Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Musical Blind Source Separation Using Ambisonics

Henry Hasti

Supervisor: Andrés Pérez López

Co-Supervisor: Xavier Serra

August 2020

Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Musical Blind Source Separation Using Ambisonics

Henry Hasti

Supervisor: Andrés Pérez López

Co-Supervisor: Xavier Serra

August 2020

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Approach and Structure of the Report 2

2 State of the Art **3**

2.1 Microphone Array Signals and Conventions 3

2.2 Ambisonics and Parametric Values Analysis 6

2.3 Musical Blind Source Separation 10

2.3.1 Spatial Filtering 10

2.3.2 Nonnegative Matrix Factorization (NMF) 13

2.3.3 Filtering in the Ambisonics Domain 17

2.3.4 Neural Network Source Separation 18

2.4 Direction of Arrival (DOA) Estimation 19

2.4.1 Parametric Value Estimation 19

2.4.2 MUSIC Estimation 19

3 Methods **22**

3.1 Proposed Method 22

3.1.1 Analyze Parametric Values 23

3.1.2 Diffuse Masking 24

3.1.3 Azimuth DOA Estimation 24

3.1.4 Elevation DOA Estimation 24

3.1.5 Beamform 24

3.2 Existing Approaches Tested . . . 25
3.2.1 NMF . . . 25
3.2.2 Ambisonic Domain Filtering . . . 25
3.2.3 MUSIC DOA Estimation . . . 25
3.3 Data Generation . . . 25
3.3.1 Two Source Case . . . 26
3.3.2 Four Source Case . . . 27

4 Experiments **29**
4.1 DOA Estimation . . . 29
4.2 Optimized Beamformer Shape . . . 29
4.3 Separation Performance . . . 29
4.4 Computation Time Performance . . . 32

5 Results and Discussion **33**
5.1 DOA Estimation . . . 33
5.2 Optimized Beamformer Shape . . . 37
5.3 Separation Performance . . . 42
5.3.1 Two Source Case . . . 42
5.3.2 Four Source Case . . . 52
5.4 Computation Time Performance . . . 61

6 Conclusions **62**
6.1 Conclusions . . . 62
6.2 This Thesis' Contributions . . . 64
6.3 Future work . . . 65

List of Figures **66**

List of Tables **71**

Abstract

This thesis deals with the musical blind source separation problem: given multiple instrument tracks recorded together, how can each be isolated from the others given no additional knowledge about the instrument locations or sounds? There are many advantages to solving the problem: live performances can convey the energy of the performers, and the sound of the room can improve the perceptual quality of the recording, while separating the instrument tracks afterwards allows for more precise equalization and mixing to improve the recording. This thesis proposes a novel approach to the problem using an ambisonic microphone array. The directionality of the microphones in the array provides information about the location of each source and allows for the implementation of a direction of arrival estimator that calculates the position of each source based on the directions that receive the most non-reverberant acoustic energy. Given the directions of arrival (DOAs), the method performs spatial filtering to virtually steer the microphone array in the desired directions. This approach is compared with a classic DOA estimator, MUSIC, a classic non-negative matrix factorization (NMF) approach, and a state of the art ambisonic domain filtering approach. The proposed method outperforms NMF and MUSIC in nearly every tested configuration, and can outperform the ambisonic filtering approach in certain high-reverberation cases. The proposed method is significantly more computationally efficient than the comparison methods, working much faster while introducing less algorithmic noise to the separated tracks. The results raise the question of how to best balance performance and computational complexity in different use cases.

Keywords: Musical Blind Source Separation; Ambisonics; Spatial Filtering; Direction of Arrival Estimation

Chapter 1

Introduction

1.1 Motivation

Humans, whether we realize it or not, perform source separation constantly. While walking down the street, we recognize that the soundscape our ears perceive can be correctly segmented according to the sources generating each sound: the car passing, the children playing, the speech from a friend. The same occurs when listening to music: provided we are familiar with the constituent instruments we recognize that, for example, the guitar, bass, drums, and vocals are distinct instruments. This effect is amplified when we are in the middle of a live performance; we can group the instruments according both to our instincts and to the directions from which we perceive sound to come.

In addition to this localization effect, live music has many advantages, both for the audience and the band. Live performances can convey to the audience more effectively the energy and acoustic context of a performance, and performers can more easily sync up, resulting in a more dynamic performance. However, a major downside of live performances and recordings is the processing afterwards - when the instruments are recorded separately as in a studio, it is easy to mix each track separately to optimize equalization and apply other filters. In live performances, generally, all tracks contain all instruments, making this impossible. We have thus

found the musical blind source separation problem: Given no outside information (about where the sources are located and what sounds they emit, for instance), how can we take a live recorded track and return a separate track for each instrument in the mix?

1.2 Approach and Structure of the Report

Many approaches to the blind musical source separation problem have been proposed. Section 2 of this report details some of the more prominent of these approaches, which use spatial information generated from a microphone array, the non-negative matrix factorization method, filtering in the ambisonics domain, and neural network approaches. Section 3 describes the novel solution proposed as part of this thesis, which combines aspects of the spatial separation methods to provide a highly time-efficient solution. The same section describes the configurations of the compared methods and of the testing data. Section 4 describes the experiments carried out to test and compare the different methods, and 5 presents and discusses the outcomes of these experiments. Finally, section 6 concludes the report and proposes future work.

Chapter 2

State of the Art

The cutting edge of the theories relevant to this thesis' work are presented here:

- Microphone array signals and conventions
- Ambisonics and parametric values analysis
- (Musical) blind source separation
- Direction of arrival estimation

2.1 Microphone Array Signals and Conventions

Let us start by considering the simplest possible recording setup: a single source playing in an anechoic (reverberation-free) room that gets recorded by a single receiver. (The lack of reverberation is termed free-field). We assume the source and receiver are infinitely small so that their presence cannot alter the recording. In this setup, if the source outputs sound signal $s(t)$, a function of time, as a spherical wave, the only changes to the received signal $x(t)$ will be due to the attenuation and delay of a traveling spherical sound wave, themselves functions of the speed of sound c (nominally 343 m/s), the distance from source to receiver d, and the physics

of spherical waves. At the receiver:

$$x(t) = \frac{1}{\sqrt{4\pi}d} s(t - \frac{d}{c}) \tag{2.1}$$

This is naturally an overly simplified model, as it ignores the effects of reverberations on the sound's path to the receiver, and is limited to a single source and single receiver. The final shortcoming, the assumption of infinitesimal sources and receivers, is beyond this thesis' scope.

To make our model more realistic, let us first maintain our single source and single receiver, but move to a reverberant environment. To simplify our calculations, we will condense the math and physics contained in (2.1) to a causal, linear time-invariant impulse response, $h(t)$, that characterizes the room acoustic response. The impulse response captures what the receiver would record if the source emitted an impulse δ, an infinitely short signal with unit amplitude (since such a compressed signal contains all frequencies, it can also effectively characterize what the receiver would record if the source emitted any given frequency. This is referred to as the frequency response). In the free-field case discussed above, the impulse response would be

$$h(t) = \frac{1}{\sqrt{4\pi}d} \delta(t - \frac{d}{c}) \tag{2.2}$$

Equation (2.2) captures the attenuation in the impulse's scaled amplitude and the sound propagation delay in its time shift. Now, moving to a reverberant room, our impulse response will contain many impulses because of the multiple paths sound can take from source to receiver: every reflection off a wall or other object will contribute another scaled and time-delayed impulse to the response. Figure 1 displays a typical room impulse response, which is made up of the direct sound, early reflections, and reverberation. The direct sound travels from the source to the receiver without making any reflections, necessarily arriving first and with the least attenuation. The early echoes arrive after making a minimal number of reflections, and can still be discerned as corresponding to the direct sound. Finally, the reverberation arrives after making many reflections, and is not easily recognized as correlating to

the original sound. Given an impulse response (either recorded in a live room or

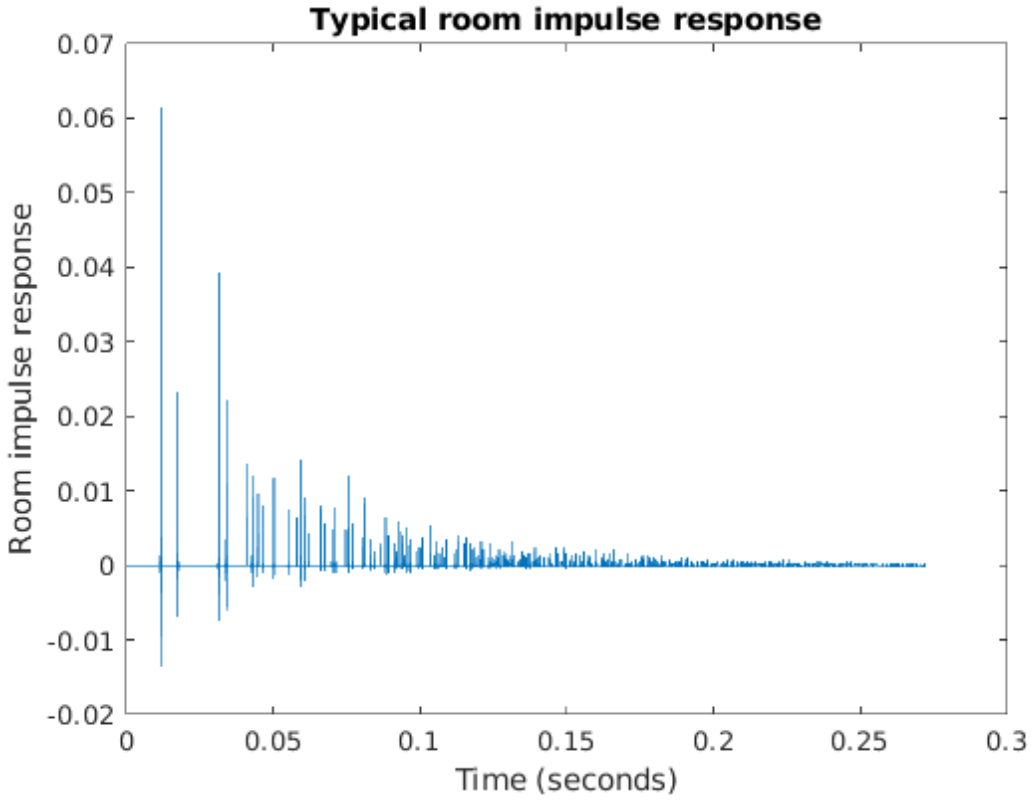

Figure 1: Typical room impulse response (RIR). The direct sound (with a height just over 0.06) reaches the receiver first, followed by the early echoes (roughly from 0.01 to 0.1 seconds), and reverberance (roughly 0.1 seconds to the end of the impulse response).

simulated through software), we can calculate the received signal $x(t)$ by breaking the source signal $s(t)$ into many impulses and adding the responses together, which is equivalent to convolution, expressed with $*$.

$$x(t) = s(t) * h(t) \tag{2.3}$$

Alternately, equation (2.3) can be viewed as taking the source signal $s(t)$ and adding it to a scaled, time-delayed version for every reflection along the sound's path.

Let us now complete our realistic model by adding more sources and receivers. Because of the linearity of sound propagation and convolution, we need only calculate a sum of convolved signals; the overall theory is unchanged. We define a vector

containing all of the I received signals (such that each column contains one received signal),

$$\mathbf{x}(t) = [x_1, x_2, ..., x_I] \tag{2.4}$$

and a vector containing all J of the emitted signals

$$\mathbf{s}(t) = [s_1, s_2, ..., s_J] \tag{2.5}$$

Since we will have an impulse response mapping every source to every receiver, we store the impulse responses in an I by J matrix

$$\mathbf{h}(t) = \begin{bmatrix} h_{1,1}(t) & ... & h_{1,J}(t) \\ ... & h_{i,j}(t) & ... \\ h_{I,1}(t) & ... & h_{I,J}(t) \end{bmatrix} \tag{2.6}$$

such that $h_{i,j}(t)$ is the impulse response from source j to receiver i. The signal at a given receiver i is then the sum of the contributions from every source, each contribution calculated by convolving the source with its impulse response:

$$x_i = \sum_j s_j * h_{i,j} \tag{2.7}$$

More detailed information about array signals and conventions is available in [1].

2.2 Ambisonics and Parametric Values Analysis

Ambisonics was invented as an efficient way to record, store, and reproduce 3-dimensional sound spaces, and has gained popularity recently for its use in virtual reality applications. Its chief advantage is storing spatial sound without regard to what recording and playback setups are used, allowing more flexibility in sound recreation. Additionally, it has been shown that converting signals to the ambisonic domain can result in more effective source separation [2]. Because of the flexibility in recording setups that ambisonics yields, we say that our array is made up of virtual

microphones that represent ambisonics up to a specified order. This thesis uses ambisonics up to first order, corresponding to the four virtual microphone channels shown in Figure 2.

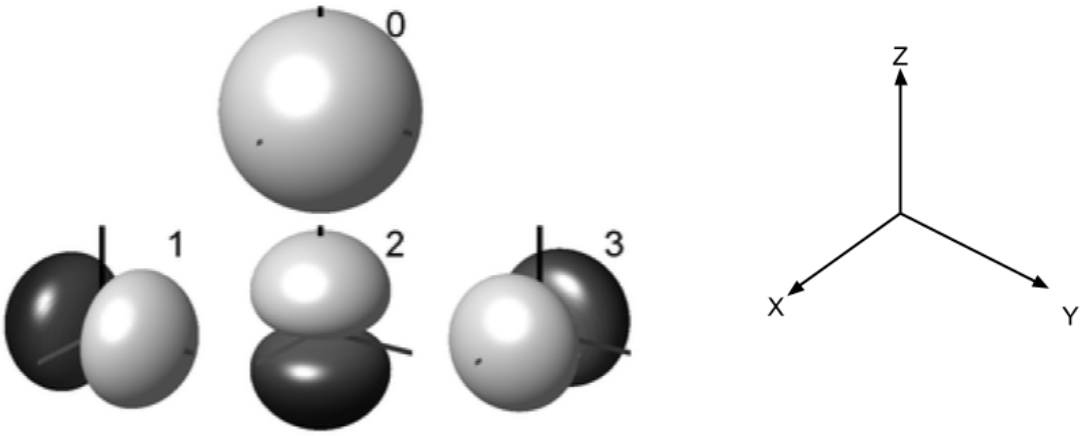

Figure 2: Virtual microphones in first-order ambisonic recording. These correspond to one omni-directional channel W (0), and three oriented channels X (3), Y (1), and Z (2). Lighter lobes indicate that recorded signals are in phase, while darker lobes indicate counterphase. Image: "File:Spherical Harmonics deg5.png" by Dr Franz Zotter <zotter@iem.at> is licensed under CC BY-SA 3.0 . With alterations.

The first channel (or zeroth order) is omni-directional, meaning that sound approaching it from any direction of arrival (DOA) will be equally amplified. The second through fourth channels (which complete the first order) are bidirectionally oriented, meaning that they amplify sounds approaching from along their preferred axes and attenuate sounds from other directions. These microphones are oriented to correspond to the 3-D cartesian coordinate system, so that one channel, X, amplifies sounds arriving along the x-axis, and likewise for Y and Z. If we imagine a listener sitting at the origin, the x-axis points in the direction the listener faces, the y-axis to the listener's left, and the z-axis up. In the figure, the sphere's distance from the origin indicates how much amplification that direction will receive. We assume these virtual microphones to be coincident and of negligible size, so that they correspond to spherical harmonic functions.

Following the process of [3], let us consider how each microphone in the array will respond to a sound reaching it, termed the pickup pattern. Relative to the listener-centric coordinate system we described above, we define ϕ as the azimuth angle (0

along the x-axis and increasing as it moves counter-clockwise when viewed from the positive z-axis) and θ as the elevation (0 in the x-y plane, increasing as it moves toward the positive z axis, and decreasing as it moves away). ϕ and θ thus form the angular basis for a spherical coordinate system. The virtual microphones' coincidence allows us to assume one signal, $x(t)$, arriving equally at every microphone. This thesis uses N3D normalization, which [4] describes as one method to normalize the spherical harmonic functions underlying the ambisonic calculations. The signals, as functions of time, received at every microphone are:

$$W(t) = \int_0^{2\pi} \int_0^{\pi} x(t, \phi, \theta) d\theta d\phi \tag{2.8}$$

$$X(t) = \sqrt{3} \int_0^{2\pi} \int_0^{\pi} x(t, \phi, \theta) cos(\phi) cos(\theta) d\theta d\phi \tag{2.9}$$

$$Y(t) = \sqrt{3} \int_0^{2\pi} \int_0^{\pi} x(t, \phi, \theta) sin(\phi) cos(\theta) d\theta d\phi \tag{2.10}$$

$$Z(t) = \sqrt{3} \int_0^{2\pi} \int_0^{\pi} x(t, \phi, \theta) sin(\theta) d\theta d\phi \tag{2.11}$$

In other words, to derive the recorded signals (consisting of the convolution of the source signals with their room impulse responses) we integrate the received signal over the spherical harmonic representing the pickup pattern of the virtual microphone. By computing the spectrogram with the short-time Fourier transform (with digital frequencies represented by k and samples by n) of each of these virtual microphone channels, we can use the resulting vector $\mathbf{B}$ to perform DirAC analysis, as Pulkki first did in [5]. This method allows us to calculate relevant sound field parameters: the intensity, the direction of arrival, the energy, and the diffuseness.

$$\mathbf{B}(k, n) = STFT([W(t), X(t), Y(t), Z(t)]) \tag{2.12}$$

The intensity refers to the amount of transmitted acoustic energy at a given frequency, time, and cartesian coordinate direction, yielding a 3-dimensional vector as a function of frequency and time. We multiply the directional channels by the complex conjugate of the omni-directional channel, and then take the real part to only look at the movement (versus circulation) of the sound field. We scale by the characteristic impedance of the medium of sound propagation, Z_0, and add a negative sign to match convention. The intensity vector indicates the direction of sound propagation for a given time and frequency.

$$\mathbf{I}(k,n) = -\frac{1}{Z_0}\Re\{[\mathbf{B}_x, \mathbf{B}_y, \mathbf{B}_z]\mathbf{B}_w*\} \tag{2.13}$$

Using the intensity information, we can calculate an estimate of the DOA for each time-frequency bin of the sound as the azimuth and elevation angles, ϕ and θ respectively, of the negative of the intensity vector; the DOA points from the receiver in the direction from which sound originates at a given time and frequency.

$$\mathbf{\Omega}(k,n) = [\phi(k,n), \theta(k,n)] = -\angle(\mathbf{I}(k,n)) \tag{2.14}$$

The energy, E, of the field considers all information of the intensity vector except direction; it indicates how much total energy is transmitted at a given time and frequency.

$$E(k,n) = \frac{|\mathbf{B}_x|^2 + |\mathbf{B}_y|^2 + |\mathbf{B}_z|^2}{2Z_0 c} \tag{2.15}$$

Finally, the diffuseness, ψ, is correlated to the ratio of direct sound to reverberance for a given time and frequency. A fully diffuse field ($\psi = 1$) is completely reverberant, potentially made up of many uncorrelated sound waves, and a fully non-diffuse field ($\psi = 0$) has no reverberation, potentially made up of a single wave in a free field. We compute ψ by weighting the directional intensity, I, of the field against the energy density. This serves to represent how direct the field is versus how reverberant it is.

We take the time average, $\langle\rangle$, of both values to get a better picture of the sound's behavior.

$$\psi(k,n) = 1 - \frac{\|\mathbf{I}(k,n)\|}{c\langle E\rangle} \tag{2.16}$$

2.3 Musical Blind Source Separation

Blind source separation is the problem of finding and isolating different sources in an audio mixture given no previous information about the physical configuration of the sources or of the audio information they carry. In most techniques with spatial information available (such as in microphone array signal processing contexts) this can be split into two tasks: estimating the directions of arrival (DOA) of the sources, and separating the sources. Because this thesis focuses on source separation, we provide in this section a deeper analysis of source separation techniques with a priori knowledge of the DOA, and provide a more cursory analysis of DOA estimation methods in Section 2.4.

2.3.1 Spatial Filtering

The most basic multichannel, spatial source separation algorithms use only directional information; they filter the incoming signals to favor the DOAs of the sources, but do not try to separate the sources given signal characteristics such as frequency or history.

Given a DOA for source j (ϕ_j, θ_j), we perform beamforming by steering virtually our ambisonic microphone array towards the desired source by weighting the channels and summing. Thus, if our channel weights are given by w, a two-entry matrix with the first entry weighting the omni-directional channel W and the second entry

weighting each of the directional channels, our source j estimate $\hat{s_j}$ will be

$$\begin{aligned} \hat{s_j} = w(1)W(t) + \\ w(2)\sqrt{3}X(t)cos(\phi_j)cos(\theta_j) + \\ w(2)\sqrt{3}Y(t)sin(\phi_j)cos(\theta_j) + \\ w(2)\sqrt{3}Z(t)sin(\theta_j) \end{aligned} \tag{2.17}$$

Depending on the w used, we achieve different beamforming patterns; Table 1 provides the weighting schemes advanced by [6] as well-tuned for spatial audio analysis: basic, MaxRE, and in-phase. However, any cardioid can be created by varying the weights.

Table 1: Beamforming patterns described in [6]

Beamformer	**w(1) (W channel weight)**	**w(2) (X, Y, and Z channel weights)**
Basic	1	1
MaxRE	0.775	0.4
In-phase	0.5	0.1

The basic beamformer, shown in Figure 3, benefits from a narrower lobe around the estimated source DOA at the expense of a larger counterphase lobe centered at 180° away from the estimated DOA. The MaxRE beamformer, shown in Figure 4, benefits from a smaller counterphase lobe at the expense of a wider centered lobe. The in-phase beamformer, shown in Figure 5, has no counterphase angles but wider main and 180° lobes. All three figures depict beamformers steered toward 0°.

