Friedrich-Alexander-Universität Erlangen-Nürnberg

Friedrich-Alexander-Universität
Erlangen-Nürnberg

Master Thesis

Sound Source Localization with the Rotating Equatorial Microphone (REM)

submitted by

Jeremy Lawrence

submitted

May 2, 2023

(revised on June 1, 2023)

Supervisor / Advisor

Prof. Dr. Nils Peters

Prof. Dr. Jens Ahrens

Reviewers

Prof. Dr. Nils Peters

AUDIO LABS

International Audio Laboratories Erlangen

A joint institution of the
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and
the Fraunhofer-Institut für Integrierte Schaltungen IIS.

Fraunhofer IIS

Erklärung

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

Erlangen, June 2, 2023

Jeremy Lawrence

Acknowledgements

I would like to express my gratitude towards the FAU LMS Chair and Dr.-ing. Heinrich Löllmann in particular for kindly allowing me to use their anechoic room. A debt of gratitude is also owed to Thomas Haubner for his personal assistance in preparing the practical experiment for this thesis. Without your help it would not have been possible to verify the theoretical concepts outlined in this thesis in practice.

Additionally, I would like to thank my ASC mentor Prof. Emanuël Habets for his kind supervision throughout my studies, as well as bringing my attention to this research topic.

Furthermore I would like to thank Prof. Jens Ahrens for laying the groundwork which sparked the ideas investigated in this thesis, as well as assisting the creation of the rotating equatorial microphone prototype.

Finally, I owe a debt of gratitude to my supervisor and mentor, Prof. Nils Peters, who not only immensely contributed towards the creation of the microphone prototype utilized in this thesis, but also greatly assisted in every step of the creation of this thesis. Without your humorous, kind and thoughtful supervision this journey would only have been half as enjoyable.

Abstract

Analyzing spatial sound fields typically necessitates a large number of microphones placed at specific spatial points. The usage of moving microphones provides an alternative method for analysis, potentially reducing the required number of microphones. To investigate the potential of a single fast moving microphone we previously developed the rotating equatorial microphone prototype. The circular motion of this microphone introduces distortions into recorded sound sources which can be used to determine their direction of arrival (DOA).

In this thesis an algorithm is derived which compensates these distortions for arbitrary sound sources given their DOA. Additionally, a metric is introduced which quantifies the distortions present in an audio recording. Subsequently, arbitrary sound sources with an unknown DOA are localized by applying the algorithm for various DOA guesses and determining the DOA with minimum distortion using the previously defined metric. With this approach we localized four simultaneous wideband audio signals in 2D space and two wideband audio signals in 3D space with $\pm 10^\circ$ accuracy in simulations. We then verified the approach in practice for one and two simultaneous audio sources. We found that accurate localization was only possible in 2D space for audio sources mainly consisting of constant tones in the mid frequency range.

Contents

Chapter 1

Introduction

Analyzing spatial sound fields to perform applications such as beamforming, source localization and blind source separation typically necessitates a large number of microphones, known as a microphone array. Conventionally these arrays consist of non-coincident microphones placed at various selected spatial points or coincident microphones with a non-omnidirectional directivity. A well known microphone array is mh acoustics' em32 Eigenmike®, which is a spherical microphone array consisting of 32 microphones placed on a rigid sphere with an 8.4 cm diameter [24]. Due to its small form factor and large number of microphones it is highly capable of a multitude of spatial audio applications. However, one major disadvantage is the high cost associated with the large number of microphones. This has lead the authors of [2] to devise their so-called equatorial microphone array (EMA), which has 17 microphones placed along the equator of a rigid sphere with a 17.5 cm diameter. Although the number of microphones is smaller than that of the Eigenmike, a spherical microphone array would require at least 81 microphones to evaluate the sound field as accurately as the EMA, as long as we can assume the sound field to be height invariant.

