results than separated processing with a linear spatial filter and a postfilter, one may conclude that the fitted distribution is not Gaussian and may speculate that the distribution has a notably larger kurtosis than a Gaussian distribution. However, examining the kurtosis of the distributions fitted to the CHiME-3 data revealed that the kurtosis is surprisingly low (Section 3). Thus, it seems that the advantage of a joint nonlinear spatial-spectral filter does not only depend on the kurtosis of the noise distribution but also on other properties. The goal of this paper is to analyze how much the *spatial structure* of the noise model impacts performance when using a joint nonlinear spatial-spectral filter instead of the traditional serial concatenation of a linear beamformer and a postfilter.

Section 2 introduces the modeling assumptions and statistical estimators that our analysis is based on. In Section 3, we show that solely the kurtosis of the noise distribution is not sufficient to characterize when the use of a nonlinear spatial filter could be worthwhile and in Section 4 we evaluate to what extent spatial properties of the noise distribution influence the gain achieved with a nonlinear spatial filter.

2. THEORETICAL BACKGROUND

2.1. Signal model

We assume that the target speech signal is disturbed by additive noise and recorded by a microphone array with D microphones. The recorded time-domain signal for every microphone-channel $\ell \in \{1, ..., D\}$ is transformed to the frequency-domain using a windowed DFT yielding DFT coefficients $Y_\ell(k, i) \in \mathbb{C}$ with frequency-bin index k and time-frame index i. As we assume the noise to be additive, we obtain the noisy DFT coefficient $Y_\ell(k, i) \in \mathbb{C}$ as the sum of the clean speech and the noise DFT coefficients $S_\ell(k, i) \in \mathbb{C}$ and $S_\ell(k, i)$, i.e.,

$$Y_\ell(k, i) = S_\ell(k, i) + N_\ell(k, i). \tag{1}$$

We model the DFT coefficents as random variables and assume that they are independent with respect to the frequency-bin index and time-frame index. As a result, we can consider every time-frequency bin separately and omit the frequency-bin and time-frame indices from the notation. Uppercase letters will denote random variables and lowercase letters will be used for their realizations. Further, we assume speech and noise to be uncorrelated and zero-mean.

The noise DFT coefficients of all channels are combined into a vector $\mathbf{N} = [N_1, ..., N_D]^T \in \mathbb{C}^D$ with correlation matrix $\mathbf{\Phi}_\mathrm{n} = \mathbb{E}[\mathbf{N}\mathbf{N}^H]$. The vector of clean speech DFT coefficients is given by $\mathbf{S} = \mathbf{d}S \in \mathbb{C}^D$ with the so-called steering vector $\mathbf{d} \in \mathbb{C}^D$ modeling the propagation path from the single target speaker to the microphones. Then, the vector $\mathbf{Y} = \mathbf{S} + \mathbf{N} \in \mathbb{C}^D$ contains the noisy DFT coefficients for every channel. The spectral power of the clean speech signal S is denoted by $\sigma_\mathrm{s}^2 = \mathbb{E}[|S|^2]$.

2.2. Estimators

In our previous work [14], we gathered theoretical results to point out that the MMSE solution can be separated into an MVDR beamformer defined as

$$T_\mathrm{MVDR}(\mathbf{y}) = \frac{\mathbf{d}^H \mathbf{\Phi}_\mathrm{n}^{-1} \mathbf{y}}{\mathbf{d}^H \mathbf{\Phi}_\mathrm{n}^{-1} \mathbf{d}} \tag{2}$$

and a single-channel postfilter *if the noise follows a multivariate complex Gaussian distribution*. However, this also implies that for the separability into a linear spatial filter concatenated with a spectral postfilter the distribution of the noise plays a decisive role. This becomes clear from the result of Hendriks et al. [7] who show that the MMSE-optimal estimator of the clean speech DFT coefficient S for noise that is distributed according to a multivariate complex Gaussian mixture distribution is in general a nonlinear and non-separable joint spatial-spectral filter.

This Gaussian mixture distribution combines M zero-mean Gaussian components with covariance matrices $\mathbf{\Phi}_\mathrm{m} \in \mathbb{C}^{D \times D}$, $m = 1, ..., M$, and the correspondig noise probability density function (PDF) is given by

$$p(\mathbf{n}) = \sum_{m=1}^{M} c_m \frac{1}{\pi^D |\mathbf{\Phi}_\mathrm{n}|} \exp\left\{ -\mathbf{n}^H \mathbf{\Phi}_\mathrm{n}^{-1} \mathbf{n} \right\} \tag{3}$$

with component weights c_m that sum to one. The estimator T_MMSE for multivariate complex Gaussian mixture distributed noise has been derived by Hendriks et al. [7] under the additional assumption that the clean speech signal amplitude follows a generalized-Gamma distribution with a shape parameter $\nu \in \mathbb{R}^+$ and that the phase $\Psi \in [0, 2\pi)$ is uniformly distributed and independent of the speech amplitude. Then, the MMSE solution is given by

$$T_\mathrm{MMSE}(\mathbf{y}) = \nu \frac{\sum_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_\mathrm{m}|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right]} \frac{\sigma_\mathrm{s}^2 T_\mathrm{MVDR}^{(m)}(\mathbf{y}) \mathcal{M}(\nu+1, 2, P_m)}{\nu(\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d})^{-1} + \sigma_\mathrm{s}^2}}{\sum_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_\mathrm{m}|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right]} \mathcal{M}(\nu, 1, P_m)} \tag{4}$$

with

$$T_\mathrm{MVDR}^{(m)}(\mathbf{y}) = \frac{\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{y}}{\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d}}, \quad Q_m = (\nu + \mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d} \sigma_\mathrm{s}^2)^{-\nu},$$

$$\text{and} \quad P_m = \frac{\sigma_\mathrm{s}^2 \mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d} \left| T_\mathrm{MVDR}^{(m)}(\mathbf{y}) \right|^2}{\nu(\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d})^{-1} + \sigma_\mathrm{s}^2}$$

and $\mathcal{M}(\cdot, \cdot, \cdot)$ being the confluent hypergeometric function [18, Sec. 9.21].

This MMSE estimator cannot be separated into a spatial filter and a spectral postfilter since the observation $\mathbf{y}$ is the input of the linear function $T_\mathrm{MVDR}^{(m)}$, which in turn depends on the summation index, and also occurs in the quadratic term $\exp\left\{-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right\}$. The latter highlights the spatial nonlinearity of the solution.

In [14], we have experimentally quantified the benefit of the nonlinear joint spatial-spectral MMSE-optimal solution T_MMSE over a separated solution $T_\mathrm{MVDR\text{-}MMSE}$ that combines an MVDR beamformer with an MMSE-optimal postfilter. We derived the MMSE postfilter under the same assumptions used to compute T_MMSE to allow for a meaningful comparison. This results in the composite estimator

$$T_\mathrm{MVDR\text{-}MMSE}(\mathbf{y}) = \\ \nu \frac{\sum_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_\mathrm{MVDR}(\mathbf{y})|^2}{\sigma_m^2}\right]} \frac{\sigma_\mathrm{s}^2 T_\mathrm{MVDR}(\mathbf{y}) \mathcal{M}(\nu+1, 2, P_m)}{\nu \sigma_m^2 + \sigma_\mathrm{s}^2}}{\sum_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_\mathrm{MVDR}(\mathbf{y})|^2}{\sigma_m^2}\right]} \mathcal{M}(\nu, 1, P_m)} \tag{5}$$

with

$$\mathbf{\Phi}_\mathrm{n} = \sum_{m=1}^{M} c_m \mathbf{\Phi}_\mathrm{m}, \quad \sigma_m^2 = \frac{\mathbf{d}^H \mathbf{\Phi}_\mathrm{n}^{-1} \mathbf{\Phi}_\mathrm{m} \mathbf{\Phi}_\mathrm{n}^{-1} \mathbf{d}}{(\mathbf{d}^H \mathbf{\Phi}_\mathrm{n}^{-1} \mathbf{d})^2},$$

$$Q_m = \left(\frac{1}{\sigma_m^2} + \frac{\nu}{\sigma_\mathrm{s}^2}\right)^{-\nu} \quad \text{and} \quad P_m = \frac{\sigma_\mathrm{s}^2 \sigma_m^{-2} |T_\mathrm{MVDR}(\mathbf{y})|^2}{\nu \sigma_m^2 + \sigma_\mathrm{s}^2}.$$

The separability into the MVDR beamformer and a single-channel postfilter can be seen from the fact that the observation is contained in this equation only as input to the MVDR beamformer.

Our previous experiments with known noise distributions indicate a dependence of the performance gain achieved by the spatially nonlinear T_{MMSE} on the kurtosis and, thus, on the heavy-tailedness of the noise distribution. For the real-world noise recordings from the CHiME-3 dataset [17] we observed a moderate improvement by using a non-linear spatial filter but did not yet investigate the kurtosis value of the fitted distribtions.

3. MULTIVARIATE KURTOSIS OF CHIME-3 NOISE DATA

The CHiME-3 dataset provides multichannel recordings obtained in different environments: on a moving bus, in a cafeteria, next to a busy street and in a pedestrian area [17]. For our analysis, we use recordings from five front-facing microphones that have been embedded in a frame around a tablet computer. To approximate the unknown and potentially time-variant distribution of the recorded noise data with a zero-mean multivariate complex Gaussian mixture distribution, we apply the expectation maximization (EM) algorithm to windows of length 750 ms that overlap by 50%.

We use the definition of the multivariate kurtosis by Mardia [19], which we extend for the complex-valued case based on the equivalence of a D-dimensional complex Gaussian distribution with a $2D$-dimensional real Gaussian distribution [20, Thm. 15.1]. Then, the kurtosis of a complex-valued random vector $\mathbf{X} \in \mathbb{C}^D$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}_{\text{x}}$ is given by

$$\kappa_{\mathbb{C}}(\mathbf{X}) = \mathbb{E}\left[(2(\mathbf{X}-\boldsymbol{\mu})^H \mathbf{C}_{\text{x}}^{-1}(\mathbf{X}-\boldsymbol{\mu}))^2\right]. \tag{6}$$

The kurtosis of a D-dimensional complex Gaussian distributed random vector $\mathbf{X} \in \mathbb{C}$ depends solely on the dimension D through

$$\kappa_{\mathbb{C}}(\mathbf{X}) = 2D(2D+2). \tag{7}$$

We now normalize all kurtosis values by the kurtosis of the Gaussian distribution with the corresponding dimensionality and name the result the *kurtosis factor* q. Thus, a kurtosis factor of one indicates a Gaussian distribution, while a larger kurtosis indicates a heavy-tailed distribution.

Figure 1 shows the histograms for the estimated kurtosis factors of the distributions that have been fitted to the CHiME-3 data using the EM algorithm. For this, a different number of mixture components M is used. The kurtosis as given in (6) is estimated by averaging over 1000 samples drawn from the distribution that we obtained with the EM algorithm. Using a single mixture component means to fit a Gaussian distribution and, as a result, we observe a peak at a kurtosis factor of 1 for the blue histogram. Estimating higher order statistics is generally difficult and this is reflected in the width of the peak, which shows that the estimate obeys some variance even when estimated from 1000 samples. If we add more components, i.e., $M \in \{2, 3, 4\}$, the peak of the histogram shifts to the right and we tend to observe larger kurtosis factors. The graphic was clipped at a kurtosis factor of 2 to improve the readability but all results are summarized in Table 1 which shows the mean and median values that confirm the observation.

In [14] we have observed that the gain obtained from T_{MMSE} in comparison $T_{\text{MVDR-MMSE}}$ reaches a value of 1.2 dB segmental SNR improvement as the number of components used to fit the noise distribution is increased to four. Here, we find that the kurtosis factor increases with the number of components and, thus, the noise distributions tend to shift towards more heavy-tailed distributions.

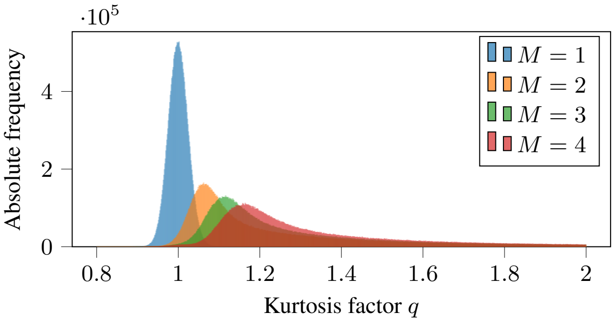


Fig. 1. Histogram of the estimated kurtosis factor for mixture distributions with M components fitted to the CHiME-3 noise data.

M	Mean	Median
1	1.00	1.00
2	1.26	1.13
3	1.36	1.20
4	1.42	1.26

Table 1. Mean and median kurtosis factor per number of components M averaged for all CHiME-3 locations (BUS, CAF, STR, PED).

However, the increase of the mean kurtosis factor up to value of 1.42 for four components is surprisingly small in comparison with the kurtosis factors that we experimented with in [14]. As a result, we conclude that the kurtosis is not the only property of the noise distribution that determines the advantage that we can expect from using the joint spatially and spectrally nonlinear estimator T_{MMSE} compared to a linear spatial filter followed by a postfilter such as $T_{\text{MVDR-MMSE}}$.

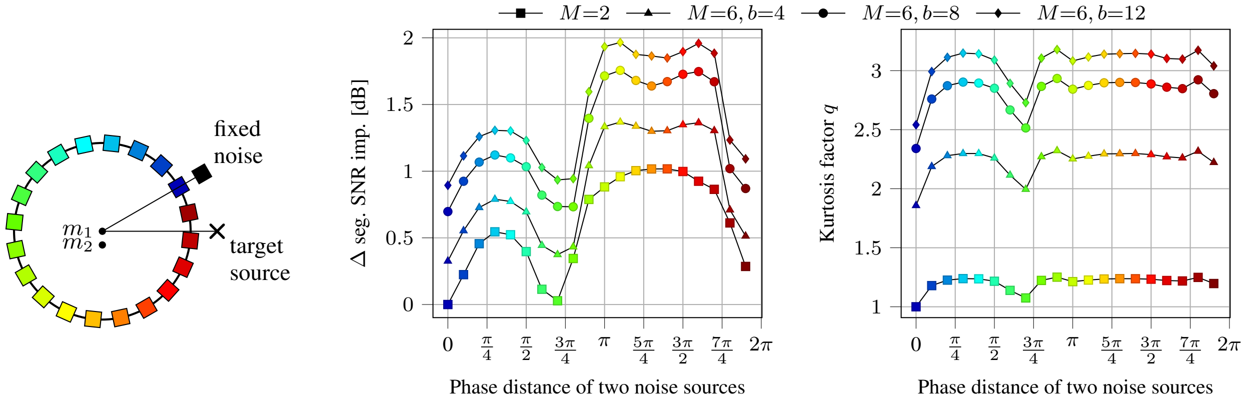
4. NONLINEAR FILTERING FOR INHOMOGENEOUS NOISE SCENARIOS

Next, we investigate the influence of spatial properties of the noise distribution on the performance of the nonlinear joint spatial-spectral T_{MMSE} compared to the concatenation of linear spatial filtering and postfiltering in $T_{\text{MVDR-MMSE}}$. For this, we set up a Gaussian mixture distribution whose Gaussian components are constructed to reassemble the spatial properties of noise point sources placed in different directions and we obtain the noise signal from sampling this multivariate complex Gaussian mixture distribution. Note that this implies that noise sources associated with this overall mixture distribution are non-Gaussian or not active for the same time-frequency bins, which is a common assumption in source separation [21].