Following beamforming, the virtual microphones that point in the direction of the source are emphasized, while other microphones are minimized. Naturally, this enhances the signal coming from the DOA. More information about beamforming, particularly in the ambisonic context, is available in [7].

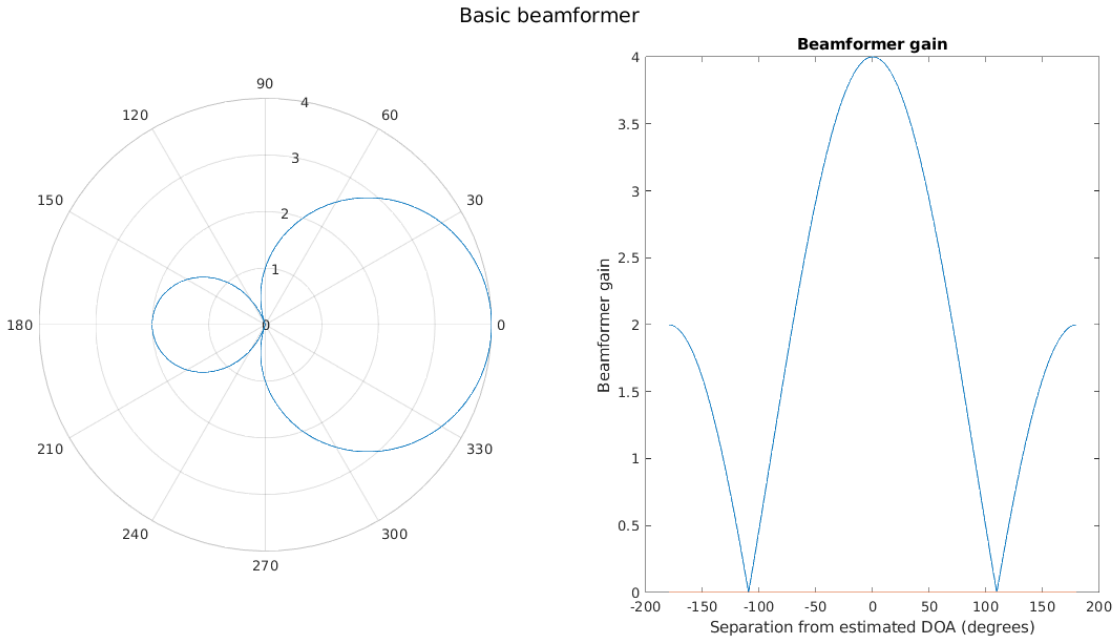

Figure 3: Left: Beampattern of basic beamformer for source at 0º. Right: Gain of basic beamformer as a function of angular distance away from the estimated DOA.

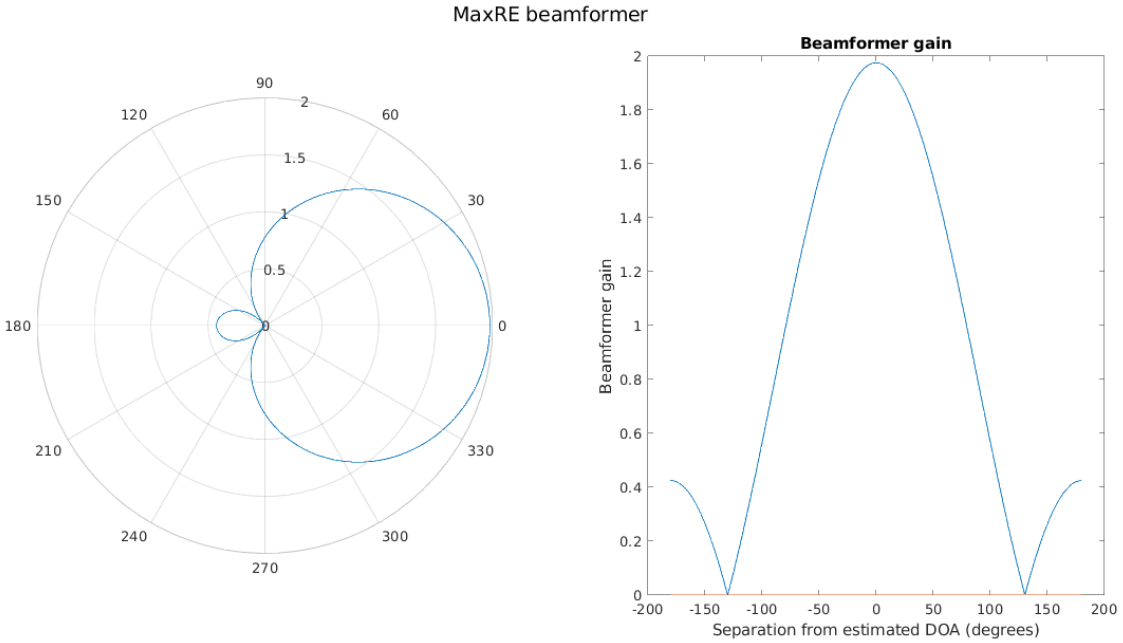

Figure 4: Left: Beampattern of MaxRE beamformer for source at 0º. Right: Gain of MaxRE beamformer as a function of angular distance away from the estimated DOA.

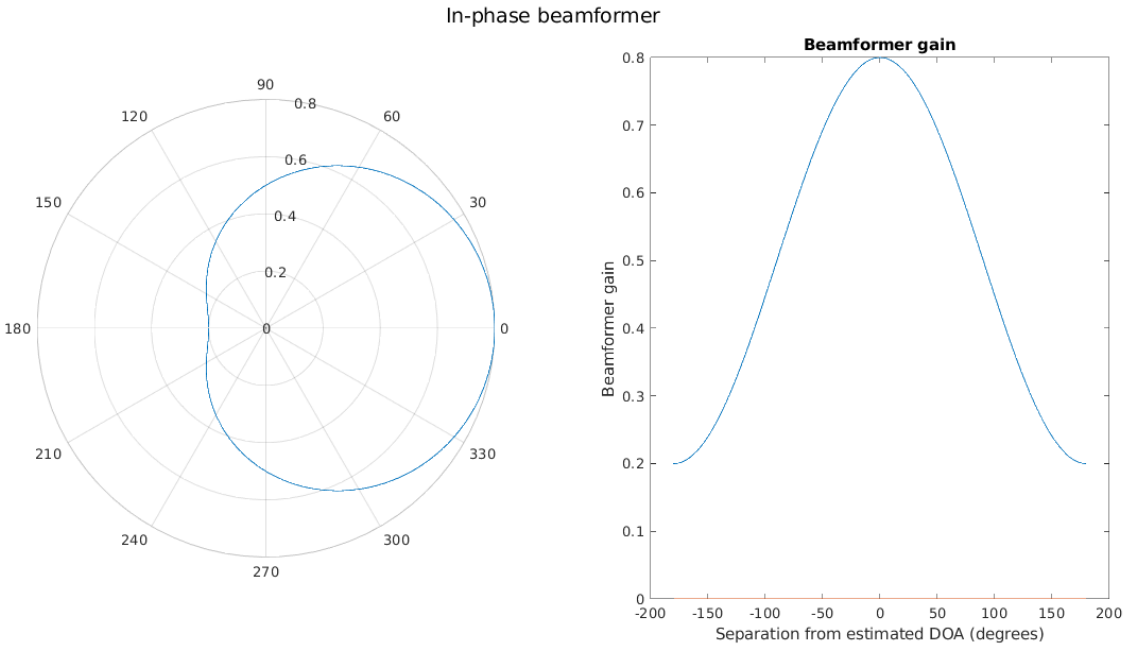

Figure 5: Left: Beampattern of in-phase beamformer for source at 0°. Right: Gain of in-phase beamformer as a function of angular distance away from the estimated DOA.

2.3.2 Nonnegative Matrix Factorization (NMF)

Gathering information beyond the spatial data of the source, such as its frequency content and temporal history, allows for a more informed separation, especially when constrained to the spatially agnostic data of a single microphone setup. This is frequently accomplished via non-negative matrix factorization (NMF).

NMF was first proposed by [8] as a way to compress matrix data such that only the most relevant features were saved. Specifically, NMF sparsely decomposes a given strictly non-negative n-by-m matrix V as Λ, the product of two strictly non-negative matrices: n-by-r matrix W and r-by-m matrix H. r is small enough to result in significant data compression.

$$V \approx \Lambda = WH \tag{2.18}$$

By requiring that all matrices be positive, the approximation matrix Λ is created

strictly additively, and the component matrices represent meaningful structures from the original data V: W can be seen to contain templates that occur frequently in the original data, and H contains the weights that determine how the templates appear in the data. Methods for calculating W and H given V, which depend on minimizing a cost function dependent on V and Λ, are given in [8].

Smaragdis first applied NMF in a musical context, to the problem of automatic music transcription [9]. The method takes the magnitude of the short time Fourier transform of the musical signal to be composed as V and decomposes it into the notes that are present (W) and when they are present (H). To illustrate for a simple case, Figure 6 displays a signal that contains two notes that randomly turn on and off.

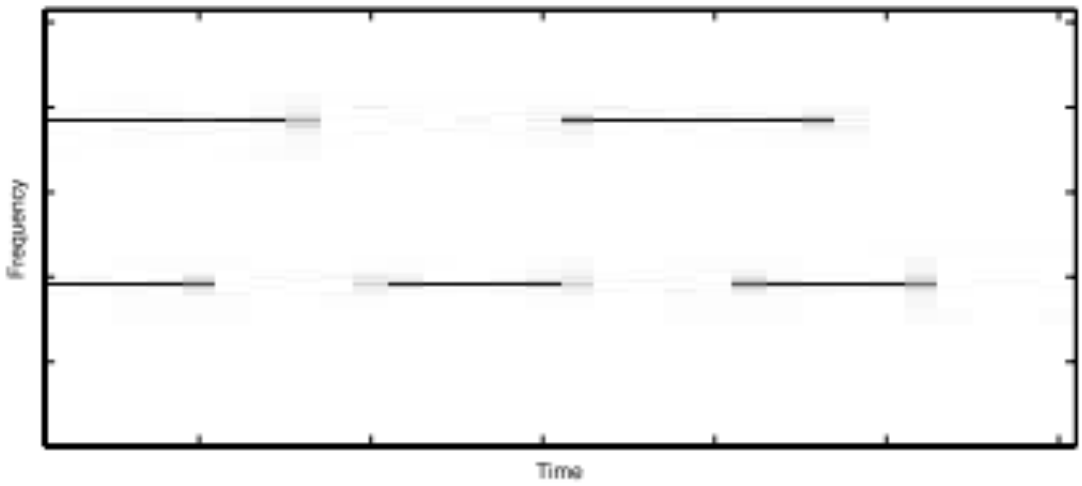

Figure 6: Spectrogram of audio signal to be transcribed; the signal consists of two notes that randomly switch on and off. Adapted from [9].

The algorithm is robust even when the number of notes present is unknown, but assuming that we know there are two notes, we can set r to 2 to compute more efficiently. Figure 7 shows our result: W contains the two notes that appear in the piece (the templates) and H indicates when either of the two notes appears (the weights). The paper continues to explain how to then transcribe the music by locating in time the notes of W according to H, but for this thesis' purposes, only the ability to detect notes is of interest.

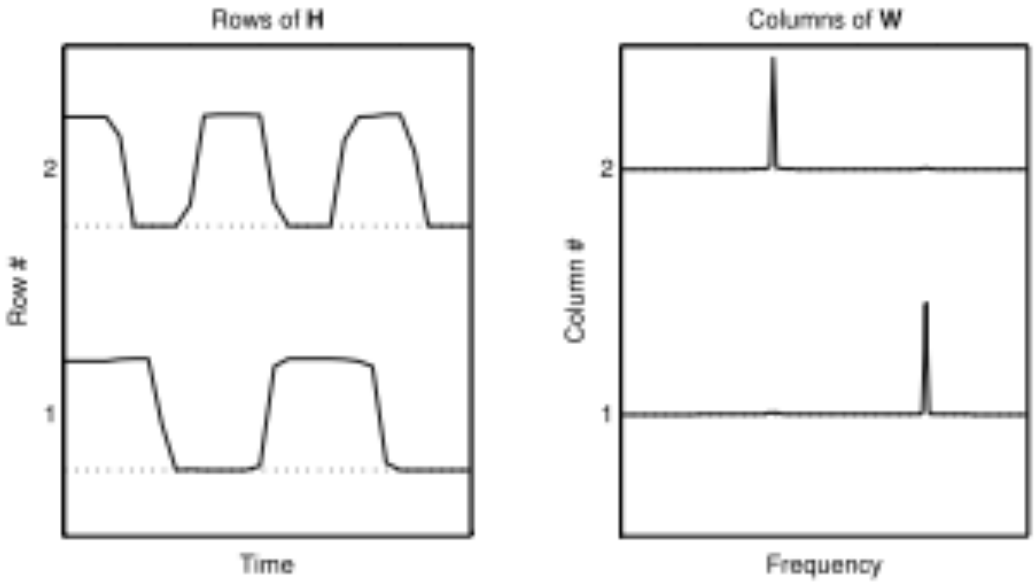

Figure 7: Result of performing NMF on the spectrogram of Figure 6. H controls when each of the two notes is active and W controls the pitch of each note. Adapted from [9].

To extend note detection to the source separation problem, we must correlate each note to its originating source by considering temporal factors. Smaragdis proposed doing this through NMF deconvolution (NMFD) in [10]. To be more realistic, we consider the signal spectrogram of Figure 8 and assume that the five lower frequency sweeps belong to one source and the higher frequency sweeps to a second source.

We set r to 2 to correspond to our two sources, add a temporal component by shifting H, and use a different W for each source. The time period in samples of the signal that each W samples is stored in T. Our decomposition is now

$$V \approx \Lambda = \sum_{t=0}^{T-1} W_t H^{t\rightarrow} \tag{2.19}$$

where W_t is the collection of templates characterizing the sources, and $t \rightarrow$ shifts

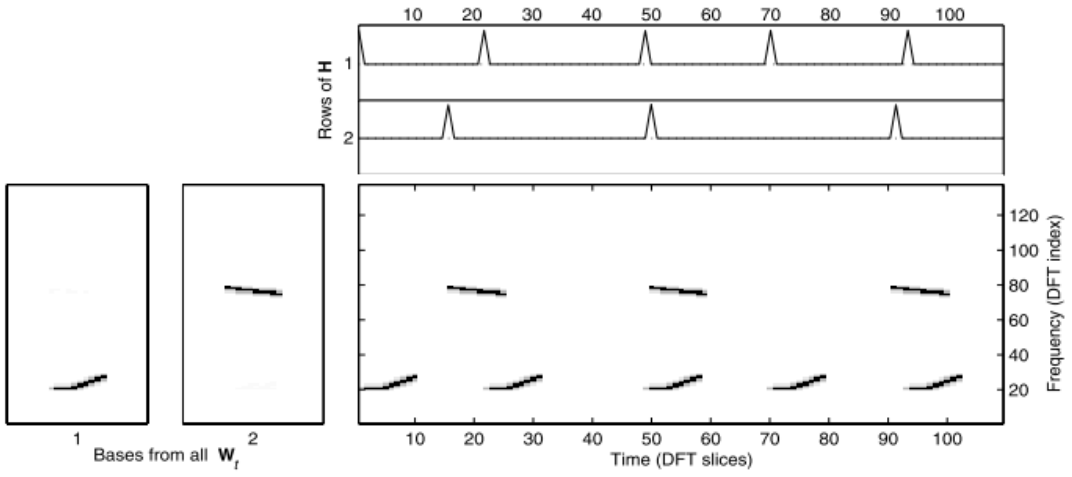

Figure 8: Result of performing NMF deconvolution on a simple music signal spectrogram. We assume the high and low frequency sweeps correspond to two separate sources, and then each template of W represents one source. Adapted from [10].

the matrix t positions to the right and back fills zeros.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 0 \end{bmatrix} 1 \rightarrow = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 4 & 5 \\ 0 & 7 & 8 \end{bmatrix} \tag{2.20}$$

Observing Figure 8, we see how our one H and r Ws are created. H stores the temporal data of when the two sources are playing as before, but W now stores the source "signature"; H defines when each W appears. From this point, we have the time and frequency bins of where each source appears, and we can extract an initial spectrum estimate for the j-th source, $\hat{S_{init_j}}$, by looking at our original sum applied to only the j-th column of W, $W_t^{(j)}$

$$\hat{S_{init_j}} \approx \sum_{t=0}^{T-1} W_t^{(j)} H^{t\rightarrow} \tag{2.21}$$

Once we have this initial estimate, we can apply Wiener or alpha-Wiener filtering to improve our accuracy, as [11] describes. Our final spectrogram estimate $\hat{S}_j$ will

appear as

$$\hat{S}_j = \frac{\hat{S_{init_j}}^{\alpha}}{\sum_{j'=1}^{J} \hat{S_{init_j}}^{\alpha}} V \tag{2.22}$$

Where α is a constant between 1 and 2 and all operations are done element-wise. This operation effectively keeps only the time-frequency bins of V that pertain to the source of interest. If the target source appears more strongly in one time-frequency bin, that bin will be weighted more heavily than a bin in which the source is less present. Finally, we perform inverse short time Fourier transforms on each separated signal spectrogram to arrive at the separated signals.

2.3.3 Filtering in the Ambisonics Domain

The current state of the art method for musical source separation, [2], relies on having the signal data in the (higher order) ambisonics (HOA) domain and applying statistical filtering. This separation method uses a multichannel Wiener filter to minimize the expected squared error between the estimate of each source j in each time-frequency bin (k, n) and the filtered version:

$$\begin{aligned} & \forall j, k, n \\ W_{j,k,n} = \underset{W}{\operatorname{argmin}} & E[\|c_{j,k,n} - W_{j,k,n} X_{k,n}\|_2^2] \end{aligned} \tag{2.23}$$

Where X is the spectrogram of one channel of the received signal and W is the Wiener filter to be applied to X. $c_{j,k,n}$ is the contribution of the j-th source to frequency bin (k, n), with covariance $\Sigma_{c_{j,k,n}}$, and which [2] approximates using a local Gaussian model (LGM) distribution,

$$c_{j,k,n} \sim \mathcal{N}(0, \Sigma_{c_{j,k,n}}) \tag{2.24}$$

Thus, to calculate W we must minimize the difference between our contribution

estimate c and filtered signal WX. As [2] indicates, the filter solution is given as a ratio of the covariance between c and X and the covariance between X and itself,

$$W_{j,k,n} = \Sigma_{c_{j,k,n},X_{k,n}} \Sigma_{X_{k,n},X_{k,n}} \tag{2.25}$$

Once we have calculated our W, we can calculate our final estimate for the contribution of each source to each time-frequency bin as

$$c_{j,k,n} = W_{j,k,n} X_{k,n} \tag{2.26}$$

and then calculate our source estimate by time-frequency bin $\hat{s_{j,k,n}}$ as

$$\hat{s_{j,k,n}} = \frac{y_{j,k}^H}{\|y_{j,k}\|^2} c_{j,k,n} \tag{2.27}$$

where $y_{j,k}$ is the spherical harmonic vector evaluated at source j's DOA and ^{H} is the Hermitian matrix operator.

2.3.4 Neural Network Source Separation

Although this thesis does not deal explicitly with neural network approaches to the source separation problem, for completeness we refer here to some representative methods. In [12], a deep neural network (DNN) method is proposed that assumes the instruments in the mixture are known; the network is then trained on recordings of performances of the solo instruments. In [13], a monaural deep convolutional neural network (CNN) method is proposed in which the network is used to estimate time-frequency soft masks. In [14], a multichannel deep neural network is used to model the source spectra, and this result is combined with a multi-channel Gaussian model to also consider spatial information. The Gaussian model parameters are then used to derive a multi-channel Wiener filter that is applied to the mixture to return the source estimates.

2.4 Direction of Arrival (DOA) Estimation

Although not a main focus of this thesis, direction of arrival estimation is an important step in many real-world source separation techniques. Here, we summarize two such methods, both of which assume the number of sources to be known.

2.4.1 Parametric Value Estimation

In this method, proposed in [15], we use the parametric values calculated in section 2.2 to create masks that indicate time-frequency bins that most likely correspond to a source rather than reverberation. To be precise, we set a threshold on our diffuseness values such that time-frequency bins with small-enough diffuseness are considered and all others are discarded. We calculate over both azimuth (ϕ) and elevation (θ) angles histograms of the DOA by only summing bins that are saved by the mask. In a single source case, this will appear as in Figure 9. We can use a peak-picking algorithm to find the DOA of the source.

