CLOSE-MICROPHONE CROSS-TALK CANCELLATION IN ENSEMBLE RECORDINGS WITH STATISTICAL ESTIMATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MUSIC
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Orchisama Das
August 2021

Abstract

While recording an ensemble of musicians, microphone cross-talk, or "bleed", is considered a nuisance by audio engineers. When two microphones pick up the same signal with a time delay, comb filtering artifacts are present. An expensive solution to the microphone bleed problem is to use acoustic isolation panels between musicians. Obviously, this is not feasible in a live setting. Simpler solutions include using directional microphones with specific polar patterns to pick up radiation from a desired direction and using the close-miking technique where microphones are placed at a distance of 5 - 50 cm from the sound source. Interference can become significant in such cases due to the effect of nearby strong reflective surfaces. The complexity lies in the fact that there is usually an arbitrary number and distribution of instruments and microphones, and results are influenced by the room acoustics of the studio where the ensemble is recorded.

In this thesis, I propose statistically optimal estimators to cancel microphone bleed offline in the mixing and production stage. First, a calibration stage is proposed, where one instrument is played at a time and recorded by all the microphones. This single-input, multiple-output (SIMO) system is used to estimate an approximate relative transfer function matrix, which represents the acoustic path from each source to each microphone and encodes the room response, as well as the mic directivity and source radiation patterns. A convex cost function is derived in the time-frequency domain that simultaneously optimizes the sources and the relative transfer function matrix, which is assumed to be time-invariant. It is shown that minimizing this cost function gives the Maximum Likelihood (ML) estimate when the microphone signals are assumed to be normally distributed. The ML estimator is extended to include *a priori* statistics of the sources, and the Maximum Aposteriori Probability (MAP) estimator is derived. The proposed methods are evaluated against a state-of-the-art Multichannel Wiener Filter based algorithm on a simulated dataset of string quartet recordings in a shoebox room, and on a drum-kit recorded in the CCRMA recording studio. Subjective results show that cross-talk cancellation is achieved while maintaining the perceptual quality of the separated sources.

Acknowledgments

Although seemingly, a doctoral program is an individual's journey, one's success is determined by the support they get along the way from supervisors, collaborators, colleagues, family and friends. I would like to thank my advisory committee - Julius Smith, Chris Chafe and Jonathan Abel, whose guidance, faith and brilliance has matured my skills and confidence as a researcher. In particular, the interactions and brain-storming sessions with Jonathan has multiplied my technical knowledge of the discipline. Likewise, the presence of the other faculty members at CCRMA - Takako Fujioka, Jonathan Berger and Ge Wang, has provided an excellent eco-system for my growth as an academic and individual. Needless to stay, my journey at CCRMA would not have been smooth without the help of the dedicated staff members - Debbie Barney, Nette Worthey, Matt Wright, Fernando Lopez-Lezcano, Carlos Sánchez García-Saavedra, Eoin Callery and Constantin Basica. John Chowning, the very first CCRMA-lite, has been every CCRMA student's cheerleader with his limitless enthusiasm. This list also feels incomplete without acknowledging the amicable company of Dave Kerr.

The CCRMA community is enriched by the presence of its current and former graduate students. I want to thank former students Blair Kaneshiro, Kurt James Werner and Romain Michon for paving the way with their achievements and mentorship. More recent graduates of the program - Elliot Kermit Canfield Dafilou, Kitty Shi and Iran Roman have been excellent TAs, colleagues and friends. Current graduate students - Doga Cavdir, Mark Rau, Jack Atherton, Noah Fram, Nolan Lem, Camille Noufi, Scott Oshiro, Mike Mulshine, Kunwoo Kim, Barbara Nerness, Vidya Rangasayee, Lloyd May and Marisse Van Zyl have complemented my journey in big and small ways. Many of them have been my friends and support system during these five years. Former MA/MSTs Megan Jurek, Prateek Murgai, Juan Sierra, Mark Hertensteiner, Yuval Adler, Elena Georgieva, Aditya Chander and Cara Turnbull have been an integral part of my early years as a graduate student. Other graduate students in the Music department - Julie Herndon, Julie Zhu, Davor Branimir Vincze, Chris Lortie, Hassan Estakhrian, Douglas McCausland, Michiko Theurer and Kelly Christensen have contributed to my time at Stanford with their spirit, talent and humour. In relation to this thesis, I'd especially like to thank Noah Fram and Takako Fujioka who have aided with data collection and listening test design.

Finally, my friends at Stanford and the Bay area, Radhika Koul, Harman Kour, Aayan Das,

Megan Jurek, Atreyi Dasmahapatra, Arindam Das and Anupriya Chakraborty, have stood as pillars of support to see me through my good and bad days. Equally significant are my long-distance friends from India (now spread all around the world) - Sneha Roy, Sunayna Chaudhury, Arunima Banerjee, Tiasha Bhattacharjee, Annesha Ganguly, Poulami Bhattacharjee, Mahasweta Chakraborti, Shrutakirti Dutta and Varun Kishore. Last and most importantly, the love and support of my parents, Ranjana Mandal and Suchikkan Das, and my grandparents, Arati Mandal and Sadananda Mandal, has allowed me to undertake and complete this doctoral degree. They have provided the most loving and comforting home to grow up in and go back to, when the burdens of the world have worn me out.

Contents

List of Figures

Chapter 1

Background

1.1 Introduction

In a common recording scenario, an audio engineer might want to record a string quartet. So, they mic each instrument in the quartet and record the musicians playing together. In this session where all the performers play together, there is too much "bleed" or cross-talk in the microphones, for e.g., the microphone on the viola also picks up the cello and the violins, especially if the studio is small and not sound-proof. Now, the engineer must record each instrument separately, because each instrument needs its own mix — the viola may be too screechy and higher frequencies need to be filtered out, the cello may need a different kind of compression than the violin. However, we lose much of the interaction and coordination among the performers when they play in isolation. Research in neuroscience shows that synchrony among musicians playing together has a significant effect on their performance [3].

Let us imagine the same situation in a live concert setting. In this case, the recording engineer does not have the luxury to record each musician separately. They usually deploy directional microphones with specific polar patterns which pick up sound from a desired direction. Another "hack" is to close-mic the sources, i.e, each mic is placed approximately $5 - 50$ cm from the source, so that it picks up radiation primarily from the source of interest. Although this does not eliminate leakage completely, it reduces it significantly. In this thesis, we focus specifically on reducing cross-talk in close-microphone recordings.

Similarly, in conferences and podcasts, multiple speakers may speak simultaneously. Each microphone picks up the primary speaker, as well as cross talk from other speakers plus background noise. It is desirable to cancel this cross talk as it affects speech intelligibility. Both of the above scenarios share a common objective — to get rid of interfering sources from the primary source.

As opposed to a linear mixture of sources, we have a convolutive mixture, where each mic picks up sources convolved with the acoustic path. It is a well known Fourier theorem that convolved signals

in the time domain become multiplicative in the frequency domain, which is why most existing algorithms work in the time-frequency domain with the short-time Fourier transform (STFT). One common method is to represent the acoustic paths with FIR filters, estimate the filter coefficients with adaptive or statistical techniques, and invert them to recover the sources separately.

Most common issues that arise in close microphone interference rejection include the effect of room acoustics and the ratio of the number of sources to the number of sensors. Most algorithms assume the number of mics to be equal to the number of sources — this is the determined case. Typically, the underdetermined case (fewer sensors than sources) is the most complicated. The overdetermined case is more common in recording scenarios, where multiple mics can be used to record the same instrument. The placement of mics and sources is also crucial to the performance of these methods. Any alteration in the location of mics or sources changes the acoustic path (hence, the filter coefficients). The acoustic transfer function also varies with fluctuations in the temperature, pressure and humidity of an acoustic space, and also with the movements of the musicians. The trade-off between leakage reduction and audible distortion also remains an open research problem.

This thesis derives statistically optimal estimators for the sources and acoustic transfer functions for each source-microphone pair in a multichannel close-microphone recording setup. The standard method [4] uses a Wiener filter (the statistically optimal Minimum Mean Squared Error estimator) to do this, and approximates the acoustic path from a source to a mic with a scalar gain and delay term. However, this is an oversimplification since strong early reflections from the room significantly impact microphone leakage [5]. In this thesis, we derive the maximum likelihood (ML) and maximum aposteriori probability (MAP) estimators for both the source and the acoustic transfer functions for each mic-source pair. For each estimator, we setup a convex objective function which converges to the global solution in a few iterations. We work in the time-frequency domain where we have a linear mixture of the sources and noisy initial measurements of the time-varying acoustic transfer functions, which we get from a *calibration stage*, where each performer plays their instrument while the others are silent. This is a practical assumption since it is common-practice to do a 'sound-check' of each microphone before recording.

The proposed estimators are compared against the state-of-the-art Multichannel Wiener filter estimator in two different scenarios — a string quartet in a virtual shoebox room and a drum kit recorded in a studio. Subjective and objective tests are conducted to compare the proposed methods against the state-of-the-art Wiener filter cross-talk canceler. The results show that the proposed methods successfully achieve interference cancellation while preserving the perceptual quality of the target signal.

1.2 Multichannel source separation methods

Broadly speaking, there are three categories of algorithms to achieve the desired result — blind source separation where each of the sources can be recovered individually, beamforming with microphone arrays which focuses on picking up signals from a desired direction only, and adaptive noise cancellation, where the interfering sources act as non-stationary noise that has to be eliminated. Each of these methods has its advantages and shortcomings.

Blind source separation does not require any *a priori* knowledge about the sources and assumes all the sources in the mixture to be statistically independent. The W-disjoint orthogonality criterion has to be satisfied, i.e., time-frequency bins cannot have overlapping sources. Two functions, $s_1(t)$ and $s_2(t)$, are W-disjoint orthogonal if the supports of their windowed Fourier transforms are disjoint, i.e., $s_1(\omega, \tau)s_2(\omega, \tau) = 0 \; \forall \omega, \tau$ [6]. Moreover, with the added effect of reverberation, BSS methods have to estimate filter coefficients of the order of thousands, which leads to speed and convergence problems. Furthermore, there are scaling and permutation issues, such as in Independent Components Analysis [7].

Beamforming techniques require microphone arrays with specific geometries. In beamforming the filter coefficients are optimized to produce a spatial pattern with a dominant response for the location of interest. Adaptive beamforming shapes the filter coefficients such that the response is minimized for the positions of interfering signals. In multipath or reverberant environments, however, the interfering signals may reach the sensor array from many directions, and so the optimization often alters the response for the region of interest, thus distorting the signal.

Adaptive noise cancellation methods adaptively tune the time-varying weights of the noise cancelling filter. They typically require a reference of the noise signal which has to be correlated with the noise corrupting the desired signal. In close microphone recordings, this is usually available since there is a microphone dedicated to each instrument/speaker. However, the reference noise signal is corrupted with the desired signal itself, which degrades performance [8]. Additionally, the noise is assumed to be broadband and its statistics are either known or need to be estimated.

Out of these three categories, noise cancellation methods have proved to be most useful in close-microphone applications [4, 9]. This is because BSS algorithms usually require identification of system transfer functions from each source to each microphone, which is complicated. Permutation problems are common in BSS methods because parameters are estimated independently in each frequency bin, and they need to be assigned so as to correspond to the same source across all bins. Beamforming requires a fixed geometry of microphones (here, the microphone locations are arbitrary). Noise cancellation methods bypass these issues, but still require some *a priori* information about the source statistics.

In Sections 1.2.1, 1.2.2 and 1.2.3, we briefly discuss the three categories of solutions available for multi-microphone cross-talk cancellation - namely BSS, beamforming and adaptive noise cancellation, before focusing on the close-microphone case in Section 1.3.

1.2.1 Blind Source Separation (BSS)

Blind source separation aims to segregate individual signals from a mixture of sources, with no *a priori* knowledge about them. It is closely related to the "cocktail party" problem, where multiple speakers speak simultaneously, and the listener is trying to follow one of the speakers. The human brain performs this with auditory streaming [10]. However, unlike computational auditory scene analysis that relies on the principles of human hearing, BSS uses signal processing methods to segregate sources. Some well-known BSS methods are given below.

- **Independent Component Analysis (ICA)** - ICA [7] tries to estimate an unmixing matrix $\boldsymbol{W}$ from the observations $\boldsymbol{x}$ (microphone signals) to recover the sources $\boldsymbol{s}$.

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{A}\boldsymbol{s} \\ \boldsymbol{s} &= \boldsymbol{W}\boldsymbol{x} \end{aligned} \tag{1.1}$$

  The sources are assumed to be independent, and the log likelihood of their joint probability distribution is maximized to get $\boldsymbol{W}$. The PDF of the individual sources has to be non-Gaussian (because the multivariate standard normal distribution is rotationally symmetric). ICA has scaling and permutation problems, i.e, the scaling factors cannot be recovered and contributions of a given source may not be assigned consistently to a single recovered signal for different frequency bins. The problem is more severe with an increasing number of sensors as the number of possible permutations increases. Additionally, it works with *instantaneous mixtures*, not *convolutive mixtures* that take into account the effect of room acoustics.

- **Multichannel NMF** - The Nonnegative matrix factorization algorithm proposed in [11] works on single channel mixtures. NMF for multichannel source separation was proposed in [12]. The multichannel model for the input at the mth microphone is:

$$x_m(k) = \sum_{n=1}^{N} a_{mn}(k) * s_m(k) + b_m(k) \tag{1.2}$$

  where b_m is some additive noise. In the STFT domain, (1.2) can be written as

$$\begin{aligned} x_{m,f\tau} &= \sum_{n=1}^{N} A_{mn,f} s_{m,f\tau} + b_{m,f\tau} \\ \boldsymbol{x}_{f\tau} &= \boldsymbol{A}_f \boldsymbol{s}_{f\tau} + \boldsymbol{b}_{f\tau} \end{aligned} \tag{1.3}$$

  The power spectrogram $|\boldsymbol{S}_m|^2$ of source m is modeled as a product of two nonnegative matrices

$$|\boldsymbol{S}_m|^2 \approx \boldsymbol{W}_m \boldsymbol{H}_m \tag{1.4}$$

The parameters $\boldsymbol{A}_f, \boldsymbol{W}_m, \boldsymbol{H}_m$ need to be estimated. Each source STFT is modeled as a sum of $\mathcal{K}_m$ latent Gaussian components, i.e., $s_{m,f\tau} \sim \mathcal{N}\left(0, \sum_{k\in\mathcal{K}_m} w_{fk}, h_{k\tau}\right)$. Two estimation methods are introduced - in the first method, the joint log likelihood of the multichannel data is maximized with the Expectation Maximization (EM) algorithm. In the second method, the sum of the individual log likelihoods of all channels is maximized using a multiplicative update algorithm - this is the NMF step. The computational load is is few hours per song, which is unsuitable for real-time applications.

- **Convolutive BSS** - The convolutive BSS method proposed by Parra and Spence in [13] optimizes an error function in the least squares sense in the STFT domain. From (1.3), we can write the forward model as:

$$R_x(f,\tau) = \boldsymbol{A}(f)\Lambda_s(f,\tau)\boldsymbol{A}(f)^H + \Lambda_b(f,\tau) \tag{1.5}$$

where $R_x(f,\tau) = \mathbb{E}(\boldsymbol{x}_{f\tau}\boldsymbol{x}_{f\tau}^H)$, and $\Lambda_s(f,\tau), \Lambda_b(f,\tau)$ are the auto-correlation matrices of the sources and noise respectively (which are assumed to be diagonal due to independence). $R_x(f,t)$ is replaced with its sample average given by :

$$\bar{R}_x(f,\tau) = \frac{1}{N}\sum_{n=0}^{N-1} \boldsymbol{x}(f,\tau+nT)\boldsymbol{x}^H(f,\tau+nT) \tag{1.6}$$

The backward model can be written as:

$$\begin{aligned} \hat{\Lambda}_s(f,\tau) &= \boldsymbol{W}(f)[\bar{R}_x(f,\tau) - \Lambda_b(f,\tau)]\boldsymbol{W}^H \\ E(f,\tau) &= \hat{\Lambda}_s(f,\tau) - \Lambda_s(f,\tau) \end{aligned} \tag{1.7}$$

The solution can be obtained by minimizing the sum of squares of the error term $E(f,\tau)$ with respect to $\boldsymbol{W}, \Lambda_s, \Lambda_b$.

$$\begin{aligned} \arg \min_{\boldsymbol{W},\Lambda_s,\Lambda_b} \sum_{\tau=1}^{T}\sum_{f=1}^{K} ||E(f,\tau)||^2 \\ \text{s.t } \boldsymbol{w}(t) = 0, t > Q << T \\ W_{ii}(f) = 1 \end{aligned} \tag{1.8}$$

The first constraint ensures that the filter length Q is less than the frame size T, which forces the solution to be smooth in the frequency domain and thus solves the frequency permutation problem. The least square solutions are found with gradient descent. A problem to be addressed in practice is that the channel is typically non-stationary as well. A slight change in the location or orientation of a source may cause drastic changes in the response characteristic of a room.

- **Spatial Covariance model** - In [14], the contribution of each source to all mixture channels in the time-frequency domain is modeled as a zero-mean Gaussian random variable whose covariance encodes the spatial characteristics of the source.

$$\begin{aligned} \boldsymbol{x}(t) &= \sum_{i=1}^{N} \boldsymbol{c}_i(t) \\ \boldsymbol{c}_i(t) &= \boldsymbol{h}_i(t) * s_i(t) \\ \boldsymbol{c}_i(f,\tau) &= \boldsymbol{h}_i(f) s_i(f,\tau) \end{aligned} \tag{1.9}$$

  $\boldsymbol{c}_i(f,\tau,) \sim \mathcal{N}(0, \boldsymbol{v}_i(f,\tau)\boldsymbol{R}_i(f))$, where $\boldsymbol{v}_i(f,\tau)$ are scalar time-varying variances encoding the spectro-temporal power of the sources, and $\boldsymbol{R}_i(f)$ are the *spatial covariance matrices*. Four specific covariance models are considered, including a full-rank unconstrained model. A family of iterative expectation-maximization (EM) algorithms are derived to estimate the parameters of each model. Suitable procedures are proposed to initialize the mixing filter and spatial covariance matrix with hierarchical clustering of the STFT bins, and to align the order of the estimated sources across all frequency bins based on their estimated directions of arrival (DOA) to solve the permutation problem. In [15], the multichannel filter is derived from the source spectra, $\boldsymbol{v}_i(f,\tau)$, which are estimated with deep neural networks, and the spatial covariance matrices, $\boldsymbol{R}_i(f)$, which are updated iteratively using EM.

Other multichannel BSS algorithms include TRINICON [16] (Triple-N ICA) exploiting non-whiteness, non-stationarity and non-gaussianity of the signal, and [17] where an efficient frequency domain algorithm for BSS is presented.

1.2.2 Beamforming with Microphone Arrays

Microphone arrays consist of a number of microphones arranged in a particular geometric pattern. They are used for source localization, source separation, binaural recordings and ambisonics. Beamforming is a spatial filtering technique that aims to enhance or attenuate signals emanating from particular directions. Beamforming can separate sources with overlapping frequency content that originate at different spatial locations. The simplest beamforming technique is delay-and-sum beamforming, where the signals at the microphones are delayed and then summed in order to combine the signal arriving from the direction of the desired source coherently, while the interference components arriving from other directions cancel to a certain extent due to destructive interference. Statistically optimum beamforming methods for source separation are covered in [18, 19], some of which are discussed below.

- **Linearly Constrained Minimum Variance** - The algorithm proposed by Frost in [20] iteratively adapts the weights of a sensor array to minimize noise power at the array output while

maintaining a chosen frequency response in the look direction. This amounts to a constrained least-mean squares problem, solved by a simple stochastic gradient descent algorithm that requires the direction of arrival and frequency band of interest to be specified *a priori*.

A processor with K sensors and J taps per sensor has KJ weights and requires J constraints to determine its look direction. Because of these constraints, minimizing the total output power is equivalent to minimizing the non-look direction power, as long as the signal and noise at the taps are uncorrelated. The constrained least squares problem becomes

$$\begin{aligned} &\min_{\boldsymbol{w}} \ \mathbb{E}[(\boldsymbol{w}^T\boldsymbol{x})^2] \\ &\min_{\boldsymbol{w}} \ \mathbb{E}[\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{w}] \\ &= \min_{\boldsymbol{w}} \boldsymbol{w}^T R_x \boldsymbol{w} \\ &\text{s.t. } \boldsymbol{C}^T\boldsymbol{w} = \mathcal{F}. \end{aligned} \tag{1.10}$$

where R_x is the autocorrelation matrix of the signal at each tap, $\boldsymbol{w}$ is the vector of weights, $\boldsymbol{C} = [c_1, \ldots, c_J]$ such that $c_j[i] = 1 \ \forall \ i = (j-1)K, \ldots, jK, \ c_j[i] = 0$ otherwise, and $\mathcal{F} = [f_1, \ldots, f_J]^T$ is the J dimensional vector of weights of the look-direction-equivalent tapped delay line. The Lagrangian and its derivative with respect to the weights is given by :

$$\begin{aligned} J(\boldsymbol{w}, \lambda) &= \frac{1}{2}\boldsymbol{w}^T R_x \boldsymbol{w} - \lambda(\boldsymbol{C}^T\boldsymbol{w} - \mathcal{F}) \\ \nabla_{\boldsymbol{w}} J(\boldsymbol{w}, \lambda) &= R_x\boldsymbol{w} - C\lambda. \end{aligned} \tag{1.11}$$

The weights are changed iteratively using stochastic gradient descent.

$$\boldsymbol{w}(k+1) = \boldsymbol{w}(k) - \mu\nabla_{\boldsymbol{w}} J(\boldsymbol{w}, \lambda) \tag{1.12}$$

The advantage of this algorithm is that it requires no prior knowledge of the signal or noise statistics.

- **Minimum Variance Distortionless Response** - The minimum variance distortionless response (MVDR) beamformer [21] minimizes the power of the output signal subject to a single constraint assuring an undistorted response for the target source (or a filtered version of it). The distortionless response constraint requires that the desired component in the output signal is equal to the target signal. It leads to the following optimization problem.

$$\min_{\boldsymbol{w}} \ \boldsymbol{w}^H R_x \boldsymbol{w} \text{ s.t. } \boldsymbol{w}^H \boldsymbol{h_0} = 1 \tag{1.13}$$

where $\boldsymbol{h_0}$ is the impulse response of the acoustic transfer function of the target source.

- **Hybrid BSS-Beamforming** - A benefit of blind source separation is that it overcomes the

conventional cross-talk or leakage problem of adaptive beamforming. Beamforming on the other hand exploits geometric information which is often readily available but not utilized in blind algorithms. In [22], a Geometric Source Separation (GSS) algorithm is proposed, which combines the method proposed in [13] with geometric constraints on the filter coefficients, similar to the linear constraints in the beamformer proposed by Frost [20]. It is found that the geometric constraints resolve some of the ambiguities inherent in the independence criterion in BSS such as frequency permutations and degrees of freedom provided by additional sensors.

Similarly, in [23], BSS is combined with ICA. First, a new subband ICA is introduced to achieve frequency domain BSS on the microphone array system, where directivity patterns of the array are explicitly used to estimate each direction of arrival (DOA) of the sound sources. This method resolves permutation problems without the assumption for interfrequency continuity of the unmixing matrices. Next, based on the DOA estimated in the ICA section, a null beamformer is constructed in which the directional null is steered to the direction of the undesired sound source, in parallel with the ICA-based BSS. There is no difficulty with respect to a low convergence of optimization because the null beamformer is determined by only DOA information without assumption of independence between sound sources.

- **Acoustic Rake Receiver** - Acoustic Rake Receivers (ARR) [24] use echoes in rooms to improve beamforming. Acoustic Raking is a multistage process comprising image source localization, image source tracking, and beamforming weight computation. ARRs can suppress interference in cases when conventional beamforming is bound to fail, for example when an interferer is occluding the desired source. The raking microphone beamformers are particularly well- suited to extracting the desired speech signal in the presence of interfering sounds, in part because they can focus on echoes of the desired sound and cancel the echoes of the interfering signals. In [25], ARRs are designed and applied in the frequency domain. ARRs require localization of the echoes, which is done by finding the *image sources* [26]. In [25], methods are proposed to find the image sources when the room geometry is either known or unknown. The results show that the signal to interference ratio (SIR) and the perceptual quality of speech with ARR is vastly improved over conventional beamforming.

In general, beamforming suffers from the drawback of being sub-optimal in reverberant conditions because the signals may reach the sensor array from many directions, so the optimization often alters the response for the region of interest also. Acoustic raking can overcome this problem by utilizing echoes. Microphone arrays are not going to be used in the close-microphone bleed cancellation problem. However, beamforming techniques are useful to study for adaptively altering the unmixing filter weights to focus on receiving signal only from a direction of interest.

1.2.3 Adaptive Noise Cancellation

Noise cancellation algorithms aim to reduce additive noise from a measured signal. The signal and the noise could be assumed to be wide-sense stationary (e.g. Wiener filter), and are generally uncorrelated. Adaptive noise cancellation is a technique to cancel noise and interfering signals by adaptively adjusting filter coefficients - the stationarity criterion is not required. Adaptive filters such as the least mean squares and recursive least squares fall in this category. In most noise cancelling algorithms, a trade-off between noise reduction and signal distortion is observed. Some common methods are discussed below.

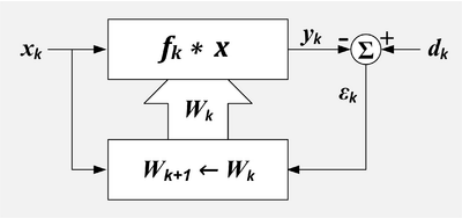


Figure 1.1: Configuration of an LMS adaptive filter.

- **Least Mean Squares (LMS)** - The famous LMS adaptive algorithm was proposed by Widrow et al. [27]. The typical configuration of an LMS adaptive filter is given in Fig. 1.1, where $\boldsymbol{x}$ is the noisy input, d is the desired input, ϵ is the error and $\boldsymbol{w}$ are the filter weights. It minimizes the mean squared error with respect to the filter weights $\boldsymbol{w}$. The cost function is

$$\begin{aligned} J(\boldsymbol{w}) &= \mathbb{E}(\epsilon^2) \\ &= \mathbb{E}[(d - \boldsymbol{w}^T\boldsymbol{x})^2] \\ &= \mathbb{E}(d^2) + \boldsymbol{w}^T R_x \boldsymbol{w} - 2\boldsymbol{w}^T\mathbb{E}(\boldsymbol{x}d) \end{aligned} \tag{1.14}$$

  The optimal weights are found by minimizing the cost function with respect to the weights.

$$\begin{aligned} \hat{\boldsymbol{w}} &= \arg\min_{\boldsymbol{w}} J(\boldsymbol{w}) \\ \nabla_{\boldsymbol{w}} J(\boldsymbol{w}) &= 2R_x\hat{\boldsymbol{w}} - 2\mathbb{E}(\boldsymbol{x}d) \end{aligned} \tag{1.15}$$

  Replacing the expectations with the sample covariances, an iterative update to the filter weights is derived using gradient descent.

$$\begin{aligned} \hat{\boldsymbol{w}}(k) &= \hat{\boldsymbol{w}}(k-1) + \mu\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) \\ \hat{\boldsymbol{w}}(k) &= \hat{\boldsymbol{w}}(k-1) + 2\mu\epsilon(k)\boldsymbol{x}(k) \end{aligned} \tag{1.16}$$

  where μ is the *convergence factor*. If μ is too large, the algorithm will not converge. If μ is too small the algorithm converges slowly and may not be able to track changing conditions.

Tracking performance of the standard LMS algorithm is improved in [28] which proposes the state-space LMS (SSLMS) which incorporates a linear time varying state-space model of the underlying environment.

- **Recursive Least Squares (RLS)** - RLS [29] is an adaptive filter algorithm that recursively finds the coefficients that minimize a weighted linear least squares cost function relating to the input signals. Contrary to LMS, RLS shows superior tracking performance and the cost of high computational complexity. The weighted cost function is formulated as

$$J(\boldsymbol{w}(k)) = \sum_{i=0}^{k} \lambda^{k-i} \epsilon(i)^2 \tag{1.17}$$

where $0 < \lambda \leq 1$ is the *forgetting factor* that gives exponentially less emphasis to older time samples. By minimizing this cost function, a recursive update is derived which is of the form

$$\hat{\boldsymbol{w}}(k) = \hat{\boldsymbol{w}}(k-1) + \boldsymbol{P}(k)\boldsymbol{x}(k)\left[d(k) - \hat{\boldsymbol{w}}(k-1)^T \boldsymbol{x}(k)\right] \tag{1.18}$$

where $\boldsymbol{P}(k) = R_x(k)^{-1}$ is the recursive update to the inverse of the autocorrelation matrix. State-space RLS (SSRLS) has been proposed in [30], that extends RLS to work with an underlying state space model.