The creation of the noise distribution is illustrated in Figure 2a. The center of the image shows a microphone array with two microphones m_1 and m_2 positioned at a distance of 5 cm. The first directional noise source n_1 stays in a fixed position 30 degrees from the target source as depicted in Figure 2a. The second noise source n_2 is placed in 20 different directions, which are indicated by the colored boxes on the circle.

For the noise sources n_1 and n_2, we can compute the steering vectors $\mathbf{d}_{n_1}$ and $\mathbf{d}_{n_2}$, which model the relative time delays of signal arrival at the microphones, based on the noise source incidence angle and the microphone array geometry. From this we construct the correlation matrices modeling the directional noise sources and some additional spatially white noise as [22]

$$\mathbf{\Phi}_{\text{n}_\text{i}} = (1-\alpha_{\text{wn}})\mathbf{d}_{n_i}\mathbf{d}_{n_i}^H + \alpha_{\text{wn}}\mathbf{I} \quad \text{with } i = \{1, 2\}. \tag{8}$$



(a) (b) (c)

Fig. 2. (a) Illustration of the creation of a multivariate Gaussian mixture distribution modelling an inhomogeneous noise field. (b) Performance gain of T_{MMSE} over $T_{\text{MVDR-MMSE}}$ for Gaussian mixture noise modeling two directional sound sources whose placement is illustrated in Figure 2a. (c) Frequency-averaged kurtosis factor of the Gaussian mixture noise modeling two directional sound sources.

The parameter α_{wn} describes the amount of spatially white noise, which we set to $\alpha_{\text{wn}} = 0.05$, and $\mathbf{I}$ denotes the identity matrix. In our first test case we use two equally weighted zero-mean Gaussian components with correlation matrices as given in (8) to construct the overall Gaussian mixture noise distribution. In a second setting, we use three Gaussian components to model one noise source n_i and obtain their correlation matrices $\mathbf{\Phi}_{\text{n}_{\text{ij}}}, j = \{1, 2, 3\}$, by scaling the matrix $\mathbf{\Phi}_{\text{n}_\text{i}}$ under the constraint $\sum_{j=1}^{J} \mathbf{\Phi}_{\text{n}_{\text{ij}}} = \mathbf{\Phi}_{\text{n}_\text{i}}$ with $J = 3$. Based on a varying scale factor $b \in \mathbb{R}^+$ we compute the component correlation matrices as

$$\mathbf{\Phi}_{\text{n}_{\text{ij}}} = \frac{b^{j-1}}{r} \mathbf{\Phi}_{\text{n}_\text{i}} \quad \text{with} \quad r = \sum_{j=1}^{J} \frac{1}{J} b^{j-1}. \tag{9}$$

The overall Gaussian mixture distribution is then scaled such that the noisy observation has an SNR of 0 dB.

For spectral analysis and synthesis we use square-root Hann windows of length 32 ms and a 50% overlap. The speech power σ_{s}^2 is estimated from the clean speech signal by time-averaging over five successive time-frequency bins. We set the speech shape parameter to $\nu = 0.25$ for both estimators and evaluate each configuration on 48 speech signals that have been taken from the WJS0 dataset [23] and balanced between male and female speakers.

Figure 2b shows the performance gain of the spatially and spectrally nonlinear estimator T_{MMSE} over the classic setup $T_{\text{MVDR-MMSE}}$ based on the segmental SNR improvement. We evaluate the segmental SNR of the signals using segments of length 10 ms in which speech is present as proposed, e.g., in [24]. The mean segmental SNR of the two noisy signals is compared to the segmental SNR of the enhanced signal to obtain a measure of the improvement. The performance results are displayed with respect to phase distance of the two noise sources' incidence angles, whereby one of the noise sources moves around the microphone array counterclockwise. The marker colors have been chosen such that they indicate the moving noise source's direction in accordance with the representation in Figure 2a.

The lowest line in Figure 2b with square markers represents the results for a Gaussian mixture distributed noise with two Gaussian components. If the two noise sources are placed in the same direction (zero phase distance, dark blue marker), the Gaussian mixture distribution reduces to a Gaussian distribution and, in accordance with the theory, we cannot observe a benefit from using the joint spatial-spectral nonlinear T_{MMSE} estimator. However, we observe a clear influence of the spatial properties of the noise field and performance gains up to 1 dB.

Our previous conjecture that the kurtosis is not the only property of the noise distribution that affects the performance gain achieved with nonlinear spatial filter is confirmed by Figure 2c. It depicts the normalized kurtosis estimate from 1500 samples which has been averaged over the frequencies on the y-axis and, again, uses the phase distance between the noise sources on the x-axis. We observe rather flat courses and for instance a small kurtosis factor of about 1.2 for the lowest line representing two mixture components ($M = 2$). In particular, the performance difference of 0.5 dB segmental SNR improvement between the first maximum, located at a phase distance of roughly $\frac{\pi}{4}$, and second maximum at a phase distance between $\frac{5\pi}{4}$ and $\frac{3\pi}{2}$ of the lowest curve in Figure 2b do not go along with an increased kurtosis.

The same observation can also be made if three scaled components are used to model each noise source ($M = 6$). A larger scaling factor leads to a higher kurtosis as can be seen in Figure 2c and as we would expect. For example, we observe a kurtosis factor of 3.2 for the scaling factor $b = 12$, but still the performance difference of 0.7 dB segmental SNR improvement for the two spatial scenarios leading to the first and second maximum cannot be predicted from the kurtosis alone.

5. CONCLUSIONS

For multivariate non-Gaussian noise, the traditional concatenation of linear beamforming and spectral postfiltering is not generally optimal. Instead, the MMSE-optimal estimator generally results in a non-separable nonlinear joint spatial-spectral filter. In this paper, we provide further insights into which properties of the multichannel noise impact the potential performance gain when replacing the traditional concatenation of linear beamforming and spectral postfiltering by a joint nonlinear spatial-spectral filter. We show that besides its heavy-tailedness also the spatial structure of the noise distribution plays an important role. In our exemplary setup, we obtain performance gains of up to 2 dB segmental SNR improvement for spatially inhomogeneous noise fields with moderate kurtosis.

6. REFERENCES

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

[2] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[3] T. Gerkmann and E. Vincent, “Spectral masking and filtering,” in *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, Ltd, 2018, ch. 5, pp. 65–85.

[4] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[5] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley, 2006.

[6] R. Balan and J. P. Rosca, “Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase,” in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, Rosslyn, Virginia, Aug. 2002, pp. 209–213.

[7] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, “On optimal multichannel mean-squared error estimators for speech enhancement,” *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, Oct. 2009.

[8] Y. Xu, J. Du, L. Dai, and C. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[9] S.-W. Fu, Y. Tsao, and X. Lu, “SNR-aware convolutional neural network modeling for speech enhancement,” in *Interspeech*, San Francisco, USA, 2016, pp. 3768–3772.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, “Multichannel signal processing with deep neural networks for automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.

[11] H. Lee, H. Y. Kim, W. H. Kang, J. Kim, and N. S. Kim, “End-to-end multi-channel speech enhancement using inter-channel time-restricted attention on raw waveform,” in *Interspeech*, Sep. 2019, pp. 4285–4289.

[12] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, G. Chen, Y. Zhang, M. Mandel, D. Yu, and M. L. Seltzer, “Deep beamforming networks for multi-channel speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5745–5749.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 196–200.

[14] K. Tesch, R. Rehr, and T. Gerkmann, “On nonlinear spatial filtering in multichannel speech enhancement,” in *Proc. Interspeech 2019*, Sep. 2019, pp. 91–95.

[15] L. T. DeCarlo, “On the meaning and use of kurtosis,” *Psychological Methods*, vol. 2, pp. 292–307, Sep. 1997.

[16] P. H. Westfall, “Kurtosis as peakedness, 1905–2014. R.I.P.” *The American Statistician*, vol. 68, no. 3, pp. 191–195, 2014.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[18] I. S. Gradshteyn and J. M. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, 2000.

[19] K. V. Mardia, “Measures of multivariate skewness and kurtosis with applications,” *Biometrika*, vol. 57, no. 3, pp. 519–530, 1970.

[20] S. M. Kay, *Fundamentals Of Statistical Signal Processing*. Pearson, 2009.

[21] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[22] C. Pan, J. Chen, and J. Benesty, “Performance study of the MVDR beamformer as a function of the source incidence angle,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, Jan. 2014.

[23] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) Complete LDC93S6A,” May 2007.

[24] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

A.3 On the Role of Spatial, Spectral, and Temporal Processing for DNN-based Non-linear Multi-channel Speech Enhancement [P4]

Abstract

Employing deep neural networks (DNN) to directly learn filters for multi-channel speech enhancement has potentially two key advantages over a traditional approach combining a linear spatial filter with an independent tempo-spectral post-filter: 1) non-linear spatial filtering allows to overcome potential restrictions originating from a linear processing model and 2) joint processing of spatial and tempo-spectral information allows to exploit interdependencies between different sources of information.

A variety of DNN-based non-linear filters have been proposed recently, for which good enhancement performance is reported. However, little is known about the internal mechanisms which turns network architecture design into a game of chance. Therefore, in this paper, we perform experiments to better understand the internal processing of spatial, spectral and temporal information by DNN-based non-linear filters.

On the one hand, our experiments in a difficult speech extraction scenario confirm the importance of non-linear spatial filtering, which outperforms an oracle linear spatial filter by 0.24 POLQA score. On the other hand, we demonstrate that joint processing results in a large performance gap of 0.4 POLQA score between network architectures exploiting spectral versus temporal information besides spatial information.

Reference

Kristina Tesch, Nils-Hendriks Mohrmann and Timo Gerkmann, “On the Role of Spatial, Spectral, and Temporal Processing for DNN-based Non-linear Multi-channel Speech Enhancement”, in *Proceedings of Interspeech*, Incheon, South Korea, 2022, pp. 2908-2912. DOI: 10.21437/Interspeech.2022-162

On the Role of Spatial, Spectral, and Temporal Processing for DNN-based Non-linear Multi-channel Speech Enhancement

Kristina Tesch, Nils-Hendrik Mohrmann, and Timo Gerkmann

Signal Processing, Universität Hamburg, Germany

firstname.lastname@uni-hamburg.de

Abstract

Employing deep neural networks (DNNs) to directly learn filters for multi-channel speech enhancement has potentially two key advantages over a traditional approach combining a linear spatial filter with an independent tempo-spectral post-filter: 1) non–linear spatial filtering allows to overcome potential restrictions originating from a linear processing model and 2) joint processing of spatial and tempo-spectral information allows to exploit interdependencies between different sources of information.

A variety of DNN-based non-linear filters have been proposed recently, for which good enhancement performance is reported. However, little is known about the internal mechanisms which turns network architecture design into a game of chance. Therefore, in this paper, we perform experiments to better understand the internal processing of spatial, spectral and temporal information by DNN-based non-linear filters.

On the one hand, our experiments in a difficult speech extraction scenario confirm the importance of non-linear spatial filtering, which outperforms an oracle linear spatial filter by 0.24 POLQA score. On the other hand, we demonstrate that joint processing results in a large performance gap of 0.4 POLQA score between network architectures exploiting spectral versus temporal information besides spatial information.

Index Terms: Multi-channel, speech enhancement, joint non-linear spatial and tempo-spectral filtering

1. Introduction

Speech enhancement algorithms are employed to improve the speech quality and speech intelligibility of speech signals recorded in noisy and often reverberant environments. Their use is indispensable for many applications that are required to work reliably in unfavorable acoustic scenarios, e.g., automatic speech recognition or hearing aids. Accordingly, research on this topic has been ongoing for decades.

Many algorithms operate in the short-term Fourier transform (STFT) domain and traditionally rely on a statistical model to derive an analytical clean speech estimator, e.g., [1–4]. However, simplifying assumptions must often be made to keep the problem tractable. For example, neighboring time-frequency-bins are often assumed to be independent. In contrast, recent state-of-the-art single-channel speech enhancement algorithms are built from DNNs [5–7], which do not require an explicit model but learn complex dependencies directly from data. It is common knowledge that correlations in the time and the frequency dimension should be exploited by DNNs for good performance [8, 9].

If the noisy signals are recorded with multiple microphones, then spatial information is available in addition to tempo-spectral information. Traditional approaches usually follow the two-step approach illustrated in Figure 1a, which first applies a linear spatial filter, a so-called beamformer [10, Sec. 12.4.2], and then employs a single-channel post-filter to exploit tempo-spectral information [11, 12]. While such a separated setup is often not considered a limitation, in our prior work [13, 14], we have demonstrated that just assuming non-Gaussian distributed noise leads to a minimum mean square error (MMSE) estimator that combines spatial and spectral processing into a single non-linear operation, which has superior performance over a linear spatial filter combined with a post-filter. While experiments with the analytic estimator show great potential for joint non-linear spatial-spectral processing, practical applicability is questionable because accurate parameter estimation of higher-order statistics required the use of oracle knowledge. However, DNNs provide a data-driven way to implement practical joint spatial and tempo-spectral non-linear filters (JNF). See Figure 1b for an illustration.

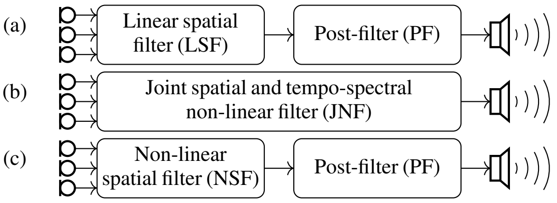

(a)
Linear spatial filter (LSF)
Post-filter (PF)
(b)
Joint spatial and tempo-spectral non-linear filter (JNF)
(c)
Non-linear spatial filter (NSF)
Post-filter (PF)

Figure 1: *(a) The traditional two-step processing using a linear spatial filter (beamformer) followed by a single-channel post-filter. (b) A joint spatial and tempo-spectral non-linear processing scheme that we implement using DNNs in this work. (c) Two-step processing scheme, however, not only the post-filter performs non-linear filtering but also the spatial filter.*

While DNN-based approaches have been dominating the single-channel speech enhancement research for a couple of years now, many publications on multi-channel speech enhancement have proposed to combine DNNs with traditional methods, e.g., [15, 16]. However, the potentially greatest advantage of using DNNs, allowing for non-linear instead of linear spatial processing and taking the interdependencies between spatial and tempo-spectral processing into account, cannot be exploited this way. This is different for the variety of data-driven multi-channel filters that have been proposed recently [17–20]. These approaches report good performance for speech enhancement tasks, but their internal mechanisms are not well understood. However, this is essential for a deliberate design of a network architecture that fully unlocks the potential of neural networks for multi-channel speech enhancement.