2.4.2 MUSIC Estimation

The MUSIC (MUltiple SIgnal Classification) method, proposed by [16], works by statistically separating the mixture into signal and noise components. We take the number of sources J as given, and assume for simplicity that we need only find the azimuth, and not elevation, angle of the DOA (the method is of course able to return both values). We start by breaking the received signal matrix x into signal and noise components:

$$x = AF + W \tag{2.28}$$

A is a mode vector that expresses how the receiver responds to sounds from different

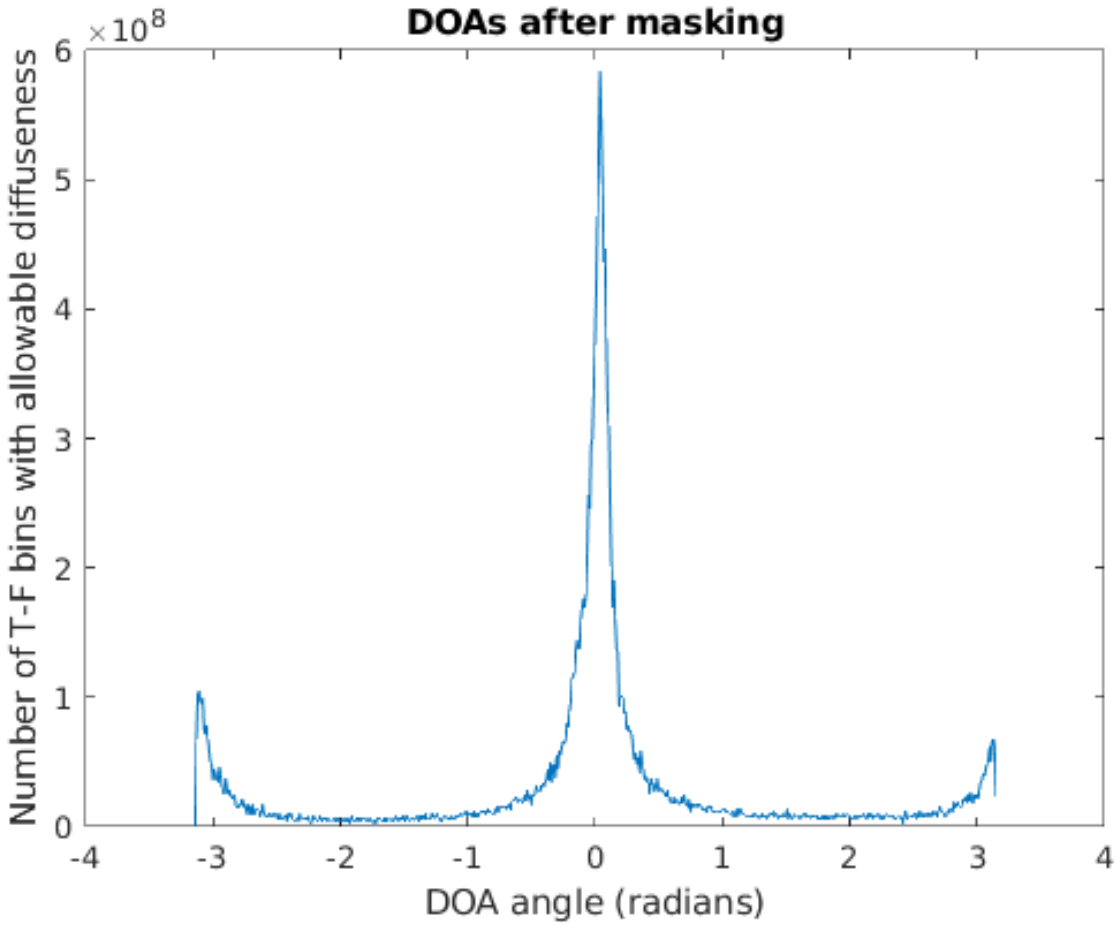

Figure 9: Direction of arrival (DOA) histogram with 0.7 diffuseness threshold. This is a single source case, and using a peak-picking algorithm would likely indicate that the source is at roughly 0 radians direction of arrival angle. The 180° peak is due to early reflections.

DOAs,

$$A = \begin{bmatrix} a(\phi_1) \\ ... \\ a(\phi_J) \end{bmatrix} \tag{2.29}$$

F expresses the signals incident on the array microphones,

$$F = \begin{bmatrix} F_1 & ... & F_J \end{bmatrix} \tag{2.30}$$

and W is the noise at each microphone,

$$W = \begin{bmatrix} W_1 & ... & W_J \end{bmatrix} \tag{2.31}$$

We then calculate S, the covariance matrix of X,

$$S = X\bar{X}^* = AF\bar{F}^* + W\bar{W}^* = APA^* + \lambda S_0 \tag{2.32}$$

where $^-$ and * are complex conjugate operators, P is defined to be $F\bar{F}^*$, λ is an eigenvalue of W, and S_0 is the eigenmatrix of W. By ordering the eigenvalues by magnitude, we determine that the J largest correspond to the signal, and span a signal subspace, and that the remaining N correspond to and span a noise subspace. We then calculate $P(\phi)$ as

$$P(\phi) = \frac{1}{a^*(\phi E_N E_N^* a(\phi))} \tag{2.33}$$

where E_N is the $N-by-I$ matrix containing the N noise eigenvectors. Plotting $P(\phi)$ against ϕ gives an angular distribution, and returning the J largest peaks produces DOA estimates that can be more accurate than methods that do not consider a separate noise subspace.

Chapter 3

Methods

This section details the novel proposed method submitted as part of this thesis, the configurations of the existing methods tested, and the generation of the testing data. Because the state of the art methods we compare against do not compute both DOA and separated signals, we test both aspects separately. All code to reproduce the methods and experiments described here is available at `https://github.com/henryhasti/master_thesis`.

3.1 Proposed Method

The novel approach works as depicted in Figure 10. It is an end-to-end combination of the approaches described in Sections 2.2, 2.3.1, and 2.4.1, and was found heuristically to work best with the configurations in Table 2. The diffuseness threshold is varied according to how diffuse the mixture is; more diffuse fields have a lower threshold so that there is enough data to compute the angular distributions. $psiSum$ is the sum of diffuseness across all time-frequency bins, and $psiMax$ is the value of $psiSum$ in a hypothetical fully diffuse sound field.

The processing blocks work as follows:

Table 2: End-to-end separation parameters

Parameter	Value (unit)
Hop size	0.01 rad
Diffuseness threshold	0.2 + psiSum/psiMax
Noise floor	0.15/1
Angular separation	0.5 rad
Beamformer weights	[0.775, 0.4] (MaxRE beamshape)

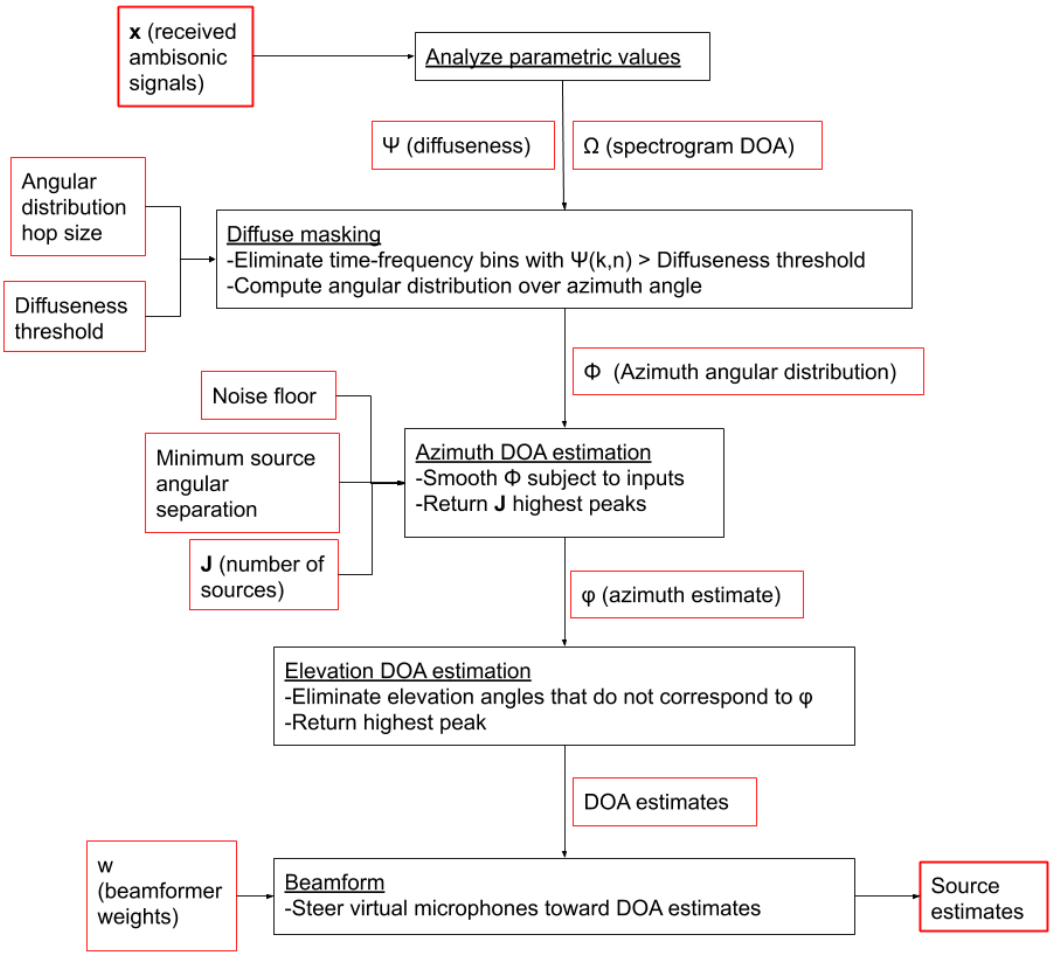

Figure 10: Flow diagram of novel approach. The combined DOA estimation and separation technique combines components from Sections 2.2, 2.3.1, and 2.4. Black boxes represent processes and red boxes represent variables.

.

3.1.1 Analyze Parametric Values

Given the ambisonic signals recorded by the receiver, the diffuseness and spectrogram DOA (the direction of arrival at each time-frequency bin) are calculated as in

[3].

3.1.2 Diffuse Masking

This algorithm calculates the azimuth angular distribution similarly to [15]. First, the time-frequency bins of the spectrogram DOA with diffuseness levels above the threshold are eliminated. Then, the DOAs of the remaining bins are sorted by magnitude into bins according to the hop size input. The output is the azimuth angle distribution as in Figure 9.

3.1.3 Azimuth DOA Estimation

This algorithm picks peaks of the azimuth distribution to estimate the azimuth component of the DOAs. First, the angular distribution is pre-processed: all values below the noise floor are set to zero and the distribution is smoothed (in this case with the built-in Matlab function). Then, the distribution is doubled to eliminate the $-\pi$ to π discontinuity: when the distribution arrives to π, a second copy is added, starting at $-\pi$. Next, all candidate peaks (the distribution's maxima that are at least as far apart as the input parameter) are calculated. Finally, one peak for each source is returned by assuming the maximum peaks most likely correspond to sources, and eliminating peaks that come from the doubled distribution.

3.1.4 Elevation DOA Estimation

This algorithm returns the elevation angle that corresponds to each azimuth angle. Given the spectrogram DOA, it eliminates the elevation time-frequency bins that do not correspond to the input azimuth angle, and then picks the highest peak after smoothing.

3.1.5 Beamform

This process steers the virtual microphones towards the input DOA as described in section 2.3.1.

3.2 Existing Approaches Tested

The blind musical source separation methods and DOA estimation methods discussed in Chapter 2 were all tested, with the following configurations:

3.2.1 NMF

NMF was conducted as in section 2.3.2, with the toolbox provided by [17] and the configurations, taken from the toolbox, as in Table 3.

Table 3: NMF implementation parameters

Parameter	**Value (unit)**
STFT window size	2048 samples
STFT hop size	512 samples
Window type	Hanning
Iterations	30
Template frames	10

3.2.2 Ambisonic Domain Filtering

This method was conducted as in section 2.3.3, using the toolbox provided by [18]. The parameters were taken from [2], with the exception of ambisonic order, which was held at one.

3.2.3 MUSIC DOA Estimation

This method was conducted as in section 2.4.2, using the Spherical-array-processing repository available at `https://github.com/polarch`.

3.3 Data Generation

For all experiments, the data was simulated using the shoebox-roomsim-master and dependent repositories available at `https://github.com/polarch`. Following the

configurations of [2], four rooms were simulated, all with dimensions $10x8x3$ meters and reverberations capped after 0.01, 0.2, 0.4, and 0.7 seconds, respectively. The 0 second reverberation room was increased to 0.01 seconds to circumvent a processing error in the shoebox-roomsim-master repository. In both test cases discussed below, the sources were held fixed at 0° elevation and 1 meter distance from the receiver, which was modeled as an infinitesimal, 1st-order ambisonic microphone array at the center of the room.

The musical segments used come from the DSD100 (demixing secrets dataset), [19], which contains anechoic recordings of songs isolated into four tracks representing bass, drums, vocals, and all remaining sources ("other"). This thesis' dataset consists of 5-second segments randomly selected from 10 randomly selected songs, displayed in Table 4. All ten segments contain all four tracks.

Table 4: DSD100 songs used

Song Name	**Artist**
Rockshow	ANiMAL
One Minute Smile	Actions
Ghost Bitch	Drumtracks
We Feel Alright	Girls Under Glass
Knockout	M.E.R.C. Music
A Reason To Leave	Patrick Talbot
What Have You Done To Me	Signe Jakobsen
Rothko	The WrongUns
Pony	Timboz
Comfort Lives In Belief	Voelund

Two room and instrument setup cases were tested:

3.3.1 Two Source Case

In this regime, illustrated in Figure 11, one source (drums) remained fixed at 0° azimuth angle, while a second source (vocals) moved in 10° increments from 0° to

180° azimuth angle. This test case highlights method performance as a function of source angular separation.

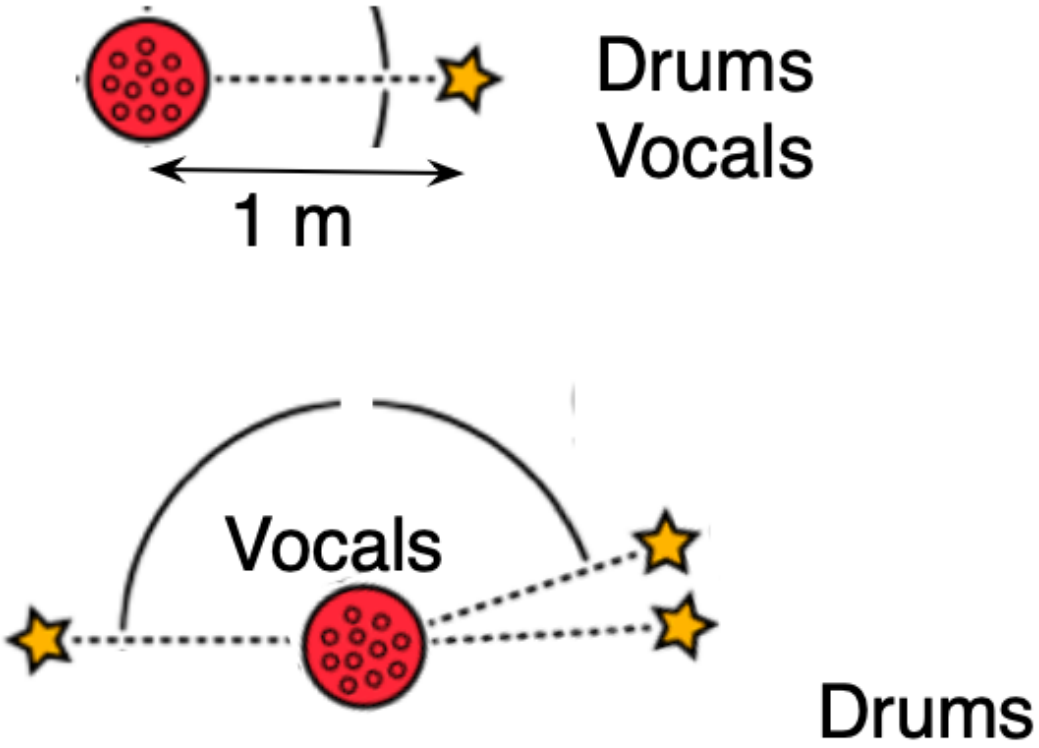

Figure 11: The two source case: the drums source remains at 0° azimuth while the vocals source moves from 0° (top image) to 180° azimuth in 10° increments (bottom image). The red circle represents the receiver and stars the sources. All elements have 0° elevation. Reproduced from [2] with alterations.

3.3.2 Four Source Case

In this regime, the four instrument configurations of [2] were used, as figure 12 displays.

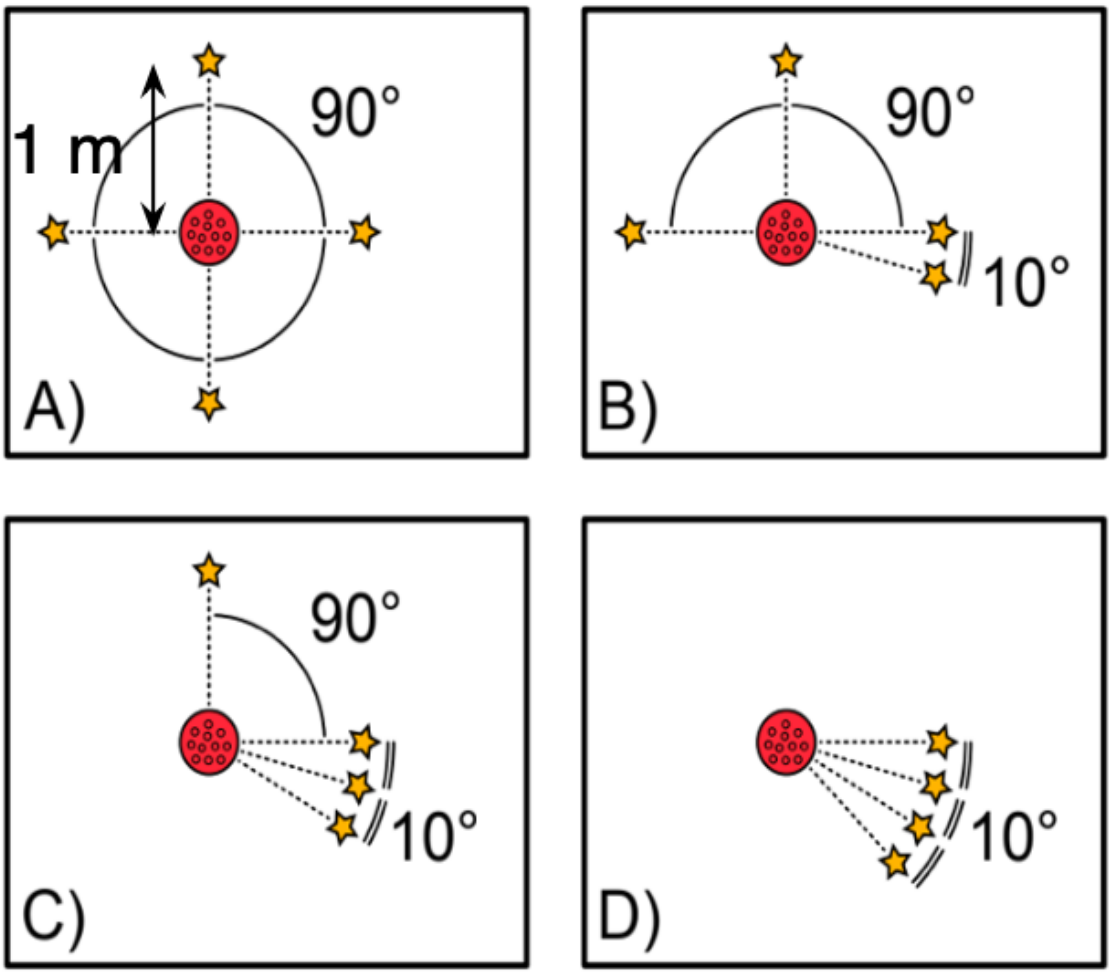

Figure 12: Source configurations as used in [2], which make up the four source test case. The red circle represents the receiver and stars the sources. All elements have 0° elevation. Reproduced from [2].

Chapter 4

Experiments

Several experiments were carried out to compare the performances of the methods defined in chapter 3.

4.1 DOA Estimation

The method implemented as part of this thesis was compared against the state of the art method detailed in section 2.4.2. The two source case was tested, and the performance was graded as the average of the great circle angular distance between the calculated and actual DOAs across the 10 song segments.

4.2 Optimized Beamformer Shape

To determine the optimal beamformer shape to use for the beamforming section of the novel method, the three beamformer shapes in Table 1 were tested with the two source case.

4.3 Separation Performance

Using the two and four source cases, the performances of all three separation methods were compared using the signal to interference (SIR) and artifact (SAR) ratios defined in [20] and now summarized. The paper additionally defines the signal to

distortion ratio (SDR), which also penalizes reverberations in the separated signals, but we do not use it since we consider reverberations to be valuable additions to the signal given that they characterize the conditions of the room and setup at the time of recording. The section in [21] on room acoustics and reverberation notes that in many cases reverberation is a desired effect that improves the perceived sound quality.

Given an estimated source $\hat{s_j}$ that we have separated from the mix of all signals, $\{s_1, ...s_J\}$, and noise recorded by each microphone, $\{n_1, ...n_I\}$, we intend to classify the effectiveness of the separation. While this is a difficult, subjective, task that is closer to psychoacoustics than to any mathematical algorithm, [20] provides a method for objectively evaluating source separations by considering ratios between different signal components of the mixture and estimated target signal.