- **Wiener Filter** - The Wiener filter minimizes the mean squared error between the estimated random process and the desired process. The noise as well as the signal statistics are assumed to be stationary. The additive noise problem can be formulated as

$$\begin{aligned} y(k) &= x(k) + n(k) \\ J(\boldsymbol{w}) &= \mathbb{E}[(\boldsymbol{w}^T \boldsymbol{y}(k) - x(k))^2] \\ \boldsymbol{w} &= [w_0, w_1, \ldots, w_P]^T \\ \boldsymbol{y}(k) &= [y(k), y(k-1), \ldots, y(k-P-1)]^T \end{aligned} \tag{1.19}$$

The optimal Wiener filter solution is:

$$W(\omega) = \frac{P_{xx}(\omega)}{P_{xx}(\omega) + P_{nn}(\omega)} \tag{1.20}$$

where P_{xx} and P_{nn} are the power spectral densities (PSDs) of the desired signal and noise respectively. The filter is not adaptive, per se. However, it can be made to work with non-stationary signals by estimating short time PSDs using the STFT, as in [4].

Speech distortion weighted multichannel Wiener filter in [31] optimizes a cost function such that trade-off is achieved between speech distortion and noise reduction.

- **Kalman Filter** - The Kalman filter [32] is an algorithm that uses a series of measurements observed over time, containing statistical noise, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each time sample. It reduces the variance of the conditional error iteratively. It is primarily used for tracking applications, which have an underlying linear dynamical system, such as, in [33, 34]. Kalman filter for speech enhancement was proposed in [35] which made use of the autoregressive model of speech production. Kalman filter tuning for speech enhancement was explored in [36]. It has also found its use in echo cancellation [37] and dereverberation [38]. It has potential to be used in microphone leakage reduction, as long as the model is linear.

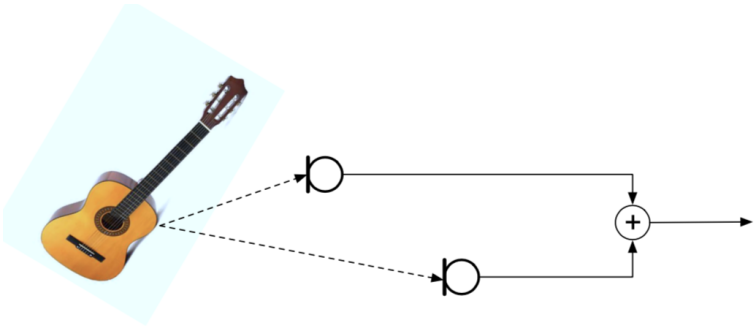
Multichannel microphone bleed cancellation can be seen as a noise cancellation problem with multiple sources of noise. The signal and noise are of course, non-stationary. In [39], a 2-channel coupled LTI system is separated using signal decorrelation. The convolutive effect of the room response makes it tricky for interfering signal statistics to be estimated in this context.

1.3 Close Microphone Bleed Cancellation

1.3.1 Background

Microphone bleed has been a nuisance in the audio engineering community for decades. When two microphones pick up the same signal with a time delay, comb filtering artifacts are present (Fig. 1.2). Furthermore, when dynamic, non-linear effects like compression are applied, the bleed from other instruments becomes even more prominent, since compression makes soft sounds louder. An expensive solution to the microphone bleed problem is to use acoustic isolation panels between musicians. For example, drums can be recorded in a separate isolation booth (acoustically treated soundproof room). Obviously, this is not feasible in a live setting. Simpler solutions include using directional microphones with specific polar patterns to pick up radiation from a desired direction, and using the close-miking technique where microphones are placed at a distance of $5 - 50$ cm from the sound source [40]. However, these techniques are unable to eliminate bleed completely. Interference can become significant in such cases due to the effect of room acoustics and nearby strong reflective surfaces. During mixing, *noise gates* can be used which allow the signal to pass through only when its amplitude is above a certain threshold. However, noise gates alter the timbre of broadband sounds, like percussion instruments, and mute subtle playing techniques, such as the scratch of a bow on the strings or the sound of a plectrum striking the strings, which adds a layer of detail and depth to the performance.

In recent years, both researchers and audio plugin companies have come up with novel solutions to this problem. Some successful products include Izotope's *De-bleed* module in RX-7 [41], *Drumatom*



(a)

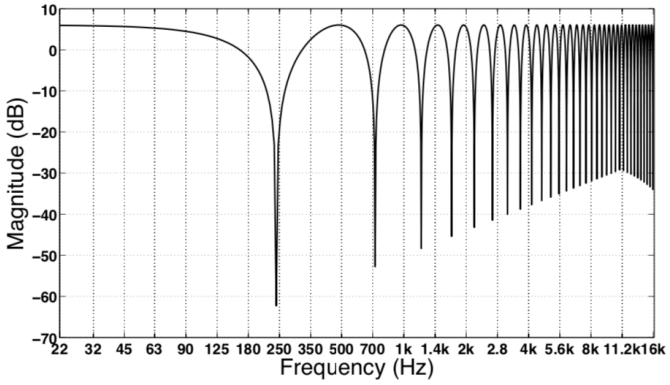


Figure 1.2: Comb filtering effects [1]

[42] for drum bleed cancellation and Wilkinson audio's *DeBleeder* plugin [43]. However, the details of the algorithms used in these products are usually not publicly available. A brief summary of the mathematical formulation of the problem, and research published by the signal processing community on this topic is given below.

1.3.2 Existing approaches

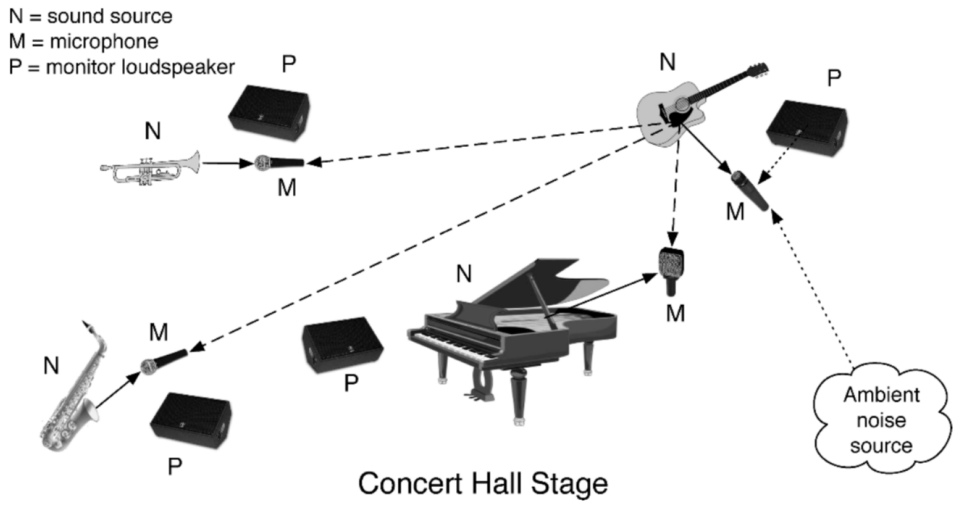


Figure 1.3: Instruments, monitors and microphones in a live concert setting [2].

Microphone and loudspeaker placement in a typical live concert setting is given in Fig. 1.3. Microphones are placed close to the instrument, typically, a few inches away. Multiple close microphones may be used to record/amplify the same instrument. We will assume that the mics are omnidirectional. The mic assigned to a particular instrument primarily picks up that instrument and leakage from all other sources plus background noise. Consider N sources, each denoted by $s_n(k)$ in a reverberant environment, and M mics picking up the signals $x_m(k)$. Following the notation used in [4], and ignoring the loudspeakers and their acoustic paths, the problem maybe formulated as:

$$x_m(k) = s_m(k) * h_{mm}(k) + \sum_{i=1, i\neq m}^{N} s_i(k) * h_{mi}(k) \tag{1.21}$$

where $h_{mn}(k)$ is the FIR filter that models the acoustic path between the nth source and the mth mic. The direct source is given as :

$$\bar{s}_{m,m}(k) = s_m(k) * h_{mm}(k) \tag{1.22}$$

and the total microphone leakage is

$$\begin{aligned} \bar{s}_{i,m}(k) &= s_i(k) * h_{mi}(k) \\ \bar{u}_m(k) &= \sum_{i=1, i\neq m}^{N} \bar{s}_{i,m}(k) \end{aligned} \tag{1.23}$$

Equation (1.21) can then be represented as a signal in additive noise problem

$$x_m(k) = \bar{s}_{m,m}(k) + \bar{u}_m(k) \tag{1.24}$$

The sources are assumed to be uncorrelated. The problem then is to estimate a correct set of filter coefficients that can recover the signal of interest while canceling leakage from each microphone signal. Let the desired filter be $w_m(k)$. Then,

$$\hat{s}_m(k) = x_m(k) * w_m(k) \tag{1.25}$$

The error signal is given as

$$e_m(k) = \bar{s}_{m,m}(k) - \hat{s}_m(k) \tag{1.26}$$

The error is used to formulate a cost function, which is minimized with respect to the coefficients of the filter w_m. Methods in the existing literature include:

- **Wiener filtering** - The state-of-the-art method described in [4] by Kokkinis et al. minimizes the mean squared error, which ultimately leads to the well-known Wiener filter, given by:

$$\begin{aligned} \hat{W}(\omega,\tau) &= \frac{P_{\bar{s}_{m,m}(\omega,\tau)}}{P_{\bar{s}_{m,m}(\omega,\tau)} + P_{\bar{u}_m(\omega,\tau)}} \\ &= \frac{P_{\bar{s}_{m,m}(\omega,\tau)}}{\sum_{i=1}^{N} P_{\bar{s}_{i,m}(\omega,\tau)}} \end{aligned} \tag{1.27}$$

  where $P(\omega,\tau)$ is the short-time power spectral density at frequency bin ω and time-frame τ. The short-time PSD for each frame needs to be estimated. A PSD estimation method is introduced based on the identification of dominant frequency bins, i.e., regions of the microphone and output PSDs that are approximately the same with that of the original source signal. A simple way to estimate the leakage PSDs is also presented, based on a set of weighting coefficients which are estimated during time intervals where only one source is active. The results show robust performance for different source-microphone distances and large reverberation times. However, the PSD estimate is adversely affected if the interfering source has high energy spread across the spectrum. A special case of the two microphone, two source problem is presented in [2].

- **Kalman based Wiener filter** - More recently, a Kalman-based Wiener filter approach has been presented in [44, 9], which uses the Kalman filter to update the interference signal's power spectra. It combines the multichannel Wiener filter proposed by Kokkinis with Multichannel Acoustic Echo Cancellation (MAEC) [45].

  The MAEC is implemented in an overlap-save (OLS) structure with a frame length of size K and a frame shift R. First, each frame of the interferer's microphone channel $\boldsymbol{x}_\mu(\ell)$ is transformed into the frequency domain and shaped into a diagonal matrix by $\underline{\boldsymbol{X}}_\mu(\ell) = \mathrm{diag}(\underline{\boldsymbol{F}}_{K\times K}\ \boldsymbol{x}_\mu(\ell))$, with $\underline{\boldsymbol{F}}_{K\times K}$ being the K-point DFT matrix. The target microphone frames are processed by the overlap-save projection matrix $\underline{\boldsymbol{Q}} = [\boldsymbol{0}_{R\times(K-R)}\ \boldsymbol{I}_{R\times R}]^\top$ as $\boldsymbol{X}_m(\ell) = \underline{\boldsymbol{F}}_{K\times K}\ \underline{\boldsymbol{Q}}\ \boldsymbol{x}_m(\ell)$, where $\boldsymbol{0}$ and $\boldsymbol{I}$ denote a zero and unity matrix, respectively.

  The prediction of the current filter coefficient state $\hat{\boldsymbol{H}}^+_{m,\mu}(\ell)$ of interferer channel μ w.r.t. the target channel m is calculated by

$$\hat{\boldsymbol{H}}^+_{m,\mu}(\ell) = A_{m,\mu}\hat{\boldsymbol{H}}_{m,\mu}(\ell-1) \tag{1.28}$$

  whereby $A_{m,\mu}$ is a first-order Markov model prediction coefficient.

  The DFT of the preliminary error vector is obtained by

$$\tilde{\boldsymbol{E}}_m(\ell) = \boldsymbol{X}_m(\ell) - \sum_{\mu\in\mathcal{I}} \underline{\boldsymbol{G}} \cdot \underline{\boldsymbol{X}}_\mu(\ell)\hat{\boldsymbol{H}}^+_{m,\mu}(\ell) \tag{1.29}$$

  with the overlap-save constraint matrix $\underline{\boldsymbol{G}} = \underline{\boldsymbol{F}}_{K\times K}\underline{\boldsymbol{Q}}\cdot\underline{\boldsymbol{Q}}^T\underline{\boldsymbol{F}}^{-1}_{K\times K}$. Subsequently, the predicted filter coefficient states are updated by

$$\hat{\boldsymbol{H}}_{m,\mu}(\ell) = \hat{\boldsymbol{H}}^+_{m,\mu}(\ell) + \underline{\boldsymbol{K}}_{m,\mu}(\ell)\tilde{\boldsymbol{E}}_m(\ell) \tag{1.30}$$

  where $\underline{\boldsymbol{K}}_{m,\mu}(\ell)$ is the Kalman gain matrix. The estimated interferer signals in channel m are then obtained by,

$$\hat{\boldsymbol{D}}_{m,\mu}(\ell) = \underline{\boldsymbol{G}} \cdot \underline{\boldsymbol{X}}_\mu(\ell)\hat{\boldsymbol{H}}_{m,\mu}(\ell). \tag{1.31}$$

  The interferer PSD is calculated from this.

- **Kernel Additive Model** - Kernel Additive Modeling for interference reduction (KAMIR) was introduced in [46]. KAMIR also minimizes the mean squared error, arriving at the same Wiener filter given in [2, 4]. It assumes the time frequency bins of the source signals to be independent and distributed normally, following $S_m(\omega,\tau) \sim \mathcal{N}(0, \lambda_{m,n}(\omega)P_m(\omega,\tau))$. the scalar $\lambda_{mn}(\omega)$ gives the amount of interference of source n into mic m at frequency ω. Therefore, its elements constitute the interference matrix $\Lambda(\omega) \in \mathbb{R}^{\mathbb{M}\times\mathbb{N}}$. All mics are assumed to share the

same latent PSD, P_i, for the ith source. The Wiener filtering step in (1.27) then becomes

$$\hat{W}_{KAM}(\omega,\tau) = \frac{\lambda_{m,m}(\omega)P_m(\omega,\tau)}{\sum_{i=1}^{N}\lambda_{m,i}(\omega)P_i(\omega,\tau)} \tag{1.32}$$

There are two steps in KAM - in the *separation step*, the Wiener filtering in (1.32) is performed. In the *parameter fitting* stage, the parameters $\Lambda(\omega), P_i \ \forall \ i = 1,\ldots,N$ are re-estimated. This procedure is repeated for a given number of iterations. The number of iterations controls the trade-off between interference reduction and distortion.

- **Cross-talk resistant adaptive noise canceler** - Typical adaptive filtering techniques such as the least mean squares (LMS) assumes the availability of a noise reference signal. In CTRANC [47], outputs of multiple adaptive filters are cascaded so that the output of one becomes the reference noise signal for the other. Centered adaptive filters try to estimate the time delay of the source reaching the microphone, and a window of filter coefficients around the delay are updated. This ensures computational efficiency and quicker convergence. Centered CTRANC is proposed in [47]. The delay estimation is done by the Generalized Cross Correlation method using the Phase Transform (GCC-PHAT) [48]. The results show poor performance in the presence of reverberation.

- **Nonnegative Signal Factorization** - Nonnegative matrix factorization (NMF) is a popular method used in source separation [11], where the mixed signal is assumed to be a weighted sum of basis functions, which are nonnegative. The task is to find the weights and the basis functions by minimizing divergence. The system proposed in [49] is composed of two main blocks: a panning matrix estimation block and the actual source separation block. Both blocks need as an input the spectrograms of the mixture microphone signals and a set of instrument basis functions calculated from an available training database. The panning matrix estimation procedure is based on the discrimination of time-frequency zones with minimum overlap between the concurrent instruments, which is performed by using the output of an automatic transcription stage. The estimated panning matrix is then fed to the NMF-based separation stage, which also uses the modeled instrument basis to estimate the magnitude spectrograms of the original sources. These spectrograms are finally used to recover the actual sources by constructing a Wiener mask that is applied over the input spectrograms.

- **IIR filters** - Taking the Z-transform of (1.21) and using matrix notation, we can write:

$$\begin{aligned}
\boldsymbol{X}(z) &= \boldsymbol{H}(z)\boldsymbol{S}(z) \\
\boldsymbol{X}(z) &= [X_1(z), X_2(z), \ldots, X_M(z)]^T \\
\boldsymbol{S}(z) &= [S_1(z), S_2(z), \ldots, S_N(z)]^T \\
\boldsymbol{H}(z) &= \begin{bmatrix} H_{11}(z) & H_{12}(z) & \ldots & H_{1N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{M1}(z) & H_{M2}(z) & \ldots & H_{MN}(z) \end{bmatrix} \\
\hat{\boldsymbol{S}}(z) &= \boldsymbol{W}(z)\boldsymbol{X}(\boldsymbol{z}) \\
\boldsymbol{W}(z) &= \boldsymbol{H}^{\dagger}(z)
\end{aligned} \tag{1.33}$$

In [50], the mixing FIR filters, $H_{mn}(z)$ are represented by a scalar gain and time delay in samples

$$H_{mn}(z) = \alpha_{mn} z^{-\tau_{mn}} \tag{1.34}$$

The inverse filters $W_{mn}(z)$ are therefore IIR feedback comb filters. Estimation of the scalar gain and the delay is done with GCC-PHAT [48]. It is observed that underestimation of gains leads to more cross talk, and delay estimation errors lead to ripple and decreasing separation in high frequency.

1.4 Goals of thesis

- To come up with a novel method for close-microphone cross-talk cancellation that takes into account the room acoustics.
- The method should be physically reasonable and mathematically optimal.
- The method should be robust and work with any number of microphones and sources in any studio with an arbitrary reverberation time.
- The method should reduce cross-talk while introducing least amount of distortion.
- The method should be computationally feasible and/or optimized for fast computation.
- The method should give results comparable to the state-of-the-art Multichannel Wiener Filter, which is the minimum mean squared error estimator.

Chapter 2

Problem Formulation

In this chapter, we mathematically derive the model for multichannel cross-talk cancellation. We rely on the captured microphone signals, as well an estimate of the acoustic path from each source to each microphone, which is typically represented with FIR filters. A few methods are suggested for estimating the relative transfer function (RTF) based on the close microphone assumption, that rely on a ***calibration stage*** during the recording setup. Based on this problem formulation, two new methods for reducing microphone-bleed will be proposed in the upcoming chapters.

2.1 Model

For N mic signals $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^\top$ and M sources $\boldsymbol{s} = [s_1, s_2, \ldots, s_M]^\top$, the n^{th} mic signal at time index k can be written as

$$x_n(k) = \sum_{i=1}^{M} s_i(k) * h_{ni}(k) + w(k) \tag{2.1}$$

where h_{ni} is the acoustic transfer function between the i^{th} source and the n^{th} mic, and $w(k)$ is additive noise, $w(k) \sim \mathcal{N}(0, \sigma_w^2)\ \forall\ k$, caused by microphone "self-noise". The acoustic transfer function contains the direct sound path, as well as frequency-dependent source radiation pattern and room acoustics. An example of this configuration is shown in Fig. 2.1. In the time-frequency domain, the convolution of the sources with the transfer function becomes multiplication of a time-invariant transfer function matrix with the source vector for each time frame, τ, and frequency bin, ω. Although in reality, the acoustic transfer function is not time-invariant due to changes in temperature and pressure, for all practical purposes, the time variation is slow enough to make such an assumption. Similarly, movements made by the musicians also affect the acoustic path, but these are usually small compared to the mean free path traveled by sound waves. Assuming the transfer

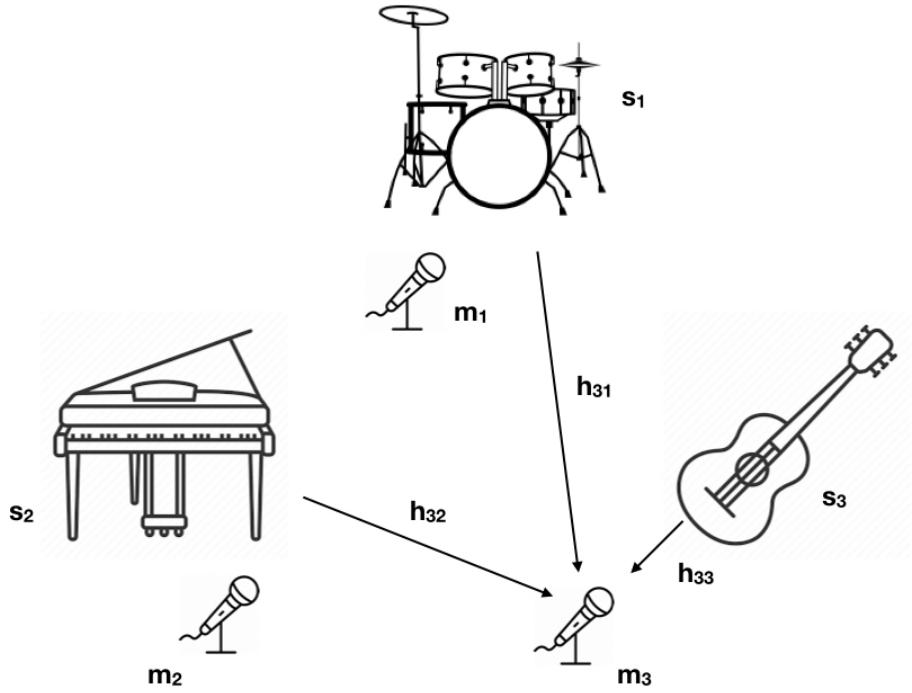


Figure 2.1: Example of a studio setup

function to be time-invariant has the advantage of considerably reducing the number of unknowns to be estimated, and gives a smoothly varying RTF.

$$\begin{aligned} \boldsymbol{x}_\tau(\omega) &= \boldsymbol{H}(\omega)\boldsymbol{s}_\tau(\omega) + \boldsymbol{w} \\ \boldsymbol{H}(\omega) &= \begin{bmatrix} h_{11}(\omega) & h_{12}(\omega) & \dots & h_{1M}(\omega) \\ h_{21}(\omega) & h_{22}(\omega) & \dots & h_{2M}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1}(\omega) & h_{N2}(\omega) & \dots & h_{NM}(\omega) \end{bmatrix} \end{aligned} \tag{2.2}$$

We want to estimate the source vector $\boldsymbol{s}_\tau(\omega)$ given the microphone signals, $\boldsymbol{x}_\tau(\omega)$. However, it is clear that without knowing $\boldsymbol{H}(\omega)$ we cannot solve this system of equations. Assuming we have some knowledge of the transfer function matrix, i.e., a noisy estimate of the transfer function matrix with each element in the noise matrix $\boldsymbol{\nu}$ independent and identically distributed with the same mean and

variance,

$$\tilde{\boldsymbol{H}}(\omega) = \boldsymbol{H}(\omega) + \boldsymbol{\nu};\ \boldsymbol{\nu} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\nu^2) \tag{2.3}$$

our goal is to jointly estimate the optimal values $\boldsymbol{H}^*(\omega)$, $\boldsymbol{s}_\tau^*(\omega)$ given $\tilde{\boldsymbol{H}}(\omega), \boldsymbol{x}_\tau^*(\omega)$.

2.1.1 Effect of Room Acoustics

Close-miking ensures a high direct-to-reverberant ratio is achieved in the captured microphone signal. This negates the effect of room acoustics to a large extent, and places prominence on the direct path, which can simply be approximated as a gain and a delay term [5]. This justifies our choice of modeling the diagonal elements of the RTF matrix as unity. When microphones are placed more than 1 m away from the source, there is a significant degree of influence of the acoustics of the space on the recorded signal.

The leakage path from the other sources, however, can be affected by room acoustics. As discussed in Chapter 2 of [1], a large presence of early reflections contributes significantly to the amount of leakage present in the microphone signal and indicates that contrary to our intuition, leakage can be significantly higher in small rooms with a large amount of reflective surfaces in proximity with the microphone, than in larger rooms with longer reverberation time. The increased amount of significant early reflections manifested in the leakage acoustic paths, increases the energy of the signals propagating through these paths, and result in interfering signals of higher energy in the microphone signals, thus increasing interference. This indicates that the off-diagonal elements of the RTF matrix can be modeled by FIR filters of the order of a few hundred samples.

2.2 Calibration

To estimate the relative transfer function matrix, $\tilde{\boldsymbol{H}}(\omega)$, we propose a ***calibration stage***. When all the microphones in the studio have been setup, a sound-check can be performed where one instrument is active at a time and picked up by all the microphones. For example, while recording drums, each drum part can be struck separately and captured simultaneously by all the mics. While recording an orchestra, each section can play at a time. This is common in actual recording setups when the individual microphone levels are adjusted after miking all the instruments. In the following sections, we discuss some methods of estimating $\boldsymbol{\tilde{H}(\omega)}$ from the recorded solo instrument sections.

2.2.1 Spectral Ratio

The diagonal elements in $\tilde{\boldsymbol{H}}(\omega)$ represent the transfer function from the desired source to the closest microphone, whereas the off-diagonal elements represent the transfer function of the interfering sources. The sound captured by the microphone closest to the m^{th} source can be approximated as the source itself. Its spectral ratio with the sound captured by the n^{th} microphone gives an

approximate value of $\tilde{h}_{nm}(\omega)$. We take the average of spectral ratios of frames that have significant energy (energy greater than or equal to -3 dB from the frame with maximum energy) to get the off-diagonal elements, $\tilde{h}_{nm}(\omega)$. The diagonal elements, $\tilde{h}_{nn}(\omega)$, are approximated to be 1 to represent a unit magnitude direct path with no delay.

$$\begin{aligned} \text{sig}_m &= \left[\tau \in 1, 2, \ldots, T : E_{\tau,m} \geq \frac{\max(E_{1,2,\ldots,T,m})}{\sqrt{2}}\right] \text{ where} \\ E_{\tau,m} &= \frac{1}{L}\sum_{l=0}^{L-1} |x_m(\tau L + l)|^2 \end{aligned} \tag{2.4}$$

$$\tilde{h}_{nm}(\omega) = \begin{cases} 1 \text{ if } n = m \\ \frac{1}{N_{\text{sig}}} \sum_{\tau \,\in\, \text{sig}_m} \frac{x_{n_\tau}(\omega)}{x_{m_\tau}(\omega)} \text{ if } n \neq m \end{cases} \tag{2.5}$$

2.2.2 M-GCC with Least Squares

The estimate of the initial transfer function obtained from the spectral ratio is not smooth. Instead, we can approximate the transfer function from the m^{th} source to the n^{th} microphone as a direct path, plus a few early reflections. This can be represented as an FIR filter with gains α_i and delays τ_i for $i = 1, 2, \ldots, p$, where p is the FIR filter order, $p-1$ is the number of early reflections we want to include in the model. We know that the strong early reflections are most prominent in determining microphone bleed [5], hence this model is a good approximation for estimating the relative transfer function from each source to each microphone.

$$\begin{aligned} \tilde{h}_{nm}(k) &= \begin{cases} \delta(k) \text{ if } n = m \\ \sum_{i=1}^{p} \alpha_i \delta(k - \tau_i) \text{ if } m \neq n \end{cases} \\ \tilde{h}_{nm}(\omega) &= \begin{cases} 1 \text{ if } n = m \\ \sum_{i=1}^{p} \alpha_i \exp\left(-j\omega\tau_i\right) \text{ if } m \neq n \end{cases} \end{aligned} \tag{2.6}$$

Here $\delta(k)$ is the Dirac-delta function. The task now is to estimate the filter gains and delays from the data.