The conception of the EMA sparked another idea, which has the potential to greatly reduce the number of required microphones even further: Instead of placing many stationary microphones around the equator of a sphere, instead it could be possible to approximate the data captured by the EMA utilizing only one microphone which rapidly moves along the equator. Naively this could be achieved by setting the sampling rate of the moving microphone to 17 times that of the EMA and performing a full rotation between every sample of the EMA. For a typical sampling rate of 48 kHz, however, this would require 48000 microphone rotations per second (RPS), which is highly infeasible in practice. Alternatively, an attempt could be made to use the data captured by the moving microphone to interpolate the sound field at the microphone positions of the EMA. This, however, requires additional information about the sound waves captured by the moving microphone, such as their direction of arrival (DOA). This motivates the topic of this thesis: A

constant, circular microphone movement introduces periodic, DOA-dependent distortions into the recorded sound. If we can estimate the DOA of incoming (plane) sound waves from these distortions, it should allow us to approximate the sound field around the equator of a rigid sphere, enabling us to use EMA sound processing algorithms while only requiring one microphone.

To investigate the potential of sound field analysis using a fast rotating microphone in practice, we previously built the rotating equatorial microphone (REM) prototype described in [22]. The microphone was designed to reach speeds of up to 1000 RPS in order to capture various points along the equator at similar points in time, however, we quickly realized that beyond 40 RPS the signal captured by the microphone becomes unusable due to motor and especially wind noise. Fortunately, inspecting the audio signals captured by the REM sparked an idea, potentially allowing for DOA estimation at much lower speeds, upon which we will expand later.

The following chapter will cover the current state of the art in sound field analysis using moving microphones, particularly in regard to DOA estimation. Chapter 3 will give a detailed description of the distortions introduced into a microphone recording due to circular movement, as well as how these distortions can be compensated using two different methods. In Chapter 4 the findings from the previous chapter will be used to perform DOA estimation in a simulated environment under various conditions. Finally, Chapter 5 will investigate the accuracy of the proposed DOA estimation algorithm in practice by performing sound source localization using the REM for various sound sources and locations in an anechoic chamber.

Chapter 2

Literature Review

Spatial sound field analysis using microphone arrays has been studied very comprehensively, whereas research using one or multiple moving microphones is scarce. The limited research that has been conducted mainly addresses the measurement of room impulse responses and head-related transfer functions with a given known source signal. There are two main approaches, those that we categorize as using slow microphone movement such that there is only a marginal distortion of the recorded sound due to the movement and those utilizing fast microphone movement where there is significant distortion. Although we do not intend to determine room impulse responses in this thesis, we will nonetheless give an overview of these approaches, since the used signal processing ideas may prove helpful for DOA estimation and future research. Section 2.1 and Section 2.2 cover room impulse response measurement for slow and fast moving microphones, respectively. In Section 2.3 we will cover the limited research that has been conducted on DOA estimation using moving microphones, as well as the limitations of these approaches.

2.1 Room Impulse Response Measurement Using Slow Microphone Movement

Impulse responses characterize a system's behaviour when presented with a very brief input signal, known as an impulse. An interesting property of impulse responses is that the output of a time-invariant system can be computed by convolving its input with its impulse response [4]. The same holds for time-variant systems, with the only difference being that the impulse response varies over time. If we want to obtain the way sound changes as it travels from one point in a room to another, we can do this by computing the impulse response between the two points and subsequently convolving it with an arbitrary input. Such an impulse response is known as a

room impulse response (RIR) as it characterizes the acoustic properties of the room. To allow us to, for example, virtually place sound sources at arbitrary points in a room and obtain the way they sound at a given listener's position we require a very large number of RIRs. These can be obtained by either placing a microphone array with potentially hundreds of microphones in the room, or utilizing a small number of microphones and moving them after obtaining each RIR. These methods are either very costly or very time-consuming, making the use of a moving microphone appealing, as it could significantly reduce the time required while keeping the cost low.