To do this, the authors outline a framework in which signals, all of length T, are represented in a space in which every sample is a dimension. For example, the three signals captured by the directional ambisonic virtual microphones will span a 3-dimensional subspace within the full T-dimensional space. In this way, a test signal can be compared to target signals by projecting it onto the subspace spanned by the target signals. The T-by-T projection matrix onto the subspace spanned by arbitrary column vectors $\{y_1, ..., y_k\}$ is

$$\prod\{y_1, ..., y_k\} = \prod\{Y\} = Y(Y^{tr}Y)^{-1}Y^{tr} \tag{4.1}$$

where tr is the matrix transpose function. We define several projection matrices relevant to our source separation problem:

$$\mathbf{P}s_j = \prod\{s_j\} \tag{4.2}$$

$$\mathbf{Ps} = \prod\{(s_{j'})1 \leq j' \leq J\} \tag{4.3}$$

$$\mathbf{Ps,n} = \prod\{(s_{j'})1 \leq j' \leq J, (n_i)1 \leq i \leq I\} \tag{4.4}$$

$\mathbf{P}s_j$ projects into the subspace of the target signal, indicating how closely an estimate matches the actual signal. $\mathbf{P}s$ projects into the subspace of all sources in the mix, indicating how closely an estimate matches the combination of all signals. **Ps,n** projects into the subspace of the combination of all sources in the mix and noise at each microphone, indicating how closely an estimate matches the total mix. Given an estimate signal , we can use these projection matrices to define several important quantities:

$$S_{target} = \mathbf{P}s_j\hat{s_j} \tag{4.5}$$

$$e_{interf} = \mathbf{P}s\hat{s_j} - \mathbf{P}s\hat{s_j} \tag{4.6}$$

$$e_{noise} = \mathbf{Ps,n}\hat{s_j} - \mathbf{P}s\hat{s_j} \tag{4.7}$$

$$e_{artif} = \hat{s_j} - \mathbf{Ps,n}\hat{s_j} \tag{4.8}$$

s_{target} is the amount of the estimate that matches the desired signal; in an ideal separation $\hat{s_j} = s_{target}$, and multiplying the projection matrix leaves the estimate unchanged. e_{interf} is the error present in the estimate due to other sources in the mix. e_{noise} is the error due to noise. e_{artif} is the component of the estimate that was never in the original mix, the artifacts from the separation algorithm. Due to how we defined our separation problem initially, our estimate is necessarily the combination of these terms:

$$\hat{s_j} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{4.9}$$

To determine the effectiveness of the separation, we define several logarithm-scale ratios between the different components of $\hat{s_j}$.

The source to distortion ratio measures how much of the desired signal was captured

versus all other components in the calculated signal.

$$SDR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{4.10}$$

The source to interference ratio measures how much of the desired signal was captured versus other signals in the mix.

$$SIR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{4.11}$$

The sources to noise ratio measures how much of any signal was captured versus noise.

$$SNR = 10log_{10}\frac{\|s_{target} + e_{interf}\|^2}{\|e_{noise}\|^2} \tag{4.12}$$

Finally, the sources to artifacts ratio measures how much of the total mix (including all sources and noise) was captured versus the artifacts introduced by the algorithms.

$$SAR = 10log_{10}\frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{noise}\|^2} \tag{4.13}$$

4.4 Computation Time Performance

The computation times of the three separation methods and two DOA estimator methods as implemented in `https://github.com/henryhasti/master_thesis` and run on an HP Notebook - 15-db0074ns personal computer were compared using the real time factor,

$$Real \quad time \quad factor = \frac{Computation \quad time}{Audio \quad length} \tag{4.14}$$

The illustrative two source case of 0^o separation and only one song was used, and the computation times were averaged across the four reverberation times.

Chapter 5

Results and Discussion

5.1 DOA Estimation

The results from this experiment are displayed in Figures 13, 14, and 15. The proposed method (abbreviated "Parametric") follows a nearly linear trajectory because of the 180° phase reflections that come from the wall opposite the stationary source; for many songs the two DOAs returned will correspond to the 0° source and the early reflections. As the second source gets closer to 180°, the total error decreases. The jump in error at 180° separation and 0.01 second reverb is due to both DOAs wrongly getting calculated at either 0° or 180°. The MUSIC algorithm displays this same early reflection bias. Comparing the two, we see that their errors are quite similar, although the proposed method outperforms MUSIC at several angles.

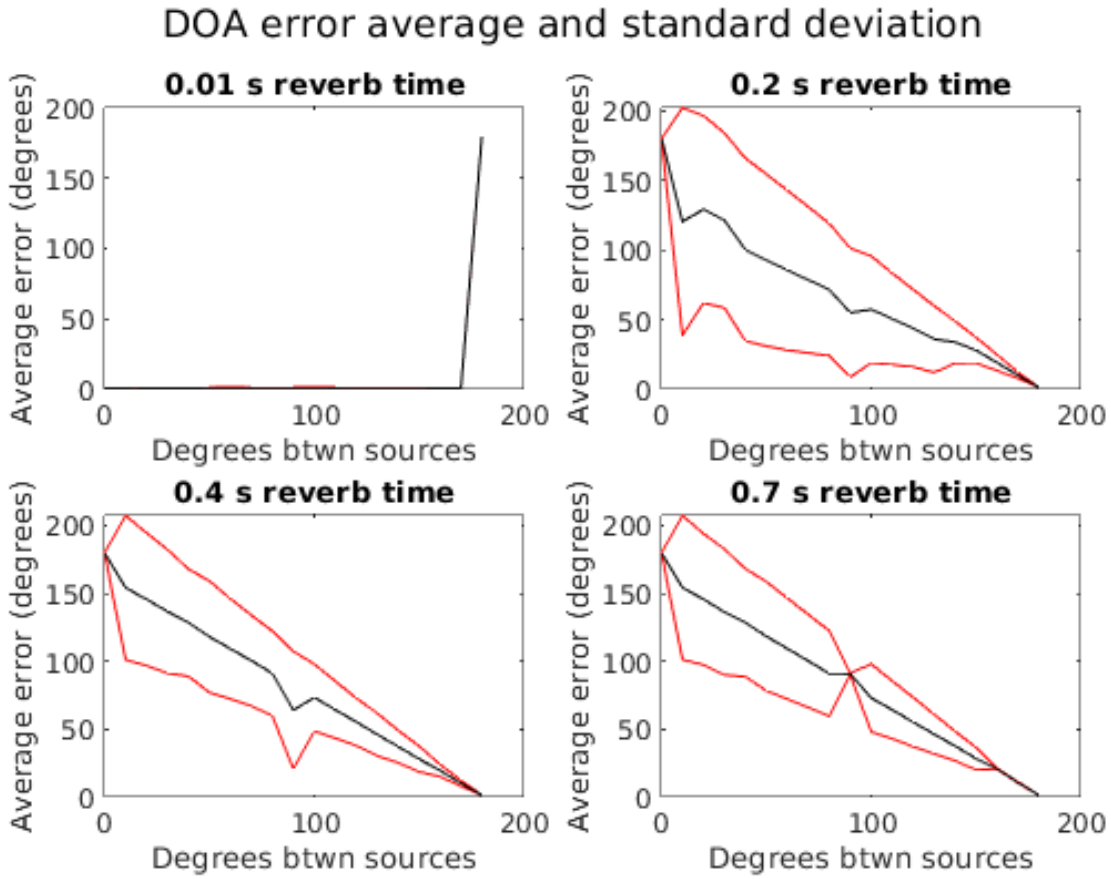

Figure 13: Average DOA error (black) and standard deviation (red) across 10 songs using the proposed parametric values estimation method.

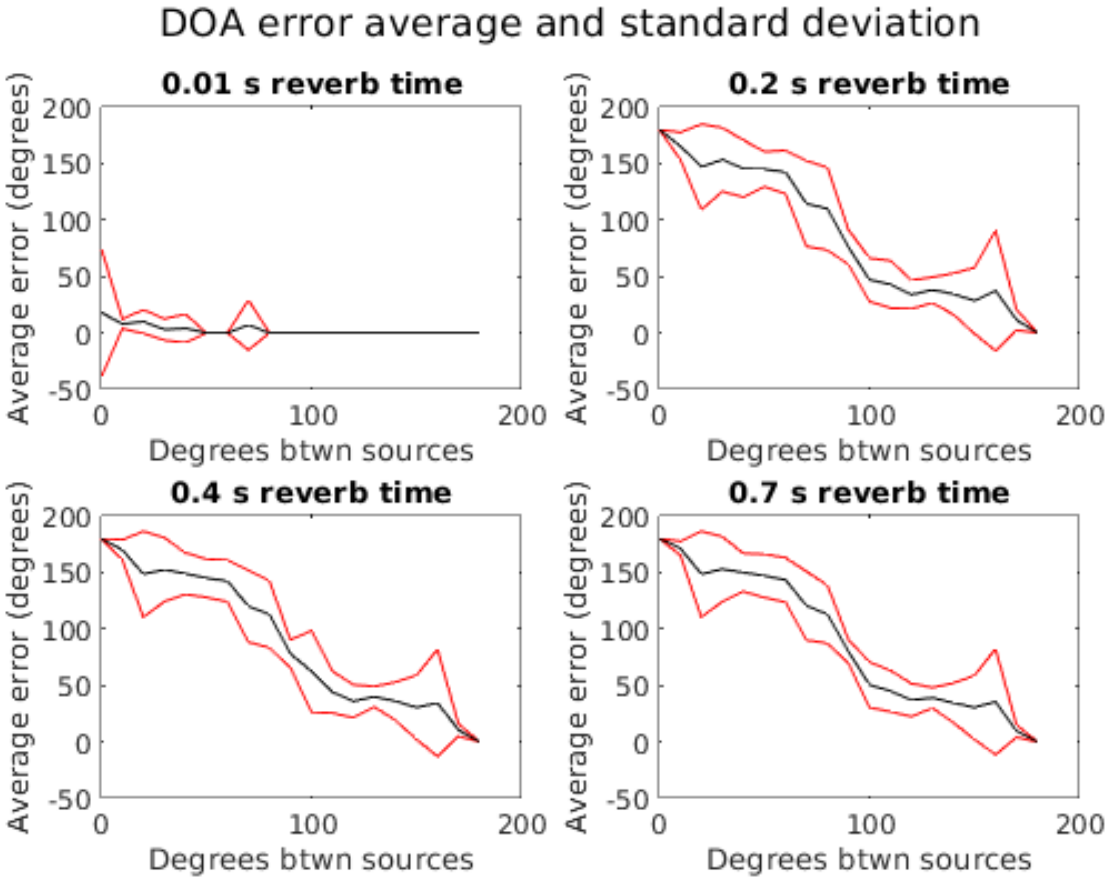

Figure 14: Average DOA error (black) and standard deviation (red) across 10 songs using MUSIC estimation method.

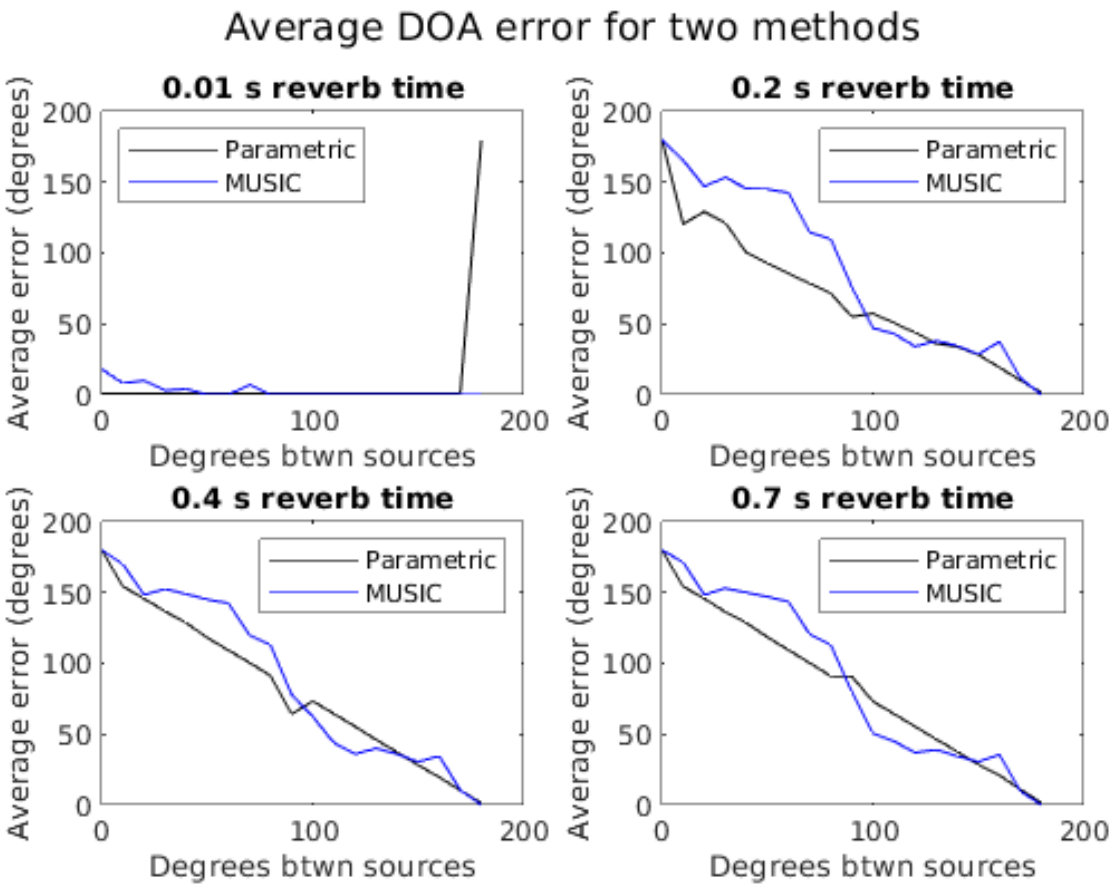

Figure 15: Average DOA errors of proposed and MUSIC estimation methods.

5.2 Optimized Beamformer Shape

Figures 16, 17, 18, and 19 display the results of this experiment. The average ratios across the 10 songs are displayed in black, and the standard deviations are in red. For reference, the beamformer's gain is shown by the dotted blue line.

By comparing the SIRs to the beampattern gain, we see that the experiment's results match what we expect: as the angular separation of the sources places the interfering source closer to the beamformer's null, separation performance improves. This is most noticeable with minimal reverb, because there are no early echoes from the interfering source.

Since the artifacts introduced by the beamformer are theoretically uncorrelated to the separation angle, we are unsurprised that SAR is generally flat. The increases in SAR at the beamformer null in the low reverb cases can be explained due to processing errors; small differences in sample alignment between the estimated and target tracks can be minimized when there is only one prominent instrument track audible.

Given these results, we conclude that the MaxRE beamformer provides the best performance, and we use it for all calculations related to spatial filtering going forward. Due to its decreased counter-phase lobe, it provides better performance when the source are separated by an angle greater than the beamformer null, and it still performs comparably to the other beamformers at closer angles.

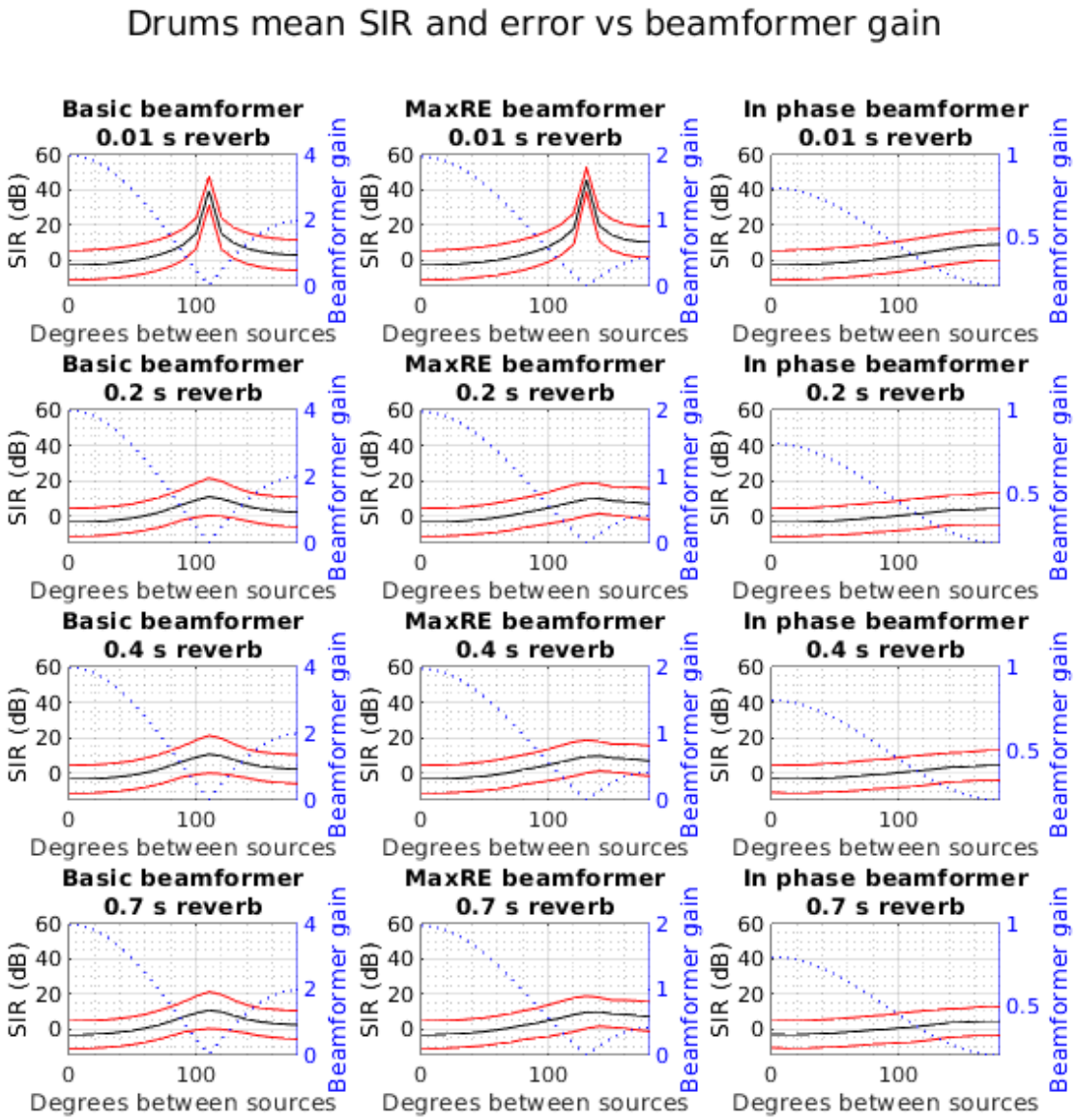

Figure 16: Signal to interference ratio of drums track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line.

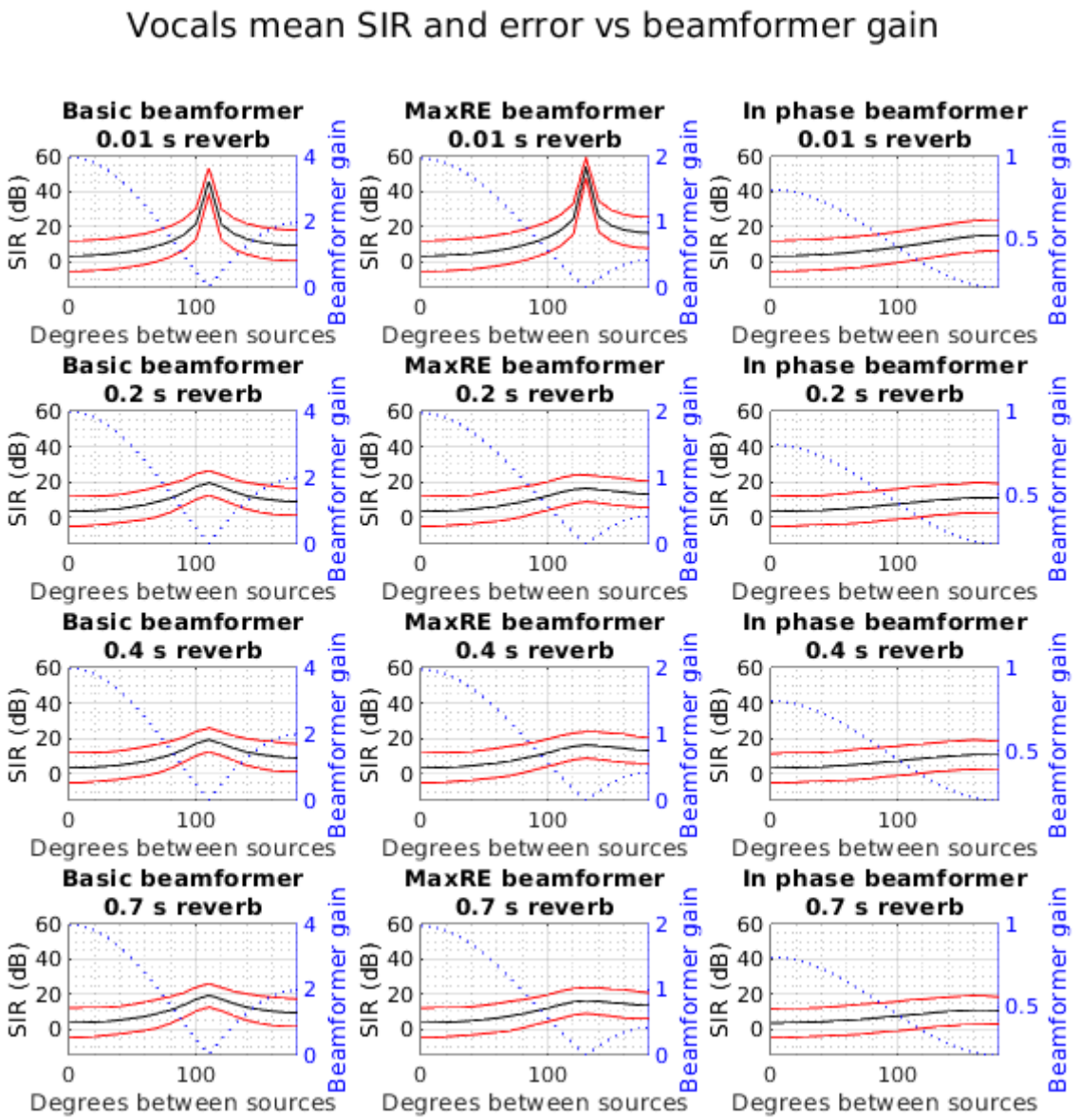

Figure 17: Signal to interference ratio of vocals track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line.