Time delay estimation with M-GCC

The relative time delay of a source arriving at two microphones can be estimated using GCC-PHAT (generalized cross-correlation with phase transform) [48]. While the time delay of arrival (TDOA) estimation of multiple sources with GCC-PHAT is a formerly investigated topic [51] and TDOA with multiple sources and multiple microphones with GCC-PHAT has been explored in [52], we expand

the analysis to show that sparse early reflections in a room impulse response can also be detected with a modified version of the GCC function. This is because the modified GCC function of such an impulse response can be approximated as a sum of Dirac-delta functions. The location of these Dirac-delta functions determines the time of arrival of the early reflections.

Let there be two microphones, $x_1(t), x_2(t)$ capturing a source, $s(t)$. $x_1(t)$ is a close mic. The time and frequency domain equations of the microphones are

$$\begin{aligned} x_1(t) &= s(t) + n_1(t) \\ x_2(t) &= \sum_{i=1}^{p} \alpha_i s(t - \tau_i) + n_2(t) \\ X_1(\omega) &= S(\omega) + N_1(\omega) \\ X_2(\omega) &= S(\omega) \sum_{i=1}^{p} \alpha_i \exp\left(-j\omega\tau_i\right) + N_2(\omega) \end{aligned} \tag{2.7}$$

where $n_1(t), n_2(t)$ are zero-mean additive white noise that are uncorrelated with each other and $s(t)$. Now, the cross correlation function of the two microphone signals (or the cross-power spectrum in the frequency domain), can be written as

$$\begin{aligned} R_{x_1,x_2}(l) &= \mathbb{E}[x_1(t)x_2(t-l)] \\ \Phi_{x_1,x_2}(\omega) &= X_1(\omega)^* X_2(\omega) \end{aligned}$$

The Generalized Cross-Correlation function with Phase Transform (GCC-PHAT) at lag l is defined as

$$\tilde{R}_{x_1,x_2}(l) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\Phi_{x_1,x_2}(\omega)}{|\Phi_{x_1,x_2}(\omega)|} e^{j\omega l} d\omega \tag{2.8}$$

If we assume that the noise variance is small enough to be ignored, and modify the GCC-PHAT function such that M-GCC (modified generalized cross-correlation) is defined as

$$\begin{aligned} \widetilde{R_{m}}_{x_1,x_2}(l) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\Phi_{x_1,x_2}(\omega)}{\Phi_{x_1,x_1}(\omega)} e^{j\omega l} d\omega \\ &\approx \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{i=1}^{p} \alpha_i \exp(-j\omega\tau_i) e^{j\omega l} d\omega \\ &\approx \frac{1}{2\pi} \sum_{i=1}^{p} \int_{-\pi}^{\pi} \alpha_i \exp(j\omega(l - \tau_i)) d\omega \\ &\approx \sum_{i=1}^{p} \alpha_i \frac{\sin(\pi(l - \tau_i))}{\pi(l - \tau_i)} \\ \arg\max_l \widetilde{R_{m}}_{x_1,x_2}(l) &= \tau_i \ \forall \ i = 1, 2, \cdots, p \end{aligned} \tag{2.9}$$

This function has peaks at the time of arrivals of the reflections.

Least squares estimation of gains

Once we determine the time delays, we can write the M-GCC function without the inverse Fourier transform as

$$\bar{G}(\omega) = \frac{\Phi_{x_1,x_2}(\omega)}{\Phi_{x_1,x_1}(\omega)} \approx \sum_{i=1}^{p} \alpha_i \exp\left(-j\omega\tau_i\right) \tag{2.10}$$

For a vector of frequencies, $\boldsymbol{\omega} = \omega_1, \ldots, \omega_N, \in [-\pi, \pi]$, eq. (2.10) can be written as the following system of equations

$$\begin{bmatrix} \bar{G}(\omega_0) \\ \vdots \\ \bar{G}(\omega_N) \end{bmatrix} = \begin{bmatrix} e^{-j\omega_0\tau_1} & \ldots & e^{-j\omega_0\tau_p} \\ \vdots & \ddots & \vdots \\ e^{-j\omega_N\tau_1} & \ldots & e^{-j\omega_N\tau_p} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}$$

The gain vector can then be estimated using linear least squares.

$$\begin{aligned} \bar{\boldsymbol{G}} &= \boldsymbol{E}\boldsymbol{\alpha} \\ \hat{\boldsymbol{\alpha}} &= (\boldsymbol{E}^{\mathsf{H}}\boldsymbol{E})^{-1}\boldsymbol{E}^{\mathsf{H}}\bar{\boldsymbol{G}} \end{aligned} \tag{2.11}$$

2.2.3 Blind Channel Identification

The aim of Blind Channel Identification (BCI) is to identify unknown system responses excited by unknown source signals. Time invariant FIR filters are used to model the unknown system responses. The order of the system (or length of the channel response) is assumed to be known. The case of Single-Input-Multi-Output (SIMO) systems are the most well-developed, although some methods have been suggested for Multi-Input-Multi-Output (MIMO) systems, which typically employ higher order statistics [53, 54]. If calibration has been done for microphone bleed cancellation, then the response from each source to all the microphones is a SIMO system.

One of the first BCI methods was the Cross-Relation method proposed in [55], which also laid down the conditions for blind channel identifiability. For channels to be uniquely identifiable, the following have to be satisfied.

- Firstly, the channels must be co-prime, i.e, the multichannel transfer functions cannot share any common zeros.
- Secondly, the Hankel matrix of the source signal cannot be rank-deficient.

Since then, many adaptive time-domain and frequency-domain BCI algorithms have been proposed - such as Multichannel LMS (MCLMS) and Multichannel Newton (MCN) [56], Normalized Multichannel Frequency-Domain LMS (NMCFLMS) [57], and more recently, robust NMCFLMS with ℓ_p-norm constraints [58] and NMCFLMS with phase constraint [59]. In this thesis, we use

the NMCFLMS algorithm that has been implemented in a `MATLAB` toolbox [60]. In the following paragraphs, we will describe the details of the algorithm.

Normalized Multichannel Frequency Domain LMS

For an M-channel SIMO system, the m^{th} impulse response with L coefficients can be denoted as

$$\boldsymbol{h}_m = \begin{bmatrix} h_{m,0} & h_{m,1} & \dots & h_{m,L-1} \end{bmatrix}^\top$$

for $m = 1, 2, \dots, M$, the m^{th} microphone signal can be expressed as

$$\begin{aligned} x_m(n) &= \sum_{j=0}^{L-1} h_{m,j} s(n-j) + b_m(n) \\ \boldsymbol{x}_m(n) &= \boldsymbol{H}_m \boldsymbol{s}(n) + \boldsymbol{b}_m(n) \end{aligned} \tag{2.12}$$

where $s(n)$ is the source signal and $\boldsymbol{s}(n) = [s(n), s(n-1), \dots, s(n-2L+2)]^\top$, $x(n)$ is the microphone signal and $\boldsymbol{x}_m(n) = [x_m(n), x_m(n-1), \dots, x_m(n-L+1)]^\top$, $b_m(n)$ is the additive noise of the same dimensions as $x(n)$, and $\boldsymbol{H}_m$ is the $L \times (2L-1)$ convolution matrix for the m^{th} channel,

$$\boldsymbol{H}_m = \begin{bmatrix} h_{m,0} & \dots & h_{m,L-1} & \dots & \dots & 0 \\ 0 & h_{m,0} & \dots & h_{m,L-1} & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & h_{m,0} & \dots & h_{m,L-1} \end{bmatrix}$$

The cross-relation method uses the fact that convolution is commutative. Therefore,

$$\boldsymbol{x}_m^\top \boldsymbol{h}_l = \boldsymbol{x}_l^\top \boldsymbol{h}_m \tag{2.13}$$

Using the Least Mean Squares algorithm (LMS) [27] , we can find iterative updates for the channel impulse response. For the n^{th} iteration, the *apriori* error in the time-domain is given by

$$e_{ml}(n) = \boldsymbol{x}_m^\top \hat{\boldsymbol{h}}_l(n-1) - \boldsymbol{x}_l^\top \hat{\boldsymbol{h}}_m(n-1)$$

The cost function, $J(n)$, is the sum of the square of the error. The update equation for the n^{th} iteration is

$$\hat{\boldsymbol{h}}(n) = \hat{\boldsymbol{h}}(n-1) - \mu \nabla J(n) \tag{2.14}$$

where μ is the step-size.

The Multichannel Frequency Domain (MCFLMS) algorithm is developed based on the derivation of cross-relation in the frequency domain. Compared to the time-domain MCLMS algorithm, MCFLMS is more computationally efficient since it uses the FFT to calculate the block convolution. We can write,

$$\tilde{\boldsymbol{y}}_{ml} = \boldsymbol{C}_{x_m}(k)\hat{\boldsymbol{h}}_l^{10}(k) \tag{2.15}$$

where k is the frame index, $\hat{\boldsymbol{h}}_l^{10}(k) = [\hat{\boldsymbol{h}}_l(k)^\top 0^\top]^\top$ is the lth channel estimate with zero padding, and $\boldsymbol{C}_{x_m}$ is a $2L \times 2L$ circulant matrix,

$$\boldsymbol{C}_{x_m}(k) = \begin{bmatrix} x_m(kL-L) & x_m(kL+L-1) & \dots & x_m(kL-L+1) \\ x_m(kL-L+1) & x_m(kL-L) & \dots & x_m(kL-L+2) \\ \vdots & \vdots & \ddots & \vdots \\ x_m(kL+L-1) & x_m(kL+L-2) & \dots & x_m(kL-L) \end{bmatrix}$$

For each $\tilde{\boldsymbol{y}}_{ml}(k)$ of length $2L$, the last L samples are retained since they correspond to the linear convolution given by $\boldsymbol{x}_m^\top(n)\hat{\boldsymbol{h}}_l(n)$. As a result, by defining two selecting matrices

$$\boldsymbol{W}_{L\times 2L}^{01} = [\boldsymbol{0}_{L\times L}\boldsymbol{I}_{L\times L}], \quad \boldsymbol{W}_{2L\times L}^{10} = [\boldsymbol{I}_{L\times L}\boldsymbol{0}_{L\times L}]^\top$$

the desired result $\boldsymbol{y}_{ml}(k)$ can be obtained.

$$\begin{aligned} \boldsymbol{y}_{ml}(k) &= \boldsymbol{W}_{L\times 2L}^{01}\tilde{\boldsymbol{y}}_{ml}(k) = \boldsymbol{W}_{L\times 2L}^{01}\boldsymbol{C}_{x_m}(k)\hat{\boldsymbol{h}}_l^{10}(k), \\ &= \boldsymbol{W}_{L\times 2L}^{01}\boldsymbol{C}_{x_m}(k)\boldsymbol{W}_{2L\times L}^{10}\hat{\boldsymbol{h}}_l(k) \end{aligned} \tag{2.16}$$

Since the matrix, $\boldsymbol{C}_{x_m}$ is circulant, it can be decomposed as

$$\boldsymbol{C}_{x_m}(k) = \boldsymbol{F}_{2L}^{-1}\underline{\mathcal{D}}_m(k)\boldsymbol{F}_{2L}$$

where $\boldsymbol{F}_{2L}$ is a $2L \times 2L$ DFT matrix and $\underline{\mathcal{D}}_m(k)$ is a diagonal matrix with the diagonal elements as the DFT coefficients of the first row of the circulant matrix. Now, the frequency-domain CR error can be written as

$$\begin{aligned} \underline{\boldsymbol{e}}_{ml}(k) &= \boldsymbol{F}_L\left[\boldsymbol{y}_{ml}(k) - \boldsymbol{y}_{lm}(k)\right] \\ &= \boldsymbol{F}_L\boldsymbol{W}_{L\times 2L}^{01}\left[\boldsymbol{C}_{x_m}(k)\boldsymbol{W}_{2L\times L}^{10}\hat{\boldsymbol{h}}_l(k) - \boldsymbol{C}_{x_l}(k)\boldsymbol{W}_{2L\times L}^{10}\hat{\boldsymbol{h}}_m(k)\right] \\ &= \mathcal{W}_{L\times 2L}^{01}\left[\underline{\mathcal{D}}_m(k)\mathcal{W}_{2L\times L}^{10}\hat{\boldsymbol{h}}_l(k) - \underline{\mathcal{D}}_l(k)\mathcal{W}_{2L\times L}^{10}\hat{\boldsymbol{h}}_m(k)\right] \end{aligned} \tag{2.17}$$

for $m, l = 1, 2, \ldots, M, m \neq l$, where

$$\begin{gathered} \mathcal{W}_{L\times 2L}^{01} = \boldsymbol{F}_L \boldsymbol{W}_{L\times 2L}^{01} \boldsymbol{F}_{2L}^{-1}, \quad \mathcal{W}_{2L\times L}^{10} = \boldsymbol{F}_{2L} \boldsymbol{W}_{2L\times L}^{10} \boldsymbol{F}_L^{-1} \\ \hat{\boldsymbol{h}}_m(k) = \boldsymbol{F}_L \hat{\boldsymbol{h}}_m(k) \end{gathered} \tag{2.18}$$

The MCFLMS algorithm is given by

$$\begin{aligned} \underline{e}_{ml}^{01}(k) =& \mathcal{W}_{2L\times L}^{01} \underline{e}_{ml}(k) = \boldsymbol{F}_{2L} \begin{bmatrix} \boldsymbol{0}_{L\times 1} \\ \boldsymbol{F}_L^{-1} \underline{e}_{ml}(k) \end{bmatrix} \\ \underline{\hat{\boldsymbol{h}}}_m^{10}(k\hat{-}1) =& \boldsymbol{F}_{2L} \hat{\boldsymbol{h}}_m^{10}(k-1) = \boldsymbol{F}_{2L} \boldsymbol{W}_{2L\times L}^{10} \hat{\boldsymbol{h}}_m(k-1), \\ \hat{\boldsymbol{h}}_m^{10}(k) =& \hat{\boldsymbol{h}}_m^{10}(k-1) - \mu \sum_{l=1}^{M} \mathcal{D}_l^*(k) \underline{e}_{ml}^{01}(k) \end{aligned} \tag{2.19}$$

where μ is the step-size and

$$\mathcal{W}_{2L\times L}^{01} = \boldsymbol{F}_{2L} \boldsymbol{W}_{2L\times L}^{10} \boldsymbol{F}_{2L}^{-1}$$

Although the MCLMS algorithm converges to the optimal solution, its convergence is slow because of nonuniform convergence rates of the filter coefficients and cross-coupling between them. In order to achieve independent and uniform convergence for each filter coefficient and, therefore, accelerate the overall convergence, the coefficient updates need to be properly normalized at each iteration, and hence, the NMCFLMS algorithm was developed. The update step in eq. (2.19) is replaced with

$$\begin{aligned} \underline{\mathcal{P}}_m(k) &= \lambda \underline{\mathcal{P}}_m(k-1) + (1-\lambda) \sum_{l=1, l\neq m}^{M} \underline{\mathcal{D}}_l^*(k) \underline{\mathcal{D}}_l(k) \\ \underline{\hat{\boldsymbol{h}}}_m^{10}(k) &= \hat{\boldsymbol{h}}_m^{10}(k-1) - \mu \left[\underline{\mathcal{P}}_m(k) + \delta \boldsymbol{I}_{2L\times 2L}\right]^{-1} \times \sum_{l=1}^{M} \underline{\mathcal{D}}_l^*(k) \underline{e}_{ml}^{01}(k), \end{aligned} \tag{2.20}$$

where $\lambda = \left(1 - \frac{1}{3L}\right)^L$ is the forgetting factor, μ is the step size and δ is the regularization parameter. To satisfy the unit-norm constraint [29], the frequency-domain coefficients of the adaptive filter are initialized as $\underline{\hat{\boldsymbol{h}}}_m^{10}(0) = \boldsymbol{1}_{2L\times 1}/\sqrt{M}$.

2.2.4 Simulation and Results

RIR with Image-Source Method

To test the methods described above, we virtually placed a linear array of $N = 5$ omnidirectional microphones in a $5\times 6\times 3$ m^3 room, with the array center at $(2.5, 2, 1.6)$ m, and a microphone spacing

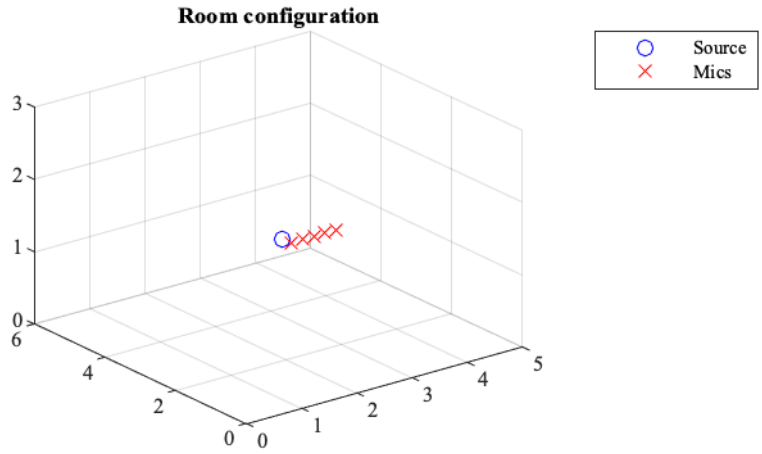


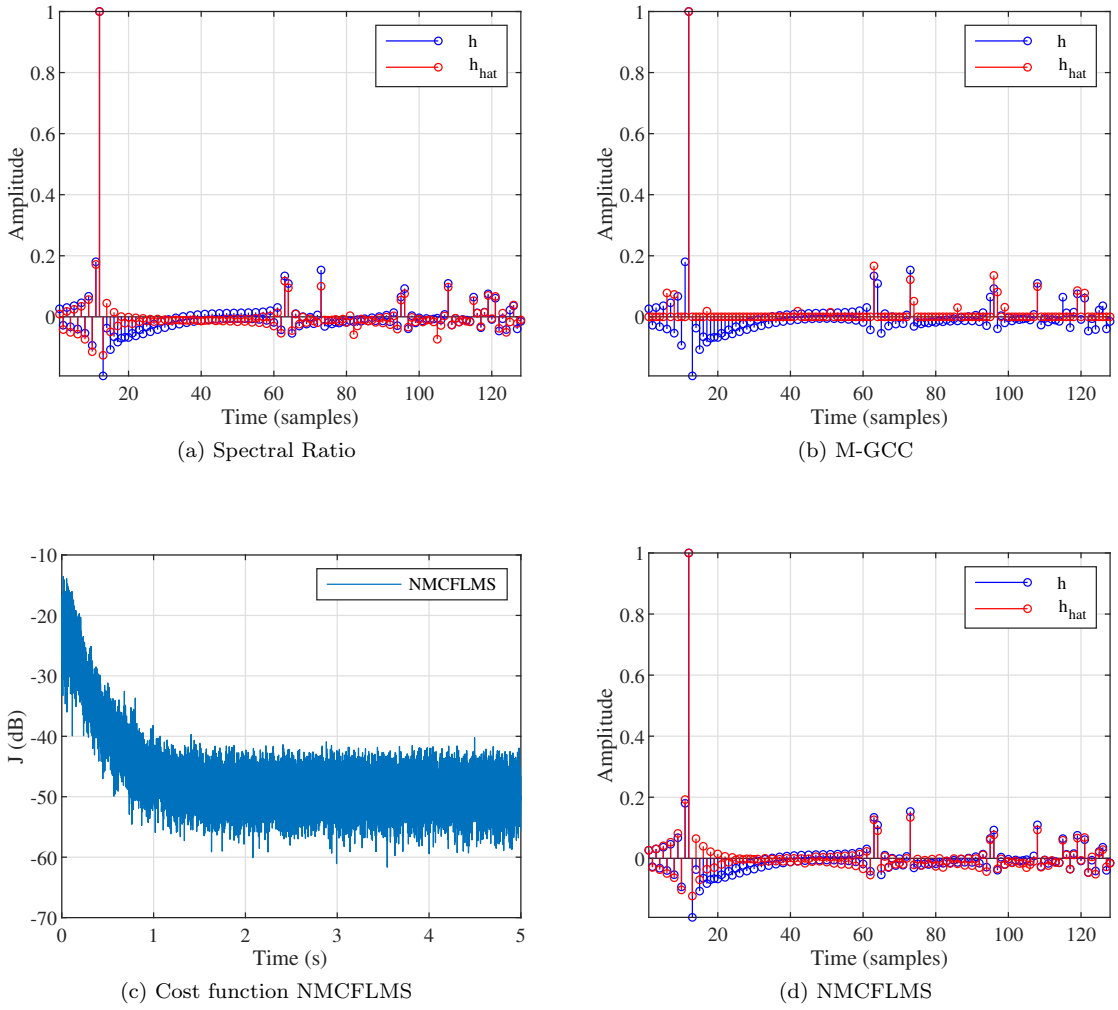
Figure 2.2: Mic and source position in virtual room

of 20 cm. The source is placed at $(2.1, 2.2, 1.6)$ m, with the closest microphone 20 cm from the source and the farthest microphone 85 cm from the source. The room has a reverberation time, T_{60}, of 300 ms. A Gaussian white noise sequence of 5 s is generated at a sampling rate of 8 kHz, and the image-source method [61] is used to generate impulse responses of channel length $L = 128$ samples from the source to each microphone. The configuration is shown in Fig. 2.2. In the following figures, we show the results of estimating the impulse response from the source to the farthest microphone, by making use of the signal captured by the closest microphone.

Spectral ratio The results of channel estimation by taking the spectral ratio of the mic closest to the source and the desired mic is shown in Fig. 2.3a. For the STFT, a frame size of 1024 samples, and a Hanning window with 50% overlap is used (for constant overlap-add). The FFT size is 4096 samples. The actual and estimated impulse responses match closely.

M-GCC The results of estimating the desired channel with M-GCC with an FIR filter order of $p = 15$ is shown in Fig. 2.3b. The low order of the filter misses many reflections but it captures the reflections with the largest amplitudes. More accurate representations can be captured with higher order filters, but this is adequate for our application, where we focus on the strong early reflections.

NMCFLMS For the NMCFLMS algorithm, we set $\mu = 0.8$, $\lambda = 0.98$, and $F = 2L$. The plots showing the measured and estimated impulse responses of the first channel, as well as the magnitude of the cost function vs data length, is shown in Fig. 2.3d. As expected, the cost function converges after processing 1 s of data. The estimated impulse response matches the actual impulse response



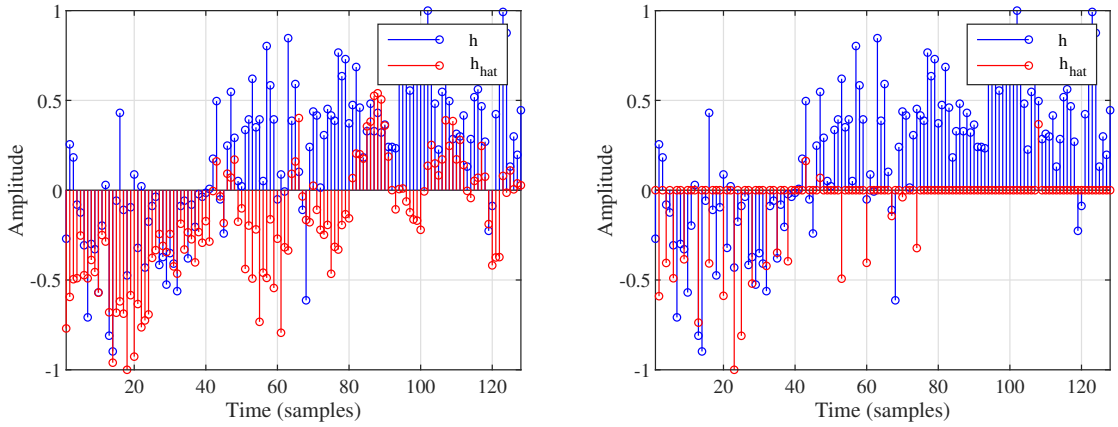
(a) Spectral Ratio

(b) M-GCC

(c) Cost function NMCFLMS

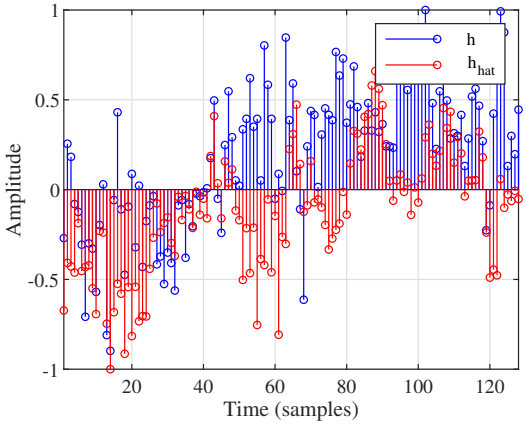
(d) NMCFLMS

Figure 2.3: Measured (blue) and estimated (red) channels.



(a) Spectral Ratio

(b) M-GCC



(c) NMCFLMS

Figure 2.4: Measured (blue) and estimated (red) channels.

very closely.

Measured RIR

Stereo Mackie CR3 (3 inch) speakers were setup near a wall in a 34 m^2 studio apartment, at a distance of approximately 1 m from each other. An AT-2020 cardioid condensor microphone was placed close to the left speaker, at a distance of approximately 12 cm. A 10 s sine sweep was played through each speaker at a time, and recorded by the microphone. From this, a set of two impulse responses were obtained, from the left speaker to the mic, and from the right speaker to the mic. For the impulse response from the left speaker to the mic (close-microphone case), the noise floor

before the direct path was truncated, and its energy was normalized to unity.

The same 5 s white noise signal from Section 2.2.4 with 40 dB SNR was convolved with the first 128 samples of the two impulse responses. Spectral ratio, M-GCC and NMCFLMS with the same parameters were used to identify the impulse response from the right speaker to the microphone. The results are shown in Fig. 2.4.

Discussion

Although the results of the discussed methods for relative transfer function estimation are promising when the image-source method is used for generating the room impulse response from the source to the receiver, this case is ideal and not representative of the RIR in an actual room. With a measured RIR, the methods fail at capturing the early reflections adequately. The spectral ratio and NMCFLMS methods perform comparably, yielding similar RIR estimates. M-GCC performs the worst. A real-life recording scenario will be more complex, and these methods are likely to under-perform in estimating the transfer function matrix accurately.

2.3 Summary

In this chapter, we have mathematically derived the model for the scenario when N microphones are used to record M sources in a studio. In the time-frequency domain, this model is linear and the microphone signals are related to the source signals via a time-invariant relative transfer function matrix. We have proposed a ***calibration stage*** that gives an initial noisy estimate of the relative transfer functions. We have discussed and compared three methods for finding this initial estimate - spectral ratio, M-GCC and SIMO BCI with Normalized Multichannel Frequency Domain LMS.

Chapter 3

Non-Bayesian Estimation : Maximum Likelihood Estimator

In this chapter, we propose a novel method for microphone cross-talk cancellation based on a maximum likelihood approach. This approach requires simultaneously optimizing the joint estimate of the sources, as well as the relative transfer functions between each source and microphone. We derive the cost function, prove that it is convex and find an optimal solution by equating its gradient to zero. We also propose methods to speed up the computation by using vectorization and parallelization on multi-core processors. The details of this algorithm have been published in [62].