In this work, we investigate this internal functioning of DNN-based non-linear filters for multi-channel speech enhancement. We aim to answer the following research questions: Is non-linear as opposed to linear spatial filtering the main factor for good performance? Or is it rather the interdependency between spatial and tempo-spectral processing? In the first case, we could independently perform the non-linear spatial and tempo-spectral processing as shown in Figure 1c, which would be advantageous for practical applications as this allows for independent optimiza-

We thank Rohde&Schwarz SwissQual AG for support with POLQA.

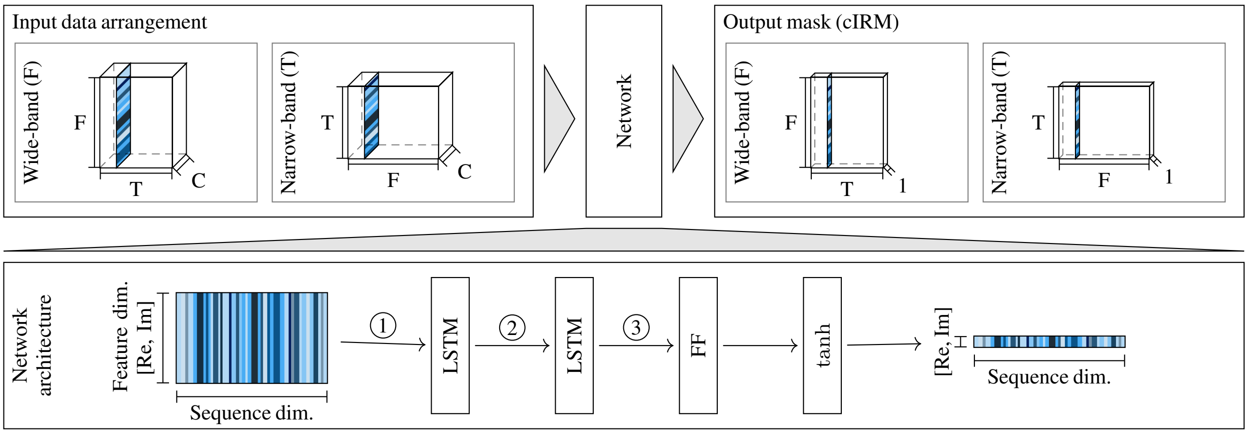


Figure 2: *Illustration of the base system architecture. The input data is arranged according to a wide-band or narrow-band input and fed into a network with two LSTM layers, an FF layer and* $\tanh$ *activation to obtain an estimate of a cIRM.*

tion of either part. Or, on the contrary, is the consideration of interdependencies between spatial and tempo-spectral information particularly important? And do temporal and spectral information have the same impact on spatial filtering performance?

We experimentally address these research questions using a set of DNN-based filter variants derived from a simple base network architecture (outlined in Section 2.1), which are then applied to a challenging speech extraction scenario (described in Section 2.2) that provides a good sense of the spatial filtering capabilities of the employed networks. We provide results for experiments on the separability of spatial and tempo-spectral processing in Section 3 and on the contribution of different information sources in Section 4.

2. DNN-based non-linear filtering for multi-channel speech enhancement

We consider a target clean time-domain signal $s(t)$, which is recorded by a microphone array with C microphone channels in a noisy reverberant room. The recording of $s(t)$ by the ℓ's microphone $x^{(\ell)}(t)$ will then not only undergo a time-shift due to the propagation delay from the target source position to the microphones but also include reverberation [12]. Transforming to the STFT domain leads to the complex-valued coefficient $X^{(\ell)}(k,i)$ with frequency-bin index k and time-frame index i. We use a bold symbol to denote a vector stacking all microphone signals, e.g., $\mathbf{X}(k,i)=[X^{(0)}(k,i),...,X^{(C-1)}(k,i)]\in\mathbb{C}^M$, and drop the indices (k,i) to denote the (multi-channel) spectrogram, e.g., $\mathbf{X}\in\mathbb{C}^{C\times F\times T}$ with F and T denoting the number of frequency-bins and time-frames respectively. We assume that speech $\mathbf{X}$ and noise signal $\mathbf{V}$ sum at the microphones to obtain the noisy signal

$$\mathbf{Y}(k,i)=\mathbf{X}(k,i)+\mathbf{V}(k,i). \tag{1}$$

In our experiments, we use multiple interfering speech sources as noise signal. Similar to the target speech signal, also the interfering signals recorded at the microphones will incorporate spatial information related to the positioning of sources and the characteristics of the room. Given the noisy recording $\mathbf{Y}\in\mathbb{C}^{C\times F\times T}$, we aim to recover the clean target speech signal $S\in\mathbb{C}^{F\times T}$ except for a time-shift caused by the propagation delay to the reference microphone, for which we pick the first channel.

2.1. Network architectures

The focus of this work is to investigate DNN-based non-linear spatial filters for multi-channel speech enhancement in order to better understand the contribution of individual sources of information (spatial, spectral, and temporal) as well as their interdependencies. For this, we develop a number of DNN-based multi-channel filters derived from an long short-term memory (LSTM) network architecture, which has been proposed by Li and Horaud [17].

2.1.1. Base LSTM architecture (F-JNF, T-JNF)

The base architecture is depicted in Figure 2. The multi-channel input (top left) is fed into a neural network (bottom) to obtain a compressed estimate ($C=K=1$ as defined in [21]) of the target speech complex ideal ratio mask (cIRM) $\mathcal{M}_S(k,i)\in\mathbb{C}$ (top right). The target speech signal estimate $\hat{S}(k,i)\in\mathbb{C}$ for every time-frequency-bin (k,i) is then obtained by multiplication of the uncompressed estimated speech mask with the reference channel's noisy recording $Y^{(0)}(k,i)$, i.e.,

$$\hat{S}(k,i)=\mathcal{M}_S(k,i)\,Y^{(0)}(k,i). \tag{2}$$

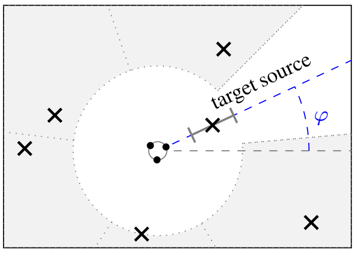
The network architecture is deliberately kept simple with only two bi-LSTM layers followed by a feed forward layer and a $\tanh$ activation. As standard LSTM layers can only process two-dimensional data (a sequence of features), slices of the three-dimensional input are processed independently with the real and imaginary parts being stacked in the channel (feature) dimension. In their work, Li and Horaud [17] propose to independently process the time-sequence of STFT coefficients $\mathbf{Y}(k,\cdot)\in\mathbb{C}^{C\times T}$ for all frequency-bins $k=0,...,F-1$. In Figure 2, this is illustrated as narrow-band input data arrangement and throughout this work we will refer to this as the temporal information based joint non-linear filter (T-JNF). This filter can utilize the fine-grained spatial information in the channel dimension and temporal information along the time axis, however, it does not have access to fine-grained spectral information. In this work, we propose a superior wideband processing scheme (F-JNF) that processes every time-step independently but combines fine-grained spatial and spectral information for mask estimation.

2.1.2. Combining temporal and spectral information (FT-JNF)

While the base architecture combines spatial information with either spectral or temporal information, we now combine all three sources of information within the same general network architecture (leaving the number of parameters unchanged). For this, we propose to simply switch the the data arrangement from wide-band to narrow-band between the two LSTM layers at the position marked with ②. This way, information in all three dimensions can be exploited and the filter is denoted as FT-JNF.

2.1.3. Non-linear spatial filtering (T-NSF, F-NSF, FT-NSF)

To investigate the non-linear spatial filtering capabilities of the DNN-based filter, we adapt the base network architecture to



Room characteristics	
Width	2.5 − 5 m
Length	3 − 9 m
Height	2.2 − 3.5 m
T60	0.2 − 0.5 s

Figure 3: *Illustration of the simulation setup. The target source is located in a fixed orientation with respect to microphone array. The five interfering sources are placed in the gray area (one per segment). Room properties are sampled from the given ranges.*

exclude the second source of fine-grained temporal or spectral information. This is done by randomly shuffling every input along the sequence dimension before feeding the input to the first LSTM at the position marked with ①. This way, only global statistics along the sequence dimension are accessible but correlations between neighboring sequence elements cannot be exploited. However, in a wide-band setup (F-NSF), we noticed that the networks have problems of exploiting spatial information if the frequency bin index is unknown to the network. The frequency-bin index is likely a very important source of information for spatial processing as spatial characteristics strongly depend on the frequency. For this reason, we append the frequency-bin index to the channel dimension such that this information is still available after shuffling along the sequence dimension. For better comparability, we also add the frequency-bin index to the narrow-band setup (T-NSF), however, this has only a small impact on the performance. The permutation of the sequence is then undone after the two LSTM layers (③). As in Section 2.1.2, we define FT-NSF, which switches from narrow-band to wide-band data arrangement at position ②, however, requiring both LSTM layers to be wrapped in permutation and inverse permutation steps.

2.1.4. *DNN-based post-filtering (PF)*

Last, we train a single-channel post-filtering scheme based on the LSTM network architecture denoted by PF. In the single-channel setup, the input has only two dimensions (time and frequency). Here we stack the real and imaginary parts in the frequency dimension, which will serve as feature dimension, while the time dimension corresponds to the sequence dimension.

2.2. Data simulation

For our analyses, we generate a simulated dataset with six spatially displaced speech sources of which one target source is to be extracted. As the target and interfering signal (five speakers) have similar tempo-spectral characteristics, the target speaker has to be identified by its spatial location. Accordingly, we place the target source in a fixed angle with respect to the microphone orientation.

An illustration of the setup simulated with pyroomacoustics [22] is depicted in Figure 3. For each sample, the room dimensions and reverberation time are uniformly sampled from the ranges listed on the right. The uniform circular microphone array has three channels and a diameter of 10 cm. Its position in the xy-plane is sampled to have a minimum distance of 1 m to the walls and placed at height 1.5 m. Furthermore, we randomly rotate the microphone array. The rotation $\varphi \in [0, 2\pi)$ is indicated by the dashed blue line in Figure 3. The target source is placed on the blue line with a minimum distance of 0.3 m and up to 1 m away from the microphone array. Five interfering sources are placed in the gray area leaving a 1 m distance to the microphone array and 20° to the position of the target source with one interfering speech source per segment as indicated by the dotted gray lines. The height of the interfering speech sources is sampled from a normal distribution with mean 1.6 m and standard deviation 0.08.

We generate 6000, 1000, and 600 samples with a sampling frequency of 16 kHz for training, validation and testing respectively using clean speech signals from the WSJ0 dataset [23]. Signals between the different sets do not overlap. The signal-to-noise ratio (SNR) is not explicitly controlled but obtained from the the simulation setup with varying distances of the sources to the microphone array. The average SNR is −4 dB and 95% of the data samples distribute between −9 dB and 2 dB.

2.3. Training details

For training the multi-channel networks (all except PF), we have access to the noisy observations $\mathbf{y}(t)$, the noise signals $\mathbf{v}(t)$ and the dry signal $s(t)$, which has been aligned with the noisy observation to include the propagation path delay. We randomly extract three seconds of audio from the utterances in each training iteration and compute the STFT using a window length of 32 ms and 50% overlap with a $\sqrt{\text{Hann}}$ window for analysis and synthesis. Using the relationship between the real part $\mathrm{Re}(\cdot)$ and the imaginary part $\mathrm{Im}(\cdot)$ of the speech and noise mask

$$\mathrm{Re}(\mathcal{M}_\mathrm{V}) = 1 - \mathrm{Re}(\mathcal{M}_\mathrm{S}), \quad \mathrm{Im}(\mathcal{M}_\mathrm{V}) = -\mathrm{Im}(\mathcal{M}_\mathrm{S}), \tag{3}$$

we obtain an estimate of the noise mask $\mathcal{M}_\mathrm{V}$. From this, an estimate of the noise signal $\hat{V}$ is computed by applying the noise mask in an analog way as the speech mask (see (2)). We use the loss function proposed by Tolooshams et al. [19], which is composed of time and frequency domain ℓ_1 loss terms:

$$L(u,\hat{u}) = \sum_{u \in \{s,v\}} \alpha \|u - \hat{u}\|_1 + \left\| |U| - |\hat{U}| \right\|_1. \tag{4}$$

We set $\alpha = 10$ to equalize the contribution of either domain in the loss term. If the ground truth for the noise signal is unknown, we only use the clean speech related parts of the loss function.

We train the networks with batch size six until convergence (max. 250 epochs) and select the best model based on the validation loss. The number of LSTM units is set to 256 and 128 for all networks, except PF, for which 256 units are used in both layers.

3. Separability of spatial processing and post-filtering

Using the DNN-based filters outlined in the previous section, we investigate if multi-channel non-linear filtering can be separated into spatial processing and single-channel post-filtering. For this, we compare the performance of the three approaches illustrated in Figure 1. The mean POLQA improvement scores [24] along with the 95% confidence interval are presented in Figure 4. The POLQA algorithm provides a measure of speech quality on a mean opinion score (MOS) scale ranging from 1 (low quality) to 5 (high quality). The blue bars in Figure 4 correspond to spatial-only filters. We compute the traditional linear minimum variance distortionless response (MVDR) [10] based on oracle parameter estimates. A time-varying noise covariance estimate is obtained via recursive averaging of the oracle data and the acoustic transfer function (ATF) is estimated by multiplying the principal eigenvector of the generalized eigenvalue problem for speech and noise covariance matrices with the speech covariance matrix [25]. As the parameters are very accurately estimated from oracle data, the displayed results achieved by the MVDR should be considered as an upper bound for the performance that is achievable with the linear processing model. Nevertheless, the oracle MVDR is outperformed by the DNN-based non-linear spatial filter (FT-NSF) evaluated on unseen test data by 0.24 POLQA score. The differences between the two estimates are clearly visible in the middle row of

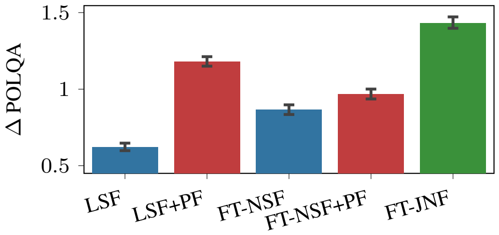


Figure 4: *A comparison of approaches combining a spatial filter (blue) with a post-filter (red) and a joint approach (green). The bars show mean POLQA scores and the 95% confidence interval.*

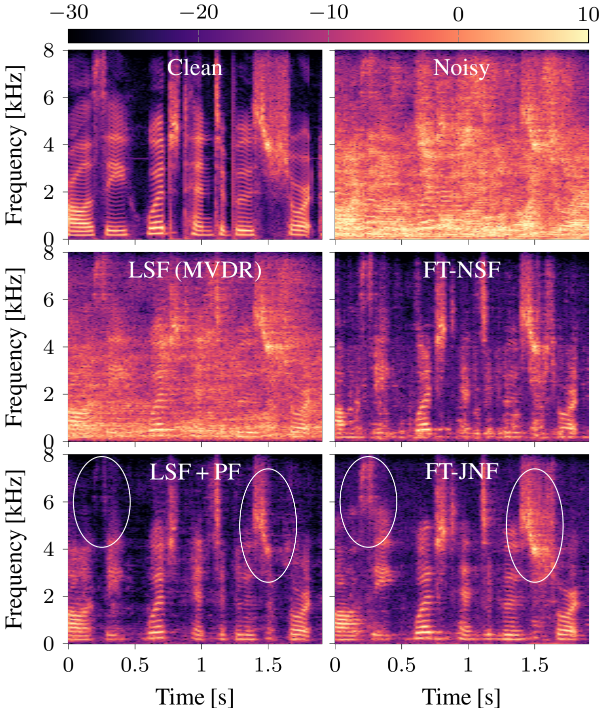


Figure 5: *Spectrogram visualization of an example utterance.*

Figure 5. While the MVDR obeys a distortionless constraint at the cost of only little noise suppression, the non-linear spatial filter provides high noise suppression at the cost of speech distortions.