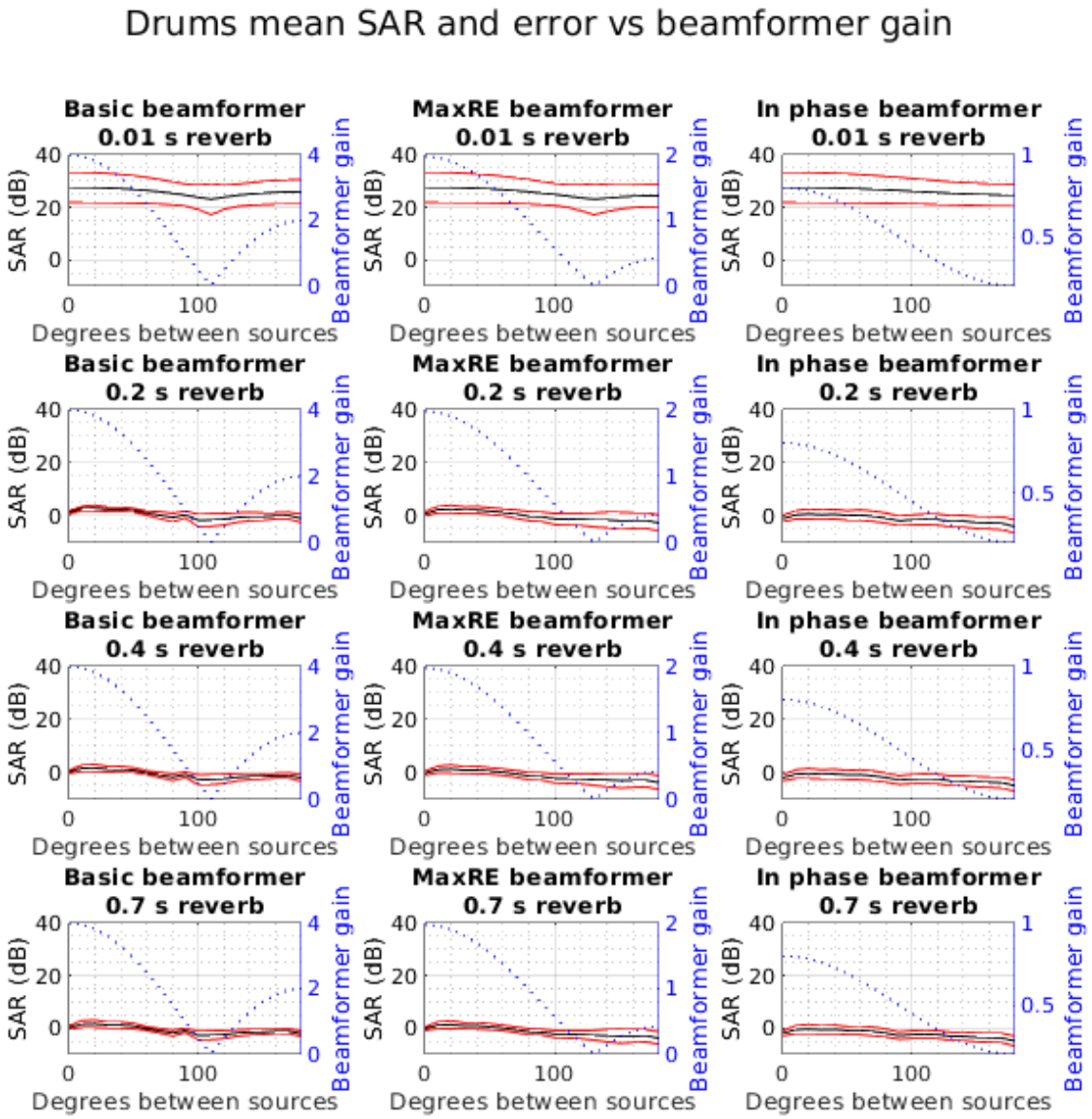

Figure 18: Signal to artifact ratio of drums track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line.

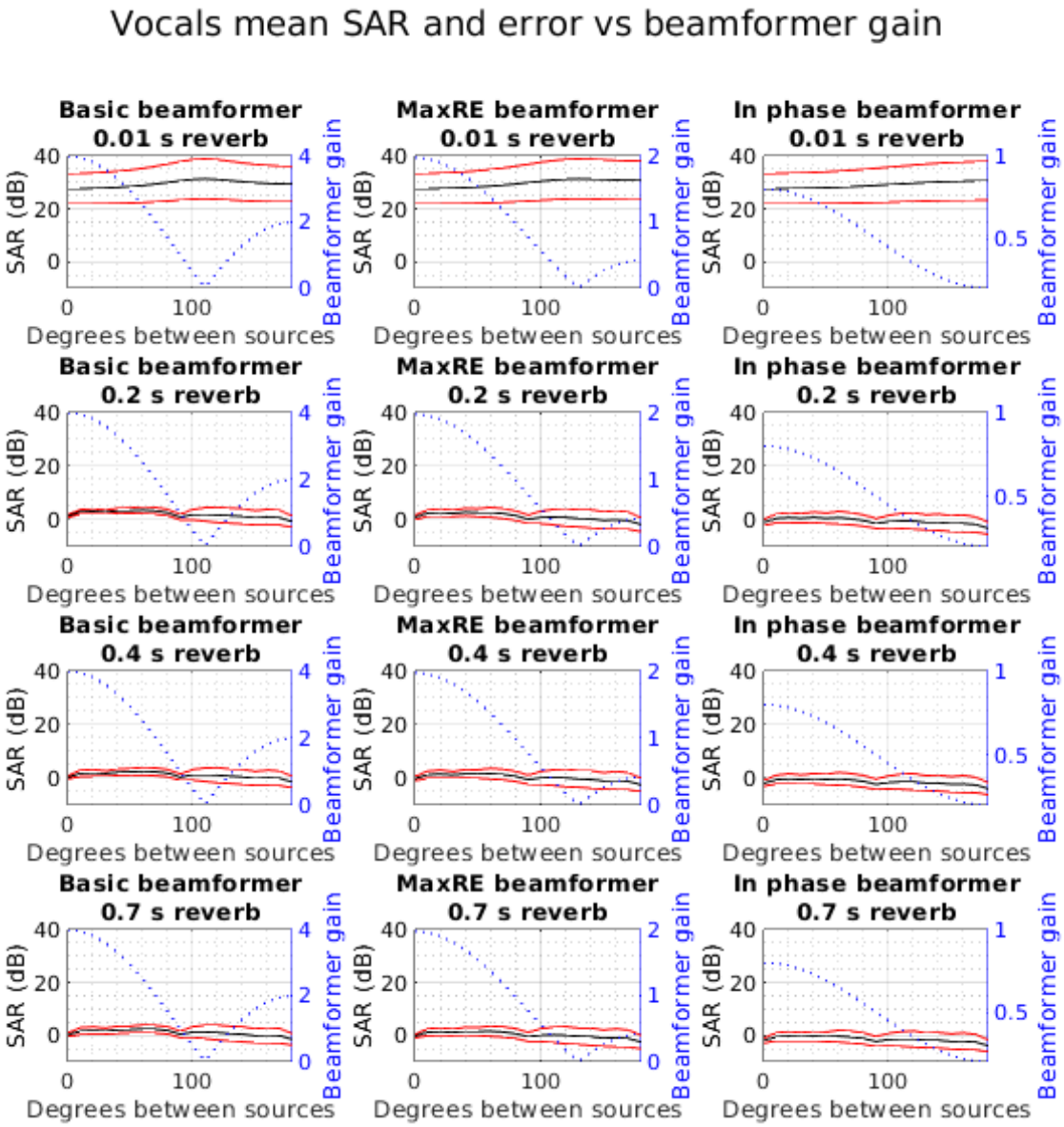

Figure 19: Signal to artifact ratio of vocals track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line.

5.3 Separation Performance

5.3.1 Two Source Case

The results from this experiment for NMF and separation in the ambisonics domain methods are presented below, along with graphics comparing all three methods. The results for spatial filtering are the same as in Section 5.2.

NMF results

The results for the two source case separated using NMF are presented in Figures 20 and 21. Averages and standard deviations across the 10 songs are represented in black and red, respectively. Of particular interest is that the SIR is not negatively affected by increasing reverb: the templates seek out the correct time and frequency bins without regard to the ambient reverberation. This does not hold true, however, for the SAR, which drops significantly with reverb. This effect is audible in estimated sources that cut in and out and sound noticeably discontinuous.

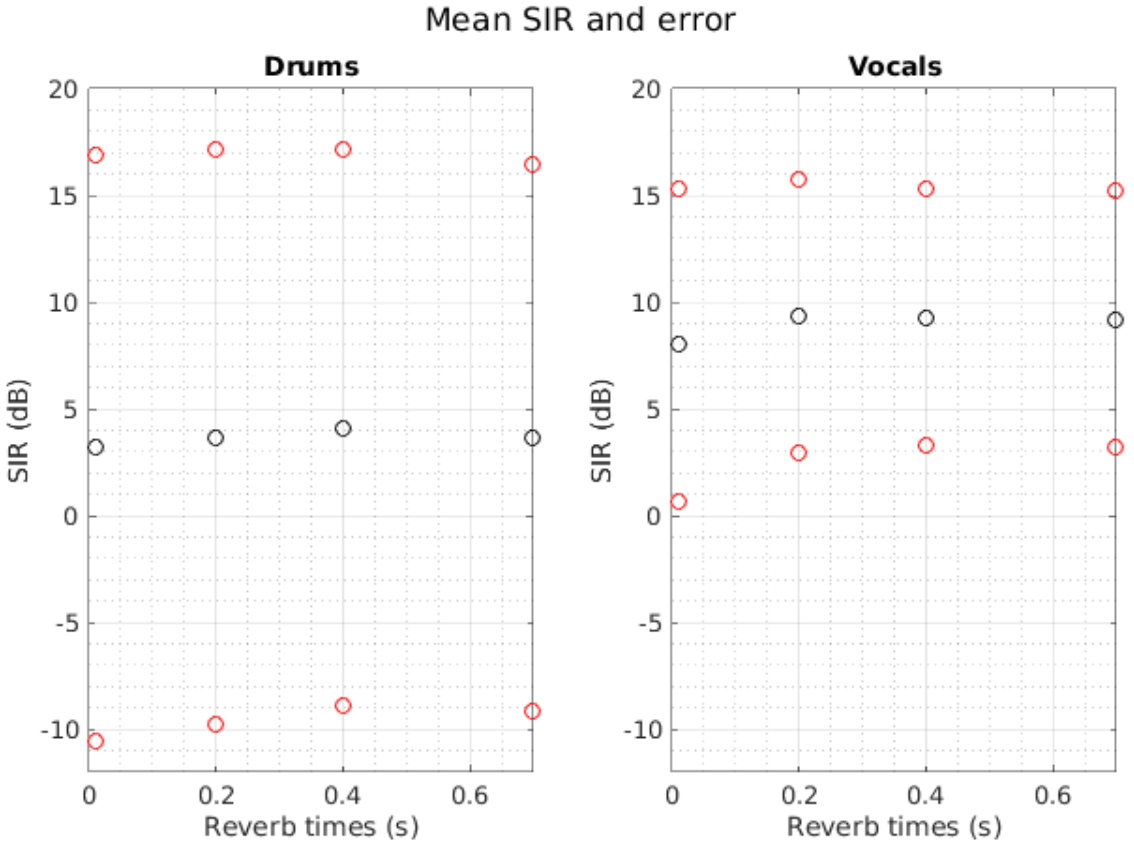

Figure 20: Signal to interference ratios of drum and vocal tracks for two source case evaluated with NMF. The average ratios in decibels are in black and the standard deviations in red.

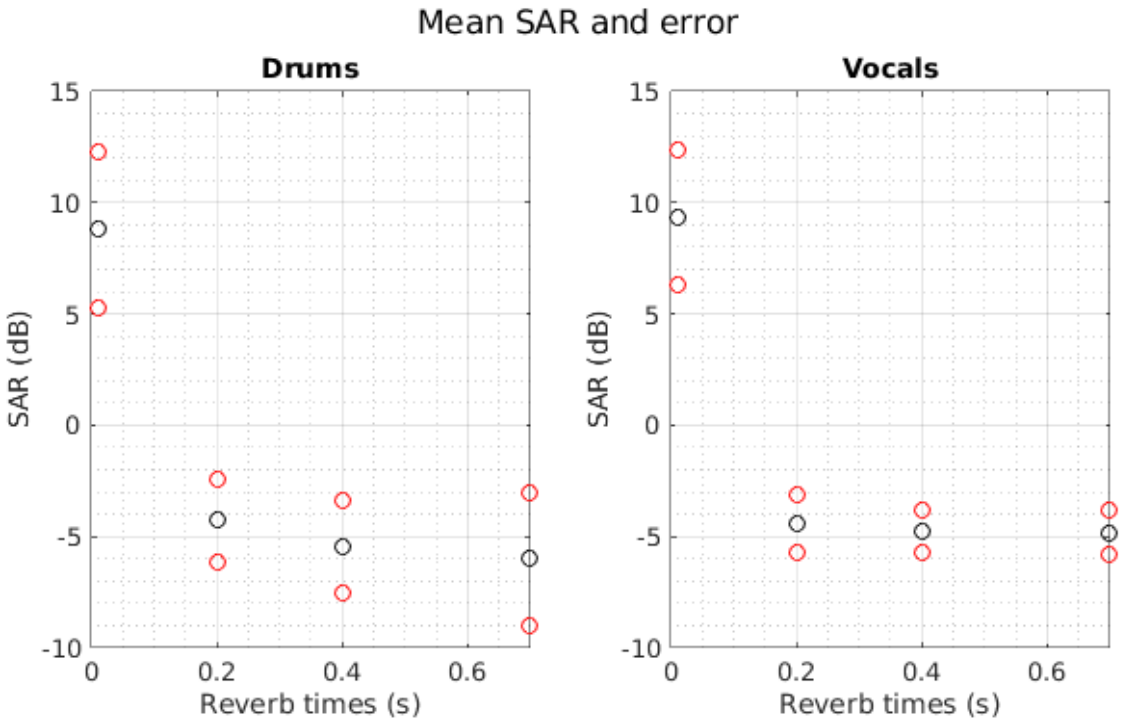

Figure 21: Signal to artifact ratios of drum and vocal tracks for two source case evaluated with NMF. The average ratios in decibels are in black and the standard deviations in red.

Ambisonic Domain Filtering

The results from the two source problem separated with ambisonic domain filtering are presented in Figures 22, 23, 24, and 25. The averages and standard deviations across the 10 songs are in black and red, respectively. With the exceptions of the performance decrease with the spatially ambiguous 0° separation case and the performance increase at 90° - likely due to the nulls of the X and Y ambisonic microphones there - performance is roughly independent of source separation angle. We can say that the algorithm is robust with respect to many source configurations, but adding reverb shows a clear performance degradation. It appears, however, that the difference going from minimal (0.01 s) to some (0.2 s) reverb is much greater than the difference going to higher reverb amounts; once any reverb is added, the system performs comparably.

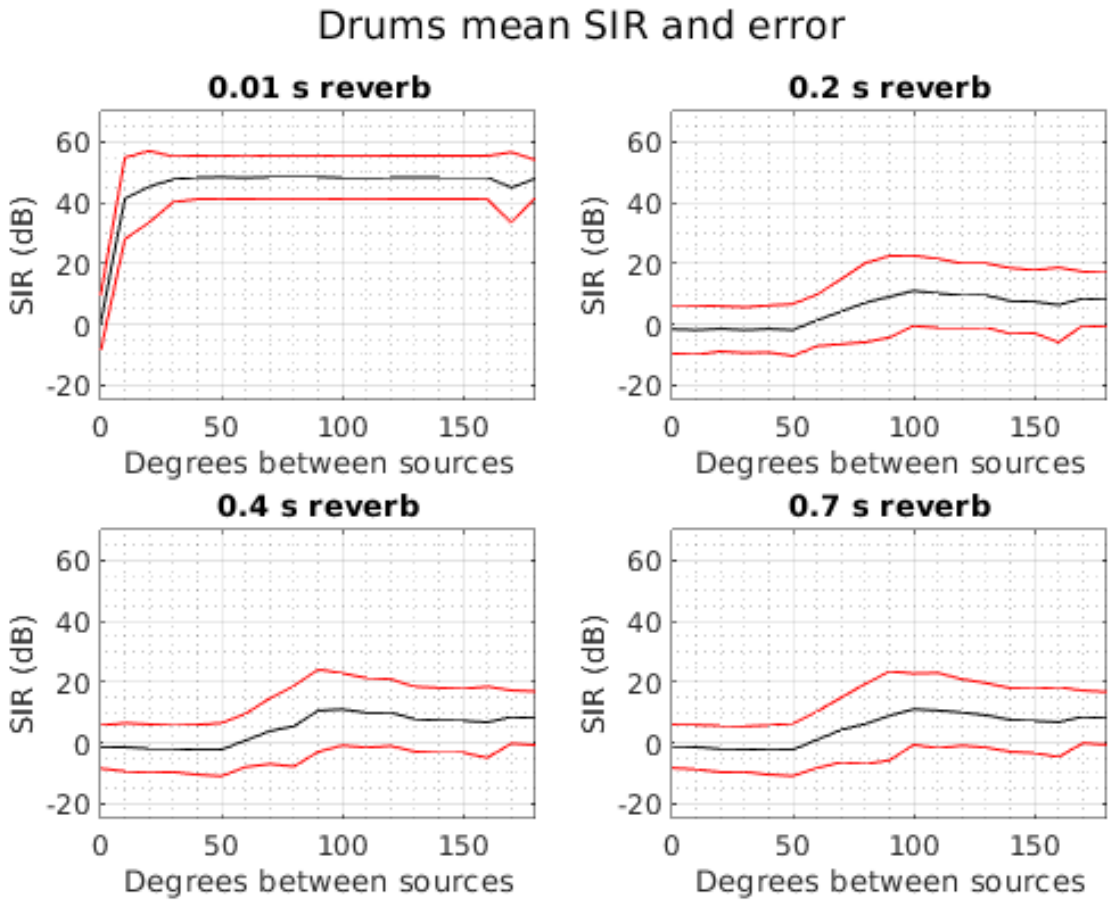

Figure 22: Signal to interference ratios of drum track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red.

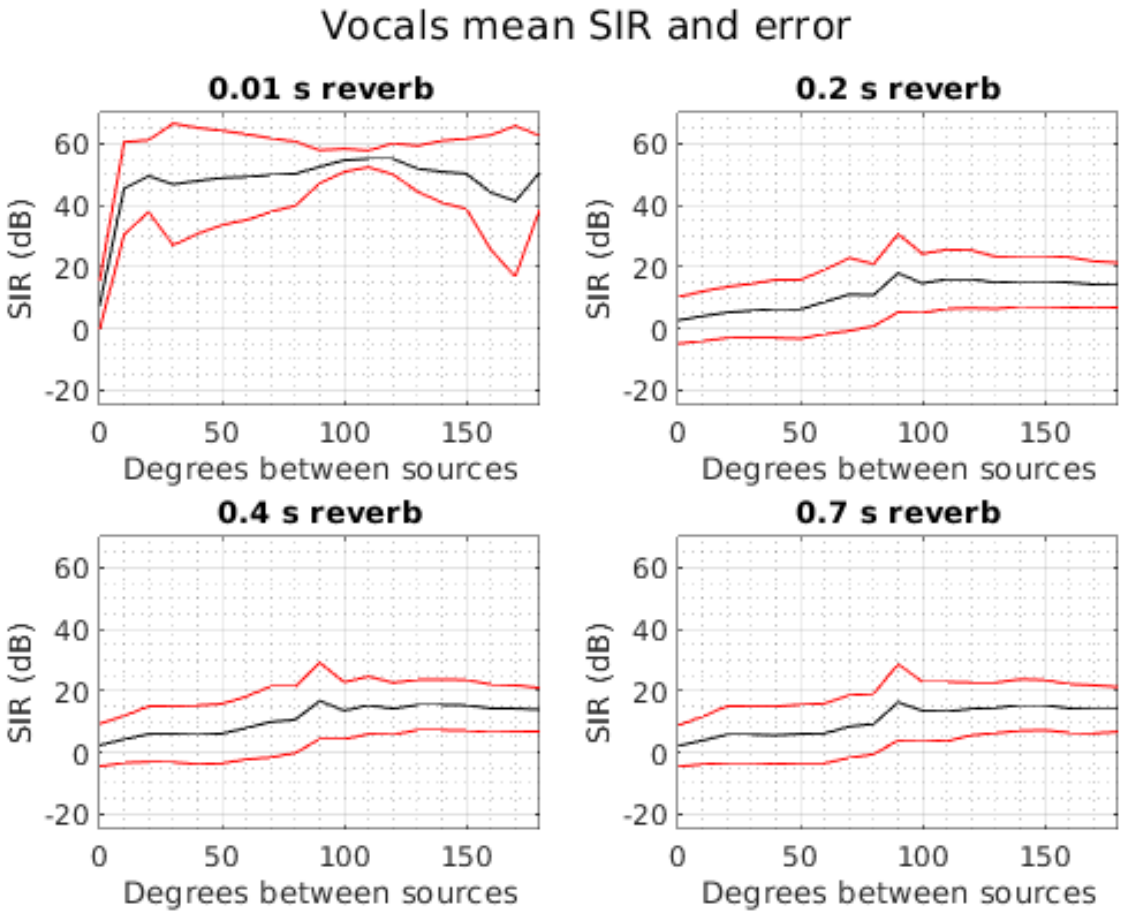

Figure 23: Signal to interference ratios of vocal track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red.