3.1 Cost function derivation

Assuming the measurement of $\tilde{\boldsymbol{h}}(\omega) = \left[\tilde{h}_{11}, \tilde{h}_{12}, \ldots, \tilde{h}_{NM}\right]^\top$ and $\boldsymbol{x}_\tau(\omega)$ to be independent, we can maximize the joint likelihood of the measured microphone signals and estimated acoustic transfer functions conditioned on the source signals and the actual transfer function (over all T frames and individually for each frequency bin). In the following derivations, the frequency index ω has been omitted for clarity, and the subscript has been used to denote a time frame.

$$\begin{aligned} J(\boldsymbol{s_1}, \ldots, \boldsymbol{s_T}, \boldsymbol{h}) &= \max p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, \tilde{\boldsymbol{h}} | \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{h}) \\ &= \max p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T | \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{h}) p(\tilde{\boldsymbol{h}} | \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{h}) \end{aligned} \tag{3.1}$$

We can assume $\boldsymbol{x}_1, \dots \boldsymbol{x}_T$ to be independent and identically distributed and the microphone signal at current frame, $\boldsymbol{x}_\tau$, to depend only on the source signal at current frame, $\boldsymbol{s}_\tau$. Similarly, the measurement of $\tilde{\boldsymbol{h}}$ is independent of $\boldsymbol{s}_\tau \forall\ \tau$. Then, we can maximize the joint likelihood, or equivalently, minimize negative log of the joint likelihood.

$$\begin{aligned} J(\boldsymbol{s_1}, \dots, \boldsymbol{s_T}, \boldsymbol{h}) &= \max \left(\prod_{t=1}^{T} p(\boldsymbol{x}_t | \boldsymbol{s}_t, \boldsymbol{h}) \right) p(\tilde{\boldsymbol{h}} | \boldsymbol{h}), \text{OR} \\ &= \min \left(-\sum_{t=1}^{T} \ln p(\boldsymbol{x}_t | \boldsymbol{s}_t, \boldsymbol{h}) - \ln p(\tilde{\boldsymbol{h}} | \boldsymbol{h}) \right) \end{aligned} \tag{3.2}$$

Since we have assumed normally distributed measurement noise, the above distributions are normal with the following statistics : $\boldsymbol{x}_t | \boldsymbol{s}_t, \boldsymbol{h} \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{s}, \sigma_w^2 \boldsymbol{I})$ and $\tilde{\boldsymbol{h}} | \boldsymbol{h} \sim \mathcal{N}(\boldsymbol{h}, \sigma_\nu^2 \boldsymbol{I})$. The quadratic cost function is

$$\begin{aligned} J(\boldsymbol{s_1}, \dots, \boldsymbol{s_T}, \boldsymbol{h}) &= \max \left[\frac{1}{\sqrt{(2\pi)^{T+1} \sigma_w^2 \sigma_\nu^2}} \exp \left(\frac{-\sum_{t=1}^{T} (\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t)^{\mathsf{H}} (\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t)}{2\sigma_w^2} - \frac{(\tilde{\boldsymbol{h}} - \boldsymbol{h})^{\mathsf{H}} (\tilde{\boldsymbol{h}} - \boldsymbol{h})}{2\sigma_\nu^2} \right) \right] \\ &= \min \frac{1}{\sigma_w^2} \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t||^2 + \frac{1}{\sigma_\nu^2} ||\tilde{\boldsymbol{h}} - \boldsymbol{h}||^2 \\ &= \min \frac{1}{\sigma_w^2} \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t||^2 + \frac{1}{\sigma_\nu^2} \operatorname{Tr} \left((\tilde{\boldsymbol{H}} - \boldsymbol{H})^{\mathsf{H}} (\tilde{\boldsymbol{H}} - \boldsymbol{H}) \right) \end{aligned} \tag{3.3}$$

where Tr(.) is trace of a matrix, and $(.)^{\mathsf{H}}$ is the Hermitian transpose. This cost function looks similar to a least squares formulation with ℓ_2-norm regularization. However, in this case, both $\boldsymbol{H}$ and $\boldsymbol{s}_\tau$ are unknown.

3.2 Proof of convexity

To prove that the cost function, J, is convex in $\boldsymbol{s}$ and $\boldsymbol{h}$, we can show that the Hessian,

$$\nabla^2 J = \begin{bmatrix} \nabla_{\boldsymbol{s}}^2 J & \nabla_{\boldsymbol{s}} (\nabla_{\boldsymbol{h}} J) \\ \nabla_{\boldsymbol{h}} (\nabla_{\boldsymbol{s}} J) & \nabla_{\boldsymbol{h}}^2 J \end{bmatrix}$$

, is positive semidefinite. We will calculate each of these block matrices individually.

For the following derivation, we will ignore the summation over frames. This is justified since convexity is additive, and the sum of convex functions is also convex [63]. It is obvious that the second derivative of the cost function J with respect to $\boldsymbol{s}$ is a quadratic term, and therefore positive-semidefinite.

$$\begin{aligned} J(\boldsymbol{s},\boldsymbol{h}) &= ||\boldsymbol{x}-\boldsymbol{H}\boldsymbol{s}||^2 + \frac{1}{\sigma^2}||\tilde{\boldsymbol{h}}-\boldsymbol{h}||^2 \\ \frac{\partial J}{\partial s_j} &= \sum_{i=1}^{N} -h_{ij}\left(x_i - \sum_{k=1}^{M} h_{ik}s_k\right) \\ \nabla^2_{\boldsymbol{s}} J &= \boldsymbol{H}^{\mathsf{H}}\boldsymbol{H} \end{aligned} \tag{3.4}$$

The second derivative of the cost function J with respect to $\boldsymbol{h}$ is a block diagonal matrix that can be expressed as a quadratic polynomial in $\boldsymbol{s}$, and therefore, is also positive-semidefinite.

$$\begin{aligned} \frac{\partial J}{\partial h_{ij}} &= -s_j(x_i - \sum_{k=1}^{M} h_{ik}s_k) - \frac{1}{\sigma^2}(\tilde{h}_{ij} - h_{ij}) \\ \frac{\partial^2 J}{\partial h_{ij}^2} &= s_j^2 + \frac{1}{\sigma^2} \\ \frac{\partial J}{\partial h_{im}\partial h_{ij}} &= s_j s_m \\ \frac{\partial J}{\partial h_{nm}\partial h_{ij}} &= 0 \;\forall\; n,m \neq i,j \\ \nabla^2_{\boldsymbol{h}} J &= \begin{bmatrix} \boldsymbol{A_1} & & & \boldsymbol{0} \\ & \boldsymbol{A_2} & & \\ & & \ddots & \\ \boldsymbol{0} & & & \boldsymbol{A_N} \end{bmatrix} \\ \boldsymbol{A_n} &= \begin{bmatrix} s_1^2 + \frac{1}{\sigma^2} & s_1 s_2 & \dots & s_1 s_M \\ s_1 s_2 & s_2^2 + \frac{1}{\sigma^2} & \dots & s_2 s_M \\ \vdots & \vdots & \ddots & \vdots \\ s_1 s_M & s_2 s_M & \dots & s_M^2 + \frac{1}{\sigma^2} \end{bmatrix} \\ &= \boldsymbol{s}\boldsymbol{s}^{\mathsf{H}} + \frac{1}{\sigma^2}\boldsymbol{I}_{M\times M} \\ \nabla^2_{\boldsymbol{h}} J &= (\boldsymbol{s}\boldsymbol{s}^{\mathsf{H}} + \frac{1}{\sigma^2}\boldsymbol{I}_{M\times M}) \otimes \boldsymbol{I}_{N\times N} \end{aligned} \tag{3.5}$$

where $\boldsymbol{I}_{M\times M} \in \mathbb{C}^{M\times M}, \boldsymbol{I}_{N\times N} \in \mathbb{C}^{N\times N}$ *in* denotes the identity matrices, and $\otimes$ is the kronecker product.

The off-diagonal matrices (partial derivatives) are hermitian symmetric and opposite in sign, i.e,

$\nabla_{\boldsymbol{s}}(\nabla_{\boldsymbol{h}} J) = -\nabla_{\boldsymbol{h}}(\nabla_{\boldsymbol{s}} J)^{\mathsf{H}}$.

$$\begin{aligned} \frac{\partial^2 J}{\partial h_{ij} \partial s_j} &= x_i - \left(\sum_{k=1}^{M} h_{ik} s_k + h_{ij} s_j \right) \\ \frac{\partial^2 J}{\partial h_{in} \partial s_j} &= -h_{ij} s_n \ \forall \ n \neq j \end{aligned} \tag{3.6}$$

$$\begin{aligned} \frac{\partial^2 J}{\partial s_j \partial h_{ij}} &= -x_i + \left(\sum_{k=1}^{M} h_{ik} s_k + h_{ij} s_j \right) \\ \frac{\partial^2 J}{\partial s_m \partial h_{ij}} &= h_{im} s_j \ \forall \ m \neq j \end{aligned} \tag{3.7}$$

To prove that the Hessian is positive semidefinite, we show that for any vector $\boldsymbol{v} = [\boldsymbol{v_1} \ \boldsymbol{v_2}] \in \mathbb{C}^{(N+1)M}$, $\boldsymbol{v}^{\mathsf{H}}(\nabla^2 J)\boldsymbol{v} \geq 0$.

$$\begin{aligned} \begin{bmatrix} \boldsymbol{v_1^{\mathsf{H}}} & \boldsymbol{v_2^{\mathsf{H}}} \end{bmatrix} \begin{bmatrix} \nabla_{\boldsymbol{s}}^2 J & \nabla_{\boldsymbol{s}}(\nabla_{\boldsymbol{h}} J) \\ \nabla_{\boldsymbol{h}}(\nabla_{\boldsymbol{s}} J) & \nabla_{\boldsymbol{h}}^2 J \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \end{bmatrix} \\ \boldsymbol{v_1^{\mathsf{H}}}(\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H})\boldsymbol{v_1} + \boldsymbol{v_2^{\mathsf{H}}}(\nabla_{\boldsymbol{h}}(\nabla_{\boldsymbol{s}} J))\boldsymbol{v_1} + \boldsymbol{v_1^{\mathsf{H}}}(\nabla_{\boldsymbol{s}}(\nabla_{\boldsymbol{h}} J))\boldsymbol{v_2} + \boldsymbol{v_2^{\mathsf{H}}}(\nabla_{\boldsymbol{h}}^2 J)\boldsymbol{v_2} \\ \boldsymbol{v_1^{\mathsf{H}}}(\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H})\boldsymbol{v_1} + \boldsymbol{v_2^{\mathsf{H}}}(\nabla_{\boldsymbol{h}}^2 J)\boldsymbol{v_2} \geq 0 \end{aligned} \tag{3.8}$$

since the block diagonal matrix $\nabla_{\boldsymbol{h}}^2 J$ is positive-semidefinite, and $\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H}$ is quadratic (hence, also positive-semidefinite).

3.3 Fisher Information Matrix

The Fisher information is the way of measuring how much information an observable random variable, y, carries about an unknown parameter, θ, of a distribution that models the variable, $y \in f(y;\theta)$. In case the observed quantities are random vectors and the parameters are multivariate, the Fisher information matrix can be calculated. The inverse of the Fisher information gives a lower bound on the variance of any unbiased estimator of $\boldsymbol{H}, \boldsymbol{s}$ (also known as the Cramer-Rao Bound [64]). This is a type of *small-error bound* [65], which depends on the probability density near its true value, and describes the variance in the case of a large SNR, where the likelihood function is unimodal (has one distinct peak, as in convex functions). We will derive the Fisher information matrix of our likelihood function in this section.

Let $\boldsymbol{y} \in \mathbb{C}^N$ be a random vector whose probability density function, $f(\boldsymbol{y};\boldsymbol{\theta})$ is characterized by an unknown parameter $\boldsymbol{\theta} \in \mathbb{C}^P$. The covariance matrix of derivative of the log likelihood of the pdf with respect to $\boldsymbol{\theta}$ is known as the Fisher information matrix.

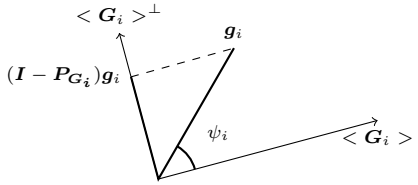


Figure 3.1: Geometric interpretation of CRB.

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \log(f(\boldsymbol{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log(f(\boldsymbol{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\mathsf{H}}\right] \tag{3.9}$$

The inverse of the Fisher information matrix, $\mathcal{I}^{-1}(\boldsymbol{\theta})$, lower bounds the error covariance matrix for any unbiased estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ of $\boldsymbol{\theta}$,

$$\mathbb{E}\left[(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{y})(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{y})^{\mathsf{H}}\right] \geq \mathcal{I}^{-1}(\boldsymbol{\theta}) \tag{3.10}$$

When $f(\boldsymbol{y};\boldsymbol{\theta})$ is a multivariate normal density $\mathcal{N}(\boldsymbol{u}(\boldsymbol{\theta}), \sigma^2\boldsymbol{I})$ with unknown mean vector $\boldsymbol{u}(\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, and known covariance $\sigma^2\boldsymbol{I}$, the Fisher information matrix is the Grammian [66],

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\boldsymbol{G}^{\top}\boldsymbol{G} \tag{3.11}$$

The ith column $\boldsymbol{g}_i$ of $\boldsymbol{G} = [\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_p]$, also known as the *sensitivity vector*, is the partial derivative $\boldsymbol{g}_i = \frac{\partial \boldsymbol{u}(\boldsymbol{\theta})}{\partial \theta_i}$, which characterizes the sensitivity of the mean vector $\boldsymbol{u}(\boldsymbol{\theta})$ to the ith parameter θ_i. The CRB for estimating $\boldsymbol{\theta}_i$ is given by

$$\left(\mathcal{I}^{-1}(\boldsymbol{\theta})\right)_{ii} = \sigma^2\left(\boldsymbol{g}_i^T\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}_i}\right)\boldsymbol{g}_i\right)^{-1} \tag{3.12}$$

where $\boldsymbol{G}_i$ consists of all columns of $\boldsymbol{G}$ except $\boldsymbol{g}_i$, and $\boldsymbol{P}_{\boldsymbol{G}_i}$ is the orthogonal projection onto the column space of $\boldsymbol{G}_i$ [67]. The projection of the i^{th} sensitivity vector $\boldsymbol{g}_i$ onto the the subspace orthogonal to $< \boldsymbol{G}_i >$ is given by $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}_i}$. The norm-squared of this projection is $\boldsymbol{g}_i^T\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}_i}\right)\boldsymbol{g}_i$, and the inverse of this norm-squared is the variance bound.

Using a different interpretation. the CRB can also be written as

$$\left(\mathcal{I}^{-1}(\boldsymbol{\theta})\right)_{ii} = \frac{\sigma^2}{\|\boldsymbol{g}_i\|_2^2 \sin^2(\psi_i)} \tag{3.13}$$

where ψ_i is the principal angle between subspaces $< \boldsymbol{g}_i >$ and $< \boldsymbol{G}_i >$.

Now, for our derivation, let $\boldsymbol{\theta} = \begin{bmatrix} s_{1\tau} & \ldots & s_{M\tau} & h_{11} & h_{12} & \ldots & h_{NM} \end{bmatrix}^{\top}, \boldsymbol{\theta} \in \mathbb{C}^{M(N+1)}$, and

$\boldsymbol{y} = \boldsymbol{x}_\tau$. We see that $\boldsymbol{u}(\boldsymbol{\theta}) = \boldsymbol{H}\boldsymbol{s}_\tau, \in \mathbb{C}^N$, and $\boldsymbol{x}_\tau|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{u}(\boldsymbol{\theta}), \sigma_w^2\boldsymbol{I})$. The matrix $\boldsymbol{G} \in \mathbb{C}^{N\times M(N+1)}$ is given by

$$\begin{aligned}
\boldsymbol{G} &= \begin{bmatrix} \frac{\partial \boldsymbol{u}}{\partial s_1} & \dots & \frac{\partial \boldsymbol{u}}{\partial s_M} & \frac{\partial \boldsymbol{u}}{\partial h_{11}} & \dots & \frac{\partial \boldsymbol{u}}{\partial h_{1M}} & \dots & \frac{\partial \boldsymbol{u}}{\partial h_{N1}} & \dots & \frac{\partial \boldsymbol{u}}{\partial h_{NM}} \end{bmatrix} \\
&= \begin{bmatrix} h_{11} & \dots & h_{1M} & s_1 & \dots & s_M & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ h_{N1} & \dots & h_{NM} & 0 & \dots & 0 & \dots & s_1 & \dots & s_M \end{bmatrix} \\
&= \left[\begin{array}{c|ccc} \boldsymbol{H} & \boldsymbol{s}^\mathsf{T} & \dots & \boldsymbol{0} \\ & \boldsymbol{0} & \dots & \boldsymbol{s}^\mathsf{T} \end{array}\right]
\end{aligned}$$

The Fisher information matrix is

$$\begin{aligned}
\mathcal{I}_{\boldsymbol{x}_\tau}(\boldsymbol{\theta}) &= \frac{1}{\sigma_w^2}\boldsymbol{G}^\mathsf{H}\boldsymbol{G} \\
&= \frac{1}{\sigma_w^2}\begin{bmatrix} \boldsymbol{H}^\mathsf{H}\boldsymbol{H} & \boldsymbol{s}^\mathsf{H} \otimes \boldsymbol{H}^\mathsf{H} \\ \boldsymbol{s} \otimes \boldsymbol{H} & \boldsymbol{s}\boldsymbol{s}^\mathsf{H} \otimes \boldsymbol{I}_{N\times N} \end{bmatrix}
\end{aligned} \tag{3.14}$$

where $\otimes$ denotes the kronecker product. We see that the microphone self-noise variance, σ_w^2, is directly proportional to the CRB. So, a smaller noise variance would give better estimates for the sources and transfer functions.

The CRB for the m^{th} source, is also inversely proportional to $||\boldsymbol{h}_m||_2^2$, (3.13). With increasing number of microphones, N, we expect $||\boldsymbol{h}_m||_2^2$ to increase, thereby reducing the error in estimation of the m^{th} source. However, this also depends on the principal angle between the subspaces. Orthogonal subspaces will give the lowest CRB, while colinear subspaces will give the highest variance in error, as shown in Fig. 3.1. Even one small principal angle can produce a large variance. The physical interpretation of this is that the addition of arbitrarily placed room microphones may yield better results than introducing additional microphones very close to the original microphone recording the desired source, as they will have similar transfer functions.

3.4 Solution

There are multiple ways of solving this optimization problem. Since the cost function in (3.3) is convex in $\boldsymbol{H}$ and $\boldsymbol{s}_\tau$, the optimal values $\boldsymbol{H}^*, \boldsymbol{s}_\tau^*$ can be found by calculating the gradient of the cost

function and finding its roots, thus giving the following solution,

$$\begin{aligned} \boldsymbol{H}^* &= \left(\tilde{\boldsymbol{H}} + \frac{\sigma_\nu^2}{\sigma_w^2}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{s}_t^{\mathsf{H}}\right)\left(\boldsymbol{I} + \frac{\sigma_\nu^2}{\sigma_w^2}\sum_{t=1}^{T}\boldsymbol{s}_t\boldsymbol{s}_t^{\mathsf{H}}\right)^{-1} \\ \boldsymbol{s}_\tau^* &= (\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{H}^*)^{-1}\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{x}_\tau \end{aligned} \tag{3.15}$$

where $\boldsymbol{I}$ is an $M \times M$ identity matrix. It is to be noted that $\frac{\sigma_\nu^2}{\sigma_w^2}$ acts as a hyperparameter. A small value of $\frac{\sigma_\nu^2}{\sigma_w^2}$ gives more weight to the initial estimate of the transfer function, $\tilde{\boldsymbol{H}}$. If the hyperparameter is zero, then $\boldsymbol{H} = \tilde{\boldsymbol{H}}$, and we converge to the least-squares solution. The two equations in (3.15) are a non-linear function of $\boldsymbol{s}_\tau$ and can be solved using any numerical root finder, such as `MATLAB`'s `fsolve`.

It is also feasible to implement gradient-based solvers, such as gradient-descent [68] or the Newton-Raphson solver [69], which makes use of the Hessian. In gradient-descent, the current iteration of $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{s}_1, \ldots \boldsymbol{s}_\tau, h_{11}, \ldots, h_{NM}\end{bmatrix}^\top$ is updated as,

$$\boldsymbol{\theta}_{n+1} := \boldsymbol{\theta}_n - \mu\nabla_{\boldsymbol{\theta}_n} J(\boldsymbol{\theta}) \tag{3.16}$$

where μ is the step size and $\nabla_{\boldsymbol{\theta}_n} J$ is the gradient of the cost function at $\boldsymbol{\theta} = \boldsymbol{\theta}_n$.

Similarly, the Newton-Raphson method can be used, which leads to quicker convergence if the starting point is close to the global minimum,

$$\boldsymbol{\theta}_{n+1} := \boldsymbol{\theta}_n - \nabla^2_{\boldsymbol{\theta}_n} J^{-1}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}_n} J(\boldsymbol{\theta}) \tag{3.17}$$

where $\nabla^2_{\boldsymbol{\theta}_n} J(\boldsymbol{\theta})$ is the Hessian matrix of the cost function.

3.4.1 Code vectorization and parallelization

Since the computation is independent over frequency bins, the optimization can be parallelized over multiple cores/clusters using `MATLAB`'s `parfor`, for example. The computations can be further sped up by replacing summation over vectors with matrix operations. More specifically, we can write the time-varying microphone signals as an $N \times T$ matrix $\boldsymbol{X}$, and the source signals as an $M \times T$ matrix $\boldsymbol{S}$, with time frames along the rows. The operations in (3.15) then become

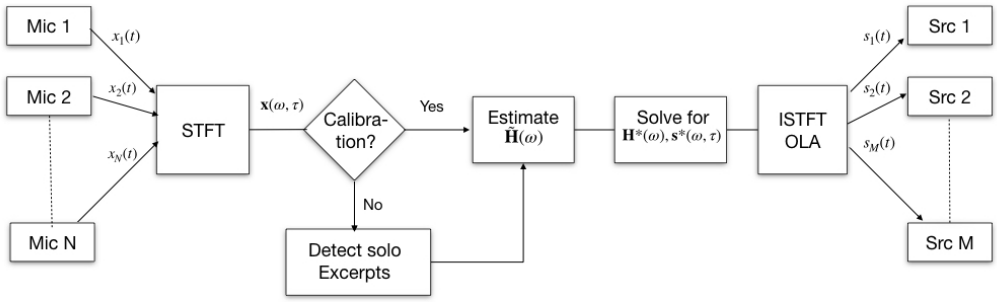


Figure 3.2: Block diagram of the proposed ML estimator.

$$
\begin{aligned}
\boldsymbol{H}^* &= \left(\tilde{\boldsymbol{H}} + \frac{\sigma_\nu^2}{\sigma_w^2}\boldsymbol{X}\boldsymbol{S}^{\mathsf{H}}\right)\left(\boldsymbol{I} + \frac{\sigma_\nu^2}{\sigma_w^2}\boldsymbol{S}\boldsymbol{S}^{\mathsf{H}}\right)^{-1} \\
\boldsymbol{S}^* &= (\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{H}^*)^{-1}\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{X} \\
\text{with } \boldsymbol{X} &= \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_T \end{bmatrix}, \\
\boldsymbol{S} &= \begin{bmatrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \cdots & \boldsymbol{s}_T. \end{bmatrix}
\end{aligned}
\tag{3.18}
$$

Regardless, the computation time depends significantly on the number of sources and microphones in the system. For a setup with M sources and N microphones, a total of $M(N+T)$ variables need to be solved for each frequency bin. So, a two microphone two-source setup of a singer-songwriter playing the guitar would compute much faster than an orchestral recording with several microphones recording the various sections.

3.5 Summary

In this chapter, we have proposed a novel solution to the microphone bleed cancellation problem by deriving a cost function that simultaneously solves for the sources and the relative transfer functions between the sources and the microphones. We have shown that minimizing this cost function gives us the Maximum Likelihood estimate when the distributions are assumed to be Gaussian. We have proved that the cost function is convex; therefore, its global minimum can be found using gradient-based methods. Finally, we have discussed methods to speed up the computation with vectorization and parallelization. A block diagram of the entire process is shown in Fig. 3.2.

Chapter 4

Bayesian Estimation : Minimum Mean Squared Error and Maximum Aposteriori Probability

Unlike Maximum Likelihood estimation, where only the conditional probability of the microphone signals given the source, $p(\boldsymbol{x}|\boldsymbol{s})$, is defined, Bayesian estimation also assumes the source to be a random variable with an *a priori* distribution, $p(\boldsymbol{s})$. In this chapter, we derive two such estimators for the microphone bleed canceller - the Minimum Mean Squared Estimator (MMSE) which gives the well-known Multichannel Wiener Filter solution, and the Maximum Aposteriori Probability (MAP) estimator, which extends the previously derived ML model. We discuss methods for estimating the *a priori* statistics of the source probability distribution.

4.1 MMSE Estimator - Multichannel Wiener Filter

Although Multichannel Wiener filter (MWF) solutions have been proposed in state-of-the-art methods for crosstalk cancellation, we discuss the Generalized Eigenvalue Decomposition based Multichannel Wiener Filter (GEVD based MWF). This method overcomes the need for using ad-hoc measures for estimating the source power spectrum, such as proposed in [4]. The GEVD based MWF requires an estimate of the noise (or interfering signal) correlation matrix, which we can estimate by making use of the *calibration stage* during setup. Typically, this solution is much faster than the maximum likelihood approach, since it has a closed-form solution and does not require any iterative

optimization.

4.1.1 Model

The model in (2.2) can also be written as the desired source for the n^{th} mic, plus a noise term that consists of the interfering sources, convolved with their respective transfer functions.

$$\begin{aligned} x_n(k) &= s_n(k) * h_{nn}(k) + \sum_{i=1, i\neq n}^{M} s_i(k) * h_{ni}(k) + w(k), \\ x_n(k) &= \bar{s}_n(k) + \bar{v}_n(k), \\ \boldsymbol{x}(\omega,\tau) &= \underbrace{\bar{\boldsymbol{s}}(\omega,\tau)}_{\text{source}} + \underbrace{\bar{\boldsymbol{v}}(\omega,\tau)}_{\text{interference}} \end{aligned} \tag{4.1}$$

4.1.2 Optimum inverse filter

For each frequency bin, ω, the inverse filtering problem is to find the optimal filter weights $\hat{\boldsymbol{W}}(\omega)$, such that

$$\hat{\boldsymbol{s}}(\omega) = \hat{\boldsymbol{W}}^{\mathsf{H}}(\omega)\boldsymbol{x}(\omega) \tag{4.2}$$

The error between $\hat{\boldsymbol{s}}$ and $\bar{\boldsymbol{s}}$ is

$$e(\omega) = \bar{\boldsymbol{s}}(\omega) - \hat{\boldsymbol{s}}(\omega) \tag{4.3}$$

The optimal MMSE solution $\hat{\boldsymbol{W}}^*$ is derived by minimizing the sum of squared errors. We are omitting the frequency variable, ω, for clarity in the derivations henceforth.

$$\begin{aligned} J(\hat{\boldsymbol{W}}) &= \mathbb{E}\left[\frac{1}{2}(\bar{\boldsymbol{s}} - \hat{\boldsymbol{s}})^{\mathsf{H}}(\bar{\boldsymbol{s}} - \hat{\boldsymbol{s}})\right] \\ &= \frac{1}{2}\mathbb{E}\left[\bar{\boldsymbol{s}}^{\mathsf{H}}\bar{\boldsymbol{s}} - 2\bar{\boldsymbol{s}}^{\mathsf{H}}\hat{\boldsymbol{W}}^{\mathsf{H}}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{W}}^{\mathsf{H}}\boldsymbol{x}\right] \\ \nabla_{\hat{\boldsymbol{W}}} J &= \frac{1}{2}\mathbb{E}\left[0 - 2\boldsymbol{x}\bar{\boldsymbol{s}}^{\mathsf{H}} + 2\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\hat{\boldsymbol{W}}^*\right] \\ 0 &= \mathbb{E}[-\boldsymbol{x}\bar{\boldsymbol{s}}^{\mathsf{H}} + \boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\hat{\boldsymbol{W}}^*] \\ \hat{\boldsymbol{W}}^* &= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}})^{-1}\mathbb{E}(\boldsymbol{x}\bar{\boldsymbol{s}}^{\mathsf{H}}) \\ &= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}})^{-1}\mathbb{E}(\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}^{\mathsf{H}}) \end{aligned} \tag{4.4}$$

where $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}})$ is the covariance matrix of the microphone signals $\boldsymbol{x}$, and $\mathbb{E}(\boldsymbol{x}\bar{\boldsymbol{s}}^{\mathsf{H}})$ is the cross-covariance between the microphone signals and the sources, which is equal to the covariance matrix of the source vector, $\mathbb{E}(\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}^{\mathsf{H}})$ if the noise, $\bar{\boldsymbol{v}}$, and the source, $\bar{\boldsymbol{s}}$, are uncorrelated. This solution is also known as the Multichannel Wiener Filter (MWF).