The authors of [3] investigated the possibility of obtaining RIRs along a given circular microphone trajectory without requiring the microphone to stop during acquisition. The main idea lies in the reconstruction of a two-dimensional spectrum for both spatial and temporal frequencies from the one-dimensional frequency spectrum of the signal captured by the microphone, given a known excitation signal. This two-dimensional spectrum can then be divided by the spectrum of the excitation signal to obtain the spectrum of the impulse responses at various spatial points, from which the impulse responses can be obtained by taking the inverse Fourier transform. To obtain the two-dimensional spectrum, the excitation signal needs to be chosen such that it contains many frequencies which are all spaced sufficiently far apart that any Doppler frequency shift due to the movement doesn't cause their spectra to overlap. The distance between higher frequencies therefore needs to be greater, since the Doppler shift is proportional to the frequency, i.e. higher frequencies get shifted by a larger amount. If this condition holds, any frequency present in the recorded signal can be uniquely projected onto the two-dimensional spectrum. This concept is illustrated in Figure 2.1, where ω represents the frequency axis, l_θ the spatial frequency axis, ω_1 and ω_2 are the two largest frequencies present in the excitation signal and $\vec{v}$ is the angular velocity of the circular movement. The recorded Doppler shifted frequencies are initially on the ω-axis and can be uniquely projected onto the true frequencies ω_1, ω_2, etc. following the direction

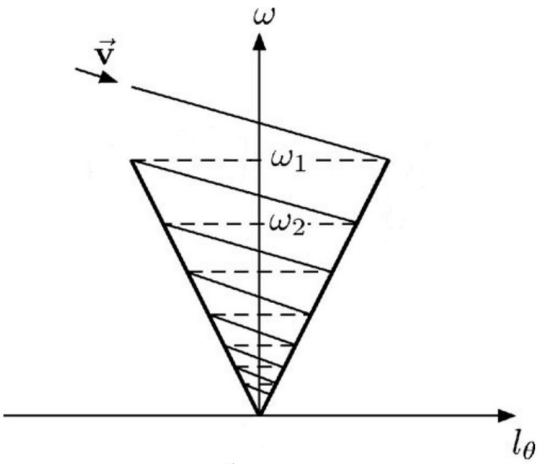


Figure 2.1: Projection of 1D spectrum onto 2D spectrum (image modified from [3]).

of $\vec{v}$. This projection removes the Doppler shift and provides us with spatial information.

For this approach there is an additional constraint regarding the maximum angular velocity of the microphone which may not exceed $v_{max} = \frac{\pi c}{r(\omega_1 T - \pi)}$, where c is the speed of sound, r the radius of the circular motion and T is the RIR length. This leads to very slow microphone movement on the order of several minutes for one microphone rotation.

Initially this approach looks promising for DOA estimation as it allows us to gather spatial information from recorded sounds by projection onto a two-dimensional spectrum. However, this is only possible if the source signals are known and their frequencies sufficiently spaced. Ideally our DOA estimation should be able to locate arbitrary, unknown sound sources, which this approach does not allow for. Furthermore the maximum allowable speed of the microphone is very limited, making it impossible to sample the sound field along the equator of a sphere at similar points in time.

The authors of [9] followed a different approach. The initial setup is similar to the approach discussed above and differs only in the utilized excitation signal, which is chosen to be a periodic perfect sequence $\psi(n)$ with period N. Perfect sequences have an ideal autocorrelation, i.e. the autocorrelation is an arbitrary value (usually 1) when the signal overlaps perfectly with itself and 0 otherwise.

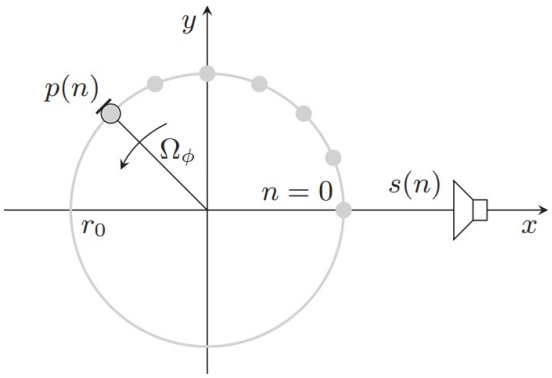


Figure 2.2: Setup for RIR measurement (image from [9]).