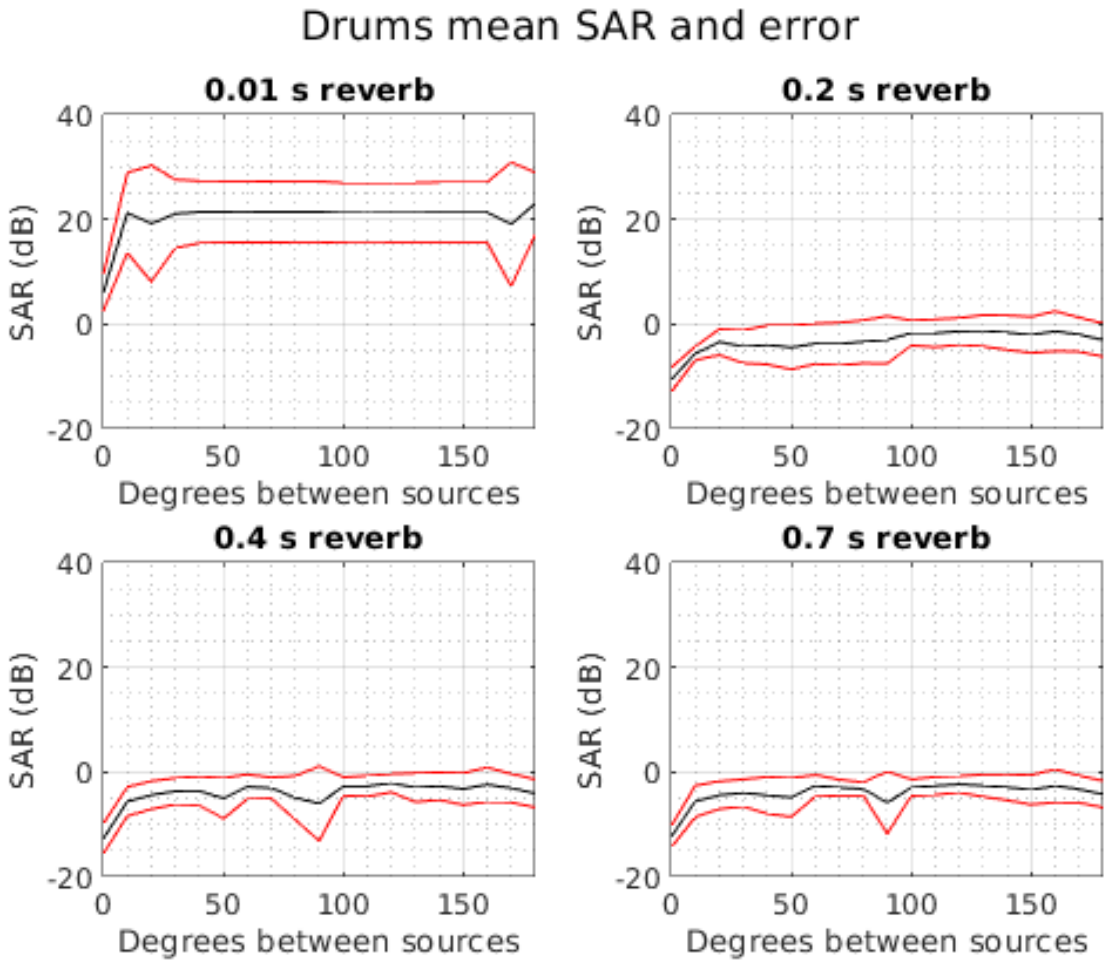

Figure 24: Signal to artifact ratios of drum track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red.

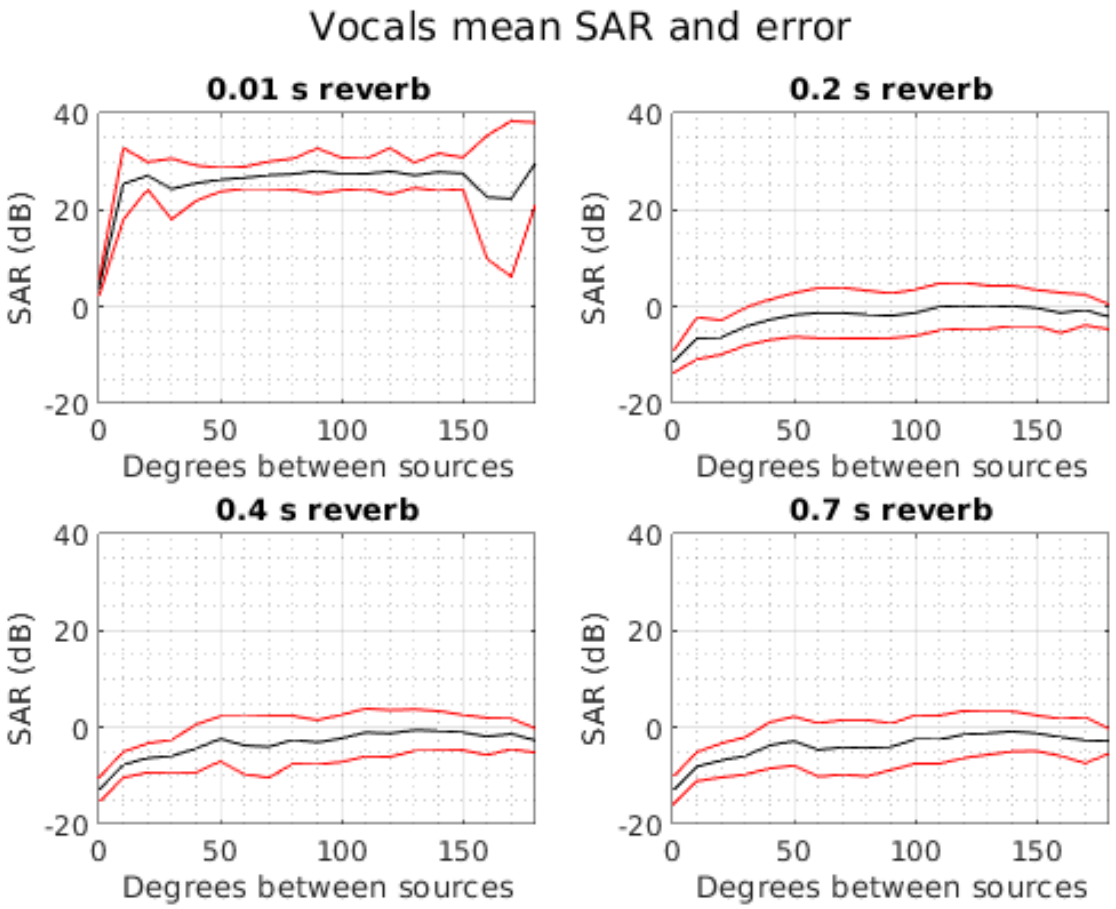

Figure 25: Signal to artifact ratios of vocals track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red.

Overall Comparison

Figures 26, 27, 28, and 29 contain a comparison of the three methods. Worth noting is that NMF is significantly outperformed by the other two methods except at the ambiguous-configuration near-0° source angular separation cases; standard NMF as implemented by [10] does not consider source positions. In terms of SIR, the ambisonic domain filtering method significantly outperforms spatial filtering at low reverb, but both perform comparably at higher reverb. In terms of SAR, spatial filtering performs better in virtually all cases; the method has very few avenues to produce extra noise in its operation.

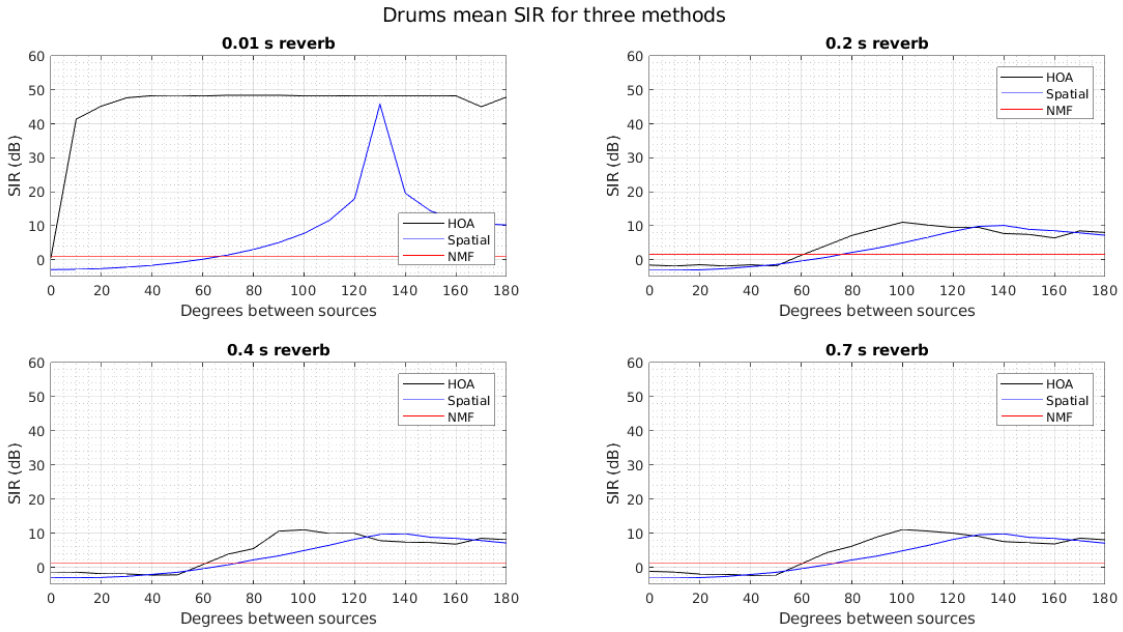

Figure 26: Signal to interference ratios of drum tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

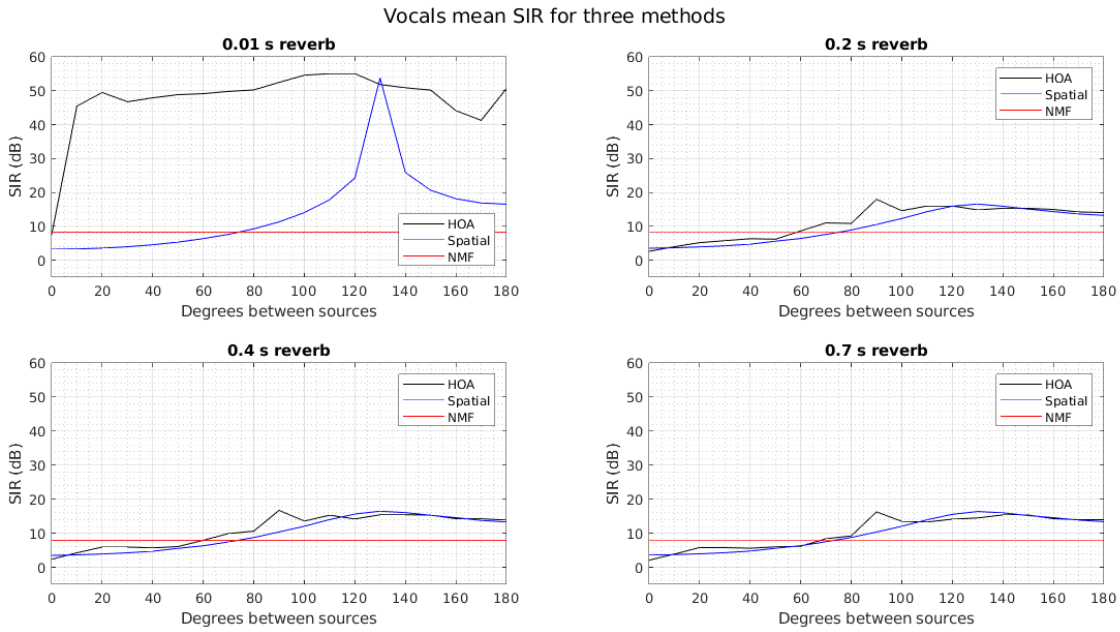

Figure 27: Signal to interference ratios of vocal tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

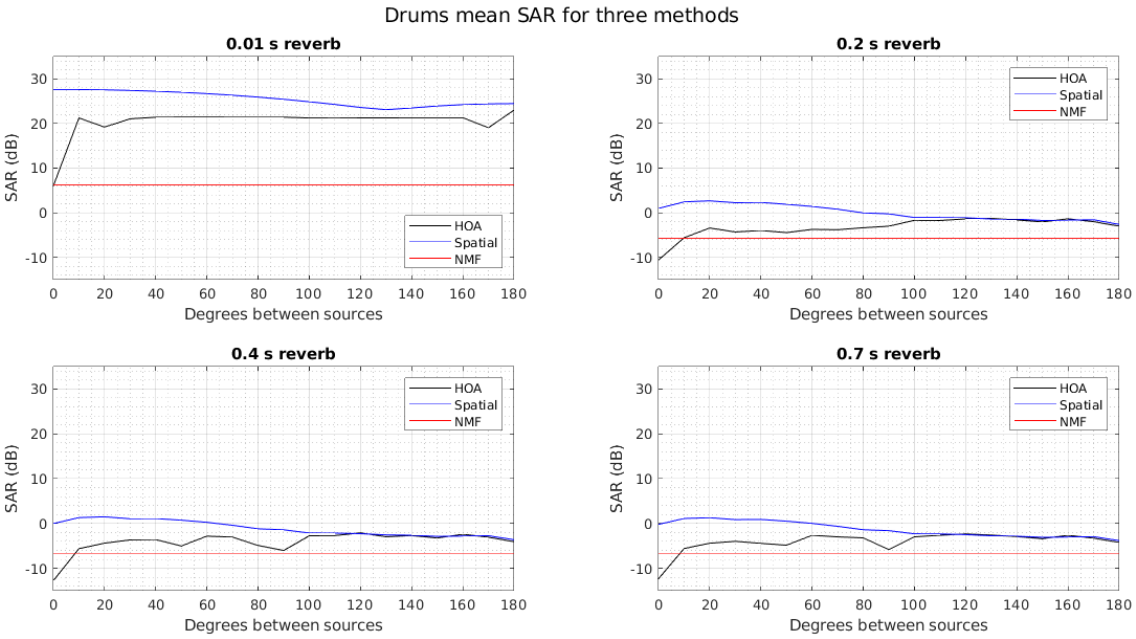

Figure 28: Signal to artifact ratios of drum tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

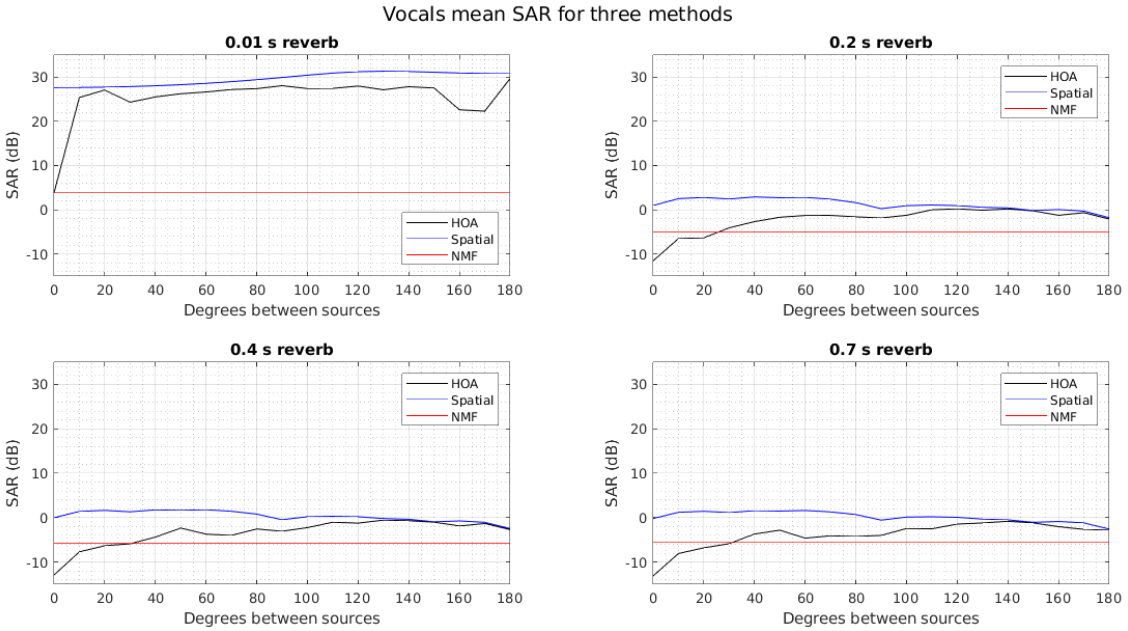

Figure 29: Signal to artifact ratios of vocal tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

5.3.2 Four Source Case

Figures 30 through 37 display the data comparing the source separation methods. For brevity, the figures deal only with averages and not standard deviations, but the code to create the complete figures is available at `https://github.com/henryhasti/master_thesis`.

In general, the results match what would be expected from extrapolating the two source case results. The ambisonic domain filtering method vastly outperforms the other two in low reverberation, but tends to perform more comparably as reverb increases. The proposed method tends to perform better in the cases when sources are more separated; its performance drops significantly when sources are only 10^{o} apart. In terms of SIR, we can conclude that the ambisonic domain method's comparative performance improves significantly as more sources are added. The SARs tend to follow the same pattern, with the proposed method virtually always outperforming: it is the least computationally complex and thus least likely to introduce algorithmic noise.

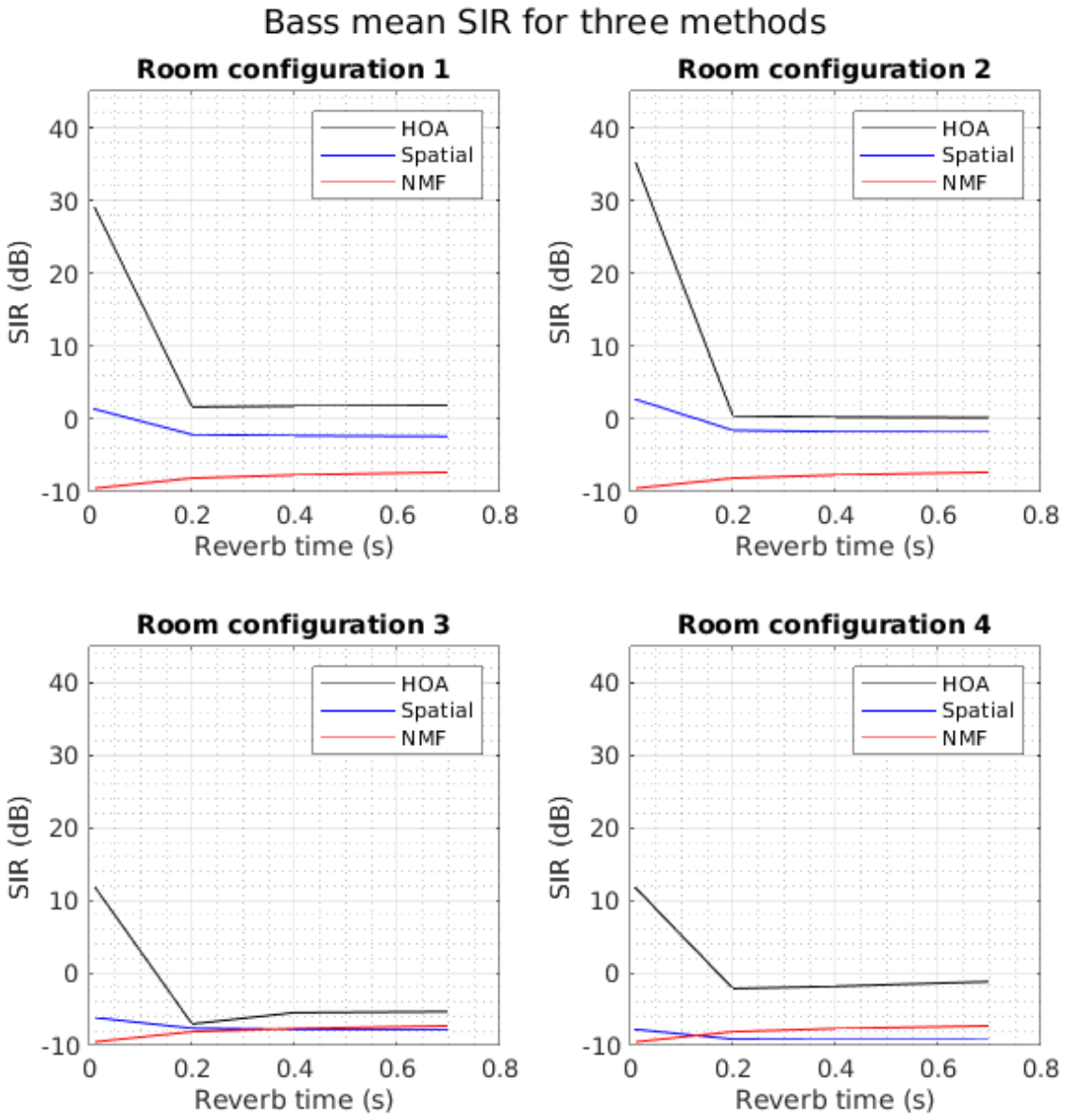

Figure 30: Source to interference ratios of bass for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