Next, we combine each spatial filter with a DNN-based post-filter that is trained independently. For this, we obtain enhancement results from the MVDR and the NSF and use this as noisy input to the single-channel network described in Section 2.1.4. The results are depicted by the red bars in Figure 4. We observe that independent post-filtering is far more effective when combined with a distortionless linear spatial filter than a non-linear spatial filter. Applying a tempo-spectral post-filter to the output of the MVDR nearly doubles the performance, which is clearly visible when comparing the two bottom left spectrograms in Figure 5. In contrast, applying an independent post-filter to the output of the DNN-based non-linear spatial filter only slightly improves the performance by about 11%. This is because speech information that was lost during spatial processing cannot be recovered by multiplication with the post-filter mask.

The overall best performance is obtained by the DNN-based joint non-linear filter that does not separate the spatial and tempo-spectral processing. This filter is represented by the green bar in Figure 4 and it outperforms the oracle MVDR with DNN-based post-filter by 0.25 POLQA score. When comparing the spectrogram of the JNF (bottom right) with the MVDR plus post-filter (bottom left), we see that the high frequency clean speech components are preserved better. Please find audio examples on our website[1]. As the joint non-linear filter clearly improves over the non-linear spatial filter plus independent post-filer, we conclude that the spectral and temporal information is used to enhance the non-linear spatial processing itself. Consequently, spatial processing should not be separated from tempo-spectral processing.

Table 1: *Impact of different sources of information included in the processing. We report mean improvements and the 95% confidence interval.*

	Δ POLQA	Δ SI-SDR [dB]
PF	-0.01 ± 0.01	-3.34 ± 0.15
F-NSF	0.78 ± 0.03	7.45 ± 0.11
T-NSF	0.46 ± 0.03	5.64 ± 0.13
FT-NSF	0.87 ± 0.03	7.70 ± 0.12
F-JNF	1.15 ± 0.04	8.99 ± 0.12
T-JNF [17]	0.74 ± 0.03	7.45 ± 0.13
FT-JNF (proposed)	1.43 ± 0.04	9.94 ± 0.13

4. Contribution of information sources

Next, we further investigate the contribution of different sources of information. As the dataset is very challenging with low SNR and many interfering speech sources having similar tempo-spectral structure as the target signal, spatial processing is critical for good performance. The same post-filter trained directly on the noisy input as opposed to the output of a spatial filter performs poorly as reported in the first row of Table 1. In contrast, all filters involving spatial processing provide a substantial improvement score.

In Table 1, we compare the performance of a non-linear spatial filter with access to global spectral, temporal or tempo-spectral information. As expected, incorporating both, the temporal and spectral information, results in higher improvement scores than complementing spatial information only with one other source of information. However, the more surprising observation is that spectral information seems to be much more valuable than temporal information as suggested by the 0.32 POLQA score and 1.8 dB SI-SDR [26] difference in performance improvement.

This finding does not only hold for the global information accessible to the non-linear spatial filter, but also for fine-grained information provided to the joint non-linear filter. The performance differences here amount to 0.41 POLQA score and 1.54 dB SI-SDR. This means that our proposed slight changes to the architecture T-JNF suggested by Li and Horaud [17], which has originally been proposed for the CHiME3 dataset, lead to drastic performance improvements of up to 0.69 POLQA score for FT-JNF on our speech extraction dataset, which requires much stronger spatial filtering capabilities for good performance.

5. Conclusions

In this paper, we have shown that non-linear spatial processing with DNNs is a key to high multi-channel speech enhancement performance. However, the potential of non-linear spatial filtering can only be fully unlocked if spatial processing is tightly integrated with tempo-spectral filtering which contradicts the traditional two-step approach of beamforming followed by post-filtering. We have furthermore shown that, in a difficult speech extraction scenario, which requires strong spatial filtering performance, spectral information is more valuable than temporal information with a difference that amounts to 0.4 POLQA score.

[1] `https://uhh.de/inf-sp-dnn-mc-filter`

6. References

[1] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Audio, Speech, Language Proc.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] T. Lotter and P. Vary, “Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model,” *EURASIP J. Adv. Signal Proc.*, vol. 2005, no. 7, pp. 1110–1126, May 2005.

[3] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors,” *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, pp. 1741–1752, Jul. 2007.

[4] M. Krawczyk-Becker and T. Gerkmann, “On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 12, pp. 2251–2262, 2016.

[5] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement.” in *ISCA Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[6] R. Giri, U. Isik, and A. Krishnaswamy, “Attention Wave-U-Net for speech enhancement,” in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2019, pp. 249–253.

[7] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, “DF-conformer: Integrated architecture of Conv-Tasnet and Conformer using linear complexity self-attention for speech enhancement,” in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2021, pp. 161–165.

[8] J. Richter, G. Carbajal, and T. Gerkmann, “Speech enhancement with stochastic temporal convolutional networks.” in *ISCA Interspeech*, 2020, pp. 4516–4520.

[9] C. Tang, C. Luo, Z. Zhao, W. Xie, and W. Zeng, “Joint time-frequency and time domain learning for speech enhancement,” in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3816–3822.

[10] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Chichester, England Hoboken, NJ: John Wiley, 2006.

[11] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,” in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 39–60.

[12] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[13] K. Tesch, R. Rehr, and T. Gerkmann, “On nonlinear spatial filtering in multichannel speech enhancement,” in *ISCA Interspeech*, Graz, Austria, 2019, pp. 91–95.

[14] K. Tesch and T. Gerkmann, “Nonlinear spatial filtering in multichannel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1795–1805, 2021.

[15] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 196–200.

[16] M. Togami, “Multi-channel Itakura Saito distance minimization with deep neural network,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2019, pp. 536–540.

[17] X. Li and R. Horaud, “Multichannel speech enhancement based on time-frequency masking using subband long short-term memory,” in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, 2019, pp. 298–302.

[18] S. Chakrabarty and E. A. P. Habets, “Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[19] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, “Channel-attention dense U-Net for multichannel speech enhancement,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2020, pp. 836–840.

[20] Z.-Q. Wang, P. Wang, and D. Wang, “Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1778–1787, 2020.

[21] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 3, pp. 483–492, 2016.

[22] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2018, pp. 351–355.

[23] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) Complete LDC93S6A,” Linguistic Data Consortium, Philadelphia, May 2007.

[24] “P.863: Perceptual objective listening quality prediction,” International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[25] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, “Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments,” in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2017, pp. 681–685.

[26] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 626–630.

A.4 Spatially Selective Deep Non-linear Filters for Speaker Extraction [P6]

Abstract

In a scenario with multiple persons talking simultaneously, the spatial characteristics of the signals are the most distinct feature for extracting the target signal. In this work, we develop a deep joint spatial-spectral non-linear filter that can be steered to an arbitrary target direction. For this we propose a simple and effective conditioning mechanism, which sets the initial state of the filter's recurrent layers based on the target direction. We show that this scheme is more effective than the baseline approach and increases the flexibility of the filter at no performance cost. The resulting spatially selective non-linear filters can also be used for speech separation of an arbitrary number of speakers and enable very accurate multi-speaker localization as we demonstrate in this paper.

Reference

Kristina Tesch and Timo Gerkmann, "Spatially Selective Deep Non-linear Filters for Speaker Extraction", in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Rhode Island, Greece, 2023. DOI: 10.1109/ICASSP49357.2023.10096098

Copyright notice

SPATIALLY SELECTIVE DEEP NON-LINEAR FILTERS FOR SPEAKER EXTRACTION

Kristina Tesch and Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany
kristina.tesch@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

ABSTRACT

In a scenario with multiple persons talking simultaneously, the spatial characteristics of the signals are the most distinct feature for extracting the target signal. In this work, we develop a deep joint spatial-spectral non-linear filter that can be steered to an arbitrary target direction. For this we propose a simple and effective conditioning mechanism, which sets the initial state of the filter's recurrent layers based on the target direction. We show that this scheme is more effective than the baseline approach and increases the flexibility of the filter at no performance cost. The resulting spatially selective non-linear filters can also be used for speech separation of an arbitrary number of speakers and enable very accurate multi-speaker localization as we demonstrate in this paper.

Index Terms— Multi-channel, speaker extraction, spatially selective non-linear filters, spatial steering

1. INTRODUCTION

In our everyday life, we are often confronted with the task of listening to a target speaker in a challenging acoustic environment containing noise, interfering human speakers, and reverberation. It is widely known that humans are able to utilize spatial information perceived with both ears to draw attention towards a particular direction of interest. Similarly, spatial information can be used in addition to tempo-spectral information for target speaker extraction in many applications since devices like hearing aids, video-conferencing systems or voice-controlled assistants are nowadays commonly equipped with multiple microphones.

Research into spatial filtering has a long-standing history, which has led to the traditional beamformers, e.g., the delay-and-sum [1] or minimum variance distortionless response (MVDR) beamformer [1,2]. While deep neural networks (DNNs) are considered the state-of-the-art in single-channel speech enhancement and separation, their integration into multi-channel techniques is a very active field of research. Here, one of the most influential ideas of the last years was to use neural networks for beamformer parameter estimation [3, 4]. Despite ease of use and demonstrated robustness of this method, the main drawback of using DNNs only for parameter estimation is that the limitations of the linear beamforming model cannot be overcome, nor can we benefit from joint processing of spatial and tempo-spectral information.

In contrast, an increasing number of recent works, trains a DNN-based filter to perform multi-channel speech enhancement, speaker extraction or separation directly with promising results [5–10]. The theoretic foundation for the potential performance improvements of DNN-based multichannel filters over traditional or DNN-driven beamforming and postfiltering is layed out in our prior work [11]. By means of statistical derivations and proof-of-concept experiments we have shown that (1) a linear beamformer will deliver optimal performance only in rare cases, namely under a multi-variate Gaussian noise assumption, and (2) that non-linear joint spatial-spectral filters may drastically outperform the beamforming plus postfiltering schemes in other cases. DNNs are a natural choice to implement such non-linear joint spatial-spectral filters for practical applications.

Consequently, we [12, 13], and also others [10], have shown that such a DNN-based joint spatial and tempo-spectral non-linear filter drastically outperforms an oracle MVDR beamformer followed by a single-channel post-filter. For this, we evaluated on a speaker extraction task with five interfering speakers. Part of the speaker extraction task is to identify the target speaker. In the literature, different cues have been investigates for this, e.g., enrollment utterances [14, 15] and video information [16, 17].

In this work, however, we focus on the spatial location of the target speaker as cue. Many previous works have explored using spatial features to aid speech separation or speaker extraction [18–22]. For example, Gu et al. [18, 19] and Chen et al. [20] have introduced so-called directional features into their speech separation and extraction systems, which indicate time-frequency bins that are dominated by signal components arriving from a particular direction and are used as additional inputs besides tempo-spectral features. Marković et al. [10] follow a different approach and define spatial regions, e.g., left and right, and train a non-linear filter that suppresses signals from the undesired region but not from the desired region. Tan et al. [7] train a non-linear spatial filter that implicitly steers towards the speech source in enhancement tasks and learns to resolve the speaker-permutation problem by implicitly sorting the speaker outputs according to their location.

In contrast, in this work, we aim for a non-linear joint filter that can be flexibly steered in a direction of choice. This is a major improvement in comparison with our previous well-performing filter [12, 13], which is restricted to a fixed look-direction and thus requires re-training for other directions. For this, we propose a simple conditioning mechanism based on an angular grid with 2° resolution. In comparison with the implicit conditioning mechanism proposed in [23], which manipulates the input signal, our proposed conditioning scheme is more explicit and does not make a far-field assumption.

The rest of this paper is structured as follows: We formally define the speech extraction problem in Section 2 and explain the non-linear filter and its conditioning on a target direction in Section 3. Section 4 describes the experimental setup including datasets and in Section 5, we present results on the effectiveness of the conditioning mechanism and the spatial selectivity of the resulting filter.

2. PROBLEM DEFINITION

This work targets the so-called cocktail-party problem: extracting the speech signal uttered by a target speaker from interfering speech. We assume that the corrupted signal is captured by a microphone array with C channels and denote with $x_\ell(t)$ the recording of the target

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 508337379. We thank Rohde&Schwarz SwissQual AG for their support with POLQA.

speech signal $s(t)$ obtained by the ℓ's microphone. The time-domain signal $x_\ell(t)$ is not only a time-shifted version of $s(t)$ caused by the propagation delay between the speaker and the microphone but also includes reverberation resulting from reflections of the signal from the surrounding walls.

We apply the short-term Fourier transform (STFT) to obtain a frequency-domain representation $X_\ell(k,i) \in \mathbb{C}$ with frequency-bin index k and time-frame index i. The spectral coefficients for all channels are stacked into a vector $\mathbf{X}(k,i) = [X_0(k,i),...,X_{C-1}(k,i)] \in \mathbb{C}^C$. We employ the same signal model to model interfering speech signals and denote the STFT representation of the sum of all interfering signals as $\mathbf{V}(k,i)$. By the additive signal model, the noisy target signal, $\mathbf{Y}(k,i)$, corrupted by interfering speakers, is then given by the sum of the target signal and interfering signal, i.e.,

$$\mathbf{Y}(k,i) = \mathbf{X}(k,i) + \mathbf{V}(k,i). \qquad (1)$$

Given the noisy recording $\mathbf{Y}(k,i)$ we aim to recover the clean target speech signal $S(k,i)$ except for a time-shift caused by the propagation delay to the chosen reference microphone, for which we pick the first channel.

3. SPATIALLY SELECTIVE NON-LINEAR FILTER

In our previous work [12, 13], we have shown that a DNN-based non-linear filter that jointly performs spatial and tempo-spectral filtering, can implicitly be steered into a specific direction, when trained on a fixed geometric setting. Here, we extend the joint non-linear filter from [12, 13], displayed on the left side of Figure 1, with a conditioning mechanism, shown on the right side of Figure 1, that allows the filter to be flexibly steered in a desired direction.