4.2 GEVD based MWF

The covariance matrix (or power spectra) of the microphone signals in eq. (4.4), $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}})$ can be calculated via sample averaging.

$$\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}) = \boldsymbol{R}_{xx} \approx \frac{1}{T-1}\sum_{\tau=1}^{T}\boldsymbol{x}_\tau\boldsymbol{x}_\tau^{\mathsf{H}} \tag{4.5}$$

To estimate the source covariance matrix in eq. (4.4), we use the Generalized Eigenvalue Decomposition based Multichannel Wiener Filter (GEVD based MWF) [70]. The source covariance matrix is the difference of the sensor covariance matrix and the interference matrix,i.e, $\boldsymbol{R}_{\bar{s}\bar{s}} = \boldsymbol{R}_{xx} - \boldsymbol{R}_{\bar{v}\bar{v}}$. It has been demonstrated in [71] that incorporating a low rank approximation based on either the eigenvalue decomposition (EVD) of $\boldsymbol{R}_{\bar{s}\bar{s}}$ or the generalized eigenvalue decomposition (GEVD) of $\boldsymbol{R}_{xx}, \boldsymbol{R}_{\bar{v}\bar{v}}$ enhances the estimation performance of the MWF, especially in low-SNR conditions. The GEVD-based rank-R approximation has been shown to deliver the best performance, as it effectively selects the R 'modes' corresponding to the highest SNR. Of course, when the number of sources is known, as it is in our case, we can assume $M = R$.

We find the generalized eigenvalues, $\boldsymbol{L} = [\lambda_1, \ldots, \lambda_N]$, and the eigenvectors, $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N]^\top$, of the matrix pair, $(\boldsymbol{R}_{xx}, \boldsymbol{R}_{\bar{v}\bar{v}})$, such that

$$\boldsymbol{R}_{\bar{v}\bar{v}}^{-1}\boldsymbol{R}_{xx} = \boldsymbol{Q}\boldsymbol{L}\boldsymbol{Q}^{-1} \tag{4.6}$$

if $\boldsymbol{R}_{vv}$ is invertible. Assuming the eigenvalues and vectors are sorted in descending order, the GEVD is equivalent to a joint diagonalization of $\boldsymbol{R}_{xx}, \boldsymbol{R}_{\bar{v}\bar{v}}$,

$$\boldsymbol{R}_{xx} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathsf{H}},\ \boldsymbol{R}_{\bar{v}\bar{v}} = \boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}^{\mathsf{H}} \tag{4.7}$$

Using eq. (4.6), it can be verified that $\boldsymbol{U} = \boldsymbol{Q}^{\mathsf{H}}$, $\boldsymbol{L} = \boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1}$, $\boldsymbol{\Sigma} = \boldsymbol{Q}^{\mathsf{H}}\boldsymbol{R}_{xx}\boldsymbol{Q}$ and $\boldsymbol{\Gamma} = \boldsymbol{Q}^{\mathsf{H}}\boldsymbol{R}_{\bar{v}\bar{v}}\boldsymbol{Q}$. If we scale the eigenvectors, $\boldsymbol{q}_n$'s, such that $\boldsymbol{Q}^{\mathsf{H}}\boldsymbol{R}_{\bar{v}\bar{v}}\boldsymbol{Q} = \boldsymbol{I}_N$, then we can write

$$\boldsymbol{R}_{xx} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{U}^{\mathsf{H}},\ \boldsymbol{R}_{\bar{v}\bar{v}} = \boldsymbol{U}\boldsymbol{U}^{\mathsf{H}} \tag{4.8}$$

Hence, the source correlation matrix becomes

$$\begin{aligned} \boldsymbol{R}_{\bar{s}\bar{s}} &= \boldsymbol{R}_{xx} - \boldsymbol{R}_{\bar{v}\bar{v}} \\ &= \boldsymbol{U}(\boldsymbol{L} - \boldsymbol{I}_N)\boldsymbol{U}^{\mathsf{H}} \end{aligned} \tag{4.9}$$

The rank-M approximation of $\boldsymbol{R}_{\bar{s}\bar{s}}$ becomes $\boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^{\mathsf{H}}$, with the last $N-M$ entries of $(\boldsymbol{L}-\boldsymbol{I}_N)$ equal to zero

$$\boldsymbol{\Delta} = \text{diag}\left[(\lambda_1 - 1), \ldots, (\lambda_M - 1),\ \underbrace{0, \ldots, 0}_{N-M \text{ zeros}}\right]$$

Thus, the optimum inverse filter may be written as

$$\begin{aligned} \boldsymbol{W}^* &= \boldsymbol{R}_{xx}^{-1}\boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{E}_N \\ &= (\boldsymbol{U}\boldsymbol{L}\boldsymbol{U}^{\mathsf{H}})^{-1}\boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{E}_N \\ &= \boldsymbol{U}^{-\mathsf{H}}\boldsymbol{L}^{-1}\boldsymbol{\Delta}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{E}_N \end{aligned} \tag{4.10}$$

where $\boldsymbol{L}^{-1}\boldsymbol{\Delta} = \text{diag}\left[1 - \frac{1}{\lambda_1}, \ldots, 1 - \frac{1}{\lambda_M}, 0, \ldots, 0\right]$ and $\boldsymbol{E}_N = [\boldsymbol{I}_{M\times M}\ \boldsymbol{0}_{M\times(N-M)}]^\top$. Thus, we can estimate the optimal filter for each frequency bin, given the estimates of the microphone and interference signal correlation matrices.

4.2.1 Estimation of interfering signal correlation matrix

To estimate the interference (or noise) signal for the n^{th} mic, $\hat{v}_n(k)$, we make use of the calibration stage where all microphones capture one instrument at a time. For the nth mic, all but the closest source contribute to the interference. We sum up all these $M-1$ sources to get the total interference for each microphone, convert it to the STFT domain and estimate $\boldsymbol{R}_{\bar{v}\bar{v}}$ by taking the outer product and averaging across all time frames, similar to (4.5).

4.2.2 Distortion vs Interference weighting

Typically in noise reduction applications, there is a tradeoff between reducing distortion in the separated sources and cancelling interference. In recording studio applications, it is imperative to get as little distortion as possible in the processed signals, even if it comes at the cost of less interference cancellation. In this section, we discuss the Speech-Distortion Weighted Multichannel Wiener filter [72] (SDW-MWF), that is parameterized to give a tradeoff between the two.

Since $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{v}}$ are uncorrelated, the cost function in eq. (4.4) can also be written as

$$\begin{aligned} J(\hat{\boldsymbol{W}}) &= \mathbb{E}\left(||\bar{\boldsymbol{s}} - \hat{\boldsymbol{W}}^{\mathsf{H}}\boldsymbol{x}||^2\right) \\ &= \mathbb{E}\left(||\bar{\boldsymbol{s}} - \hat{\boldsymbol{W}}^{\mathsf{H}}(\bar{\boldsymbol{s}} + \bar{\boldsymbol{v}})||^2\right) \\ &= \underbrace{\mathbb{E}\left(||\bar{\boldsymbol{s}} - \hat{\boldsymbol{W}}^{\mathsf{H}}\bar{\boldsymbol{s}}||^2\right)}_{\epsilon_{\bar{s}}^2} + \underbrace{\mathbb{E}\left(||\hat{\boldsymbol{W}}^{\mathsf{H}}\bar{\boldsymbol{v}}||^2\right)}_{\epsilon_{\bar{v}}^2} \end{aligned} \tag{4.11}$$

The first term, $\epsilon_{\bar{s}}^2$, is a source distortion term, while the second term, $\epsilon_{\bar{v}}^2$, is a noise reduction term. The Speech-Distortion Weighted Multichannel Wiener filter (SDW-MWF) [72] provides a tradeoff between these two errors, with a trade-off parameter $\mu \geq 0$.

$$\begin{aligned} J_{\text{SDW}}(\hat{\boldsymbol{W}}) &= \mathbb{E}\left(||\bar{\boldsymbol{s}} - \hat{\boldsymbol{W}}^{\mathsf{H}}\bar{\boldsymbol{s}}||^2\right) + \mu\mathbb{E}\left(||\hat{\boldsymbol{W}}^{\mathsf{H}}\bar{\boldsymbol{v}}||^2\right) \\ \hat{\boldsymbol{W}}_{\text{SDW}}^{*} &= (\boldsymbol{R}_{\bar{s}\bar{s}} + \mu\boldsymbol{R}_{\bar{v}\bar{v}})^{-1}\boldsymbol{R}_{\bar{s}\bar{s}} \\ &= (\boldsymbol{R}_{xx} + (\mu-1)\boldsymbol{R}_{\bar{v}\bar{v}})^{-1}(\boldsymbol{R}_{xx} - \boldsymbol{R}_{\bar{v}\bar{v}}) \end{aligned} \tag{4.12}$$

Using eq. (4.8), the optimal weight filter can be simplified as

$$\begin{aligned} \hat{\boldsymbol{W}}_{\text{SDW}}^{*} &= \boldsymbol{U}^{-\mathsf{H}}\left(\boldsymbol{L} + (\mu-1)\boldsymbol{I}_M\right)^{-1}(\boldsymbol{L} - \boldsymbol{I}_M)\boldsymbol{U}^{\mathsf{H}}\boldsymbol{E}_N \\ &= \boldsymbol{U}^{-\mathsf{H}}\boldsymbol{\Phi}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{E}_N \end{aligned} \tag{4.13}$$

where $\boldsymbol{\Phi} = \text{diag}\left[\frac{\lambda_1 - 1}{\lambda_1 + \mu - 1}, \ldots, \frac{\lambda_M - 1}{\lambda_M + \mu - 1}, 0, \ldots, 0\right]$. When $\mu = 1$, the two solutions are equivalent, i.e., $\hat{\boldsymbol{W}}_{\text{SDW}} = \hat{\boldsymbol{W}}$ and both source distortion and interference reduction get equal weighting. It has been proved that the output SNR after noise reduction with the speech-distortion weighted multichannel Wiener filter is always larger than or equal to the input SNR, for any filter length, for any value of the tradeoff parameter between noise reduction and speech distortion, and for all possible speech

and noise correlation matrices [31].

4.3 MAP Estimator - Maximum Aposteriori Probability

The MAP estimator is closely related to the ML estimator, but employs an optimization objective with a prior distribution. It can be seen as a regularization of the ML estimator. For deriving the MAP estimator used in this problem, we use the same model as in (2.2), and assume that the sources are independent and identically distributed across frames with zero mean and a covariance matrix, known *a priori*, i.e., $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{P}_s)$. To estimate the source covariance matrix, we can estimate the covariance matrices of the microphone signals and the interfering signals, as described in Section 4.10, find their generalized eigenvalues, and then use (4.9).

4.3.1 Cost function derivation

Using the same definitions of the vectors as in Section 3.1, we maximize the joint probability of the source and the true transfer function, given the microphone signals over all frames, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ and a noisy estimate of the initial transfer function, $\tilde{\boldsymbol{h}}$. Then, we use Bayes' theorem to write the conditional probability of the source given the data as the conditional probability of the data given the source, multiplied with the source probability.

$$\begin{aligned} J(\boldsymbol{s_1}, \ldots, \boldsymbol{s_T}, \boldsymbol{h}) &= \max\ p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{h} | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, \tilde{\boldsymbol{h}}) \\ &= \max\ p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, \tilde{\boldsymbol{h}} | \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{h})\ p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{h}) \\ &= \max\ p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T | \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T)\ p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T)\ p(\boldsymbol{h} | \tilde{\boldsymbol{h}}) \\ &= \max\ \left(\prod_{t=1}^{T} p(\boldsymbol{x}_t | \boldsymbol{s}_t)\ p(\boldsymbol{s}_t) \right)\ p(\boldsymbol{h} | \tilde{\boldsymbol{h}}) \end{aligned} \tag{4.14}$$

As before, we minimize the negative of the log likelihood which gives us the following cost function

$$\begin{aligned} J(\boldsymbol{s_1}, \ldots, \boldsymbol{s_T}, \boldsymbol{h}) = \min \sum_{t=1}^{T} \left(\frac{1}{\sigma_w^2} ||\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t||^2 + \boldsymbol{s}_t^{\mathsf{H}} \boldsymbol{P}_s^{-1} \boldsymbol{s}_t \right) \\ + \frac{1}{\sigma_\nu^2} \operatorname{Tr}\left((\tilde{\boldsymbol{H}} - \boldsymbol{H})^{\mathsf{H}} (\tilde{\boldsymbol{H}} - \boldsymbol{H}) \right) \end{aligned} \tag{4.15}$$

Here, $\boldsymbol{P}_s^{-1}$, is an $M \times M$ matrix that can be estimated as

$$\hat{\boldsymbol{P}}_s^{-1} = \boldsymbol{E}_N^{\top} \left(\boldsymbol{Q} \boldsymbol{\Delta}^{-1} \boldsymbol{Q}^{\mathsf{H}} \right) \boldsymbol{E}_N \tag{4.16}$$

where $\boldsymbol{\Delta}^{-1} = \text{diag}\left[\frac{1}{\lambda_1 - 1}, \ldots, \frac{1}{\lambda_M - 1}, \underbrace{0, \ldots, 0}_{N-M \text{ zeros}} \right]$ and λ_i's are the first M generalized eigenvalues of the microphone and interfering signal covariance matrix pair, sorted in descending order of magnitude.

4.3.2 Proof of convexity

In general, the sum of convex functions is convex (see Appendix A). Since we have already proved that the MLE cost function is convex in Sec. 3.2, we only need to prove that $J_0 = \sum_{t=1}^{T} \boldsymbol{s}_t^{\mathsf{H}} \boldsymbol{P}_s^{-1} \boldsymbol{s}_t$ is convex in $\boldsymbol{s}_t$. The second derivative of J_0 with respect to $\boldsymbol{s}_t$ is $\nabla_{\boldsymbol{s}_t}^2 J_0 = \boldsymbol{P}_s^{-1}$. Therefore, to prove convexity, we only need to prove that the inverse of the covariance matrix, $\boldsymbol{P}_s^{-1}$, is positive-semidefinite.

We know that $\boldsymbol{P}_s$ is an autocorrelation matrix, hence it is positive-semidefinite with non-negative eigenvalues, β, and orthonormal eigenvectors, $\boldsymbol{Q}$ (see Appendix B).

$$\boldsymbol{P}_s = \boldsymbol{Q} \boldsymbol{\beta} \boldsymbol{Q}^{\mathsf{H}}$$

where $\boldsymbol{\beta} = \text{diag}(\beta_1, \ldots, \beta_M)$ The inverse of $\boldsymbol{P}_s$ is

$$\boldsymbol{P}_s^{-1} = \boldsymbol{Q}^{\mathsf{H}} \boldsymbol{\beta}^{-1} \boldsymbol{Q}$$

where $\boldsymbol{\beta}^{-1} = \text{diag}\left(\frac{1}{\beta_1}, \ldots, \frac{1}{\beta_M} \right)$. Since $\boldsymbol{P}_s$ is rank-M, all β_i's are positive, so are $\frac{1}{\beta_i}$'s. Hence, the matrix $\boldsymbol{P}_s^{-1}$ exists and is positive definite. Therefore, the MAP cost function is also convex.

4.3.3 Solution

Similar to Section 3.4, we can analytically calculate the gradient of the cost function with respect to $\boldsymbol{s}_t$ and $\boldsymbol{H}$, and equate it to zero to find the optimal solution. This non-linear equation can be solved using a numerical root finder.

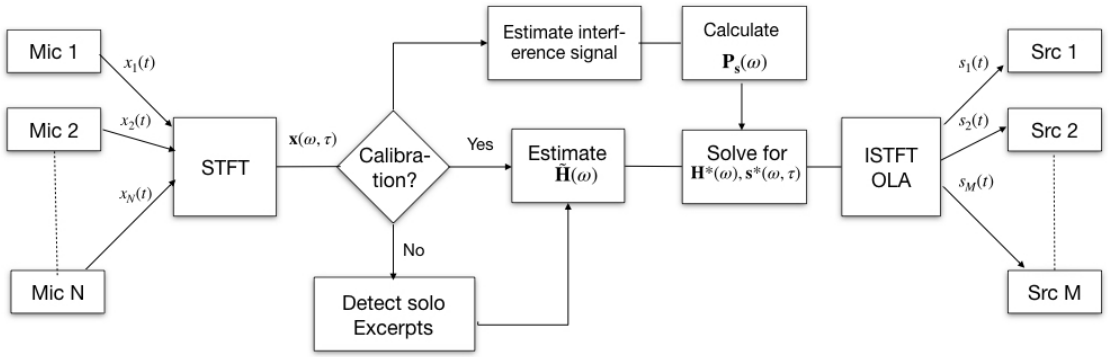


Figure 4.1: Block diagram of the proposed MAP estimator.

$$\begin{aligned} \boldsymbol{H}^* &= \left(\tilde{\boldsymbol{H}} + \frac{\sigma_\nu^2}{\sigma_w^2}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{s}_t^{\mathsf{H}}\right)\left(\boldsymbol{I} + \frac{\sigma_\nu^2}{\sigma_w^2}\sum_{t=1}^{T}\boldsymbol{s}_t\boldsymbol{s}_t^{\mathsf{H}}\right)^{-1} \\ \boldsymbol{s}_\tau^* &= (\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{H}^* + \sigma_w^2\boldsymbol{P}_s^{-1})^{-1}\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{x}_\tau \end{aligned} \tag{4.17}$$

With the source and microphone vectors stacked frame-wise, the optimal solution is

$$\begin{aligned} \boldsymbol{H}^* &= \left(\tilde{\boldsymbol{H}} + \frac{\sigma_\nu^2}{\sigma_w^2}\boldsymbol{X}\boldsymbol{S}^{\mathsf{H}}\right)\left(\boldsymbol{I} + \frac{\sigma_\nu^2}{\sigma_w^2}\boldsymbol{S}\boldsymbol{S}^{\mathsf{H}}\right)^{-1} \\ \boldsymbol{S}^* &= (\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{H}^* + \sigma_w^2\boldsymbol{P}_s^{-1})^{-1}\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{X} \end{aligned} \tag{4.18}$$

When $\sigma_w = 0$ (when there is no microphone measurement noise), the MAP solution is equal to the MLE solution. The value of σ_w^2 determines how much we trust the *a priori* covariance matrix estimate.

4.4 Summary

In this chapter, we have proposed Bayesian methods to cancel microphone bleed. First, we have derived the MMSE estimator (Wiener filter), and discussed ways of estimating the source signal covariance matrix with Generalized Eigenvalue Decomposition. Then, we have extended the analysis of Chapter 3 to derive the MAP estimator, which makes use of the *a priori* statistics of the source

signals, i.e, the source covariance matrix. Finally, we have shown that the MAP cost function is also convex, and hence the minimum is at the point where its gradient vanishes. The block diagram of the proposed MAP estimator is given in Fig. 4.1.

Chapter 5

Example: String Quartet in a Virtual Studio

Now that we have laid down the mathematical foundations of the microphone bleed cancellation problem, we need to test the proposed methods against a state-of-the-art algorithm in a controlled experimental setting. The experimental setup to gather data for analysis consists of two parts — i) simulation of a virtual recording setup using anechoic audio samples and synthesized room impulse responses, so that the room dimensions, materials, reverberation time and instrument and sensor positions can be controlled accurately, and ii) recordings made in the CCRMA recording studio for bleed cancellation in multichannel drum recordings. In this chapter, we will describe the virtual studio setup.

5.1 Synthesized data

For a realistic recording scenario, we used a dataset of anechoic string quartet recordings from TU Berlin [73]. The experimental setup of the anechoic recording of the quartet, with each musician placed in each corner of an anechoic chamber, is shown in Fig. 5.1. We placed the instruments in a virtual shoebox room of dimensions $3 \times 4 \times 3.25$ m^3. For the following analysis, we primarily worked with two instruments – the viola (Va) was placed at $(1.9, 2.5, 1.0)$ m and the violoncello (Vcl) was placed at $(1.7, 2.8, 0.8)$ m. Omnidirectional microphones were placed directly in front of the instruments on the same axis. RIR Generator [61] was used to simulate the room acoustics with the image-source method [26]. The reverberation time (RT_{60}) was fixed to be 0.8 s and the length of the impulse response generated was 128 samples at a sampling rate of 48 kHz. The short

Figure 5.1: Recordings made in the anechoic chamber at TU Berlin.

length of the impulse response is adequate to capture the early reflections, which primarily affect the cross-talk between microphones. Close-miking ensures that the effect of the room acoustics is largely negated in the mid and high frequency range. However, the lower frequency range is still affected by the early reflections in a room. The impulse response for each source-microphone pair was convolved with the anechoic recordings to generate the captured microphone signals. While the image-source method is not fool-proof, it simulates microphone cross-talk adequately because it reconstructs the early-reflections of a shoebox room correctly.

5.2 Evaluation metrics

The evaluation was done with the the Perceptual Evaluation methods for Audio Source Separation (PEASS) toolbox [74] which gives perceptually motivated scores for the separated sources. In the first step, the estimation error $\hat{s}(t)$ - $s(t)$ is split into three components: target distortion, $e_{\text{target}}(t)$, interference, $e_{\text{interf}}(t)$ and artifacts, $e_{\text{artif}}(t)$.

$$\hat{s}(t) - s(t) = e_{\text{target}}(t) + e_{\text{interf}}(t) + e_{\text{artif}}(t) \tag{5.1}$$

The use of objective measures based on energy ratios between the signal components, i.e., source to distortion ratio (SDR), the source to interference ratio (SIR) and the source to artifacts ratio (SAR), has been the standard approach in the specialized scientific community to test the quality of extracted signals [75]. The ratios are defined as:

$$
\begin{aligned}
\hat{s} &= s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \\
\text{SDR} &= 10\log_{10}\frac{||s||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2} \\
\text{SIR} &= 10\log_{10}\frac{||s||^2}{||e_{\text{interf}}||^2} \\
\text{SAR} &= 10\log_{10}\frac{||s + e_{\text{interf}} + e_{noise}||^2}{||e_{\text{artif}}||^2}.
\end{aligned}
\tag{5.2}
$$

However, a better perceptual metric is to calculate the PEMO-Q scores [76, 77]. Four new metrics are defined in [74] and used in this thesis for objective evaluation.

$$
\begin{aligned}
\text{OPS} &= \text{PEMO-Q}(\hat{s}, s) \\
\text{TPS} &= \text{PEMO-Q}(\hat{s}, \hat{s} - e_{\text{target}}) \\
\text{IPS} &= \text{PEMO-Q}(\hat{s}, \hat{s} - e_{\text{interf}}) \\
\text{APS} &= \text{PEMO-Q}(\hat{s}, \hat{s} - e_{\text{artif}})
\end{aligned}
\tag{5.3}
$$

The overall perceptual score (OPS) gives an indication of how close the separated signal is to the target signal. Target-related perceptual score (TPS) gives an indication of the target distortion. Interference-related perceptual score (IPS) is an indication of how much interference has been eliminated from the separated signal, and Artifact-related perceptual score (APS) is an indication of the artifacts in the separated signal. The mean scores (averaged over all instruments) are reported in this chapter.

5.3 Experimental details

For the Short-Time Fourier Transform (STFT), we used a frame size of 1024 samples with a Hann window, a hop size of 512 samples, and FFT length of 4096 samples. We tested our proposed MLE and MAP estimators with three different values of the hyperparameter, $\sigma_w^2/\sigma_\nu^2 = [0, 1, 100]$, as well as with the known transfer function. Three different methods of transfer function estimation were used - spectral ratio (2.2.1), M-GCC (2.2.2) and BCI (2.2.3). The first 15 reflections were calculated for M-GCC. The NMFLMS method was used for calibration with BCI, with $\lambda = 0.98$ and $\rho = 0.2$. The MCWF parameters used were the same as in [4], with the aforementioned frame size and hop size. For estimation of $\widetilde{\boldsymbol{H}}$, a different 2 s excerpt from the same recording was used. We provide sound examples for an informal listening test [78]. It is to be noted that the gains of these audio files have been normalized, so that the amplitudes lie within ± 1.

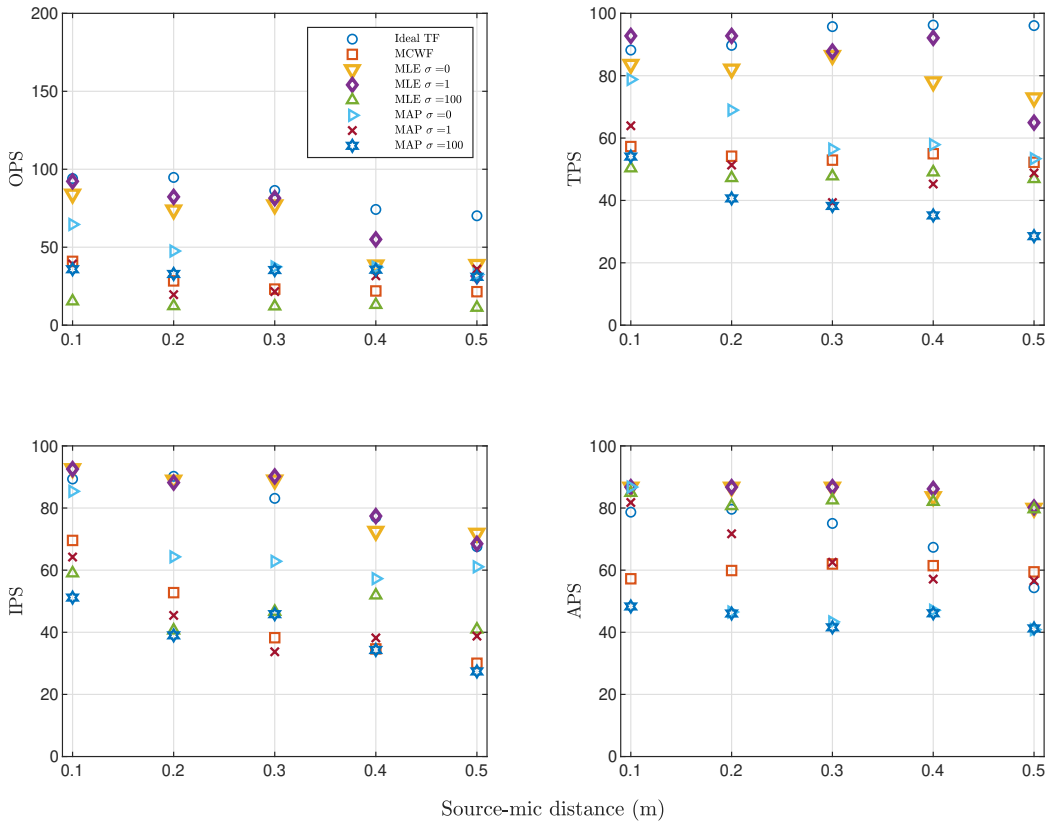


Figure 5.2: PEASS scores with spectral-ratio initialization for varying source-microphone distance.

5.4 Results

5.4.1 Effect of source-microphone distance

For the determined case, ($N = M = 2$), we varied the distance between the source and the microphones linearly from $10 - 50$ cm in steps of 10 cm.