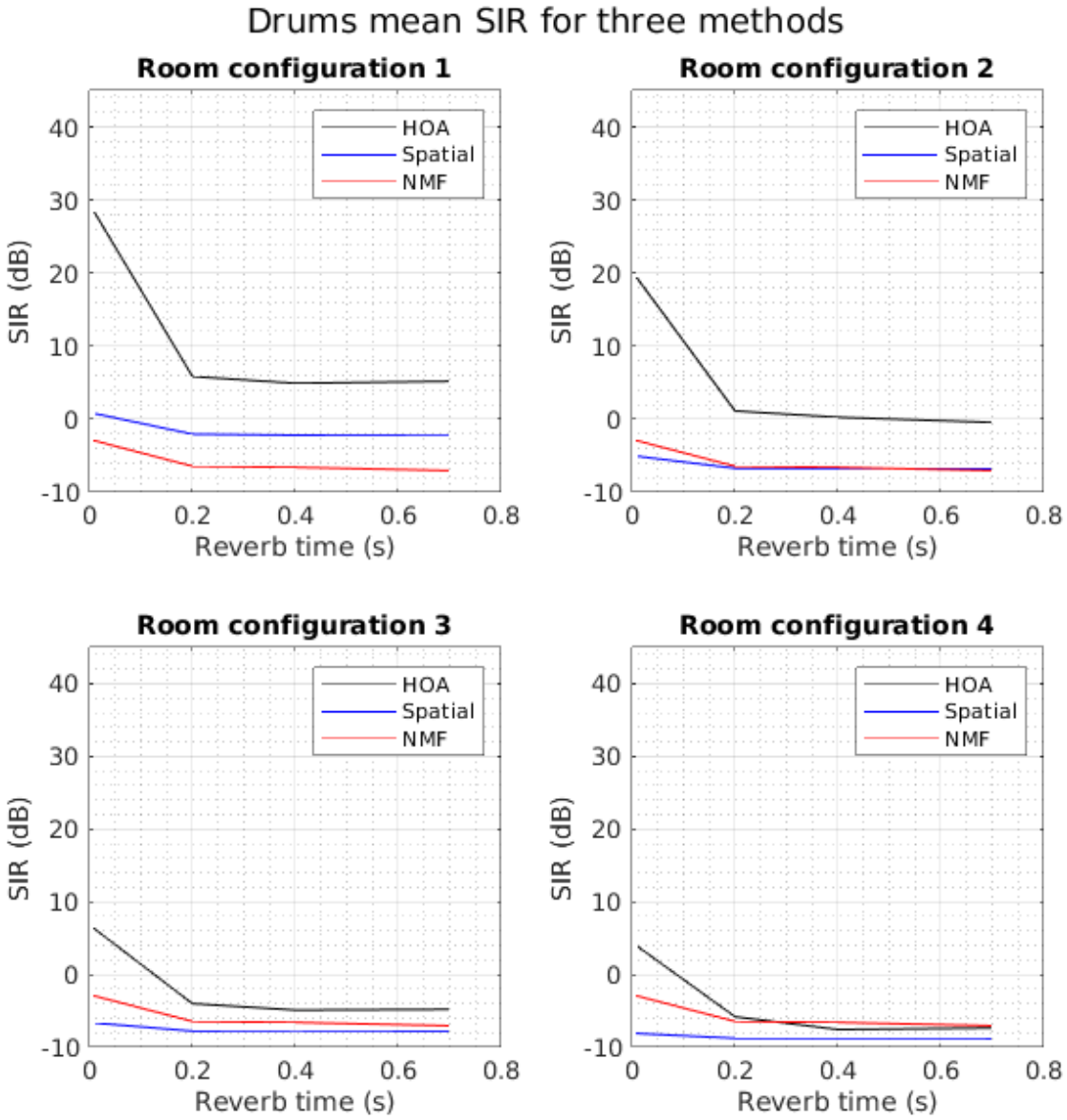

Figure 31: Source to interference ratios of drums for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

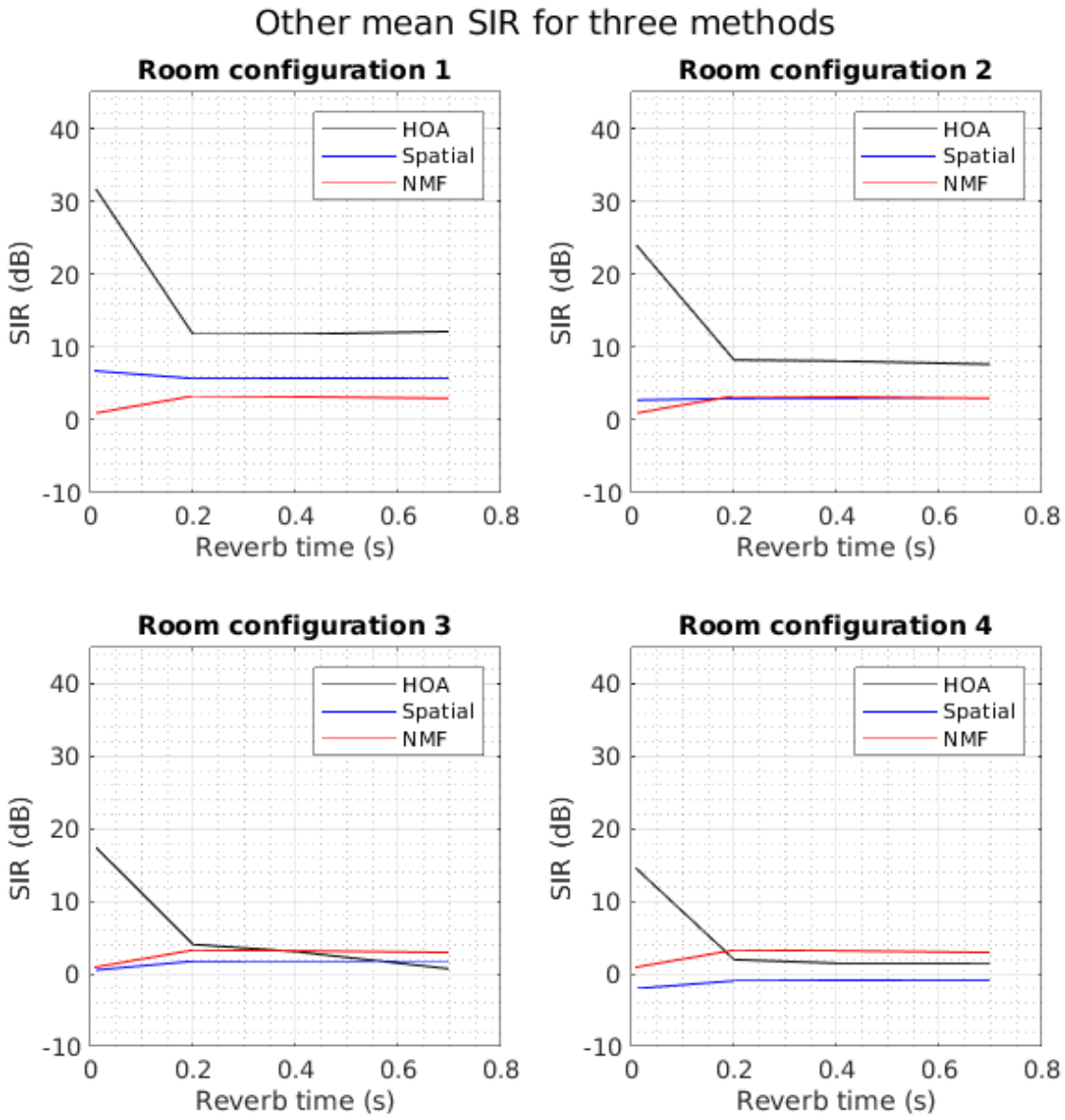

Figure 32: Source to interference ratios of "other" for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

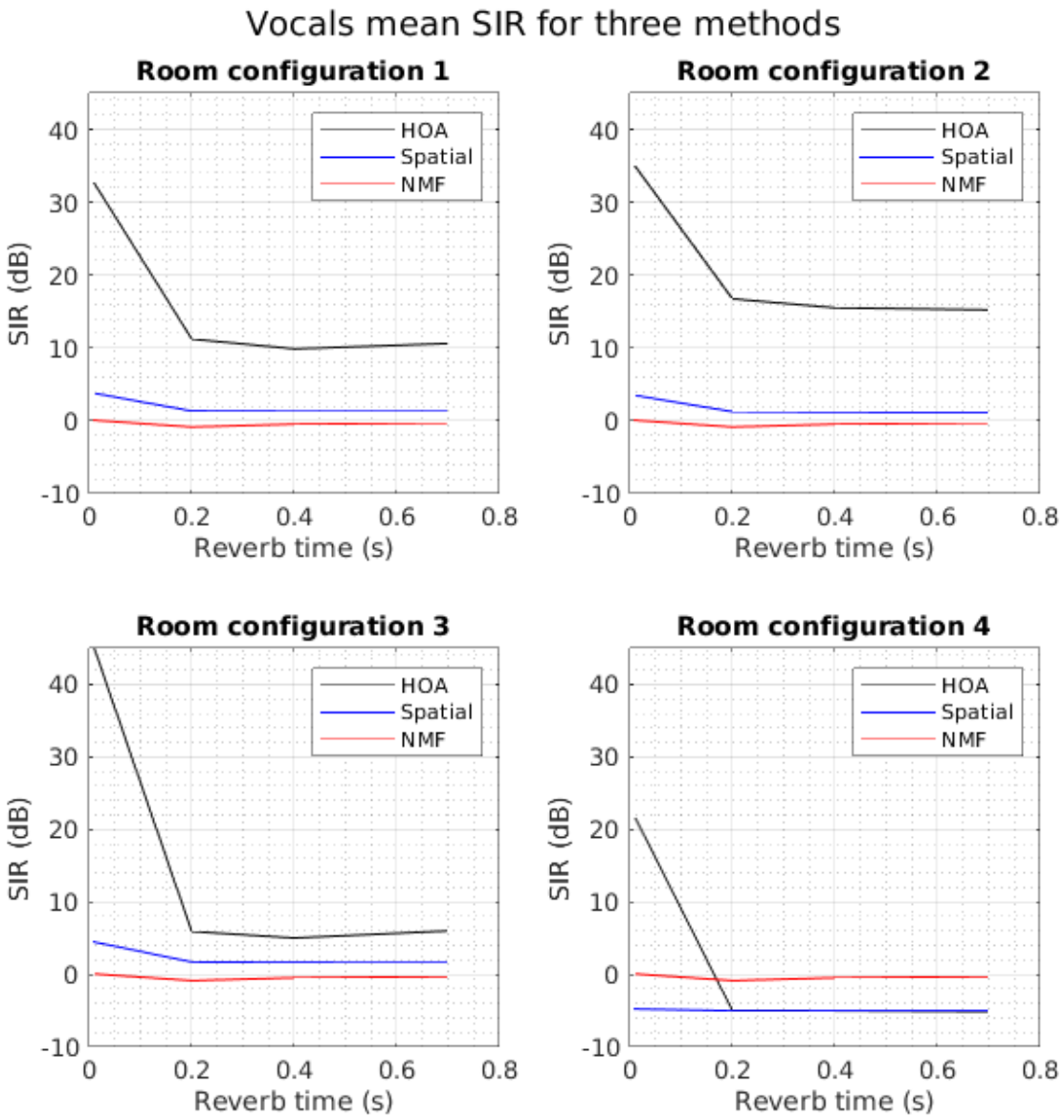

Figure 33: Source to interference ratios of vocals for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

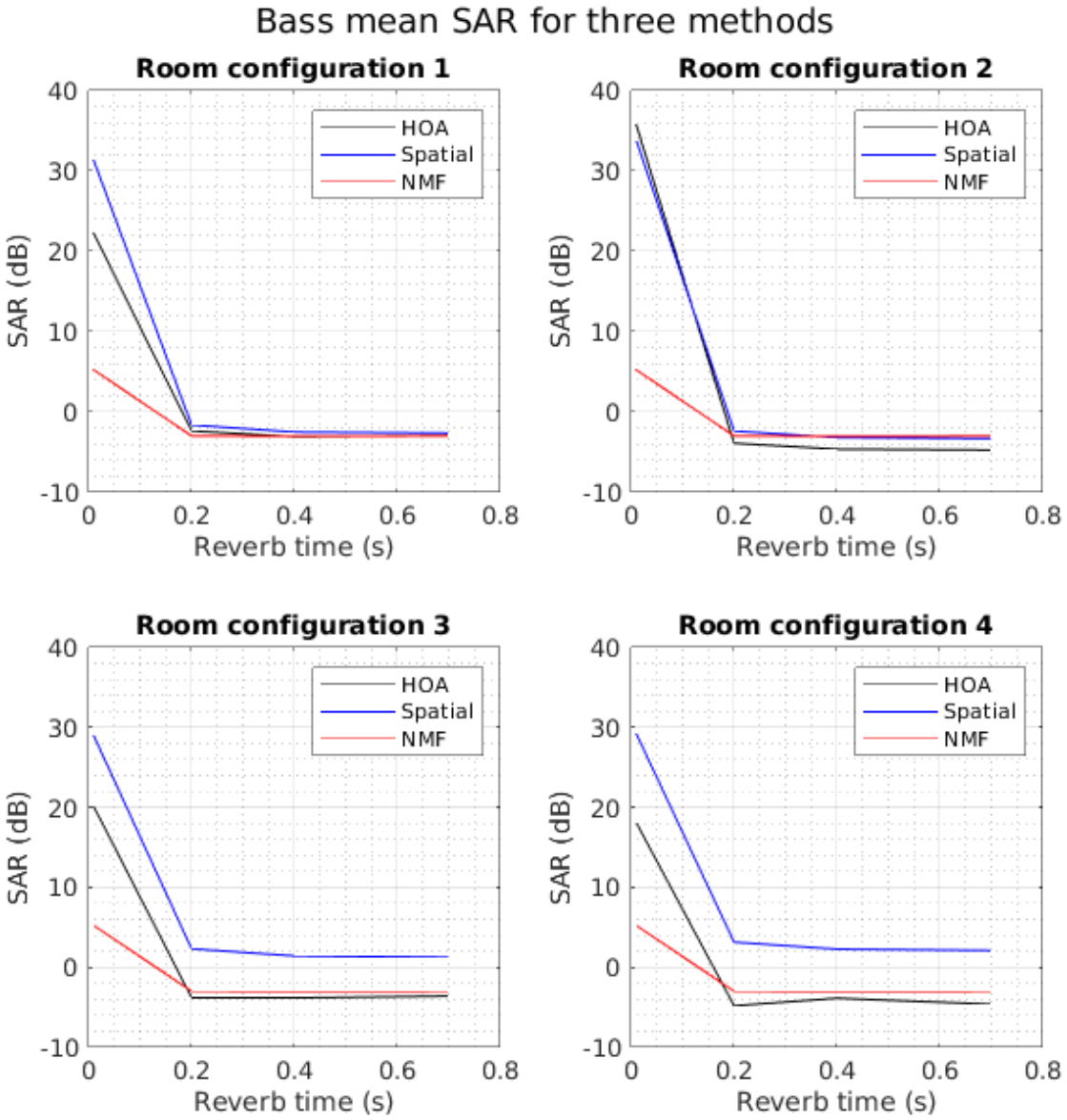

Figure 34: Source to artifact ratios of bass for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

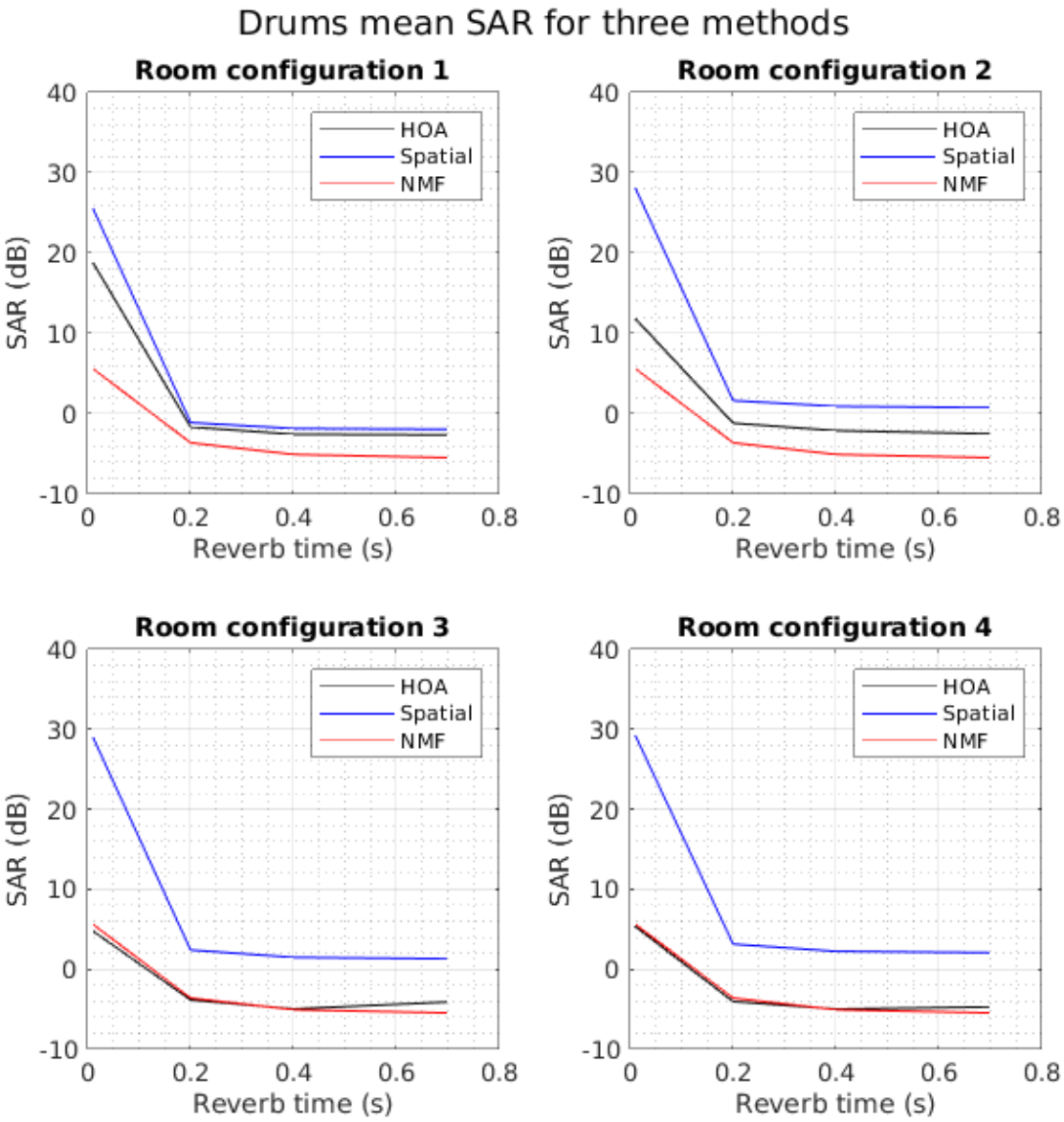

Figure 35: Source to artifact ratios of drums for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

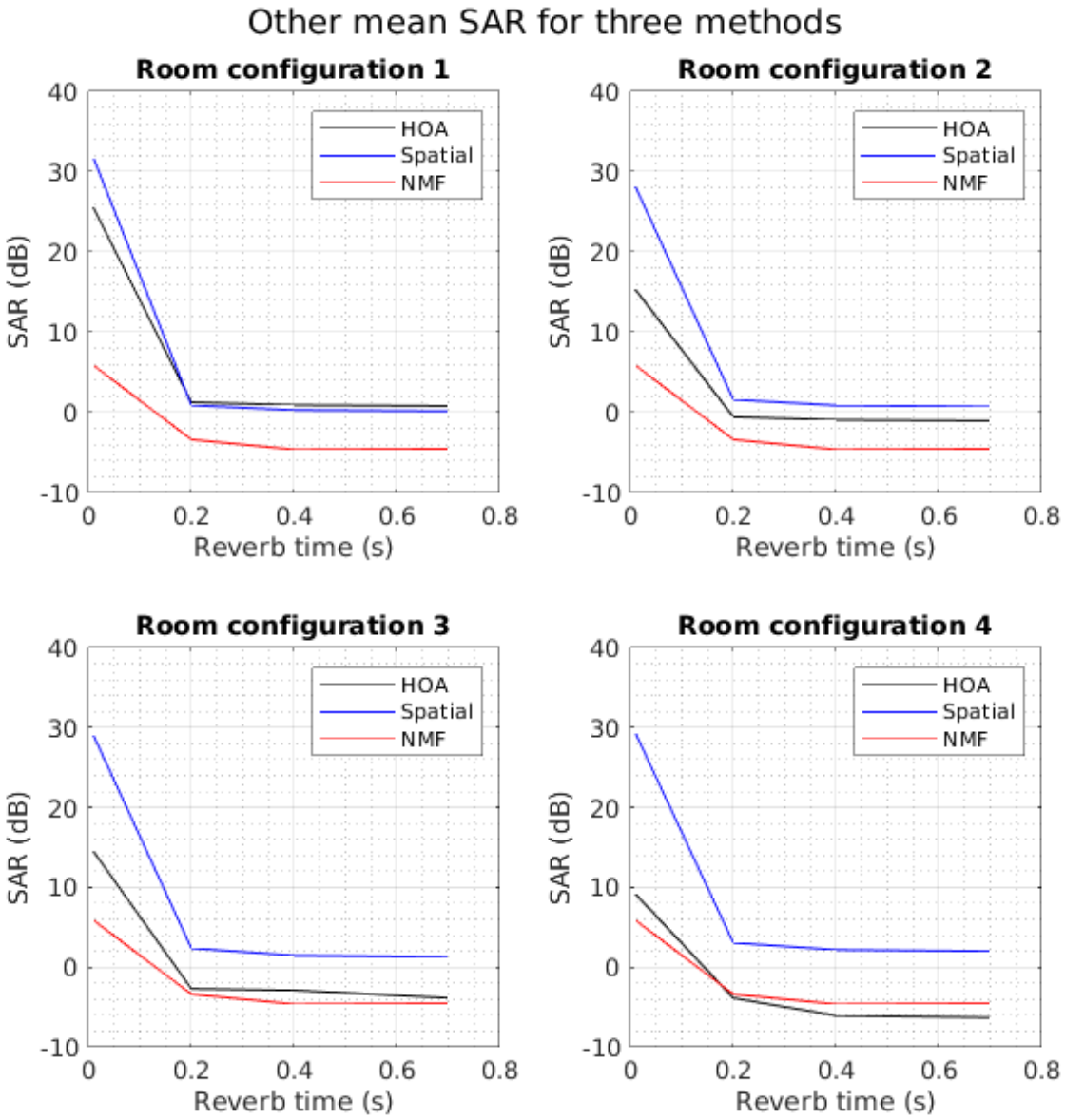

Figure 36: Source to artifact ratios of other for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