The locations at which the RIRs are to be determined are the grey points shown in Figure 2.2, which are also the sampling positions of the microphone. Here $s(n)$ is the source signal, $p(n)$ the recorded signal at discrete time and position n, r_0 is the radius and Ω_ϕ the angular velocity of the circular microphone movement. Using a periodic perfect sequence $\psi(n)$ as a source signal allows for the k-th coefficients of the impulse responses at all positions n to be expressed as $h(k,n) = \sum_{m=0}^{N-1} a_m(n)\psi(-k+m)$, where $a_m(n)$ is an expansion coefficient and the impulse responses are assumed to be shorter than N. The expansion coefficients can be obtained as $p(n) = a_{(n \bmod N)}(n)$. If we choose $N = 4$ this means we only sample $\frac{1}{4}$-th of the required expansion coefficients, as shown in Figure 2.3. This, however, does not pose a problem if the

spatial sampling points are placed sufficiently close to each another, i.e. if the microphone moves slowly enough, since we can obtain the missing samples from neighbouring coefficients using linear or sinc interpolation.

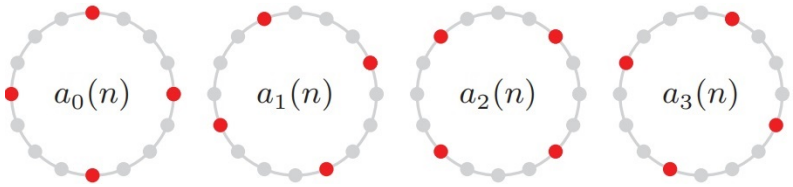


Figure 2.3: Observed expansion coefficients for $N = 4$ (image modified from [9]).

The authors of this approach also investigated the case in which the microphone does not move uniformly in [10], as it is difficult to achieve perfectly uniform movement in practice. Non-uniform movement results in uneven spacing of the sampling points from Figure 2.3, making interpolation more challenging. This is circumvented by using Lagrange interpolation instead of sinc interpolation.

Interestingly, the maximum allowable speed of this approach is very similar to the approach from [3], even though the methodology is quite different, once again making it impossible to sample different points around the equator of a sphere at similar points in time. Additionally, this approach requires a very specific, known excitation signal, making it challenging to use any of its signal processing ideas for DOA estimation.

2.2 Room Impulse Response Measurement Using Fast Microphone Movement

Contrary to the approaches from the previous section, the approach from [16] imposes no constraints regarding the excitation signal and the speed of the microphone movement. Here a Cartesian grid G with equidistant sampling points is defined where the RIRs are computed at each point of the grid. The spacing Δ between the points is chosen sufficiently small to prevent spatial aliasing, i.e. $\Delta < \frac{c}{2f_c}$, where c is the speed of sound and f_c is the maximum frequency to be considered. The microphone moves along an arbitrary tracked path $\boldsymbol{r}(n)$ and captures the signal

$$x(n) = \sum_{k=0}^{L-1} h(\boldsymbol{r}(n), k) \cdot s(n-k) + \eta(n) \ , \tag{2.1}$$

where $s(n)$ is the source signal, $\eta(n)$ is the measurement noise and $h(\boldsymbol{r}(n), k)$ is the k-th coefficient

of the RIR of length L between the source and position $\boldsymbol{r}(n)$. The main idea is to interpolate the RIRs $h(\boldsymbol{r}(n), k)$ from the RIRs on the grid as