The results with the spectral ratio calibration are shown in Fig. 5.2. For the ML estimator, the OPS, TPS and IPS of all calibration methods decline as the source-microphone distance is increased, since the close-microphone assumption starts failing. The proposed ML estimator with $\sigma_\nu^2/\sigma_w^2 \in \{0, 1\}$ outperforms MCWF by a large margin. However, performance deteriorates when the hyperparameter value is large ($\sigma_\nu^2/\sigma_w^2 = 100$). This is because the initial transfer function matrix estimated with the spectral ratio is fairly accurate, and not trusting it leads to overfitting. In this case, choosing a large hyperparameter value hurts us. The advantages of using optimization is not clear in the figure; however, listening to the sound examples provided when the source-microphone

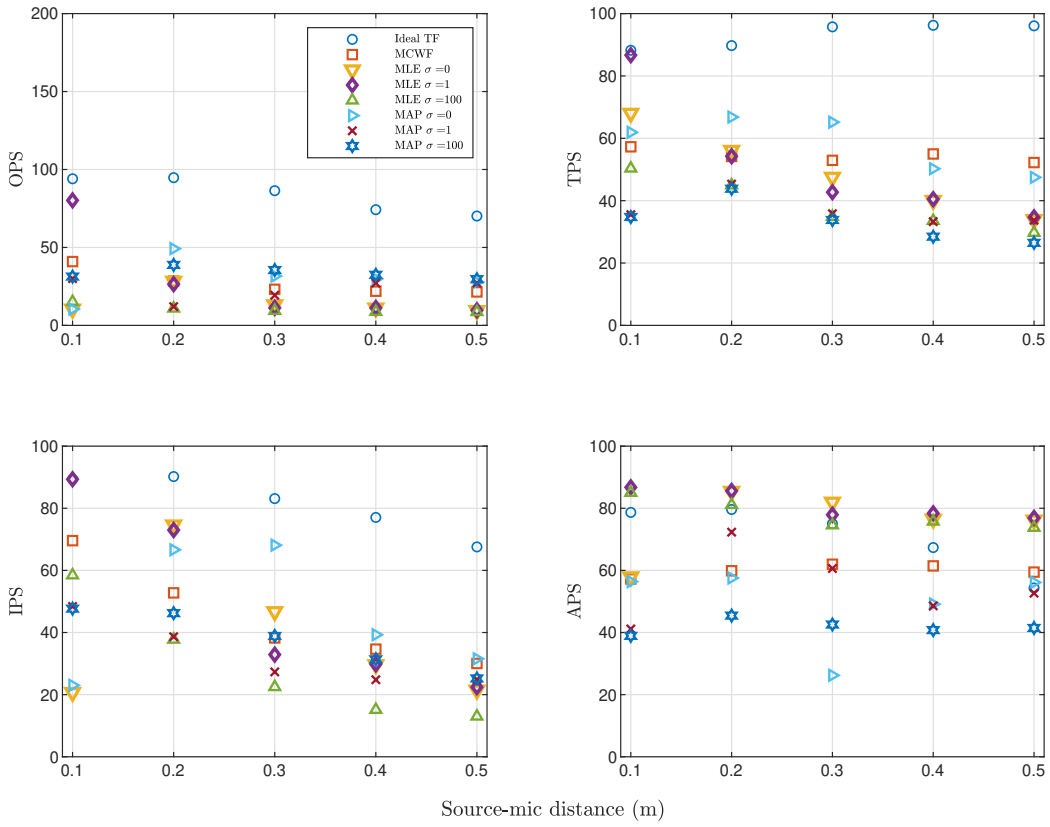


Figure 5.3: PEASS scores with M-GCC initialization for varying source-microphone distance.

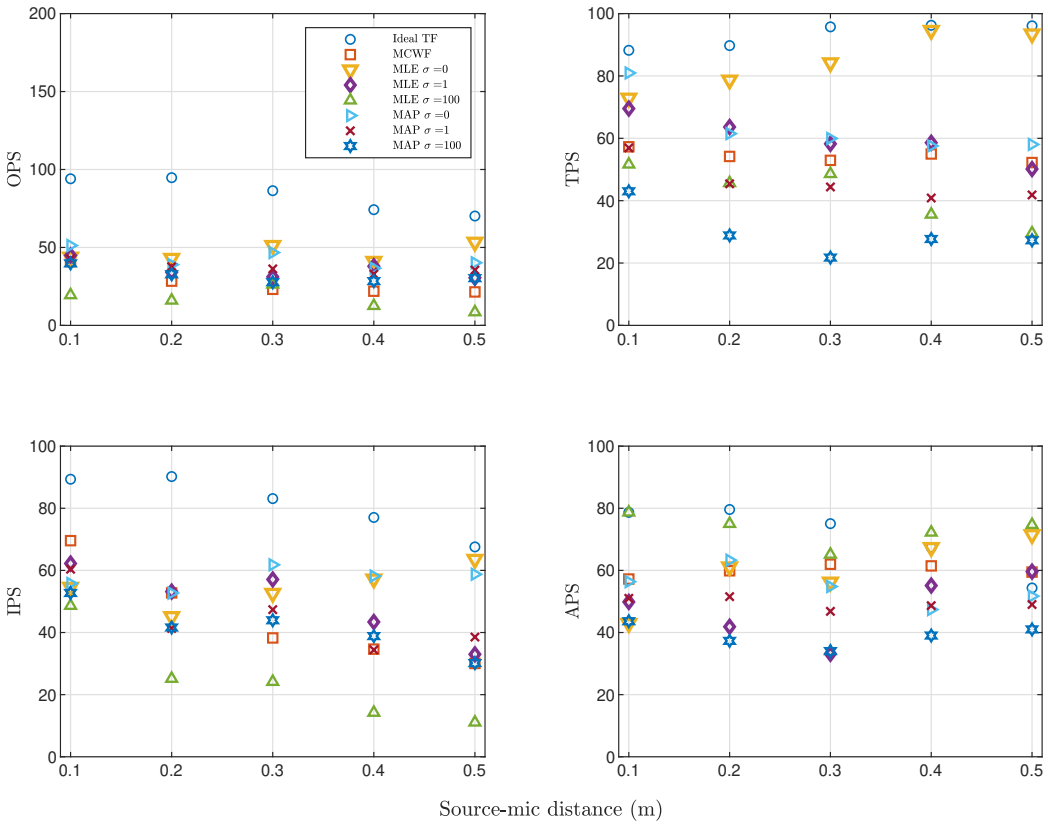


Figure 5.4: PEASS scores with BCI initialization for varying source-microphone distance.

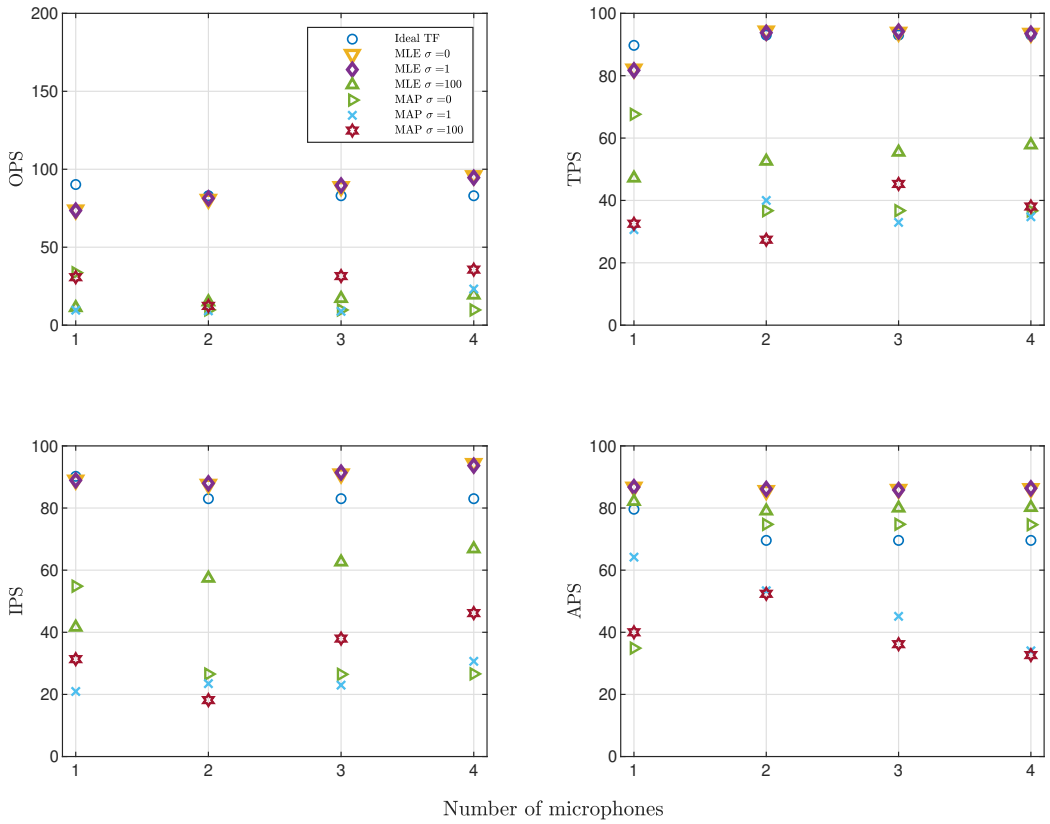


Figure 5.5: PEASS scores with spectral-ratio initialization for varying number of microphones.

distance is 0.3 m, makes it obvious. Optimization gets rid of artifacts in the separated sources.

The MAP estimator performance is considerably worse than the ML estimator. This is likely because the statistical independence criterion of the sources for calculation of the source covariance matrix is not satisfied for the quartet. Still, when the hyperparameter is zero, the OPS, TPS and IPS are better than those of MCWF. Performance further deteriorates as the source-microphone distance increases.

In Fig. 5.3, the M-GCC initialization also performs well. The results for the MLE estimator when $\sigma_\nu^2/\sigma_w^2 = 1$ and the source-microphone distance is 10 cm are exceptionally good. The variance in the scores for the MAP estimator is quite high. In Fig. 5.4, with BCI initialization, the APS is poor. Optimization hurts the TPS and IPS.

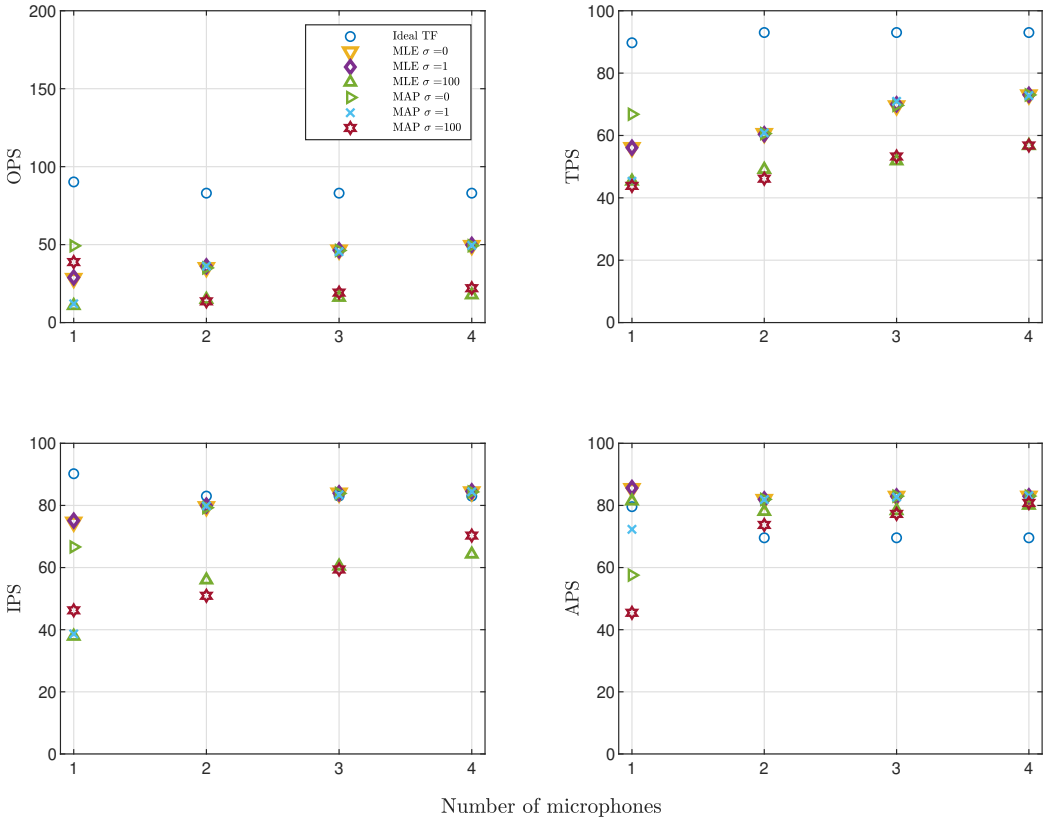


Figure 5.6: PEASS scores with M-GCC initialization for varying number of microphones.

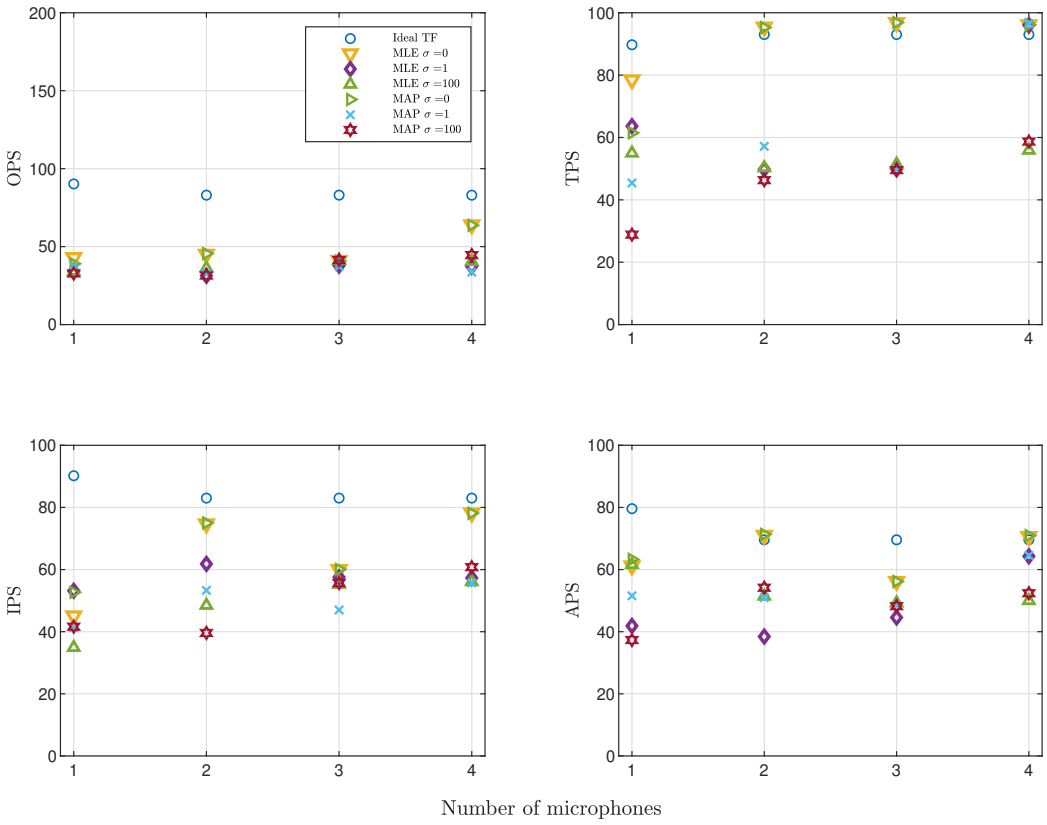


Figure 5.7: PEASS scores with BCI initialization for varying number of microphones.

5.4.2 Effect of number of microphones

We varied the number of microphones capturing each instrument ($N \in \{1, 2, 3, 4\}, M = 2$) to simulate an overdetermined system. The closest mic was kept on the same axis as the source at a distance of 20 cm, and the other microphones were distributed uniformly on an arc of angle 60° of the same radius.

The results with spectral ratio calibration are shown in Fig. 5.5. For the ML estimator, the OPS, TPS and IPS improve as the number of microphones is increased. This is expected in overdetermined systems. The APS remains constant. Again, the performance of the ML estimator with $\sigma_\nu^2/\sigma_w^2 \in \{0, 1\}$ are closely matched. For the MAP estimator and $\sigma_\nu^2/\sigma_w^2 = 0$, the APS score improves as number of mics increases, but the OPS, TPS and IPS decrease. Optimization with $\sigma_\nu^2/\sigma_w^2 = 1$ improves the APS significantly for the determined system ($N = 1$). As in the case of varying microphone distance, the ML estimator performs better than MAP.

In Fig. 5.6, an improvement in the scores for both ML and MAP estimators are observed as the number of microphones increases. The scores with $\sigma_\nu^2/\sigma_w^2 = 0$ and $\sigma_\nu^2/\sigma_w^2 = 1$ are very closely matched. In Fig. 5.7, the improvement in scores when $\sigma_\nu^2/\sigma_w^2 = 0$ is significant as the number of microphones increase. The optimization performance is subpar in this case, as in the case with varying source-microphone distance.

5.4.3 Effect of number of sources

For a determined system ($N = M$), we varied the number of sources, $M \in \{2, 3, 4\}$ to include all four instruments in the quartet. The four instruments — viola, violoncello and two violins were virtually placed at locations $(1.9, 2.5, 1.0)$, $(1.7, 2.8, 0.8)$, $(1.3, 2.8, 1.0)$, $(1.1, 2.5, 1.0)$ respectively. The virtual microphones were placed at a distance of 20 cm from the sources.

The mean perceptual evaluation scores of the separated sources with the spectral ratio calibration are shown in Fig. 5.8. The OPS, TPS and IPS of the ML estimator show a decreasing trend as the number of sources increases, which is expected. However, the APS score remains fairly constant, which is desired. MCWF outperforms the MLE in terms of OPS when $M \in \{3, 4\}$, but the rest of its scores are sub-par. Similarly, a decreasing trend in the scores is seen with the MAP estimator as the number of sources increases. The optimization improves the APS in all cases.

In Fig. 5.9, with the ML estimator and M-GCC calibration, the TPS and APS outperform those of the MCWF. Again, for the MLE estimator, the scores with $\sigma_\nu^2/\sigma_w^2 = 0$ and $\sigma_\nu^2/\sigma_w^2 = 1$ are very closely matched. With BCI initialization, the IPS shows a surprising improvement with increasing sources when $\sigma_\nu^2/\sigma_w^2 = 0$. The performance is especially good when the number of sources is 4.

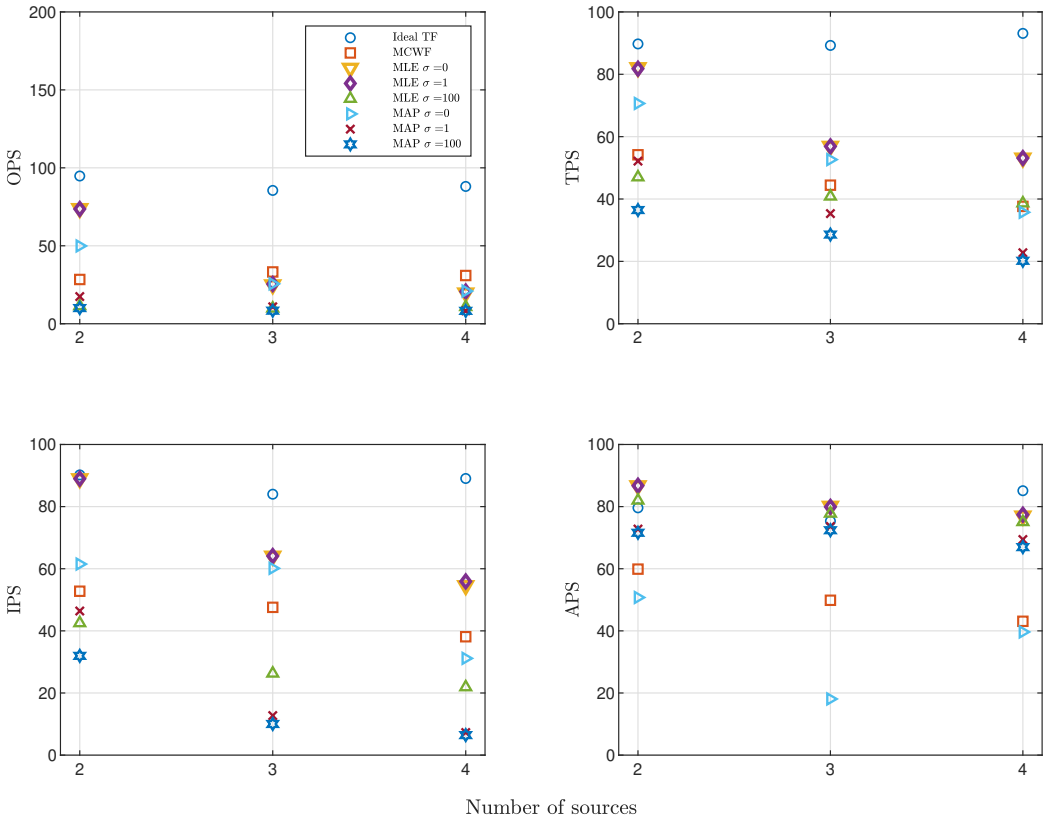


Figure 5.8: PEASS scores with spectral ratio initialization for varying number of sources.

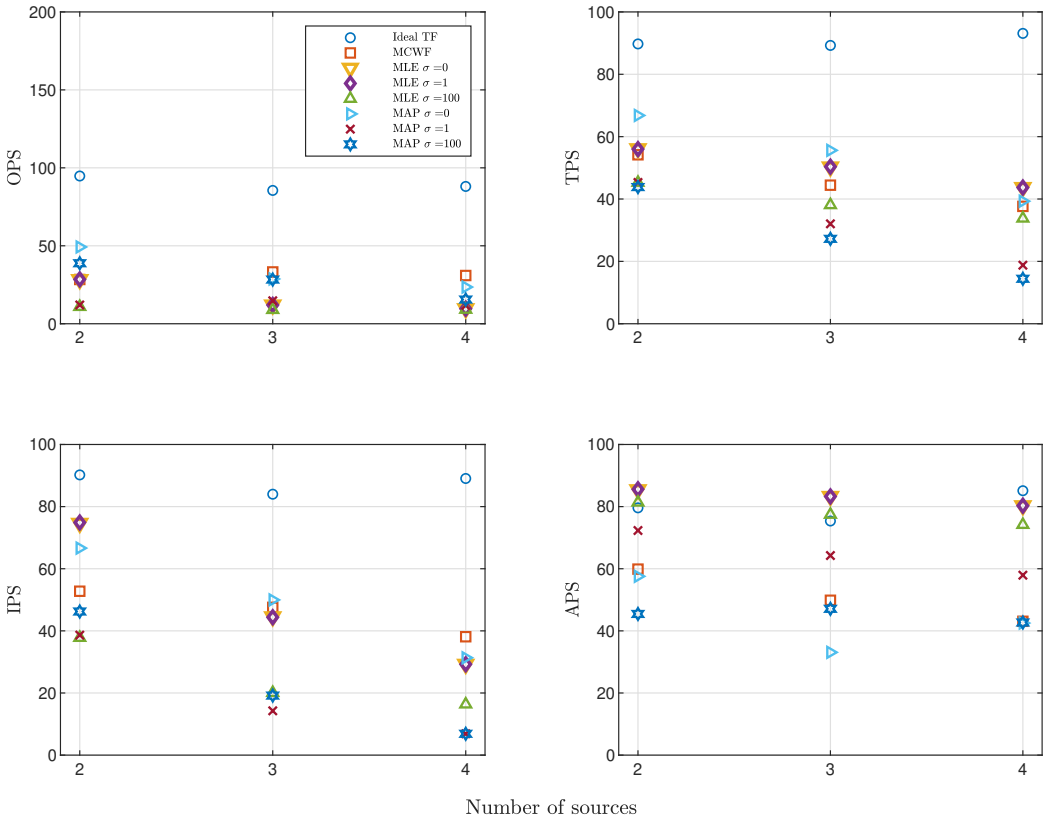


Figure 5.9: PEASS scores with M-GCC initialization for varying number of sources.

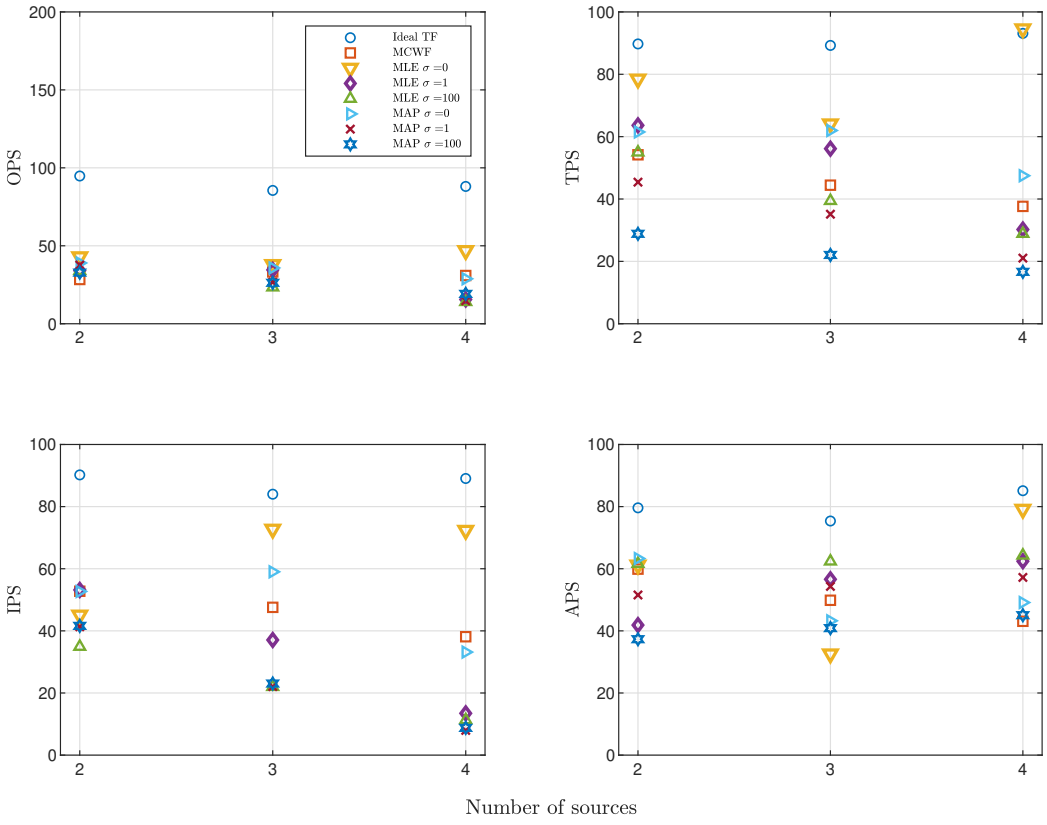


Figure 5.10: PEASS scores with BCI initialization for varying number of sources.

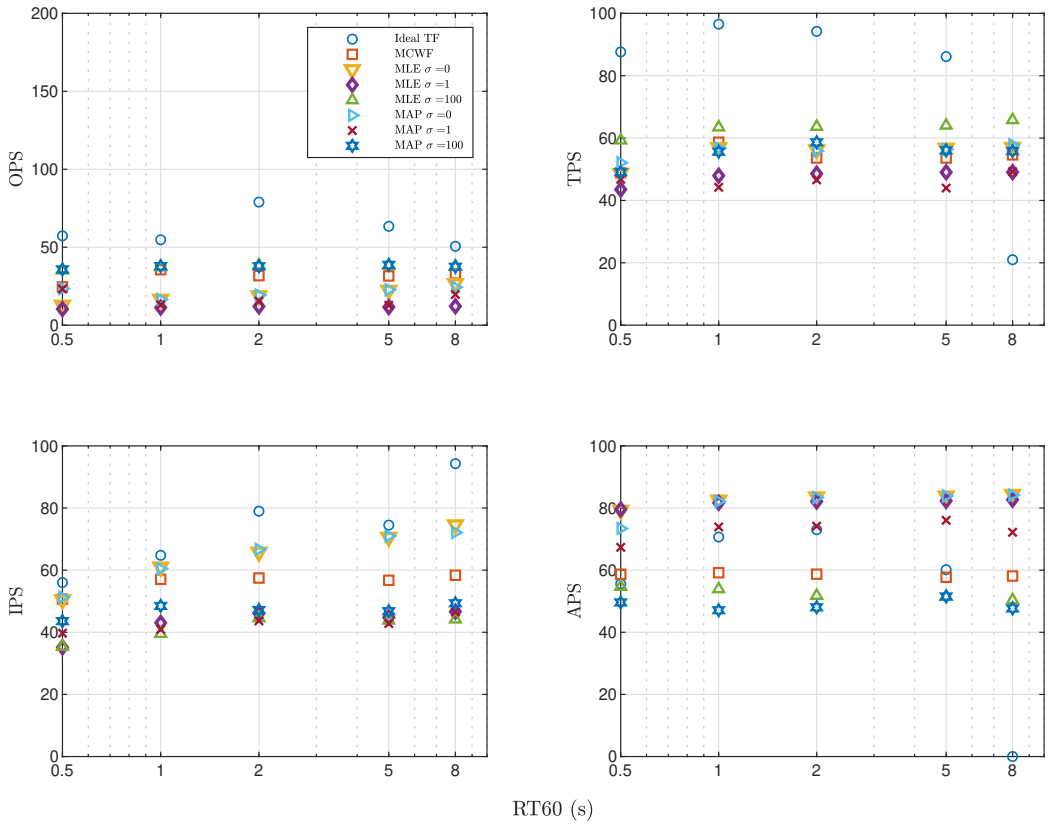


Figure 5.11: PEASS scores with spectral ratio initialization for varying volume and reverberation times.