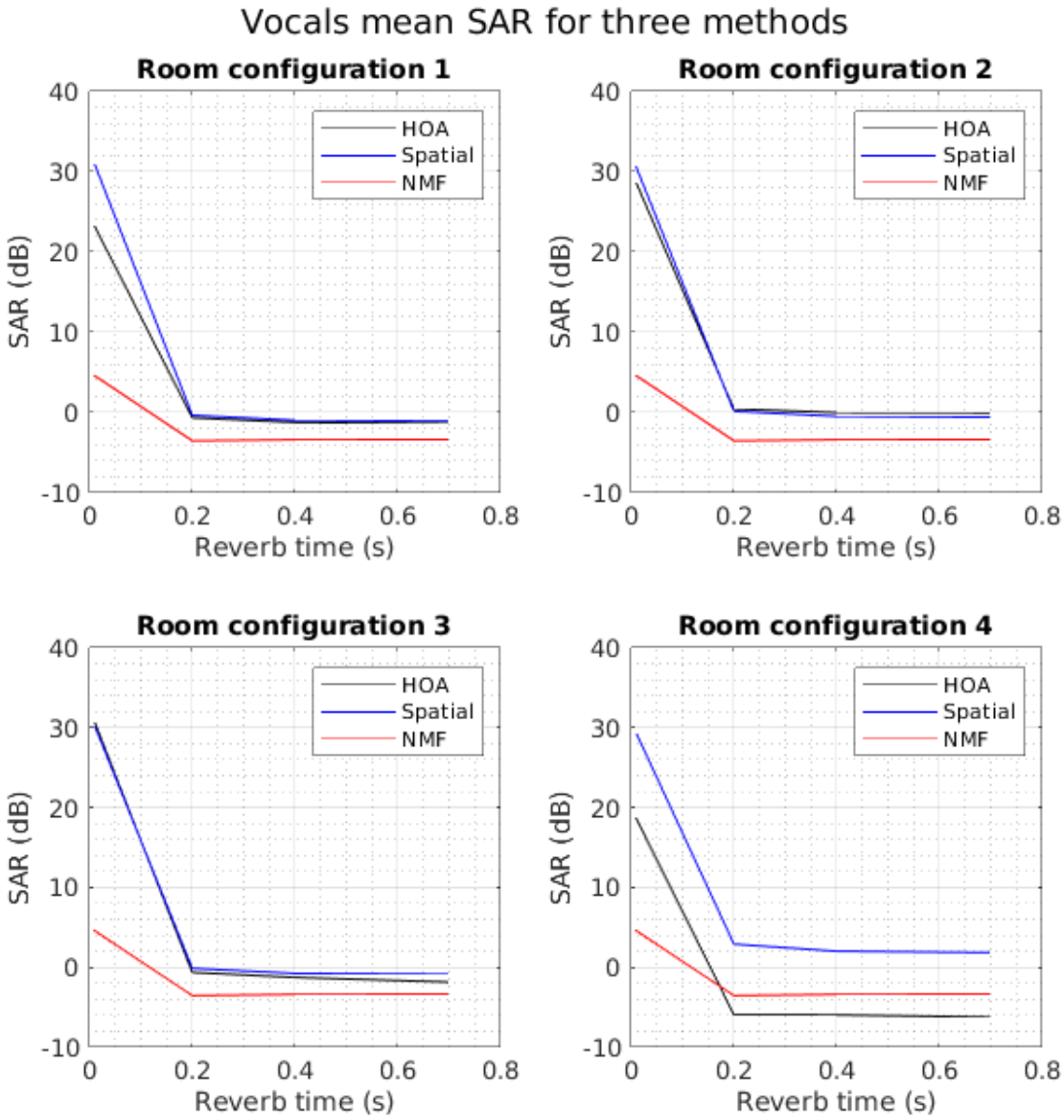

Figure 37: Source to artifact ratios of vocals for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method.

5.4 Computation Time Performance

The real time factors of the three separation and two DOA estimation methods are presented in Table 5.

Table 5: Real time factors of the methods

Separation method	**Real time factor**
Ambisonic domain	29.5978
NMF	1.0638
Spatial	0.0097572
DOA estimation method	**Real time factor**
MUSIC	1.3324
Intensity	0.086803

Of particular note is the worse performance of the ambisonic domain filtering method; although it tends to (but does not always) outperform the other two in terms of separation scores, it does so at the cost of computation time. In the given testing configuration, the NMF and MUSIC algorithms both work at roughly (but slower than) real time, but it is worth noting that a real-world system requires both DOA estimation and source separation; an implementation combining the two would work at roughly double real time. Finally, the proposed method combining spatial filtering and intensity vector statistics DOA estimation are both much faster than real time; even in combination their real time factor is only 0.0966.

Chapter 6

Conclusions

The low-level discussions related to the individual experiments are presented in the previous section; we move directly to the high-level conclusions.

6.1 Conclusions

Looking at the methods this thesis investigated and the results they delivered, it seems fair for us to define a continuum based on computational complexity along which DOA estimation and source separation algorithms (or any algorithm, for that matter) lie. At the far end (most computationally intensive) clearly lies the ambisonic domain approach of [2], and at the opposite end the novel methods proposed in this thesis. In evaluating the trade-offs along this continuum, we must consider algorithm performance, obviously, but also application to the real-world setting. In terms of performance, we willingly concede that the ambisonic domain approach is best: while NMF and spatial filtering outperform it in select cases, it never performs vastly worse, and at many points performs significantly better. While its SAR scores are lower than those of the spatial approach, this tradeoff is both mathematically and subjectively compensated by its SIR scores.

However, computation time can be important: in a real-time application, such as applying separation to live signals before mixing and transmitting them, for instance,

the user would expect near real-time performance. In evaluating DOA estimation and separation methods, we must then ask what level of performance we truly require, and how much computation time we will sacrifice to achieve it. Figures 38 and 39 illustrate this point with the methods tested in this thesis.

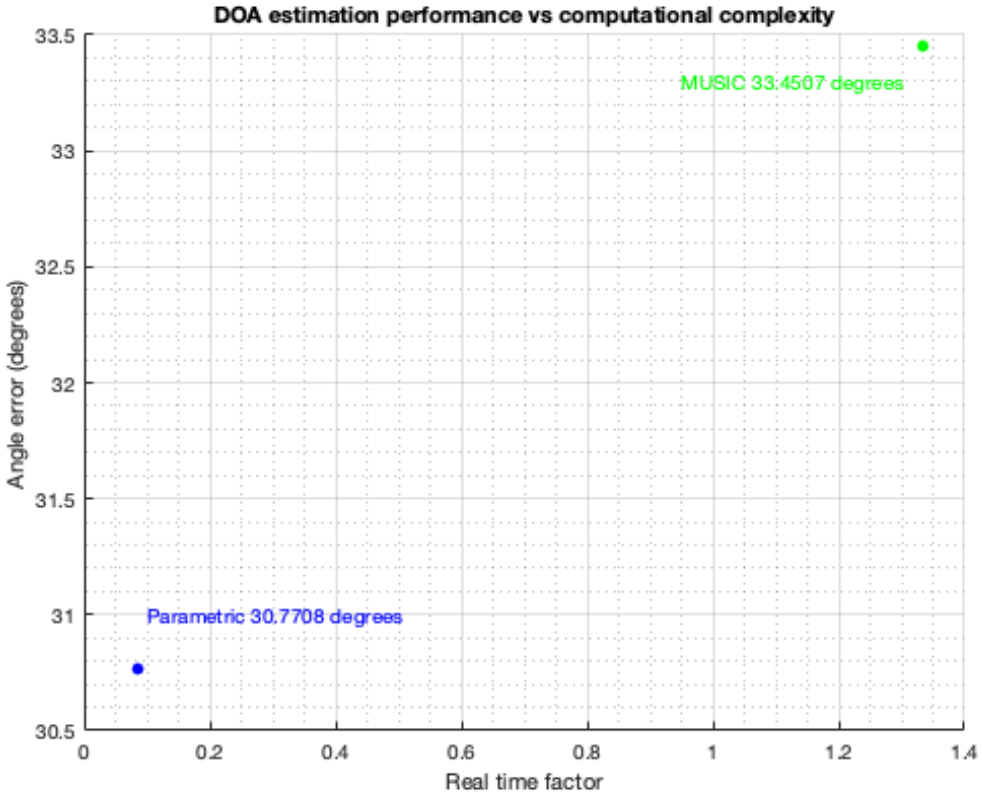

Figure 38: Average performance of the DOA estimation algorithms on the two source case as a function of their computation times.

.

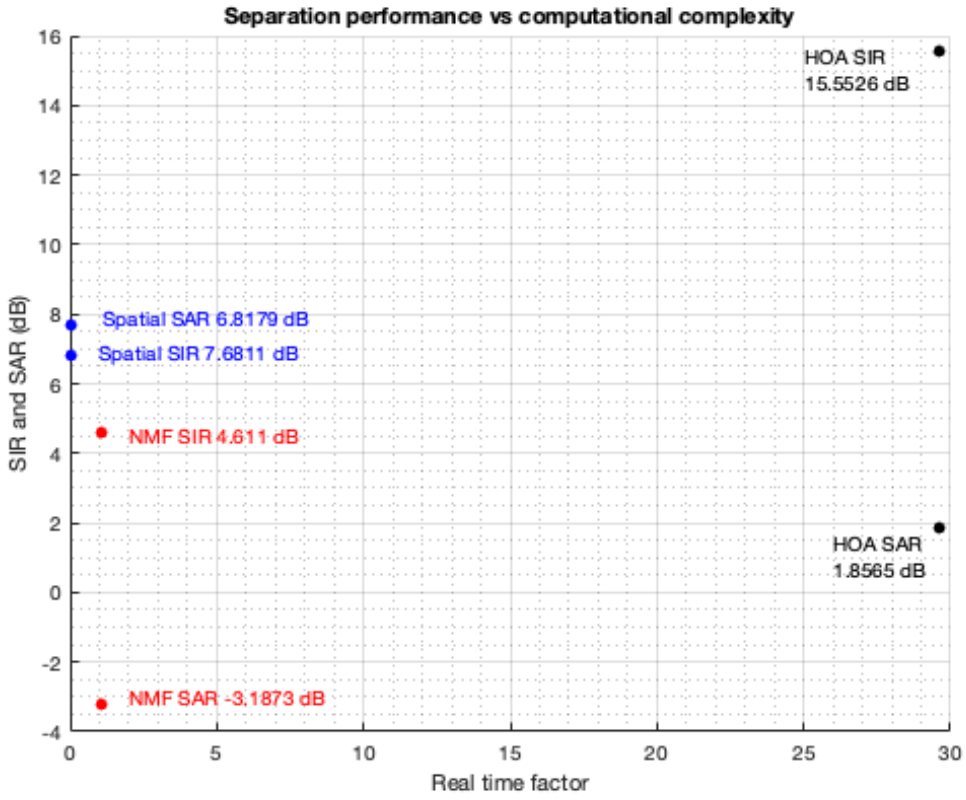

Figure 39: Average performance of the source separation algorithms on the two source case as a function of their computation times.

.

6.2 This Thesis' Contributions

This thesis proposes a novel method to solve the musical blind source separation problem using ambisonics in a time-efficient approach. In terms of computational complexity and SAR, it vastly outperforms the existing methods of NMF and ambisonic domain separation. In terms of SIR, it outperforms NMF except in cases when sources are close together, and outperforms ambisonic domain separation in some higher reverberation cases. The proposed DOA estimation algorithm is significantly less complex than the MUSIC algorithm it is compared to, and performs roughly comparably at all angular separations; its average error is superior to that of the MUSIC method. Overall, this thesis' method provides an extremely computationally efficient DOA estimation and source separation at the cost of decreased performance.

6.3 Future work

Several areas of the proposed method could be improved with further fine-tuning. Most notably, the DOA estimator should be improved to be more robust to the 180° error induced by early reflections off of the opposite wall. The parameters to the DOA estimator could also be further refined. More diverse beamformer shapes could be tested to potentially improve performance in specific cases. Additionally, a larger testing dataset should be explored, moving beyond pop music and considering the higher reverberation times common in orchestra halls and other spaces.

In terms of larger scale research questions that this thesis raises, one of the thesis' goals is that the question of computational complexity versus performance be brought more into the spotlight. It is increasingly common to see approaches proposed that can be highly accurate, but that require significant computation time (or even unaddressed computation time in the case of many papers). In the context of musical blind source separation, this thesis has shown that computationally simpler approaches can reach and, in certain circumstances, exceed the performance levels of state of the art approaches without the computation time cost. As a concluding thought, this thesis proposes that these considerations play a larger role in future algorithms both for source separation and for general computation tasks.

List of Figures

1 Typical room impulse response (RIR). The direct sound (with a height just over 0.06) reaches the receiver first, followed by the early echoes (roughly from 0.01 to 0.1 seconds), and reverberance (roughly 0.1 seconds to the end of the impulse response). 5

2 Virtual microphones in first-order ambisonic recording. These correspond to one omni-directional channel W (0), and three oriented channels X (3), Y (1), and Z (2). Lighter lobes indicate that recorded signals are in phase, while darker lobes indicate counterphase. Image: "File:Spherical Harmonics deg5.png" by Dr Franz Zotter <zotter@iem.at> is licensed under CC BY-SA 3.0 . With alterations. . . . 7

3 Left: Beampattern of basic beamformer for source at 0°. Right: Gain of basic beamformer as a function of angular distance away from the estimated DOA. 12

4 Left: Beampattern of MaxRE beamformer for source at 0°. Right: Gain of MaxRE beamformer as a function of angular distance away from the estimated DOA. 12

5 Left: Beampattern of in-phase beamformer for source at 0°. Right: Gain of in-phase beamformer as a function of angular distance away from the estimated DOA. 13

6 Spectrogram of audio signal to be transcribed; the signal consists of two notes that randomly switch on and off. Adapted from [9]. 14

7 Result of performing NMF on the spectrogram of Figure 6. H controls when each of the two notes is active and W controls the pitch of each note. Adapted from [9]. 15

8 Result of performing NMF deconvolution on a simple music signal spectrogram. We assume the high and low frequency sweeps correspond to two separate sources, and then each template of W represents one source. Adapted from [10]. 16

9 Direction of arrival (DOA) histogram with 0.7 diffuseness threshold. This is a single source case, and using a peak-picking algorithm would likely indicate that the source is at roughly 0 radians direction of arrival angle. The 180° peak is due to early reflections. 20

10 Flow diagram of novel approach. The combined DOA estimation and separation technique combines components from Sections 2.2, 2.3.1, and 2.4. Black boxes represent processes and red boxes represent variables. 23

11 The two source case: the drums source remains at 0° azimuth while the vocals source moves from 0° (top image) to 180° azimuth in 10° increments (bottom image). The red circle represents the receiver and stars the sources. All elements have 0° elevation. Reproduced from [2] with alterations. 27

12 Source configurations as used in [2], which make up the four source test case. The red circle represents the receiver and stars the sources. All elements have 0° elevation. Reproduced from [2]. 28

13 Average DOA error (black) and standard deviation (red) across 10 songs using the proposed parametric values estimation method. . . . 34

14 Average DOA error (black) and standard deviation (red) across 10 songs using MUSIC estimation method. 35

15 Average DOA errors of proposed and MUSIC estimation methods. . 36

16 Signal to interference ratio of drums track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line. 38

17 Signal to interference ratio of vocals track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line. 39

18 Signal to artifact ratio of drums track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line. 40

19 Signal to artifact ratio of vocals track in optimized beamformer shape experiment. The average ratios in decibels across the 10 songs are displayed in black, and the standard deviations in red. For reference, the beamformer's gain is shown by the dotted blue line. 41

20 Signal to interference ratios of drum and vocal tracks for two source case evaluated with NMF. The average ratios in decibels are in black and the standard deviations in red. 43

21 Signal to artifact ratios of drum and vocal tracks for two source case evaluated with NMF. The average ratios in decibels are in black and the standard deviations in red. 44

22 Signal to interference ratios of drum track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red. 45

23 Signal to interference ratios of vocal track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red. 46

24 Signal to artifact ratios of drum track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red. 47

25 Signal to artifact ratios of vocals track for two source case evaluated with ambisonic domain filtering method. The average ratios in decibels are in black and the standard deviations in red. 48

26 Signal to interference ratios of drum tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. . . . 49

27 Signal to interference ratios of vocal tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. . . . 50

28 Signal to artifact ratios of drum tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 50

29 Signal to artifact ratios of vocal tracks for two source case evaluated with all three methods. The average ratios in decibels are in black and the standard deviations in red. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 51

30 Source to interference ratios of bass for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 53

31 Source to interference ratios of drums for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 54

32 Source to interference ratios of "other" for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 55

33 Source to interference ratios of vocals for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 56

34 Source to artifact ratios of bass for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 57

35 Source to artifact ratios of drums for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 58

36 Source to artifact ratios of other for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 59

37 Source to artifact ratios of vocals for the four source case evaluated with all methods. "HOA" is the ambisonic domain filtering method and "Spatial" is the proposed method. 60

38 Average performance of the DOA estimation algorithms on the two source case as a function of their computation times. 63

39 Average performance of the source separation algorithms on the two source case as a function of their computation times. 64

List of Tables

1 Beamforming patterns described in [6] 11

2 End-to-end separation parameters 23

3 NMF implementation parameters 25

4 DSD100 songs used 26

5 Real time factors of the methods 61

Bibliography

[1] Gannot, S., Vincent, E., Markovich-Golan, S. & Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Transactions on Audio Speech and Language Processing* **25**, 692–730 (2017).

[2] Hafsati, M. *et al.* Sound source separation in the higher order ambisonics domain. In *International Conference on Digital Audio Effects*, 1–7 (HAL, Birmingham, United Kingdom, 2019).

[3] Pulkki, V., Politis, A., Del Galdo, G. & Kuntz, A. Parametric spatial audio reproduction with higher-order B-format microphone input. *134th Audio Engineering Society Convention 2013* 830–839 (2013).

[4] Carpentier, T. Normalization schemes in ambisonic: Does it matter? *142nd Audio Engineering Society International Convention 2017, AES 2017* **0** (2017).

[5] Pulkki, V. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc* **55**, 503–516 (2007). URL `http://www.aes.org/e-lib/browse.cfm?elib=14170`.

[6] Daniel, J. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Ph.D. thesis, l'Université Paris (2001).

[7] Günel, B., Hachabiboğlu, H. & Kondoz, A. M. Acoustic source separation of convolutive mixtures based on intensity vector statistics. *IEEE Transactions on Audio, Speech and Language Processing* **16**, 748–756 (2008).

[8] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

[9] Smaragdis, P. & Brown, J. C. Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* **2003-Janua**, 177–180 (2003).

[10] Smaragdis, P. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **3195**, 494–499 (2004).

[11] Liutkus, A. & Badeau, R. Generalized Wiener filtering with fractional power spectrograms. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2015-Augus**, 266–270 (2015).

[12] Uhlich, S., Giron, F. & Mitsufuji, Y. Deep neural network based instrument extraction from music. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2015-Augus**, 2135–2139 (2015).

[13] Chandna, P., Miron, M., Janer, J. & Gómez, E. Monoaural audio source separation using deep convolutional neural networks. In Tichavský, P., Babaie-Zadeh, M., Michel, O. J. & Thirion-Moreau, N. (eds.) *Latent Variable Analysis and Signal Separation*, 258–266 (Springer International Publishing, Cham, 2017).

[14] Nugraha, A. A., Liutkus, A. & Vincent, E. Multichannel Audio Source Separation With Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **24**, 1652–1664 (2016).

[15] Politis, A., Delikaris-Manias, S. & Pulkki, V. Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015).

[16] Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation* **34**, 276–280 (1986).

[17] Dittmar, C., Audio, I. & Erlangen, L. NMF Toolbox : Music Processing Applications of Nonnegative Matrix Factorization 1–8 (2019).

[18] Ozerov, A., Vincent, E. & Bimbot, F. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech and Language Processing* **20**, 1118–1133 (2012).

[19] Liutkus, A. *et al.* The 2016 Signal Separation Evaluation Campaign. In Tichavsky, P., Babaie-Zadeh, M., Michel, O. J. & Thirion-Moreau, N. (eds.) *SiSEC16*, 323–332 (Springer International Publishing, Cham, 2017).

[20] Vincent, E., Gribonval, R. & Fevotte, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 1462–1469 (2006).

[21] Miron, M. *Source Separation Methods for Orchestral Music: Timbre-Informed and Score-Informed Strategies*. Ph.D. thesis, Universitat Pompeu Fabra (2017).