$$h(\boldsymbol{r}(n), k) = \sum_{\boldsymbol{g} \in G} \varphi(\boldsymbol{r}(n), \boldsymbol{r_g}) \cdot h(\boldsymbol{g}, k) \ , \tag{2.2}$$

where $\varphi(\boldsymbol{r}(n), \boldsymbol{r_g})$ is the interpolation coefficient for the displacement $\boldsymbol{r}(n) - \boldsymbol{r_g}$ between the microphone position and the grid positions $\boldsymbol{r_g}$. If $s(n)$ is known and $\boldsymbol{r}(n)$ is perfectly tracked then a system of linear equations can be constructed, allowing the reconstruction of all the RIRs, given enough samples $x(n)$.

The authors found that the accuracy of this approach degraded close to the boundary of the grid, hence they utilized Lissajous trajectories in conjunction with Lagrange interpolation to improve the performance. Lissajous curves were chosen due to their high density at the boundaries of the grid, as shown in Figure 2.4.

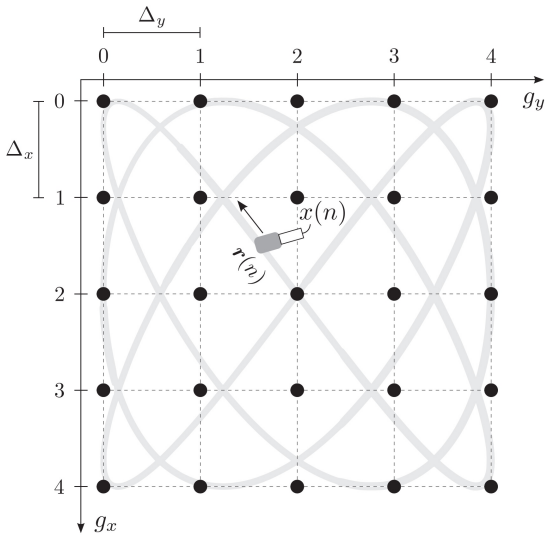


Figure 2.4: Lissajous trajectory and RIR grid (image from [18]).

The authors further improved upon this approach in [17], where they utilized multiple grid spacings Δ. Since the required grid density is dependent on the highest considered frequency, this approach allowed for faster reconstruction of lower frequency by utilizing a sparser grid. Additionally, they showed that under noisy conditions this improved the RIR recovery quality.

The described approach looks promising for DOA estimation, as it allows us to characterize a sound field without any imposed requirements regarding the sound source or the microphone trajectory and speed. The main problem is that we require precise knowledge of the sound source. This could possibly be circumvented by utilizing an additional stationary microphone as a

reference, however, since we would like to perform DOA estimation using only one microphone we will set aside this idea for future research. Another problem inherent with this approach is that it is difficult to implement Lissajous trajectories at high speed in practice. Although different paths are also permissible, they may decrease the accuracy of the sound field reconstruction.

Another modification of the previously described approach was performed in [15]. Here, rather than interpolating $h(\boldsymbol{r}(n), k)$ from Equation (2.1) using the RIRs on the grid as shown in Equation (2.2), it is instead approximated with P multidimensional basis functions $f_p(\cdot)$ as

$$h(\boldsymbol{r}(n), k) \approx \sum_{p=1}^{P} a_p f_p(\boldsymbol{r}(n), k) \ , \tag{2.3}$$

where $f_p(\cdot)$ are chosen to be spherical harmonics basis functions and a_p are the spherical harmonic coefficients to be determined. An example of the first few spherical harmonics are depicted in Figure 2.5, where blue and yellow regions represent positive and negative function values, respectively.