3.1. Joint spatial and tempo-spectral non-linear filter

As indicated by the top left yellow box, the filter takes the frequency-domain raw multi-channel observations as input. Including the batch dimension denoted by B, the input is four-dimensional with T being the number of time-steps, and F the number of the frequency-bins. The real and imaginary parts for all C microphone channels are stacked resulting in the last dimension being $2C$. The filter is composed of only three layers represented by dark green boxes and outputs an estimate of a compressed complex ideal ratio mask (cIRM). We use compression parameters $\mathcal{K} = \mathcal{C} = 1$ as defined in [24] that comply with the range of the tanh activation function used in the last layer. The estimate of the target speech signal $\hat{S}(k,i)$ is then obtained by multiplying the uncompressed mask $\mathcal{M}(k,i) \in \mathbb{C}$ with the reference channel's noisy recording $Y_0(k,i)$, i.e.,

$$\hat{S}(k,i) = \mathcal{M}(k,i) \cdot Y_0(k,i). \qquad (2)$$

The network design is inspired by the work of Li and Horaud [25], who proposed a narrow-band multi-channel speech enhancement scheme. Their core idea is to use a simple network structure (two bi-directional long short-term memory (LSTM) layers and one linear layer) and process all frequency-bins independently while sharing the network parameters between all frequencies. This processing scheme puts a focus on spatial and temporal information and neglects the information present in the frequency dimension. However, our previous work [12, 13] has shown that spectral information, including the correlations between neighboring frequency-bins, should be included in the processing to obtain a filter with high spatial selectivity. Therefore, we rearrange the data such that the first LSTM layer (F-LSTM) focuses on spatial and spectral information and the second LSTM layer (T-LSTM) focuses on spatial and temporal information. The data arrangement is shown in the light green boxes in Figure 1. Before feeding the data into the first LSTM layer, the time-dimension is pulled into the batch dimension, which means that all time-steps are processed independently by the first layer, while the second layer processes all frequency bins independently. This simple change enables capturing spectral correlations and gives rise to state-of-the-art multi-channel speaker extraction and enhancement performance as shown in [13].

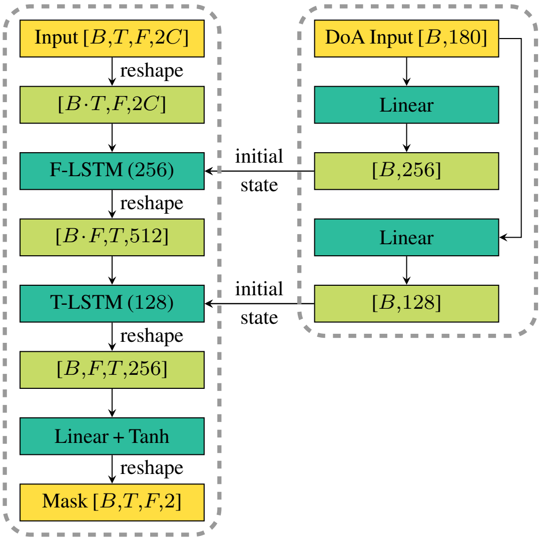


Fig. 1. Illustration of the network architecture. The left part shows the mask estimation network that performs joint spatial and tempo-spectral filtering and the right part shows the conditioning mechanism that enables the filter to be steered towards a chosen direction.

3.2. Directional conditioning

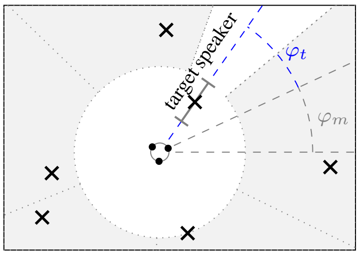
The right part of Figure 1 shows the proposed conditioning mechanism, which enables flexible steering of the filter, which was not possible before. The input is a one hot encoding of the target steering direction. The yellow box shows the dimension for a two degree angle resolution, which results in 180 possible steering directions. Two linear layers are used to map the one-hot encoded input to a dimension that matches in the number of LSTM units, which we set to 256 for the first and 128 for the second layer. The encoded inputs are then used as initial state for the forward and reverse direction of the bi-directional LSTM layers.

This conditioning mechanism, also used by Vinyals et al. [26] for image caption generation, introduces only little overhead as no explicit fusion of input and condition is required. Furthermore, in contrast to [23], which is the only other conditioning scheme for steering a DNN-based filter that we are aware of, it does not make a far-field assumption and can thus easily be trained also for larger microphone distances and/or close speakers.

4. EXPERIMENTAL SETUP

4.1. Datasets

We generate a simulated dataset using pyroomacoustics [27], which implements the source-image model [28]. For each sample, we randomly select width, length, height and reverberation time from the value ranges given in Figure 2. The left side of Figure 2 shows an illustration of the geometric setup of our speaker extraction task. We



Room characteristics	
Width	2.5 − 5 m
Length	3 − 9 m
Height	2.2 − 3.5 m
T60	0.2 − 0.5 s

Fig. 2. Illustration of the simulation setup. The target source is located on the dashed blue line at a random angle φ_t relative to the microphone orientation in the room described by φ_m. Five interfering sources are placed in the gray area (one per segment). Room properties are uniformly sampled from the given ranges.

use a circular microphone array, which has three omni-directional microphones and a 10 cm diameter. The microphone array is placed at a random location for each example, but with at least one meter distance to the walls and at a fixed height of 1.5 m above the floor. For each sample, the microphone array is randomly rotated by $\varphi_m \in [0°,360°]$ as indicated by the dashed gray lines.

4.1.1. *Fixed target speaker location*

The unconditioned joint non-linear filter in [12, 13] learns to steer towards a specific direction based on a fixed target location in the dataset. That is, the target speaker is located in the same direction relative to the microphone orientation (position and rotation) in all samples. In Figure 2, for example, the target speaker is located at a $\varphi_t = 30°$ angle as indicated by the dashed blue lines. The distance between the microphone array and the target speaker ranges from 30 cm to 1 m. The height of the speakers are sampled from a normal distribution with mean 1.6 m and standard deviation 0.08 m. Five interfering sources are placed in the gray area, each of them at least 1 m away from the microphone array and one per segment as illustrated by the dotted gray lines. As indicated by the white area, we leave a side-room of 15° on either side of the target speaker free of interfering sources.

For training the joint non-linear filter, we generate 6000 training examples at 16 kHz sampling frequency with the target speaker located at the chosen direction φ_t. The clean speech utterances are selected from the WSJ0 dataset respecting its train, test and validation split. The SNR (target speech vs mixture of interfering speakers) of the generated samples distributes in the range from −14 dB to 0 dB. For validation and testing, we generate 1000 and 600 utterances respectively.

4.1.2. *Variable target speaker location*

To train a joint non-linear filter that can flexibly steer towards a selected direction, we create a dataset with a variable target speaker location. For this, we discretize the target speaker location $\varphi_t \in [0°, 360°]$ using a 2° resolution, which results in 180 target speaker locations in the training dataset. We generate a dataset with 300 utterances per direction, which results in a total of 54000 training examples. The validation set has 15 examples per direction.

4.1.3. *Multiple target speakers (speech separation)*

In addition to the speaker extraction task, we also evaluate on a speech separation task with multiple target speakers to investigate the spatial selectivity of the filter. The speakers are placed at a distance of 0.8 to 1.2 m away from the microphone array. For speaker angle sampling, we split the circle in as many segments as there are speakers and uniformly place each speaker in one of the segments. Consequently, the speaker angles are likely to not lie on the 2° grid used in training. A minimum angluar distance of 10° is enforced for sources in neighboring segments. We use 1800 utterances with two, three and five mixed speakers for evaluation.

	0°	15°	30°	60°	90°	120°
JNF (fixed)	1.38	1.36	1.34	1.37	1.38	1.39
EaBNet [8] (fixed)	1.16	1.15	1.19	1.20	1.18	1.19
JNF (proposed)	1.38	1.36	1.35	1.36	1.37	1.39
JNF (CoS [23])	1.26	1.27	1.25	1.20	1.26	1.25

Table 1. ΔPOLQA scores for a fixed training scheme (re-training filter for each angle (φ_t) with 6000 examples per direction) in the upper part and the filter conditioned on the given direction in the bottom part. Thus, all results in the bottom rows are obtained with the same non-linear filter, which has been trained with 300 examples per direction.

4.2. Training details

The joint non-linear filters are trained based on an ℓ_1 loss [5], i.e.,

$$L(s,\hat{s}) = \alpha \|s - \hat{s}\|_1 + \Big\| |S| - |\hat{S}| \Big\|_1, \tag{3}$$

with α set to 10 to approximately equalize the contribution of time and frequency-domain loss terms. We train using the Adam [29] optimizer with an initial learning rate of 0.001 and reducing the learning rate by a factor of $\gamma = 0.75$ every 50 epochs. We train with a maximum of 300 epochs using a batch size of eight and select the best network based on the validation loss. For computing the STFT, we use 32 ms windows with 50% overlap and a $\sqrt{\text{Hann}}$ window for synthesis and analysis.

5. RESULTS: SPEAKER EXTRACTION

5.1. Fixed geometry vs conditional training

Our first experiment compares the speech extraction performance of a filter trained for a fixed speaker location and a filter that has been trained for variable target speaker locations using the proposed directional conditioning method (Section 3.2). The first row of Table 1 shows the perceptual objective listening quality analysis (POLQA) [30] mean opinion score (MOS) improvement for six joint non-linear filters, each trained on its own dataset with the target speaker placed at the same respective angle in all 6000 training samples. The improvement performance is very similar for each tested angle, which means that the filters can learn to steer in every direction equally well. As can be seen by the comparison with the Embedding-and-Beamforming Network (EaBNet) in the second row, the learned filters deliver very good state-of-the-art performance. A detailed comparison of more architectures for the 0° fixed case can be found in [13].

In contrast, all results displayed in the third row of Table 1 have been obtained with the same non-linear filter trained on the dataset with variable speaker locations, and conditioned on the respective target angle using the approach proposed in Section 3.2 to obtain the result. As before, we do not observe any major deviations for the different angles and, more importantly, we also do not see a performance degradation in comparison with the non-linear filters in the first row that have been explicitly trained to focus on a fixed spatial location. This is quite remarkable considering that it is a network with only three layers,

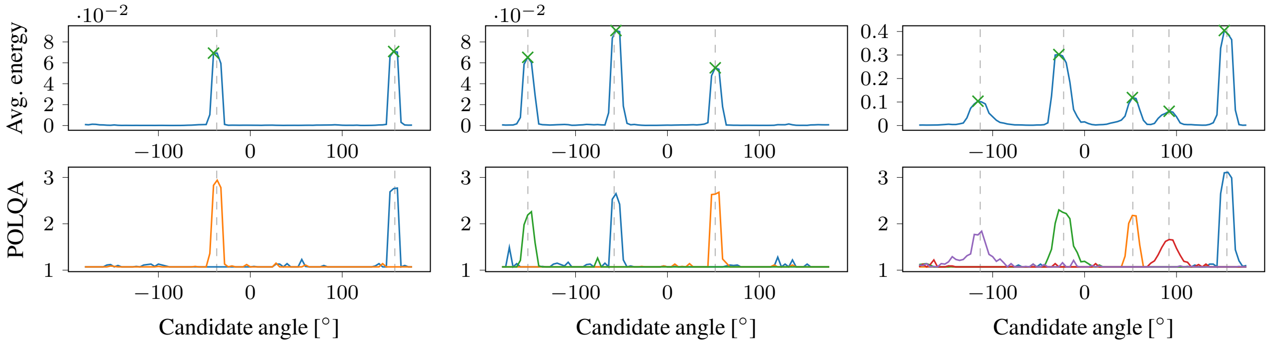


Fig. 3. Examples for blind speaker separation and localization for a mixture of two, three and five speakers using non-linear filters steered in all candidate directions. The vertical dashed gray lines indicate the true positions of the speakers and the green cross marks the estimates speaker location based on the energy peaks in the results.

which is now capable of learning not only one spatial filter but 180 with much fewer (300 instead of 6000) training examples per direction. We also compare with the conditioning mechanism proposed in the cone-of-silence (CoS) paper [23]. For a fair comparison, we use the same network for the non-linear filter and only replace the conditioning mechanism [23] as follows: using knowledge of the array geometry, the channels of the input signal are shifted such that the signals arriving from a given target direction should align according to a far-field assumption. To allow for fractional time-shifts, we perform the alignment in the frequency domain. The idea is that the network learns to extract the speaker signal, which is phase-aligned in the input. Given the results in Table 1, we find that this seems to be a valid cue for extracting the right target, but that it performs approximately 0.1 POLQA score worse than our proposed direct conditioning. We assume that this is mainly related to the limiting far-field assumption in [23].

5.2. Spatial selectivity of the steered filter

Next, we examine the spatial sectivity of the steered filter. Figure 3 shows examples for mixtures of two, three or five speakers. We evaluate the filter on the noisy mixture, generated as described in Section 4.1.3, conditioned on a set of candidate locations using a 4° resolution. For each candidate location, we obtain an estimate of the signal arriving from that particular direction. The plots in the top row show the average segmental energy for this resulting signal $\hat{s}$. We compute the average energy for non-overlapping segments of 10 ms length, in which speech is active, defined analogous to the segmental SNR in [31].

The vertical dashed gray lines indicate the true locations of the speakers in the mixture. In particular for mixtures of two and three speakers, we observe distinct peaks of the energy at the target speaker locations. The small width of the peaks shows high spatial selectivity of the learned filter and proves that it can be steered very accurately towards a specific location. For a mixture of five speakers, we can still see peaks that correspond to the speaker locations, but with a greater width. Likely this is due to the much increased difficulty by a larger overlap of the signals in the time-frequency plane and more reflections arriving from all directions. Still, the POLQA results in the bottom row of Figure 3 show that even five speakers can be separated quite well by the spatially selective filters, which is remarkable given the difficulty of the problem. Audio examples can be found on our website [1].

The green crosses in the top row mark the estimated speaker locations that have been found using `scipy.signal.find_peaks`. We normalize the highest peak to 1 and initially use a prominence of 0.009 and a height of 0.05, which are decreased until enough peaks have been identified. We then merge close-by peaks less then 12° apart and with similar height, which are likely to correspond to the same speaker. If more peaks than speakers are detected, we select the highest peaks. Comparing the distance of the green crosses to the dashed gray line with respect to the x-dimension, shows that the location of the target speakers can be estimated from the steered filter's results quite accurately. In Table 2, we compare the estimated speaker locations using a 4° resolution with the true speaker locations on 1800 mixtures. For two speakers, the average error is only 2.2° including an average quantization error of 1° (as the speaker locations are not limited to the 4° test grid). The error increases for more speakers mainly due to a higher number of errors in the peak-finding heuristic and is still fairly accurate considering the difficulty of the task using only three microphones in a reverberant room. This difficulty is also visible from the fact that the classic SRP-PHAT algorithm [32] is not able to solve the problem in most cases even for only two speakers. As for the extraction task in Table 1, we observe that our proposed conditioning scheme outperforms the baseline CoS approach in all configurations.

# speakers	mean angular error [°]		
	proposed	CoS [23]	SRP-PHAT [32]
2	2.17±0.13	3.72±0.46	17.74±1.04
3	2.47±0.16	3.72±0.36	20.47±0.79
5	3.50±0.21	5.85±0.35	25.62±0.59

Table 2. The speaker localization accuracy for mixtures of two, three and five speakers in a reverberant room. We report the mean angular error and the 95% confidence interval.