Optimization hurts the MAP estimator scores.

5.4.4 Effect of reverberation time

We varied the reverberation time, RT_{60}, of our virtual studio logarithmically from 0.5 s to 8 s. This, in turn, affects the reflection coefficients of the wall materials according to Sabine's formula [79],

$$RT_{60} = \frac{24 \ln 10 V}{c \sum_{i=1}^{6} S_i (1 - \beta_i^2)} \tag{5.4}$$

where V is the volume of the room, c is the speed of sound in air and and β_i and S_i denote the reflection coefficient and the surface of the ith wall, respectively. The scores were observed to be independent of the RT_{60}. This is because for a fixed volume, a change in RT_{60} only affects the late reverberation tail, but the early reflections are unaffected.

However, when the volume of the room is changed along with the RT_{60}, the gains and times of arrival of the early reflections are impacted. The reflections are now more sparsely distributed in time. To replicate this scenario, we simulated cube-shaped rooms ranging from $20--200$ m^3 having RT_{60}s on a log scale, from $0.5--8$ s. A longer sequence (512 samples) of the RIR was convolved with the anechoic viola and violoncello recordings. The MAP, MLE and MCWF estimators were then applied to the data, with the spectral-ratio method used for initialization, since it was observed to give the best results. The results are shown in Fig. 5.11. The performance of all estimators remains fairly consistent, with a slight increase in scores as reverberation time increases. MAP/ML estimators with $\sigma_\nu^2/\sigma_w^2 = 100$ outperform MCWF in terms of TPS and IPS, whereas the estimators with $\sigma_\nu^2/\sigma_w^2 = \{0, 1\}$ outperform MCWF in terms of APS. A trade off between interference cancellation and quality preservation is inevitable, as in the other cases.

5.5 Takeaways

- MLE consistently performs better than MCWF in regards to the APS. It performs worse than MCWF with respect to the IPS as the number of sources increase.
- MAP performs worse than MLE in terms of the IPS. As discussed before, this is due to the statistical dependence of the quartet instruments.
- There is always a trade-off between IPS and APS. More interference cancellation is achieved at the cost of quality.
- The OPS, TPS and IPS of the ML estimator are closely matched for $\sigma_\nu^2/\sigma_w^2 = 0, 1$. This is only because the initialization methods work well for the image-source generated RIRs. In a real room, estimation with $\sigma_\nu^2/\sigma_w^2 = 0$ will underperform.
- Optimization improves the APS for all estimators at the cost of computation time.
- Performance worsens with increasing source-mic distance.
- Performance improves with increasing number of microphones capturing each source.
- Performance worsens with increasing number of sources.
- Performance is largely independent of room reverberation time.

Chapter 6

Example: Drum Bleed Suppression

In this chapter, the proposed methods for microphone bleed cancellation are tested on a far more complex real-life recording scenario. We record a drum kit and try to isolate each part of the kit. This particular scenario is challenging because drum spectra are very broad, spanning from the low-frequency kick drum to the high-frequency crash cymbals. The transients need to be reconstructed without any time-smearing. Moreover, all the different parts are located very close to each other. As a result, even with directional microphones, complete isolation is not possible. Some popular methods to reduce drum bleed include NMF based algorithms, such as the commercially available *Drumatom* [42] and [80]. Live drum separation using probabilistic spectral clustering has been explored in [81].

6.1 Studio recordings

A drum-kit was recorded at the CCRMA recording studio. Microphone placement guidelines for recording drum kits are suggested in [82]. Our recording setup is shown in Fig. 6.1. The microphones used to record the various drum parts are listed in Table 6.1. The overhead mics record the crash cymbals, as well as the room response.

Drum part	**Mic**	**Dynamic/Condensor**	**Directivity**
Floor tom, Rack tom	Rhode NT55	Condensor	Cardioid
Crash cymbals	AKG C414 pair	Condensor	Cardioid
Snare, Hi hat	Shure SM57	Dynamic	Cardioid
Kick	Sennheiser MD421	Dynamic	Cardioid

Table 6.1: Microphones used to record drum kit.

Figure 6.1: Microphone setup for recording drum kit.

6.2 Calibration

For estimation of the initial transfer function matrix, $\tilde{\boldsymbol{H}}(\omega)$, a single-input multi-output scenario is required. After the microphone gains were adjusted, the drummer was asked to strike each drum part separately as it was captured by all the mics. The process was repeated for all the drum parts.

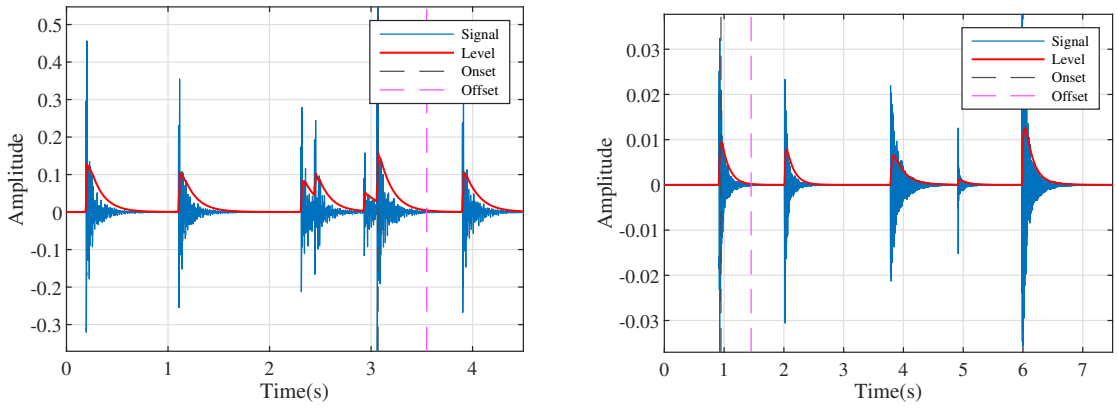
A single drum hit was extracted for each part of the kit from the signal captured by its closest microphone. First, a leaky-integrator was used to detect the signal level [83], with a fast attack time to detect sharp transients, and a slow decay time. The signal level, $\hat{\lambda}$, is

$$\begin{aligned} &\text{if } x(n) > \hat{\lambda}_n : \\ &\qquad \hat{\lambda}_{n+1} = \hat{\lambda}_n + (1 - e^{\frac{-1}{\tau_a f_s}})(|x(n)| - \hat{\lambda}_n) \\ &\text{else :} \\ &\qquad \hat{\lambda}_{n+1} = \hat{\lambda}_n + (1 - e^{\frac{-1}{\tau_r f_s}})(|x(n)| - \hat{\lambda}_n) \end{aligned} \tag{6.1}$$

where $x(n)$ is the current signal sample, $\tau_a = 10$ ms is the attack time constant, $\tau_r = 100$ ms is the release time constant and f_s is the sample rate. The detected level was smoothed by a 3 point moving average filter. The peak positions of the smoothed level were marked as the onset times. The offset times were marked as the time taken for the peak level to fall to $\exp\left(-\frac{1}{2}\right)$ of its starting value (2 time constants away from the peak). The onsets detected for the kick drum and snare drum are shown in Fig. 6.2. The same onset and offset times were used to extract the source from all the other microphones. Repeating the process for each drum part gave us 7 calibration files for each of the 7 different parts. These files were used for initial transfer function calculation, as well as for interference correlation matrix estimation for the MAP estimator.

6.3 Results

The MAP and ML estimators were initialized with spectral ratio, M-GCC and BCI calibration for three different hyperparameter values each, $\sigma_\nu^2/\sigma_w^2 = \{0, 1, 100\}$. This yielded a total of 9 different outcomes for each estimator. The optimization was run for a few iterations with a maximum function count of 10^4 on a 4 core Intel i5 CPU. The recorded and separated audio files are available at [84]. The proposed estimators with optimization are good at preserving the quality of the target signal without introducing any audible artifacts. They successfully reduce high frequency bleed from the cymbals and the hi-hat. Although the MCWF estimator does a better job at suppressing interference, it does so by compromising the quality of the separated signals. From a practical standpoint, studio engineers and producers almost always prefer separated signals with the least possible distortion.



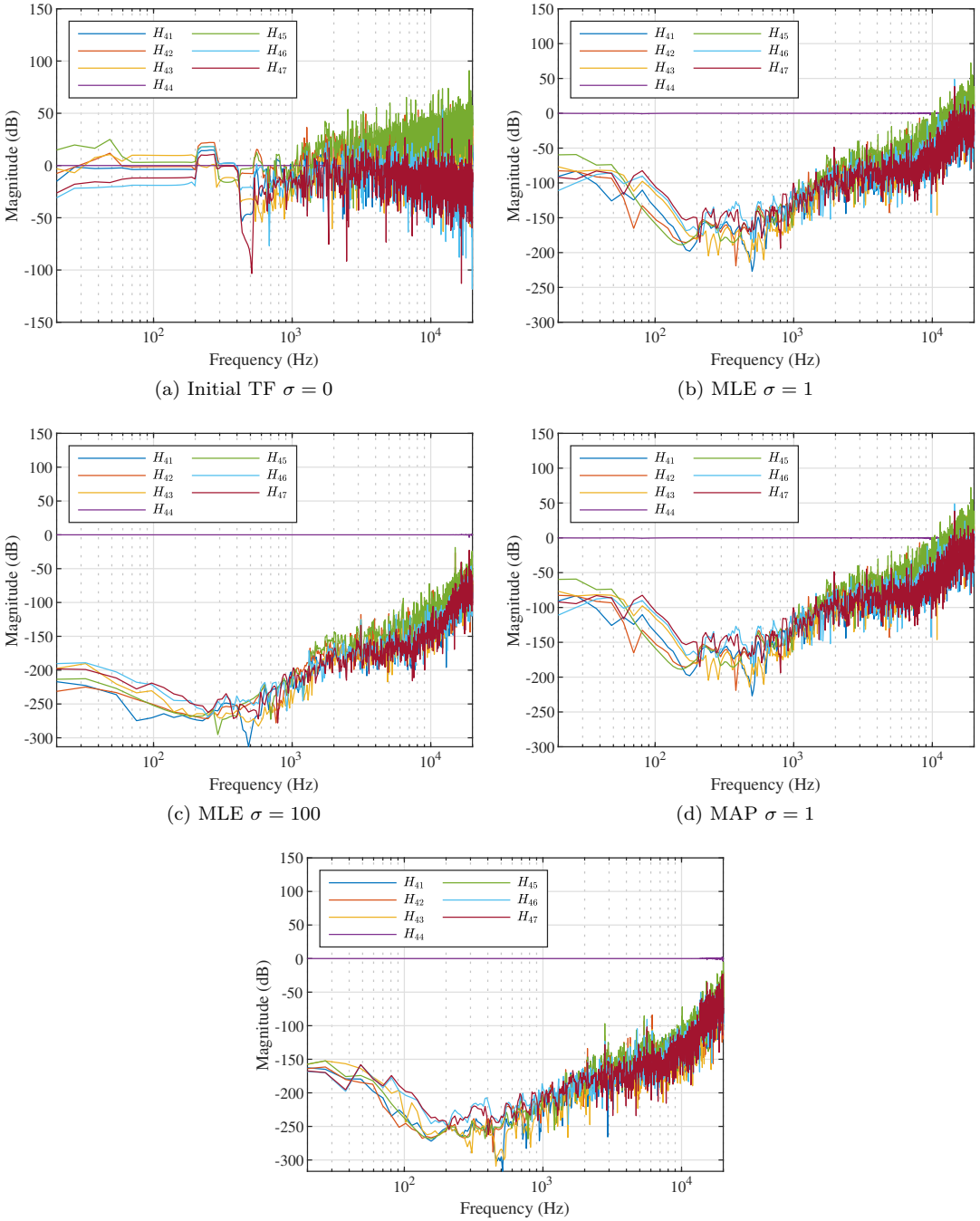
(a) Kick drum (b) Snare drum

Figure 6.2: Onset and offset detection.

Herein, we think that our proposed method outperforms the state of the art MCWF estimator.

The initial and optimized transfer functions for the rack tom, for various hyperparameter values, are shown in Figs. 6.3, 6.5, 6.7. For optimization with $\sigma_\nu^2/\sigma_w^2 = 1$, the transfer functions share a similar band-stop filter shape, with a stop band from $100 - 1000$ Hz, regardless of the method used for initialization. This shape is speculated to represent the room response, as it remains largely independent of the source and microphone positions. The MAP and MLE transfer functions for this hyperparameter value are nearly identical. For optimization with $\sigma_\nu^2/\sigma_w^2 = 100$, there is a dip in the lower frequency gains. The response from the source to the closest microphone remains flat post optimization.

The spectrograms of the recorded and separated signals for $\sigma_\nu^2/\sigma_w^2 = \{0, 1\}$, normalized to -14 dB LUFS [85], are shown in Figs. 6.4, 6.6, 6.8. The figures show that the high frequency bleed from the cymbals and hi-hat is suppressed. Mid-frequency bleed from the snare drum is mitigated, but still audible. This is because the spectrum of the rack tom overlaps significantly with that of the snare drum. No audible distortion is introduced in any the signals. The optimization results are consistent over the three different methods of initialization. Using the $\tilde{\boldsymbol{H}}$ directly for estimation of the sources produces severely distorted results. Hence, optimization is essential in any real-life MIMO recording scenario.



(a) Initial TF $\sigma = 0$

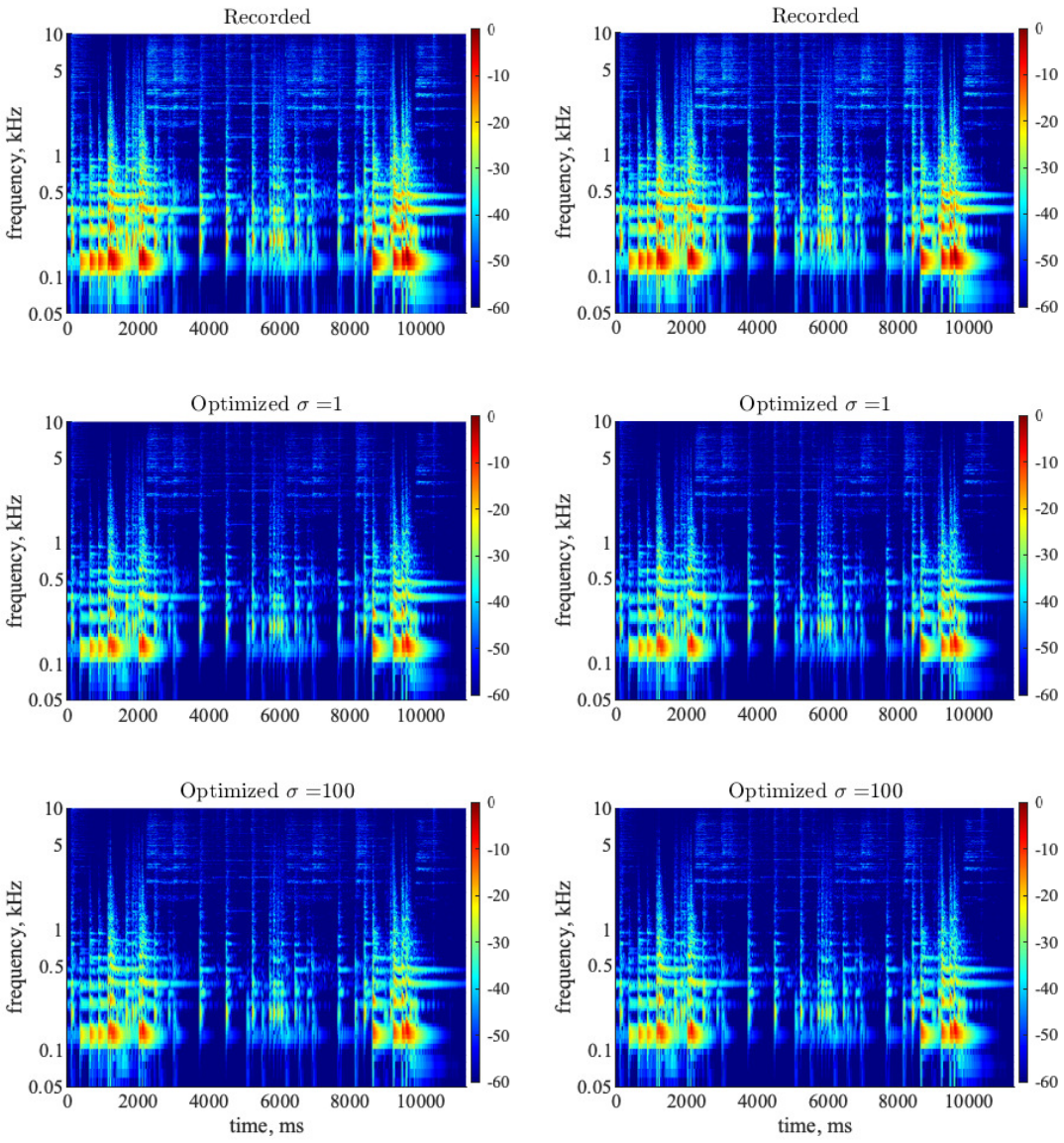
(b) MLE $\sigma = 1$

(c) MLE $\sigma = 100$

(d) MAP $\sigma = 1$

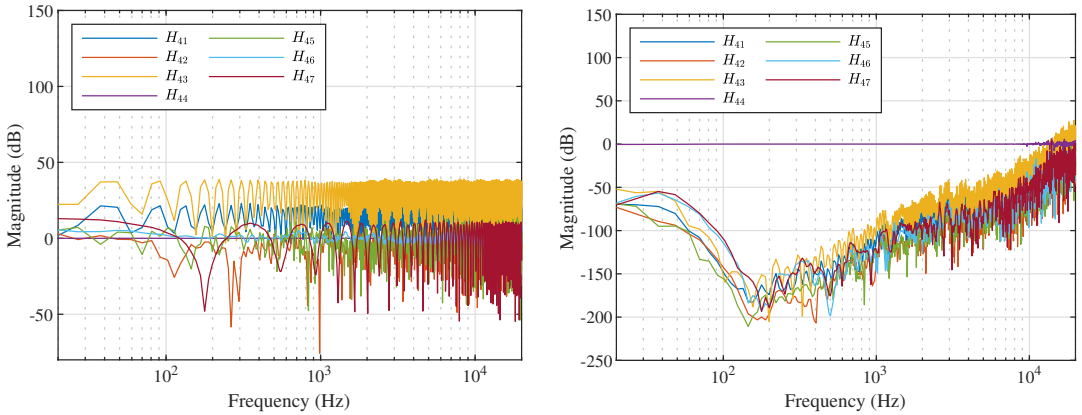
(e) MAP $\sigma = 100$

Figure 6.3: Initial and optimized transfer functions for the rack tom with spectral ratio calibration.



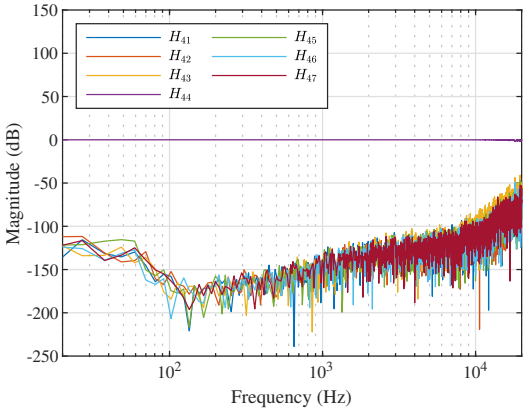
(a) MLE (b) MAP

Figure 6.4: Spectrograms for the recorded and separated rack tom with spectral ratio calibration.

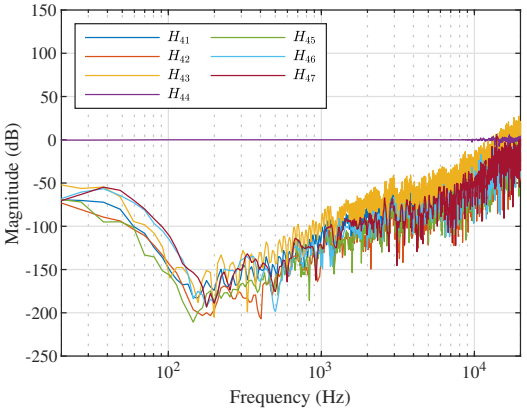


(a) Initial TF $\sigma = 0$

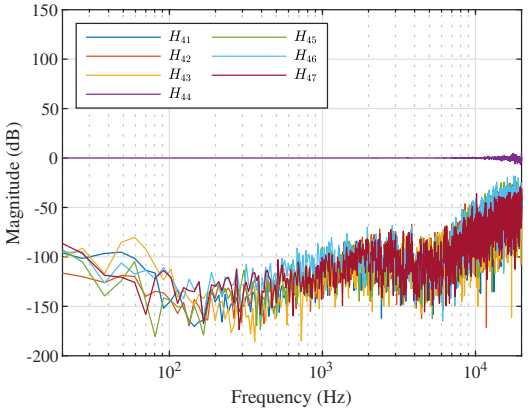
(b) MLE $\sigma = 1$



(c) MLE $\sigma = 100$

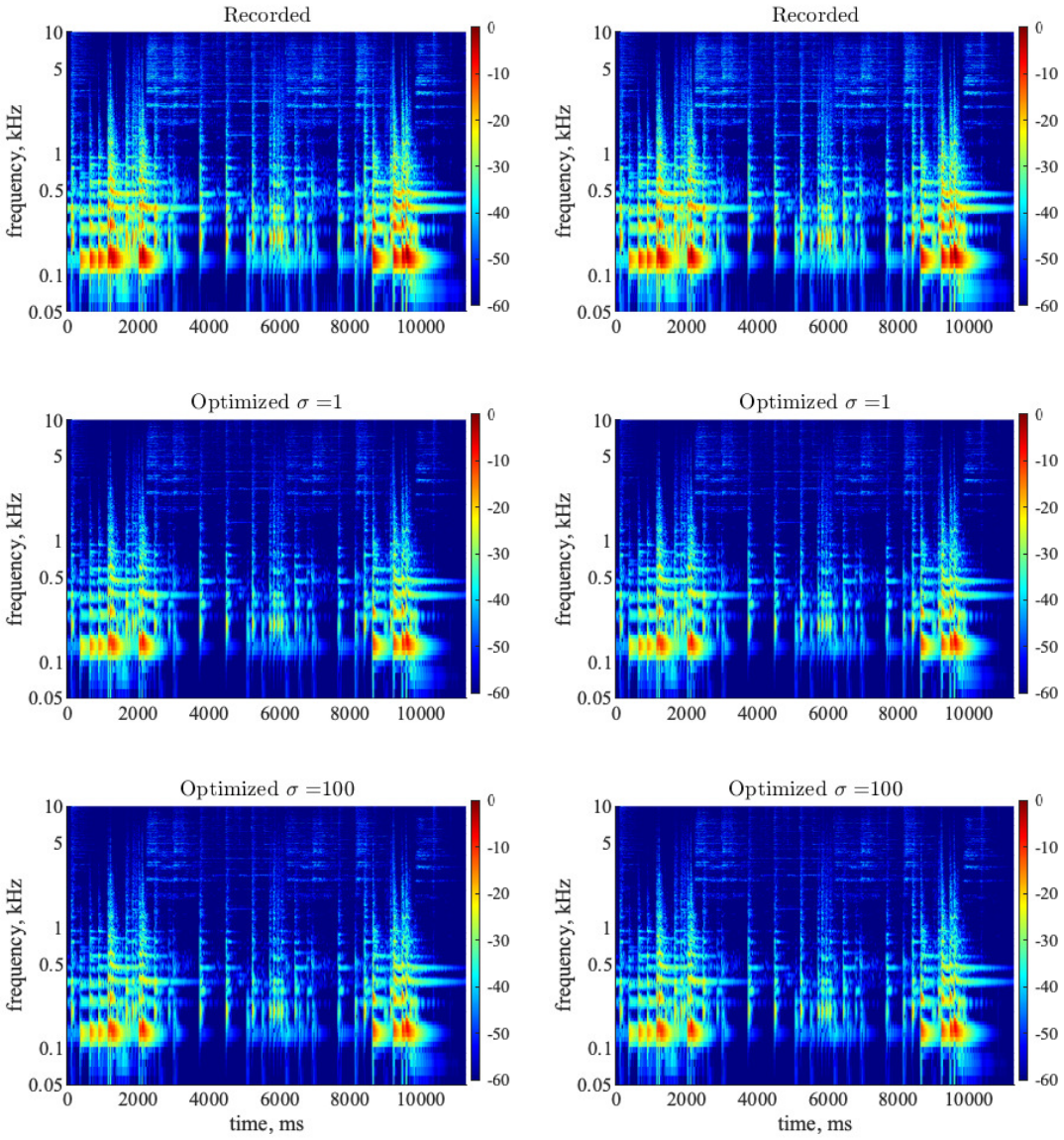


(d) MAP $\sigma = 1$



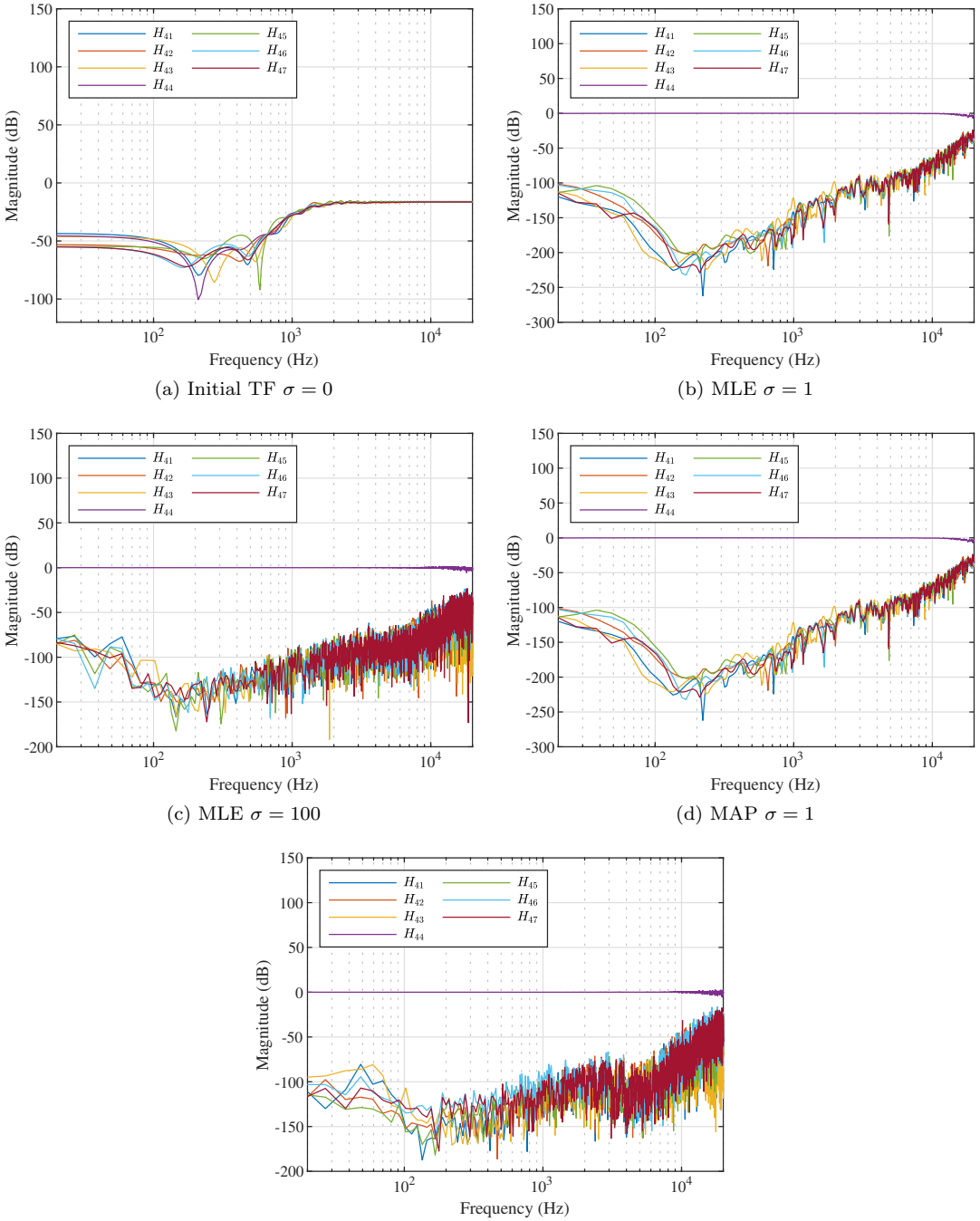
(e) MAP $\sigma = 100$

Figure 6.5: Initial and optimized transfer functions for the rack tom with M-GCC calibration.