6. CONCLUSION

In this paper, we have presented a simple but very effective conditioning mechanism to train a non-linear filter that can be steered in any direction of choice. The conditioning is performed by modifying the initial state of the LSTM layers in the non-linear filter and, thus, introduces only minimal overhead, while achieving the same state-of-the-art performance as a filter with fixed look-direction and also outperforming the baseline cone-of-silence approach. We show that the resulting spatially selective filters can be used for speech separation with an arbitrary number of speakers and can also be employed for accurate multi-speaker localization.

[1] https://uhh.de/inf-sp-spatially-selective

7. REFERENCES

[1] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Chichester, England Hoboken, NJ: John Wiley, 2006.

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Proc. Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[3] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *IEEE Workshop Autom. Speech Recog. and Underst. (ASRU)*, Dec. 2015, pp. 444–451.

[4] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Mar. 2016, pp. 5745–5749.

[5] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense U-net for multichannel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2020, pp. 836–840.

[6] X. Li and R. Horaud, "Narrow-band Deep Filtering for Multichannel Speech Enhancement," *arXiv preprint arXiv:1911.10791*, 2019. [Online]. Available: http://arxiv.org/abs/1911.10791

[7] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 30, pp. 605–621, 2022.

[8] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2022, pp. 6487–6491.

[9] M. M. Halimeh and W. Kellermann, "Complex-valued spatial autoencoders for multichannel speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2022, pp. 261–265.

[10] D. Markovic, A. Defossez, and A. Richard, "Implicit Neural Spatial Filtering for Multichannel Source Separation in the Waveform Domain," in *Interspeech*, Sep. 2022, pp. 1806–1810.

[11] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1795–1805, 2021.

[12] K. Tesch, N.-H. Mohrmann, and T. Gerkmann, "On the role of spatial, spectral, and temporal processing for dnn-based non-linear multi-channel speech enhancement," in *Interspeech*, Sep. 2022, pp. 2908–2912.

[13] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 563–575, 2023.

[14] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018, pp. 5554–5558.

[15] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Interspeech*, Sep. 2019, pp. 2728–2732.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," in *Interspeech*, Sep. 2018, pp. 3244–3248.

[17] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 29, pp. 1368–1396, 2021.

[18] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information." in *Interspeech*, Sep. 2019, pp. 4290–4294.

[19] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Proc. Letters*, vol. 28, pp. 1370–1374, 2021.

[20] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2018, pp. 558–565.

[21] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018, pp. 1–5.

[22] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 27, no. 2, pp. 457–468, 2019.

[23] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, Dec. 2020, pp. 20 925–20 938.

[24] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, no. 3, pp. 483–492, 2016.

[25] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *IEEE Workshop Applications Signal Proc. Audio, Acoustics (WASPAA)*, Oct. 2019, pp. 298–302.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. of the IEEE Conf. on Comp. Vision and Pattern Recog. (CVPR)*, Jun. 2015.

[27] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2018, pp. 351–355.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Repr. (ICLR)*, May 2015.

[30] "P.863: Perceptual objective listening quality prediction," International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[31] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[32] J. H. DiBiase, "A high-accuracy, low -latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

Derivations of the Statistical Estimators

B

The time-frequency bin indices have been dropped for improved readability. Random variables are denoted with a capital letter, and lower-case letters are used for their respective realizations.

B.1 MMSE Estimator Under Gaussian Mixture Noise Assumption

Result B.1 (MMSE Estimator Under Gaussian Mixture Noise Assumption) *Let $\mathcal{M}$ denote the confluent hypergeometric function. Assume the noise discrete Fourier transform (DFT) coefficients to follow a multivariate complex Gaussian mixture distribution with M zero-mean components. The mixing coefficients are denoted by c_i, $i = 1, ..., M$ and the components' covariance matrices are denoted by $\boldsymbol{\Phi}_m$, $i = 1, ..., M$. Further, assume the amplitude of the clean speech coefficient $S = A \cdot e^{j\Psi}$ to be generalized-Gamma distributed with $\gamma = 2$ and $\beta = \nu/\sigma_s^2$, where σ_s^2 denotes the PSD of the speech signal. For phase $\Psi \in [0, 2\pi)$ a uniform distribution is assumed and also that A and Ψ are independent. The vector $\boldsymbol{d}$ denotes the steering vector and C denotes the number of microphones. Then the MMSE estimator is given by*

$$\widetilde{T}_{MMSE}(\boldsymbol{y}) = \nu \frac{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{|\boldsymbol{\Phi}_m|} \exp\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\} \frac{\sigma_s^2 T_{MVDR}^{(m)}(\boldsymbol{y}) \mathcal{M}(\nu+1, 2, P_m)}{\nu (\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d})^{-1} + \sigma_s^2}}{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{|\boldsymbol{\Phi}_m|} \exp\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\} \mathcal{M}(\nu, 1, P_m)} \tag{B.1}$$

with

$$T_{MVDR}^{(m)}(\boldsymbol{y}) = \frac{\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}}{\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}}, \tag{B.2}$$

$$Q_m = (\nu + \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \sigma_s^2)^{-\nu}, \tag{B.3}$$

and

$$P_m = \frac{\sigma_s^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \left| T_{MVDR}^{(m)}(\boldsymbol{y}) \right|^2}{\nu (\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d})^{-1} + \sigma_s^2}. \tag{B.4}$$

Proof. We compute the MMSE estimator using the fact that it equals the mean of the posterior distribution $p(S|\boldsymbol{y})$. For better clarity, the integrals that are to be solved next are given names as can be seen in the following equation:

$$\begin{aligned}
\mathbb{E}[S|\boldsymbol{y}] &= \int_{\mathbb{C}} s \cdot p(s|\boldsymbol{y}) ds \\
&= \int_0^\infty \int_0^{2\pi} a \cdot e^{j\psi} p(a, \psi|\boldsymbol{y}) d\psi da \\
&= \frac{\int_0^\infty \int_0^{2\pi} a \cdot e^{j\psi} p(\boldsymbol{y}|a, \psi) p(a, \psi) d\psi da}{p(\boldsymbol{y})} \\
&= \frac{\int_0^\infty \int_0^{2\pi} a \cdot e^{j\psi} p(\boldsymbol{y}|a, \psi) p(a, \psi) d\psi da}{\int_0^\infty \int_0^{2\pi} p(\boldsymbol{y}|a, \psi) p(a, \psi) d\psi da} \\
&= \frac{\int_0^\infty \int_0^{2\pi} a \cdot e^{j\psi} p(\boldsymbol{y}|a, \psi) p(a) d\psi da}{\int_0^\infty \int_0^{2\pi} p(\boldsymbol{y}|a, \psi) p(a) d\psi da} \\
&= \frac{\int_0^\infty \left(\int_0^{2\pi} e^{j\psi} p(\boldsymbol{y}|a, \psi) d\psi\right) a \cdot p(a) da}{\int_0^\infty \left(\int_0^{2\pi} p(\boldsymbol{y}|a, \psi) d\psi\right) p(a) da} \\
&= \frac{\int_0^\infty I_n^{(1)} \cdot a \cdot p(a) da}{\int_0^\infty I_d^{(1)} \cdot p(a) da}.
\end{aligned} \tag{B.5}$$

1. Solving integrals $I_n^{(1)}$ and $I_d^{(1)}$ over the phase variable Ψ

First, the integral $I_d^{(1)}$ in the denominator of B.5 is computed, which requires the likelihood function. The likelihood is another Gaussian mixture distribution with mean $\boldsymbol{d}s$ for all mixture components and covariance matrices from the noise distribution. We set

$$z_m = |z_m| \cdot e^{j\varphi_{z_m}} = \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y},$$

which simplifies the following expression of the likelihood function:

$$\begin{aligned}
p(\boldsymbol{y}|s) &= p(\boldsymbol{y}|a, \psi) \\
&= \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\left(\boldsymbol{y} - \boldsymbol{d}(a \cdot e^{j\psi})\right)^H \boldsymbol{\Phi}_m^{-1} \left(\boldsymbol{y} - \boldsymbol{d}(a \cdot e^{j\psi})\right)\right\} \\
&= \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} + 2\,\mathrm{Re}\left\{a \cdot e^{-j\psi} \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\right\} \\
&= \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} + 2a|z_m| \cdot \cos(\varphi_{z_m} - \psi) - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\right\}.
\end{aligned} \tag{B.6}$$

To simplify the integrals, we rely on the fact that it is integrated over a full period of the periodic integrand $(*_1)$ and that the cosine function is an even function $(*_2)$. The last step requires the following identity [121, Eq. 3.339]

$$\int_0^\pi \exp\{z\cos(x)\} dx = \pi \mathcal{I}_0(z) \tag{B.7}$$

with $\mathcal{I}_n$ being the modified Bessel function of the first kind and order n. We compute

$$\begin{aligned}
I_d^{(1)} &= \int_0^{2\pi} p(\boldsymbol{y}|a,\psi)d\psi \\
&= \int_0^{2\pi} \left(\sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} + 2a|z_m| \cdot \cos(\varphi_z - \psi) - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} \right) d\psi \\
&= \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} \int_0^{2\pi} \exp\left\{ 2a|z_m| \cdot \cos(\varphi_z - \psi) \right\} d\psi \\
&\overset{*_1}{=} \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} \int_0^{2\pi} \exp\left\{ 2a|z_m| \cdot \cos(\psi) \right\} d\psi \\
&\overset{*_2}{=} \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} 2 \int_0^{\pi} \exp\left\{ 2a|z_m| \cdot \cos(\psi) \right\} d\psi \\
&\overset{(B.7)}{=} \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} 2\pi \mathcal{I}_0(2a|z_m|).
\end{aligned} \tag{B.8}$$

Next, the integral $I_n^{(1)}$ in the numerator of B.5 is computed. Using Euler's formula

$$e^{j\beta} = \cos(\beta) + j\sin(\beta), \tag{B.9}$$

we can divide the integral into two parts ($*_1$). Substitution of $\psi - \varphi_{z_m} = \theta$ ($*_2$) and exploitation of fact that the integrands are 2π-periodic and even just like the cosine function ($*_3$) allow for some simplifications. We can solve one of the integrals directly by computing the antiderivative ($*_4$). The other requires to use the equality

$$\pi \mathcal{I}_n(z) = \int_0^{\pi} e^{z\cos(\theta)} \cos(n\theta) \tag{B.10}$$

that can be found in [122, Eq. 10.32.3]. Overall we compute the result as follows

$$\begin{aligned}
I_n^{(1)} &= \int_0^{2\pi} e^{j\psi} p(\boldsymbol{y}|a,\psi)d\psi \\
&= \sum_{m=1}^{M} \underbrace{\frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\}}_{=\tau_m} \int_0^{2\pi} e^{j\psi} \exp\left\{ 2a|z_m| \cdot \cos(\varphi_{z_m} - \psi) \right\} d\psi \\
&= \sum_{m=1}^{M} \tau_m e^{j\varphi_{z_m}} \int_0^{2\pi} e^{j(\psi - \varphi_{z_m})} \exp\left\{ 2a|z_m| \cdot \cos(\varphi_{z_m} - \psi) \right\} d\psi \\
&\overset{*_1}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_{z_m}} \left(\int_0^{2\pi} \cos(\psi - \varphi_{z_m}) \exp\left\{ 2a|z_m| \cdot \cos(\varphi_{z_m} - \psi) \right\} d\psi \right. \\
&\qquad \left. + j \int_0^{2\pi} \sin(\psi - \varphi_{z_m}) \exp\left\{ 2a|z_m| \cdot \cos(\varphi_{z_m} - \psi) \right\} d\psi \right) \\
&\overset{*_2}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_{z_m}} \left(\int_{-\varphi_{z_m}}^{2\pi - \varphi_{z_m}} \cos(\theta) \exp\left\{ 2a|z_m| \cdot \cos(-\theta) \right\} d\theta \right. \\
&\qquad \left. + j \int_{-\varphi_{z_m}}^{2\pi - \varphi_{z_m}} \sin(\theta) \exp\left\{ 2a|z_m| \cdot \cos(-\theta) \right\} d\theta \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=1}^{M} \tau_m e^{j\varphi_{z_m}} \left(\int_{-\varphi_{z_m}}^{2\pi-\varphi_{z_m}} \cos(\theta) \exp\left\{2a|z_m| \cdot \cos(\theta)\right\} d\theta \right. \\
&\qquad \left. + j \int_{-\varphi_{z_m}}^{2\pi-\varphi_{z_m}} \sin(\theta) \exp\left\{2a|z_m| \cdot \cos(\theta)\right\} d\theta \right) \\
&\overset{*3}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_{z_m}} \left(\int_{0}^{2\pi} \cos(\theta) \exp\left\{2a|z_m| \cdot \cos(\theta)\right\} d\theta \right. \\
&\qquad \left. + j \int_{0}^{2\pi} \sin(\theta) \exp\left\{2a|z_m| \cdot \cos(\theta)\right\} d\theta \right) \\
&\overset{*4}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_{z_m}} \left(2 \int_{0}^{\pi} \cos(\theta) \exp\left\{2a|z_m| \cdot \cos(\theta)\right\} d\theta \right. \\
&\qquad \left. + j \left[-\frac{e^{2a|z_m| \cdot \cos(\theta)}}{2a|z_m|} \right]_0^{2\pi} \right) \\
&= \sum_{m=1}^{M} \tau_m e^{j\varphi_{z_m}} 2 \int_{0}^{\pi} \cos(\theta) \exp\left\{2a|z_m| \cdot \cos(\theta)\right\} d\theta \\
&\overset{(B.10)}{=} \sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} e^{j\varphi_{z_m}} 2\pi \mathcal{I}_1(2a|z_m|).
\end{aligned}
\tag{B.11}
$$

Substitution of the results for $I_n^{(1)}$ and $I_d^{(1)}$ into B.5 yields

$$
\begin{aligned}
\mathbb{E}[S|\boldsymbol{y}] &= \frac{\int_0^\infty I_n \cdot a \cdot p(a) da}{\int_0^\infty I_d \cdot p(a) da} \\
&= \frac{\int_0^\infty \left(\sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} e^{j\varphi_{z_m}} 2\pi \mathcal{I}_1(2a|z_m|) \right) \cdot a \cdot p(a) da}{\int_0^\infty \left(\sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} - a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} 2\pi \mathcal{I}_0(2a|z_m|) \right) \cdot p(a) da} \\
&= \frac{\sum_{m=1}^{M} \frac{c_m \cdot e^{j\varphi_{z_m}}}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} \right\} \left(\int_0^\infty \exp\left\{ -a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} \mathcal{I}_1(2a|z_m|) \cdot a \cdot p(a) da \right)}{\sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} \right\} \left(\int_0^\infty \exp\left\{ -a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \right\} \mathcal{I}_0(2a|z_m|) \cdot p(a) da \right)} \\
&= \frac{\sum_{m=1}^{M} \frac{c_m \cdot e^{j\varphi_{z_m}}}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} \right\} I_n^{(2)}}{\sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{ -\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y} \right\} I_d^{(2)}}.
\end{aligned}
\tag{B.12}
$$

2. Solving integrals $I_n^{(2)}$ and $I_d^{(2)}$ over the amplitude variable A

In order to compute the integrals $I_n^{(2)}$ and $I_d^{(2)}$ in B.12, the probability density function (PDF) of the speech amplitude is plugged in and the following identity [121, Eq. 6.643.2]

$$\int_0^\infty x^{m-\frac{1}{2}} \cdot e^{-kx} \cdot \mathcal{I}_{2n}(2\ell\sqrt{x}) = \frac{\Gamma(m+n+\frac{1}{2})}{\Gamma(2n+1)} \cdot \ell^{-1} \cdot e^{\dfrac{\ell^2}{2k}} \cdot k^{-m} \cdot \mathcal{W}_{-m,n}\left(\frac{\ell^2}{k}\right) \tag{B.13}$$

is used with $\mathcal{W}_{\lambda,\mu}$ being a Whittaker function. This requires a substitution of $a^2 = x$ $(*_1)$. With setting

$$q_m = \frac{\nu}{\sigma_s^2} + \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}$$

the integrals are computed as

$$\begin{aligned}
I_n^{(2)} &= \int_0^\infty \exp\left\{-a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\right\} \mathcal{I}_1(2a|z_m|) \cdot a \cdot p(a) da \\
&= \int_0^\infty \exp\left\{-a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\right\} \mathcal{I}_1(2a|z_m|) \cdot a \cdot \frac{2\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} a^{2\nu-1} \exp\left\{-\frac{\nu}{\sigma_s^2} a^2\right\} da \\
&= \frac{2\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \int_0^\infty \mathcal{I}_1(2a|z_m|) \cdot a^{2\nu} \exp\left\{-a^2 \underbrace{\left(\frac{\nu}{\sigma_s^2} + \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\right)}_{=q_m}\right\} da \\
&\overset{*1}{=} \frac{2\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \int_0^\infty \mathcal{I}_1(2\sqrt{x}|z_m|) \cdot x^\nu \exp\left\{-x \cdot q_m\right\} \frac{1}{2\sqrt{x}} dx \\
&= \frac{\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \int_0^\infty \mathcal{I}_1(2\sqrt{x}|z_m|) \cdot x^{\nu-\frac{1}{2}} \exp\left\{-x \cdot q_m\right\} dx \\
&\overset{(B.13)}{=} \frac{\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \frac{\Gamma(\nu+1)}{\Gamma(2)} \cdot |z_m|^{-1} \cdot \exp\left\{\frac{|z_m|^2}{2q_m}\right\} \cdot q_m^{-\nu} \cdot \mathcal{W}_{-\nu,\frac{1}{2}}\left(\frac{|z_m|^2}{q_m}\right)
\end{aligned} \tag{B.14}$$

and

$$
\begin{aligned}
I_d^{(2)} &= \int_0^\infty \exp\left\{-a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\right\} \mathcal{I}_0(2a|z_m|) \cdot p(a) da \\
&= \int_0^\infty \exp\left\{-a^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\right\} \mathcal{I}_0(2a|z_m|) \cdot \frac{2\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} a^{2\nu-1} \exp\left\{-\frac{\nu}{\sigma_s^2} a^2\right\} da \\
&= \frac{2\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \int_0^\infty \mathcal{I}_0(2a|z_m|) \cdot a^{2(\nu-\frac{1}{2})} \exp\left\{-a^2 q_m\right\} da \\
&\overset{*1}{=} \frac{2\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \int_0^\infty \mathcal{I}_0(2\sqrt{x}|z_m|) \cdot x^{\nu-\frac{1}{2}} \exp\left\{-x \cdot q_m\right\} \frac{1}{2\sqrt{x}} dx \\
&= \frac{\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \int_0^\infty \mathcal{I}_0(2\sqrt{x}|z_m|) \cdot x^{\nu-1} \exp\left\{-x \cdot q_m\right\} dx \\
&\overset{(B.13)}{=} \frac{\left(\dfrac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)} \frac{\Gamma(\nu)}{\Gamma(1)} \cdot |z_m|^{-1} \cdot \exp\left\{\frac{|z_m|^2}{2q_m}\right\} \cdot q_m^{-\nu+\frac{1}{2}} \cdot \mathcal{W}_{-\nu+\frac{1}{2},0}\left(\frac{|z_m|^2}{q_m}\right).
\end{aligned}
\tag{B.15}
$$

3. Rearranging the formulas to match

After insertion of the results for the integrals $I_n^{(2)}$ and $I_d^{(2)}$ in B.12, some more computations have to be performed to reach the final result. The identities that are applied, are

$$
\mathcal{W}_{\lambda,\mu}(z) = z^{\mu+\frac{1}{2}} e^{-\frac{z}{2}} \mathcal{M}(\mu - \lambda + \frac{1}{2}, 2\mu + 1, z) \tag{B.16}
$$

from [121, Eq. 9.220.2],

$$
\frac{|z_m|^2}{q_m} = \frac{|\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{y}|^2}{\dfrac{\nu}{\sigma_s^2} + (\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d})} = \frac{\sigma_s^2(\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d})^{-1}|\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{y}|^2}{\nu(\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d})^{-1} + \sigma_s^2} = \frac{\sigma_s^2 \boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d} \left|T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y})\right|^2}{\nu(\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d})^{-1} + \sigma_s^2} = P_m \tag{B.17}
$$

and

$$
\frac{z_m}{q_m} = \frac{\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{y}}{\dfrac{\nu}{\sigma_s^2} + \boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d}} = \frac{\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d} \cdot T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y})}{\dfrac{\nu}{\sigma_s^2} + \boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d}} = \frac{\sigma_s^2 T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y})}{\nu(\boldsymbol{d}\boldsymbol{\Phi}_m^{-1}\boldsymbol{d})^{-1} + \sigma_s^2}. \tag{B.18}
$$

Finally, the result can be obtained as follows

$$
\begin{aligned}
\mathbb{E}[S|\boldsymbol{y}] &= \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot e^{j\varphi_{z_m}}}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} I_n^{(2)}}{\displaystyle\sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} I_d^{(2)}} \\
&= \nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot e^{j\varphi_{z_m}}}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot |z_m|^{-1} \cdot \exp\left\{\frac{|z_m|^2}{2q_m}\right\} \cdot q_m^{-\nu} \cdot \mathcal{W}_{-\nu,\frac{1}{2}}\left(\frac{|z_m|^2}{q_m}\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot |z_m|^{-1} \cdot \exp\left\{\frac{|z_m|^2}{2q_m}\right\} \cdot q_m^{-\nu+\frac{1}{2}} \cdot \mathcal{W}_{-\nu+\frac{1}{2},0}\left(\frac{|z_m|^2}{q_m}\right)} \\
&\overset{(B.3)}{=} \nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m \cdot e^{j\varphi_{z_m}}}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot |z_m|^{-1} \cdot \exp\left\{\frac{|z_m|^2}{2q_m}\right\} \cdot \mathcal{W}_{-\nu,\frac{1}{2}}\left(\frac{|z_m|^2}{q_m}\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot |z_m|^{-1} \cdot \exp\left\{\frac{|z_m|^2}{2q_m}\right\} \cdot \sqrt{q_m} \cdot \mathcal{W}_{-\nu+\frac{1}{2},0}\left(\frac{|z_m|^2}{q_m}\right)} \\
&\overset{(B.16)}{=} \nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m \cdot e^{j\varphi_{z_m}}}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \frac{|z_m|}{q_m} \cdot \mathcal{M}\left(\nu+1, 2, \frac{|z_m|^2}{q_m}\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \mathcal{M}\left(\nu, 1, \frac{|z_m|^2}{q_m}\right)} \\
&\overset{(B.17)}{=} \nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m \cdot e^{j\varphi_{z_m}}}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \frac{|z_m|}{q_m} \cdot \mathcal{M}\left(\nu+1, 2, P_m\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \mathcal{M}\left(\nu, 1, P_m\right)} \\
&= \nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \frac{z_m}{q_m} \cdot \mathcal{M}\left(\nu+1, 2, P_m\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \mathcal{M}\left(\nu, 1, P_m\right)} \\
&\overset{(B.18)}{=} \nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \frac{\sigma_s^2 T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y})}{\nu(\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d})^{-1} + \sigma_s^2} \cdot \mathcal{M}\left(\nu+1, 2, P_m\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi^C |\boldsymbol{\Phi}_m|} \exp\left\{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}\right\} \cdot \mathcal{M}\left(\nu, 1, P_m\right)}.
\end{aligned}
\tag{B.19}
$$

B.2 MMSE Estimator Under Gaussian Noise Assumption

Result B.2 (MMSE Post-filter at the Ouput of the MVDR Beamformer Under Gaussian Mixture Noise Assumption) *Assume the noise DFT coefficients to follow a multivariate complex Gaussian mixture distribution with M components and mixing coefficients c_i, $i = 1, ..., M$. Further assume the amplitude of the clean speech coefficient $S = A \cdot e^{j\Psi}$ to be generalized-Gamma distributed with $\gamma = 2$ and $\beta = \nu/\sigma_s^2$, the phase $\Psi \in [0, 2\pi)$ to be uniformly distributed and amplitude and phase to be independent. Let $\mathcal{M}$ denote the confluent hypergeometric function. Then the MMSE estimator that is constrained to use the MVDR beamformer as spatial filter is given by*

$$T_{MVDR\text{-}MMSE}(\boldsymbol{y}) = \nu \frac{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} \exp\left\{-\frac{|T_{MVDR}(\boldsymbol{y})|^2}{\sigma_m^2}\right\} \frac{\sigma_s^2 T_{MVDR}(\boldsymbol{y}) \mathcal{M}(\nu+1, 2, P_m)}{\nu\sigma_m^2 + \sigma_s^2}}{\sum\limits_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} \exp\left\{-\frac{|T_{MVDR}(\boldsymbol{y})|^2}{\sigma_m^2}\right\} \mathcal{M}(\nu, 1, P_m)} \tag{B.20}$$

with

$$\boldsymbol{\Phi}_n = \sum_{m=1}^{M} c_m \boldsymbol{\Phi}_m, \tag{B.21}$$

$$\sigma_m^2 = \frac{\boldsymbol{d}^H \boldsymbol{\Phi}_n^{-1} \boldsymbol{\Phi}_m \boldsymbol{\Phi}_n^{-1} \boldsymbol{d}}{(\boldsymbol{d}^H \boldsymbol{\Phi}_n^{-1} \boldsymbol{d})^2}, \tag{B.22}$$

$$Q_m = \left(\frac{1}{\sigma_m^2} + \frac{\nu}{\sigma_s^2}\right)^{-\nu} \tag{B.23}$$

and

$$P_m = \frac{\sigma_s^2 \sigma_m^{-2} |T_{MVDR}(\boldsymbol{y})|^2}{\nu\sigma_m^2 + \sigma_s^2}. \tag{B.24}$$

Proof. Under the multivariate Gaussian mixture noise assumption, the distribution of the output of the MVDR beamformer output given the clean speech signal S is a single-channel Gaussian mixture distribution with all components centered around the clean speech signal and variances [123, Appx. 15B].

$$\sigma_m^2 = \frac{\boldsymbol{d}^H \boldsymbol{\Phi}_n^{-1} \boldsymbol{\Phi}_m \boldsymbol{\Phi}_n^{-1} \boldsymbol{d}}{(\boldsymbol{d}^H \boldsymbol{\Phi}_n^{-1} \boldsymbol{d})^2}. \tag{B.25}$$

Thus, the likelihood function is

$$p(T_{\mathrm{MVDR}}(\boldsymbol{y})|s) = \sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{-\frac{1}{\sigma_m^2} |T_{\mathrm{MVDR}}(\boldsymbol{y}) - s|^2\right\}. \tag{B.26}$$

The second part of the proof is to derive the MMSE estimator for the clean speech signal that operates on the MVDR beamformer output

$$T_{\text{MVDR-MMSE}}(\boldsymbol{y}) = \mathbb{E}[S|T_{\text{MVDR}}(\boldsymbol{y})]. \tag{B.27}$$

To improve the readability we introduce the complex-valued random variable X as the output of the MVDR beamformer

$$T_{\text{MVDR}}(\boldsymbol{y}) = x = |x| \cdot e^{\varphi_x}.$$

We rewrite the PDF of X given the amplitude A and phase Ψ of the clean speech signal as

$$p(x|a,\psi) = \sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{\frac{1}{\sigma_m^2}\left(-|x|^2 + 2a|x|\cos(\varphi_x - \psi) - a^2\right)\right\}. \tag{B.28}$$

Furthermore, the following calculations are carried out in the same way as it has been done in the proof of B.1 for multivariate complex random variables. The integrals that need to be solved are shown in B.29.

$$\begin{aligned}
\mathbb{E}[S|x] &= \frac{\int_0^\infty \int_0^{2\pi} a \cdot e^{j\psi} p(x|a,\psi)p(a)d\psi da}{\int_0^\infty \int_0^{2\pi} p(x|a,\psi)p(a)d\psi da} \\
&= \frac{\int_0^\infty \left(\int_0^{2\pi} e^{j\psi} p(x|a,\psi)d\psi\right) a \cdot p(a)da}{\int_0^\infty \left(\int_0^{2\pi} p(x|a,\psi)d\psi\right) p(a)da} \\
&= \frac{\int_0^\infty I_n^{(1)} \cdot a \cdot p(a)da}{\int_0^\infty I_d^{(1)} \cdot p(a)da}.
\end{aligned} \tag{B.29}$$

For better clarity, the integrals that are to be solved next are given names as can be seen above.