(a) MLE (b) MAP

Figure 6.6: Spectrograms for the recorded and separated rack tom with M-GCC calibration.



(a) Initial TF $\sigma = 0$

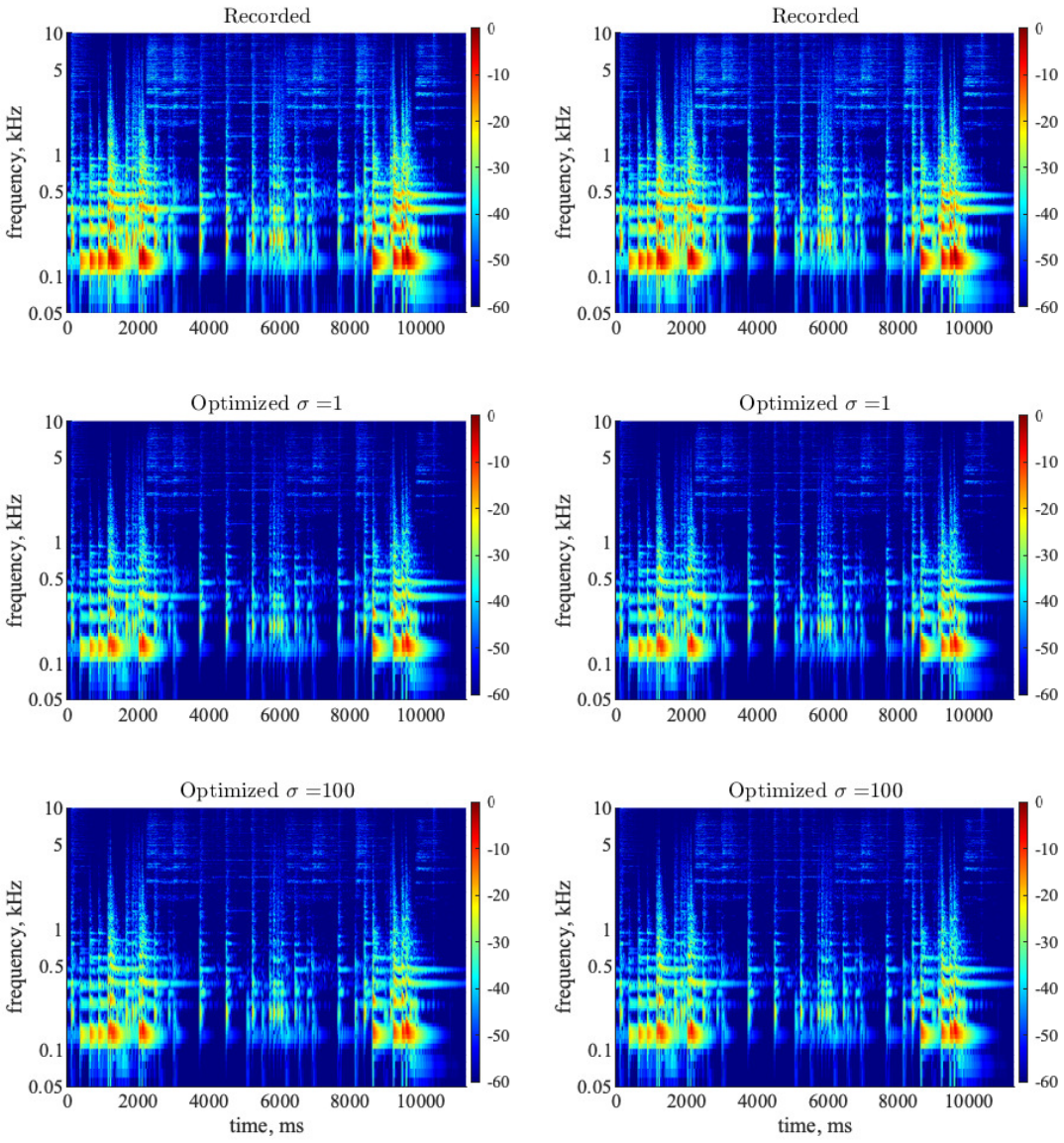
(b) MLE $\sigma = 1$

(c) MLE $\sigma = 100$

(d) MAP $\sigma = 1$

(e) MAP $\sigma = 100$

Figure 6.7: Initial and optimized transfer functions for the rack tom with BCI calibration.



(a) MLE (b) MAP

Figure 6.8: Spectrograms for the recorded and separated rack tom with BCI calibration.

6.3.1 Listening test

An online listening test was conducted over the *Qualtrics* survey platform (because of remote work due the the COVID-19 pandemic). Participants were given an instruction sheet that explained the

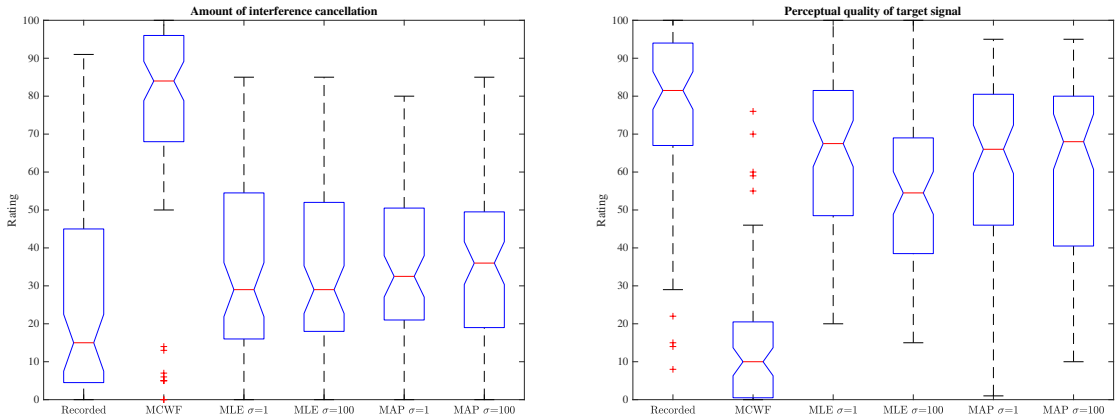


Figure 6.9: Listening test results.

goal of the experiment, and were asked to be wear over-the-ear headphones while taking the test in a quiet listening environment. They could adjust the volume to a level that felt comfortable for them.

Participants were presented with 4 sets of drum sounds – floor tom, kick, snare and rack tom. Each set contained an audio clip of the target signal, i.e., what the clean source itself would sound like without any cross-talk, (obtained from the calibration files). All files were adjusted to have a loudness of -14dB LUFS. Then, two questions were asked, one pertaining to the amount of cross-talk/interference reduction, and the other on the perceptual quality of the target signal. These questions correlate to the IPS and APS scores from the PEASS toolbox. Participants were asked to give a rating of 0 to 100, 0 being the lowest score. Six options were presented to the participant : the signal recorded by the closest microphone, MCWF, MLE with $\sigma_\nu^2/\sigma_w^2 = 0, 1$ and MAP with $\sigma_\nu^2/\sigma_w^2 = 0, 1$. The options were randomly shuffled for each question. For MLE and MAP, the spectral ratio method was used for calibration. Participants were given a training question, with instructions on how to rate.

At the end of the test, participants' responses were anonymized and saved, along with information about their age, gender, years of experience with audio engineering, hearing impairment, experience with a listening test and model of headphones used. Only a nickname was used as a unique identifier, and the participants were renumerated with gift cards once they contacted me via email with their nicknames and unique completion code. The experiment was approved by the university's Internal Review Board.

A total of 18 subjects participated in the test, 9 males and 9 females in an age range of $18 - 54$ years. No one reported any hearing impairment, and everyone took the test over headphones in a

quiet listening environment. All but 2 participants had former experience with sound recording and production. To analyze the results, a one-way ANOVA test was conducted with the different estimators as groups. The results had a p-value < 0.01, showing statistical significance. The box-plots of the ratings for all 4 test samples are shown in Fig. 6.9. As expected, the recorded signal (hidden reference) has the minimum amount of cross-talk cancellation and the best perceptual quality. MCWF performs the best in terms of bleed suppression but does so at the cost of compromising the signal's perceptual quality. Among the other estimators, MAP with $\sigma_\nu^2/\sigma_w^2 = 100$ performs the best in terms of both cross-talk cancellation and perceptual quality.

Chapter 7

Conclusions

7.1 Summary

This thesis has proposed novel DSP methods for suppressing microphone cross-talk or 'bleed' in multi-microphone ensemble recordings. While cross-talk cancellation methods based on Non-negative Matrix Factorization [12, 49], adaptive filtering [47] and Wiener filtering [4, 9] exist in the literature, we explored the simultaneous estimation of the relative transfer function matrix from each source to each microphone, and the interference-free, desired source signal, within an optimization framework. We showed that the cost functions we derive are statistically optimal, in a Bayesian and a non-Bayesian sense. Furthermore, we showed that the cost functions are convex, and the computations can be performed on a multi-core GPU or CPU. We proposed a *calibration* stage during sound-check where one instrument is active at a time, thereby reducing a complicated MIMO system to SIMO for initial transfer function estimation. The proposed methods were tested against a state-of-the-art multichannel Wiener filter algorithm on a simulated dataset of anechoic string quartet recordings placed in a virtual studio, as well as on drums recorded in the CCRMA studio. The major findings of this thesis are being prepared for publication as a journal manuscript [86]. The chapter-wise breakdown of this thesis is as follows:

- **Chapter 1** - We introduced the microphone bleed cancellation problem and why it is important to the audio engineering community. We discussed three broad categories of algorithms for audio source separation — blind source separation, beamforming and adaptive noise cancellation. Then, we discussed in detail some existing methods for close-microphone bleed cancellation. To conclude the chapter, we delineated the goals of the thesis.
- **Chapter 2** - We introduced the mathematical model and explained the physically motivated

reasons for our modeling choices. We proposed a *calibration* stage for the estimation of a noisy initial transfer function matrix. This is a novel addition to the existing literature which often does not make use of this readily available information while doing source separation. We discussed three different methods of estimating this matrix from a SIMO system — spectral-ratio, Modified GCC and NMFLMS. We tested the discussed methods to identify the early response of an RIR created with the image-source method. Both the spectral ratio and NMFLMS methods performed identically.

- **Chapter 3** - We derived the Maximum Likelihood estimator assuming the microphone and initial RTF to be normally distributed random vectors. This estimator simultaneously solves for the sources and the relative transfer function matrix. We discussed the role of a hyperparameter that decides how much we trust the initial estimate of the transfer functions. We proved that the ML cost function is convex by showing that the Hessian matrix is positive semi-definite. We also derived the Fisher information matrix, and showed how the CRB is determined by the number of microphones. We estimated the optimal source by finding the roots of the gradient of the cost function using a numerical root finder and showed that the computations can be vectorized and parallelized over the frequency bins for improvements in speed. The contributions of this chapter are all novel and have been published in [62].

- **Chapter 4** - We started this chapter by deriving the MMSE estimator, the multichannel Wiener filter (MWF), and discussed the Generalized Eigenvalue Decomposition (GEVD) based MWF. Then, we expanded the previously proposed ML estimator to have a Bayesian prior. This prior, the source covariance matrix, was derived using GEVD principals. Following the derivations from Chapter 3, we derived the Maximum Aposteriori Probability (MAP) estimator when the source vector is normally distributed with a given covariance matrix. The MAP estimator and the GEVD based source covariance matrix estimation are the main contributions of this chapter.

- **Chapter 5** - We placed a virtual string quartet in a shoebox room, and simulated microphone bleed by varying a number of parameters, such as, the source-microphone distance, the number of microphones recording each source, the number of sources and room reverberation time. We tested the proposed estimators against the MCWF for a range of hyperparameter values and report the overall (OPS), target-related (TPS), interference-related (IPS) and artifact-related perceptual scores (APS) using the PEASS toolbox. The ML estimator gave high values of IPS and APS.

- **Chapter 6** - We tested our algorithms on a drum kit recorded in the CCRMA studio. Each part of the kit was miked separately. There was significant bleed from the snare, hi-hat and cymbals. We explained how we calibrate the recordings to find the initial transfer functions, and presented the results of our experiments. We conducted a subjective listening test with

18 participants to evaluate the interference-related and artifact-related scores in the separated signals. The results showed that our method outperforms MCWF in retaining the perceptual quality of the processed signal, while reducing a few decibels of cross-talk.

7.2 Future work

- While the time-invariant transfer function assumption yields smooth results, in reality, there are small variations in the transfer function with time which needs to be captured by a slowly time-varying model.
- An improvement needs to be made to the amount of cross-talk cancellation while maintaining the same level of perceptual clarity in the separated sources.
- A faster C++ implementation of the algorithm that runs on a GPU can shed light on its performance capabilities. Currently, it is implemented in MATLAB, which is slow.
- The MAP estimator performs better on the drum-kit than the string quartet recordings. This is because the drums are broadband, and the covariance matrix of each drum part is uncorrelated with the others. Therefore, the source covariance matrix derivation in Section 4.2, which assumes that the source and interference matrices are independent, is correct. The quartet, on the other hand, has overlapping harmonic sources which do not satisfy this assumption. Here, a pre-processing stage consisting of a multi-pitch detector can improve the results if pitch-based harmonic constraints are imposed on the different sources' frequencies.
- The proposed methods can be tested on more ensemble recordings to check for performance robustness with different genres of music and instruments. Furthermore, the listening test should be re-done in a controlled environment.
- Exploration of cross-talk cancellation with ambisonics microphones is an interesting direction for future research. The increasing popularity of these microphones can lead to new research directions by exploiting the inherent orthogonality of the spherical harmonic coefficients to do direction-based source separation.

Appendix A

Proof that sum of convex functions is convex

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x_1, x_2 \in \mathbb{R}^n$ in its domain, and for all $0 \leq \alpha \leq 1$,

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2) \tag{A.1}$$

Geometrically, this means the line segment connecting $(x_1, f(x_1))$ to $(x_2, f(x_2))$ must sit above the graph of f. Let there be another convex function $g : \mathbb{R}^n \to \mathbb{R}$ for all $x_1, x_2 \in \mathbb{R}^n$ in its domain, we also have

$$g(\alpha x_1 + (1-\alpha)x_2) \leq \alpha g(x_1) + (1-\alpha)g(x_2) \tag{A.2}$$

Now, for the function, $h = f + g : \mathbb{R}^n \to \mathbb{R}$,

$$\begin{aligned} h(\alpha x_1 + (1-\alpha)x_2) &= f(\alpha x_1 + (1-\alpha)x_2) + g(\alpha x_1 + (1-\alpha)x_2) \\ &\leq \alpha f(x_1) + (1-\alpha)f(x_2) + \alpha g(x_1) + (1-\alpha)g(x_2) \\ &\leq \alpha(f(x_1) + g(x_1)) + (1-\alpha)(f(x_2) + g(x_2)) \\ &\leq \alpha h(x_1) + (1-\alpha)h(x_2) \end{aligned} \tag{A.3}$$

Therefore, we see that $h = f + g$ is also convex.

Appendix B

Proof that eigenvalues of positive-semidefinite matrices are non-negative

A matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, is positive-semidefinite if it is symmetric and if $\forall\ \boldsymbol{v} \in V, \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \geq 0$. Let λ be an eigenvalue of $\boldsymbol{A}$. Then, for eigenvector $\boldsymbol{u}$,

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u} \tag{B.1}$$

Pre-multiplying both sides by $\boldsymbol{u}^\top$, we get

$$\begin{aligned} \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} &= \boldsymbol{u}^\top \lambda \boldsymbol{u} \\ \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} &= \lambda \boldsymbol{u}^\top \boldsymbol{u} \end{aligned} \tag{B.2}$$

By definition, $\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} \geq 0$. Therfore, $\lambda \boldsymbol{u}^\top \boldsymbol{u} \geq 0$. The quadratic term, $\boldsymbol{u}^\top \boldsymbol{u}$ is non-negative by default. This implies that $\lambda \geq 0$ for the inequality to hold. We can prove this for all the eigenvalues of $\boldsymbol{A}$.

Appendix C

Publications at CCRMA

1. "Grouped Feedback Delay Networks for Modeling of Coupled Spaces" - **O. Das**, J.S Abel in J. Audio Eng. Soc. - JAES, vol 69 no. 7/8, pp.486-496, 2021.

2. "Room Impulse Response Interpolation from a Sparse Set of Measurements Using a Modal Architecture" - **O. Das**, P. Calamia, S.V.A Gari in Proc. of IEEE 46th Int. Conf. on Acoust., Speech, Signal Process., ICASSP 2021.

3. "Microphone Cross-talk Cancellation in Ensemble Recordings with Maximum Likelihood Estimation" - **O. Das**, J.O Smith, J.S Abel in Proc. of 150th Audio Eng. Soc. Conv., AES 2021.

4. "Delay Network Architectures for Room and Coupled Space Modeling" - **O. Das**, J.S Abel, E.K Canfield-Dafilou in Proc. of 23rd Int. Conf. Digit. Audio Effects, DAFx 2020.

5. "Improved Real-time Monophonic Pitch Tracking with the Extended Complex Kalman Filter" - **O. Das**, J.O. Smith, C. Chafe in J. Audio Eng. Soc. - JAES, vol 68 no.1/2, pp.78-86, 2020.

6. "On the Behavior of Delay Network Reverberator Modes" - **O. Das**, E.K Canfield-Dafilou, J.S Abel in Proc. of IEEE Workshop Appl. Signal Process. Audio, Acoust., WASPAA 2019.

7. "Improved Carillon Synthesis" - M.Rau, **O. Das**, E.K Canfield-Dafilou in Proc. of 22nd Int. Conf. Digit. Audio Effects, DAFx 2019.

8. "FAST MUSIC - An efficient implementation of the MUSIC algorithm for frequency estimation of approximately periodic signals" - **O. Das**, J.S Abel, J. O. Smith in Proc. of 21st Int. Conf. on Digit. Audio Effects, DAFx 2018.

9. "Analyzing and Classifying Guitarists from Rock Guitar Solo Tablature" - **O. Das**, B. Kaneshiro, T. Collins in Proc. of 15th Int. Conf. on Sound, Music Comput., SMC 2018.

10. "An Infinite Sustain Effect Pedal Designed for Live Guitar Performance" - M. Rau, **O. Das** in Proc. of 143rd Audio Eng. Soc. Conv., AES 2017.

11. "Real-time Pitch Tracking in Audio Signals with the Extended Complex Kalman Filter"- **O. Das**, J.O. Smith, C. Chafe in Proc. of 20th Int. Conf. Digit. Audio Effects, DAFx 2017.

Bibliography

[1] E. K. Kokkinis, *Blind signal processing methods for microphone leakage suppression in multi-channel audio applications*. PhD thesis, University of Patras, 2012.

[2] E. K. Kokkinis and J. Mourjopoulos, "Unmixing acoustic sources in real reverberant environments for close-microphone applications," *Journal of the Audio Engineering Society*, vol. 58, no. 11, pp. 907–922, 2010.

[3] A. Washburn, I. Román, M. Huberth, N. Gang, T. Dauer, W. Reid, C. Nanou, M. Wright, and T. Fujioka, "S7m: Symposium 7-shared musical experience as shaping and shaped by interpersonal dynamics," in *15th International Conference on Music Perception and Cognition 10th triennial conference of the European Society for the Cognitive Sciences of Music*, vol. 17, p. 68, 2018.

[4] E. K. Kokkinis, J. D. Reiss, and J. Mourjopoulos, "A Wiener filter approach to microphone leakage reduction in close-microphone applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 767–779, 2011.

[5] E. K. Kokkinis, E. Georganti, and J. Mourjopoulos, "Statistical properties of the close-microphone responses," in *Audio Engineering Society Convention 132*, Audio Engineering Society, 2012.

[6] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–529, IEEE, 2002.

[7] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[8] A. Singh, "Adaptive noise cancellation," *Dept. of Electronics & Communication, Netaji Subhas Institute of Technology*, vol. 1, 2001.

[9] P. Meyer, S. Elshamy, and T. Fingscheidt, "Multichannel speaker interference reduction using frequency domain adaptive filtering," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–17, 2020.

[10] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[11] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

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

[14] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

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

[16] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. iii–889, IEEE, 2004.

[17] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 4–pp, IEEE, 1997.

[18] H. Adel, M. Souad, A. Alaqeeli, and A. Hamid, "Beamforming techniques for multichannel audio signal separation," *arXiv preprint arXiv:1212.6080*, 2012.

[19] 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, 2015.

[20] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[21] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2009.

[22] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[23] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, p. 569270, 2003.

[24] P. Annibale, F. Antonacci, P. Bestagini, A. Brutti, A. Canclini, L. Cristoforetti, E. Habets, J. Filos, W. Kellermann, K. Kowalczyk, *et al.*, "The scenic project: Space-time audio processing for environment-aware acoustic sensing and rendering," in *131st AES Convention*, Audio Engineering Society, 2011.

[25] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the cocktail party," *IEEE journal of selected topics in Signal Processing*, vol. 9, no. 5, pp. 825–836, 2015.

[26] 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.

[27] B. Widrow, J. McCool, and M. Ball, "The complex LMS algorithm," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 719–720, 1975.

[28] M. B. Malik and M. Salman, "State-space least mean square," *Digital Signal Processing*, vol. 18, no. 3, pp. 334–345, 2008.

[29] M. H. Hayes, "9.4: Recursive least squares," *Statistical Digital Signal Processing and Modeling*, vol. 541, p. 445, 1996.

[30] M. B. Malik, "State-space RLS," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 6, pp. VI–645, IEEE, 2003.

[31] S. Doclo and M. Moonen, "On the output SNR of the speech-distortion weighted multichannel wiener filter," *IEEE Signal Processing Letters*, vol. 12, no. 12, pp. 809–811, 2005.

[32] R. E. Kalman *et al.*, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[33] O. Das and C. Chafe, "Real-time pitch tracking in audio signals with the extended complex Kalman filter," in *International Conference on Digital Audio Effects, DAFx*, vol. 20, 2017.

[34] O. Das, J. O. Smith III, and C. Chafe, "Improved real-time monophonic pitch tracking with the extended complex Kalman filter," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 78–86, 2020.

[35] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 177–180, IEEE, 1987.

[36] O. Das, B. Goswami, and R. Ghosh, "Application of the tuned Kalman filter in speech enhancement," in *IEEE First International Conference on Control, Measurement and Instrumentation (CMI)*, vol. 1, pp. 62–66, 2016.

[37] C. Paleologu, J. Benesty, and S. Ciochină, "Study of the general Kalman filter for echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1539–1549, 2013.

[38] S. Braun and E. A. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.

[39] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 405–413, 1993.

[40] B. Owsinski, *The Mixing Engineer's Handbook*. Nelson Education, 2013.

[41] "Izotope RX-7 debleed." `https://www.izotope.com/en/products/repair-and-edit/rx.html`. Online; Accessed: 2020-05-34.

[42] E. Kokkinis, A. Tsilfidis, T. Kostis, and K. Karamitas, "A new dsp tool for drum leakage suppression," in *Audio Engineering Society Convention 135*, Audio Engineering Society, 2013.

[43] "Wilkinson debleeder." `https://wilkinsonaudio.com/products/debleeder`. Online; Accessed: 2020-05-34.

[44] P. Meyer, S. Elshamy, and T. Fingscheidt, "A multichannel Kalman-based Wiener filter approach for speaker interference reduction in meetings," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 451–455, IEEE, 2020.

[45] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[46] T. Prätzlich, R. M. Bittner, A. Liutkus, and M. Müller, "Kernel additive modeling for interference reduction in multi-channel music recordings," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 584–588, IEEE, 2015.

[47] A. Clifford, J. D. Reiss, *et al.*, "Microphone interference reduction in live sound," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[48] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[49] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodríguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, p. 184, 2013.

[50] C. Uhle and J. Reiss, "Determined source separation for microphone recordings using iir filters," in *129th AES Convention*, Audio Engineering Society, 2010.

[51] B. Kwon, Y. Park, and Y.-S. Park, "Analysis of the GCC-PHAT technique for multiple sources," in *International Conference on Control, Automation and Systems (ICCAS)*, pp. 2070–2073, IEEE, 2010.

[52] A. Clifford and J. Reiss, "Calculating time delays of multiple active sources in live sound," in *129th AES Convention*, Audio Engineering Society, 2010.

[53] B. Chen and A. P. Petropulu, "Frequency domain blind MIMO system identification based on second-and higher order statistics," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1677–1688, 2001.

[54] J. Liang and Z. Ding, "Blind MIMO system identification based on cumulant subspace decomposition," *IEEE Transactions on Signal Processing*, vol. 51, no. 6, pp. 1457–1468, 2003.

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

[56] Y. A. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Processing*, vol. 82, no. 8, pp. 1127–1138, 2002.

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

[58] H. He, J. Chen, J. Benesty, and T. Yang, "Noise robust frequency-domain adaptive blind multichannel identification with lp-norm constraint," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1608–1619, 2018.

[59] B. Jo and P. Calamia, "Robust blind multichannel identification based on a phase constraint and different lp-norm constraints," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1966–1970, IEEE, 2021.

[60] E. A. Habets and P. A. Naylor, "Adaptive blind system identification and equalization toolbox for MATLAB." `https://github.com/patrickanaylor/BSIE_toolbox`, 2011.

[61] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[62] O. Das, J. O. Smith, and J. S. Abel, "Microphone cross-talk cancellation in studio ensemble recordings with maximum likelihood estimation," in *150th AES convention*, Audio Engineering Society, 2021.

[63] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[64] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Prentice Hall PTR, 1993.

[65] J. S. Abel, "A bound on mean-square-estimate error," *IEEE Transactions on Information Theory*, vol. 39, no. 5, pp. 1675–1680, 1993.

[66] P. Pakrooh, L. L. Scharf, A. Pezeshki, and Y. Chi, "Analysis of Fisher information and the Cramér-Rao bound for nonlinear parameter estimation after compressed sensing," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, pp. 6630–6634, IEEE, 2013.

[67] L. L. Scharf and L. T. McWhorter, "Geometry of the Cramer-Rao bound," *Signal Processing*, vol. 31, no. 3, pp. 301–311, 1993.

[68] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[69] T. J. Ypma, "Historical development of the Newton–Raphson method," *SIAM review*, vol. 37, no. 4, pp. 531–551, 1995.

[70] A. Hassani, A. Bertrand, and M. Moonen, "GEVD-based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2557–2572, 2015.

[71] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[72] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech enhancement*, pp. 199–228, Springer, 2005.

[73] C. Böhm, D. Ackermann, and S. Weinzierl, "A multi-channel anechoic orchestra recording of Beethoven's symphony no. 8 op. 93," *Journal of the Audio Engineering Society*, vol. 68, no. 12, pp. 977–984, 2021.

[74] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[75] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[76] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective"signal processing in the auditory system. i. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[77] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[78] O. Das, "Supplemental materials : Microphone cross-talk cancellation in ensemble recordings with statistical estimation." `https://ccrma.stanford.edu/~orchi/Mic-bleed/sound_examples_quartet.html`, 2021.

[79] W. C. Sabine, *Collected papers on acoustics*. Los Alto, CA: Peninsula Publishing, 1993.

[80] J. Janer, R. Marxer, and K. Arimoto, "Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 281–284, IEEE, 2012.

[81] E. Battenberg, V. Huang, and D. Wessel, "Live drum separation using probabilistic spectral clustering based on the Itakura-Saito divergence," in *AES 45th Conference on Time-Frequency Processing in Audio*, Audio Engineering Society, 2012.

[82] Shure, "Microphone techniques for live sound reinforcement."

[83] U. Zolzer, *DAFX: Digital Audio Effects, 2nd Edition*. John Wiley & Sons Ltd, Hamburg, Germany, 2011.

[84] O. Das, "Supplemental materials : Microphone cross-talk cancellation in ensemble recordings with statistical estimation." `https://ccrma.stanford.edu/~orchi/Mic-bleed/sound_examples_drums.html`, 2021.

[85] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES convention*, Audio Engineering Society, 2021.

[86] O. Das, J. O. Smith, and J. S. Abel, "Close microphone cross-talk cancellation in multichannel recordings with statistical estimation." In preparation.