1. Solving integrals $I_n^{(1)}$ and $I_d^{(1)}$ over the phase variable Ψ

First, the integral $I_d^{(1)}$ in the denominator of B.29 is computed. For this purpose, the definition of $p(x|s)$ given in B.28 is used. To simplify the integral, we rely on the fact that it is integrated over a full period of the periodic integrand $(*_1)$ and that the cosine function is an even function $(*_2)$. The last step requires the following identity [121, Eq. 3.339]

$$\int_0^\pi \exp\{z\cos(x)\}dx = \pi\mathcal{I}_0(z) \tag{B.30}$$

with $\mathcal{I}_n$ being the modified Bessel function of the first kind and order n. We find

$$\begin{aligned}
I_d^{(1)} &= \int_0^{2\pi} p(x|a,\psi)d\psi \\
&= \int_0^{2\pi} \left(\sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{\frac{1}{\sigma_m^2}\left(-|x|^2 + 2a|x|\cos(\varphi_x - \psi) - a^2\right)\right\}\right) d\psi \\
&= \sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{\frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\} \int_0^{2\pi} \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\varphi_x - \psi)\right\} d\psi
\end{aligned}$$

$$
\begin{aligned}
&\overset{*_1}{=} \sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{\frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\} \int_0^{2\pi} \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\varphi_x)\right\} d\psi \\
&\overset{*_2}{=} \sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{\frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\} 2\int_0^{\pi} \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\varphi_x)\right\} d\psi \\
&\overset{(B.30)}{=} \sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{\frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\} 2\pi\mathcal{I}_0\left(\frac{2a|x|}{\sigma_m^2}\right). \qquad \text{(B.31)}
\end{aligned}
$$

Next, the integral $I_n^{(1)}$ in the numerator of B.29 is computed. Using Euler's formula $e^{j\beta} = \cos(\beta) + j\sin(\beta)$, we can divide the integral into two parts ($*_1$). Substitution of $\psi - \varphi_{z_m} = \theta$ ($*_2$) and exploitation of fact that the integrands are 2π-periodic and even just like the cosine function ($*_3$, $*_4$) allow for some simplifications. We can solve one of the integrals directly by computing the antiderivative ($*_4$). The other requires to use the equality

$$
\pi\mathcal{I}_n(z) = \int_0^{\pi} e^{z\cos(\theta)}\cos(n\theta) \qquad \text{(B.32)}
$$

that can be found in [122, Eq. 10.32.3]. Overall we compute the result as follows

$$
\begin{aligned}
I_n^{(1)} &= \int_0^{2\pi} e^{j\psi} p(x|a,\psi) d\psi \\
&= \sum_{m=1}^{M} \underbrace{\frac{c_m}{\pi\sigma_m^2} \exp\left\{\frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\}}_{=\tau_m} \int_0^{2\pi} e^{j\psi} \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\varphi_x - \psi)\right\} d\psi \\
&= \sum_{m=1}^{M} \tau_m e^{j\varphi_x} \int_0^{2\pi} e^{j(\psi-\varphi_x)} \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\varphi_x - \psi)\right\} d\psi \\
&\overset{*_1}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_x} \left(\int_0^{2\pi} \cos(\psi - \varphi_x) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\varphi_x - \psi)\right\} d\psi \right. \\
&\qquad \left. + j\int_0^{2\pi} \sin(\psi - \varphi_x) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\varphi_x - \psi)\right\} d\psi \right) \\
&\overset{*_2}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_x} \left(\int_{-\varphi_x}^{2\pi-\varphi_x} \cos(\theta) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(-\theta)\right\} d\theta \right. \\
&\qquad \left. + j\int_{-\varphi_x}^{2\pi-\varphi_x} \sin(\theta) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(-\theta)\right\} d\theta \right) \\
&= \sum_{m=1}^{M} \tau_m e^{j\varphi_x} \left(\int_{-\varphi_x}^{2\pi-\varphi_x} \cos(\theta) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\theta)\right\} d\theta \right. \\
&\qquad \left. + j\int_{-\varphi_x}^{2\pi-\varphi_x} \sin(\theta) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\theta)\right\} d\theta \right) \\
&\overset{*_3}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_x} \left(\int_0^{2\pi} \cos(\theta) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\theta)\right\} d\theta \right. \\
&\qquad \left. + j\int_0^{2\pi} \sin(\theta) \exp\left\{\frac{2a|x|}{\sigma_m^2}\cos(\theta)\right\} d\theta \right)
\end{aligned}
$$

$$\begin{aligned}
&\overset{*_4}{=} \sum_{m=1}^{M} \tau_m e^{j\varphi_x} \left(2\int_0^{\pi} \cos(\theta) \exp\left\{ \frac{2a|x|}{\sigma_m^2}\cos(\theta)\right\} d\theta \right.\\
&\qquad \left. + j \left[- \frac{\exp\left\{ \frac{2a|x|}{\sigma_m^2}\right\}\cos(\theta)}{\frac{2a|x|}{\sigma_m^2}} \right]_0^{2\pi} \right)\\
&= \sum_{m=1}^{M} \tau_m e^{j\varphi_x} 2\int_0^{\pi} \cos(\theta) \exp\left\{ \frac{2a|x|}{\sigma_m^2}\cos(\theta)\right\} d\theta\\
&\overset{(B.32)}{=} \sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{ \frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\} e^{j\varphi_x} 2\pi \mathcal{I}_1\left(\frac{2a|x|}{\sigma_m^2}\right).
\end{aligned} \tag{B.33}$$

Substitution of the results for $I_n^{(1)}$ and $I_d^{(1)}$ into B.29 yields

$$\begin{aligned}
\mathbb{E}[S|x] &= \frac{\int_0^\infty I_n \cdot a \cdot p(a)da}{\int_0^\infty I_d \cdot p(a)da}\\
&= \frac{\int_0^\infty \left(\sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{ \frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\} e^{j\varphi_x} 2\pi \mathcal{I}_1\left(\frac{2a|x|}{\sigma_m^2}\right)\right) \cdot a \cdot p(a)da}{\int_0^\infty \left(\sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{ \frac{1}{\sigma_m^2}\left(-|x|^2 - a^2\right)\right\} 2\pi \mathcal{I}_0\left(\frac{2a|x|}{\sigma_m^2}\right)\right) \cdot p(a)da}\\
&= \frac{\sum_{m=1}^{M} \frac{c_m \cdot e^{j\varphi_x}}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \left(\int_0^\infty \exp\left\{-\frac{a^2}{\sigma_m^2}\right\} \mathcal{I}_1\left(\frac{2a|x|}{\sigma_m^2}\right) \cdot a \cdot p(a)da \right)}{\sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \left(\int_0^\infty \exp\left\{-\frac{a^2}{\sigma_m^2}\right\} \mathcal{I}_0\left(\frac{2a|x|}{\sigma_m^2}\right) \cdot p(a)da \right)}\\
&= \frac{\sum_{m=1}^{M} \frac{c_m \cdot e^{j\varphi_x}}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} I_n^{(2)}}{\sum_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} I_d^{(2)}}.
\end{aligned} \tag{B.34}$$

2. Solving integrals $I_n^{(2)}$ and $I_d^{(2)}$ over the amplitude variable A

In order to compute the integrals $I_n^{(2)}$ and $I_d^{(2)}$ in B.34, the PDF of the speech amplitude is plugged in and the following identity [121, Eq. 6.643.2]

$$\int_0^\infty x^{m-\frac{1}{2}} \cdot e^{-kx} \cdot \mathcal{I}_{2n}(2\ell\sqrt{x}) = \frac{\Gamma(m+n+\frac{1}{2})}{\Gamma(2n+1)} \cdot \ell^{-1} \cdot e^{\frac{\ell^2}{2k}} \cdot k^{-m} \cdot \mathcal{W}_{-m,n}\left(\frac{\ell^2}{k}\right) \tag{B.35}$$

is used with $\mathcal{W}_{\lambda,\mu}$ being a Whittaker function. This requires a substitution of $a^2 = u$ $(*_1)$. With setting

$$q_m = \frac{1}{\sigma_m^2} + \frac{\nu}{\sigma_s^2}$$

the integrals are computed as

$$\begin{aligned}
I_n^{(2)} &= \int_0^\infty \exp\left\{-\frac{a^2}{\sigma_m^2}\right\}\mathcal{I}_1\left(\frac{2a|x|}{\sigma_m^2}\right)\cdot a\cdot p(a)da\\
&= \int_0^\infty \exp\left\{-\frac{a^2}{\sigma_m^2}\right\}\mathcal{I}_1\left(\frac{2a|x|}{\sigma_m^2}\right)\cdot a\cdot \frac{2\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}a^{2\nu-1}\exp\left\{-\frac{\nu}{\sigma_s^2}a^2\right\}da\\
&= \frac{2\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\int_0^\infty \mathcal{I}_1\left(\frac{2a|x|}{\sigma_m^2}\right)\cdot a^{2\nu}\exp\left\{-a^2\underbrace{\left(\frac{1}{\sigma_m^2}+\frac{\nu}{\sigma_s^2}\right)}_{=q_m}\right\}da\\
&\overset{*1}{=} \frac{2\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\int_0^\infty \mathcal{I}_1\left(\frac{2\sqrt{u}|x|}{\sigma_m^2}\right)\cdot u^{\nu}\exp\left\{-u\cdot q_m\right\}\frac{1}{2\sqrt{u}}du\\
&= \frac{\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\int_0^\infty \mathcal{I}_1\left(\frac{2\sqrt{u}|x|}{\sigma_m^2}\right)\cdot u^{\nu-\frac{1}{2}}\exp\left\{-u\cdot q_m\right\}du\\
&\overset{(B.35)}{=} \frac{\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\frac{\Gamma(\nu+1)}{\Gamma(2)}\cdot\left(\frac{|x|}{\sigma_m^2}\right)^{-1}\cdot\exp\left\{\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{2q_m}\right\}\cdot q_m^{-\nu}\cdot\mathcal{W}_{-\nu,\frac{1}{2}}\left(\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right)
\end{aligned}\tag{B.36}$$

and

$$\begin{aligned}
I_d^{(2)} &= \int_0^\infty \exp\left\{-\frac{a^2}{\sigma_m^2}\right\}\mathcal{I}_0\left(\frac{2a|x|}{\sigma_m^2}\right)\cdot p(a)da\\
&= \int_0^\infty \exp\left\{-\frac{a^2}{\sigma_m^2}\right\}\mathcal{I}_0\left(\frac{2a|x|}{\sigma_m^2}\right)\cdot \frac{2\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}a^{2\nu-1}\exp\left\{-\frac{\nu}{\sigma_s^2}a^2\right\}da\\
&= \frac{2\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\int_0^\infty \mathcal{I}_0\left(\frac{2a|x|}{\sigma_m^2}\right)\cdot a^{2(\nu-\frac{1}{2})}\exp\left\{-a^2q_m\right\}da\\
&\overset{*1}{=} \frac{2\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\int_0^\infty \mathcal{I}_0(2\sqrt{u}|z_m|)\cdot u^{\nu-\frac{1}{2}}\exp\left\{-u\cdot q_m\right\}\frac{1}{2\sqrt{u}}du\\
&= \frac{\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\int_0^\infty \mathcal{I}_0(2\sqrt{u}|z_m|)\cdot u^{\nu-1}\exp\left\{-u\cdot q_m\right\}du\\
&\overset{(B.35)}{=} \frac{\left(\frac{\nu}{\sigma_s^2}\right)^\nu}{\Gamma(\nu)}\frac{\Gamma(\nu)}{\Gamma(1)}\cdot\left(\frac{|x|}{\sigma_m^2}\right)^{-1}\cdot\exp\left\{\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{2q_m}\right\}\cdot q_m^{-\nu+\frac{1}{2}}\cdot\mathcal{W}_{-\nu+\frac{1}{2},0}\left(\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right).
\end{aligned}\tag{B.37}$$

3. Rearranging the formulas

After insertion of the results for the integrals $I_n^{(2)}$ and $I_d^{(2)}$ in B.34, some more computations have to be performed to reach the final result. The identities that are applied, are

$$\mathcal{W}_{\lambda,\mu}(z) = z^{\mu+\frac{1}{2}} e^{-\dfrac{z}{2}} \mathcal{M}(\mu-\lambda+\frac{1}{2}, 2\mu+1, z) \tag{B.38}$$

from [121, Eq. 9.220.2],

$$\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m} = \frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{\frac{1}{\sigma_m^2}+\frac{\nu}{\sigma_s^2}} = \frac{|x|^2}{\sigma_m^2+\frac{\sigma_m^4\nu}{\sigma_s^2}} = \frac{\sigma_s^2\sigma_m^{-2}|x|^2}{\nu\sigma_m^2+\sigma_s^2} = \frac{\sigma_s^2\sigma_m^{-2}\left|T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y})\right|^2}{\nu\sigma_m^2+\sigma_s^2} = P_m \tag{B.39}$$

and

$$\frac{x}{q_m\sigma_m^2} = \frac{x}{\left(\frac{1}{\sigma_m^2}+\frac{\nu}{\sigma_s^2}\right)\sigma_m^2} = \frac{\sigma_s^2 x}{\sigma_s^2+\nu\sigma_m^2} = \frac{\sigma_s^2 T_{\mathrm{MVDR}}(\boldsymbol{y})}{\sigma_s^2+\nu\sigma_m^2}. \tag{B.40}$$

Finally, the result can be obtained as follows

$$\mathbb{E}[S|x] = \frac{\sum\limits_{m=1}^{M} \frac{c_m\cdot e^{j\varphi_x}}{\pi\sigma_m^2}\exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} I_n^{(2)}}{\sum\limits_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2}\exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} I_d^{(2)}}$$

$$= \nu\frac{\sum\limits_{m=1}^{M} \frac{c_m\cdot e^{j\varphi_x}}{\pi\sigma_m^2}\exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\}\cdot\left(\frac{|x|}{\sigma_m^2}\right)^{-1}\cdot\exp\left\{\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{2q_m}\right\}\cdot q_m^{-\nu}\cdot\mathcal{W}_{-\nu,\frac{1}{2}}\left(\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right)}{\sum\limits_{m=1}^{M} \frac{c_m}{\pi\sigma_m^2}\exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\}\cdot\left(\frac{|x|}{\sigma_m^2}\right)^{-1}\cdot\exp\left\{\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{2q_m}\right\}\cdot q_m^{-\nu+\frac{1}{2}}\cdot\mathcal{W}_{-\nu+\frac{1}{2},0}\left(\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right)}$$

$$\overset{(B.23)}{=} \nu\frac{\sum\limits_{m=1}^{M} \frac{c_m\cdot Q_m\cdot e^{j\varphi_x}}{\pi\sigma_m^2}\exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\}\cdot\left(\frac{|x|}{\sigma_m^2}\right)^{-1}\cdot\exp\left\{\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{2q_m}\right\}\cdot\mathcal{W}_{-\nu,\frac{1}{2}}\left(\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right)}{\sum\limits_{m=1}^{M} \frac{c_m\cdot Q_m}{\pi\sigma_m^2}\exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\}\cdot\left(\frac{|x|}{\sigma_m^2}\right)^{-1}\cdot\exp\left\{\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{2q_m}\right\}\cdot\sqrt{q_m}\cdot\mathcal{W}_{-\nu+\frac{1}{2},0}\left(\frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right)}$$

$$\overset{(B.38)}{=}\nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m \cdot e^{j\varphi_x}}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \cdot \frac{|x|}{q_m\sigma_m^2} \cdot \mathcal{M}\left(\nu+1, 2, \frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \cdot \mathcal{M}\left(\nu, 1, \frac{\left(\frac{|x|}{\sigma_m^2}\right)^2}{q_m}\right)}$$

$$\overset{(B.39)}{=}\nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \cdot \frac{x}{q_m\sigma_m^2} \cdot \mathcal{M}\left(\nu+1, 2, P_m\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \cdot \mathcal{M}\left(\nu, 1, P_m\right)}$$

$$\overset{(B.40)}{=}\nu \frac{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \cdot \frac{\sigma_s^2 x}{\sigma_s^2+\nu\sigma_m^2} \cdot \mathcal{M}\left(\nu+1, 2, P_m\right)}{\displaystyle\sum_{m=1}^{M} \frac{c_m \cdot Q_m}{\pi\sigma_m^2} \exp\left\{-\frac{|x|^2}{\sigma_m^2}\right\} \cdot \mathcal{M}\left(\nu, 1, P_m\right)}. \tag{B.41}$$