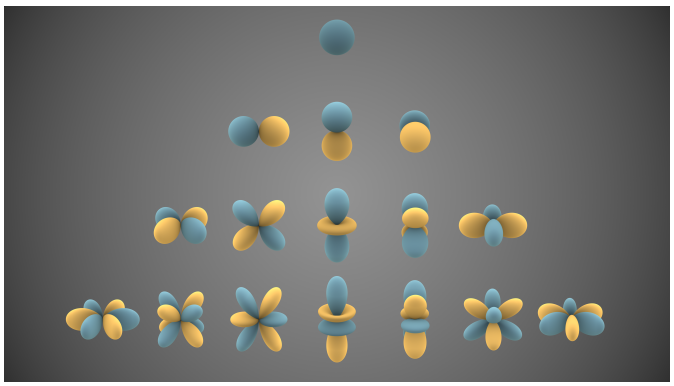


Figure 2.5: Visual representations of the first few real spherical harmonics (image from [27]).

Equation (2.3), in essence, represents a spherical harmonics decomposition of the RIRs. Much like a Fourier transform decomposes an arbitrary signal into a weighted sum of sinusoidal basis functions with different frequencies and phases, a spherical harmonics decomposition decomposes a three-dimensional sound field into a weighted sum of spherical harmonics basis functions.

Side note: This is one way spherical microphone arrays and the EMA are able to characterize a sound field. In the case of spherical microphone arrays the spherical harmonic coefficients are obtained by evaluating a weighted integral of the sound pressure over the surface of the sphere, which can be performed more accurately the more microphones we have on the sphere's surface. In the case of the EMA a simplified formula is obtained, only requiring a weighted integration over the equator of the sphere, given that the sound field is height invariant. As the required

formulae are rather cumbersome, they are omitted here and can be found in [2]. The formulae for determining the parameters a_p will also not be included here and can be found in reference [15].

Since the ultimate goal of DOA estimation using the REM is to be able to approximate the data captured by the EMA or spherical microphone array, this approach looks very promising, since it skips the requirement for DOA estimation altogether and characterizes the sound field using a spherical harmonics decomposition. However, as was the case with the previous approach, we require knowledge of the excitation signal and the microphone trajectory is difficult to implement in practice. Workarounds could be to use a simpler trajectory and an additional stationary reference microphone, but, as before, we set aside this idea for future research.

2.3 Direction of Arrival Estimation Using Moving Microphones

We will now cover the research that has been conducted specifically investigating DOA estimation using moving microphones. The authors of [31] devised a DOA estimation algorithm utilizing a static circular microphone array with up to 8 microphones which is sampled in a round robin fashion, i.e. a signal is constructed by taking the first sample from the first microphone, the second sample from the neighbouring microphone and so on in a circular fashion. Although this research technically did not utilize a moving microphone, the constructed signal is equivalent to a signal captured by a (very) fast rotating microphone.

The circular sampling introduces a periodic Doppler shift into the signal which can be utilized for DOA estimation. Further elaboration on this frequency shift will be provided in Section 3.1. Due to this effect we observe a periodically shifted instantaneous frequency $f_{obs}(n)$ at discrete time n when recording a sound source emitting a tone of frequency f_0 arriving at azimuth angle φ relative to the position of the first microphone, given by

$$f_{obs}(n) = f_0 \cdot \left(1 - \frac{2\pi r f_s}{Mc} \cdot \sin\left(\frac{2\pi f_s}{M} \cdot n - \varphi\right)\right) , \tag{2.4}$$

where M is the number of microphones of the circular array with radius r, f_s is the sampling rate and c is the speed of sound. As it can be seen, the instantaneous frequency periodically shifts around f_0 in a sinusoidal manner with phase offset φ. Therefore, determining the phase of this sinusoid provides us with the DOA of the source signal. The instantaneous frequency is estimated using the Teager-Kaiser Energy Operator (TKEO) from [14], which is given by $\psi(x(n)) = x^2(n) - x(n-1) \cdot x(n+1)$. It is an estimate of the instantaneous energy of $x(n)$ and approximately equal to the squared product of the frequency and amplitude of the signal. The DOA is subsequently estimated by

$$\varphi = \frac{\pi}{2} - \angle \left\{ \sum_{n=0}^{N-1} \psi_{\text{diff}}(n) \cdot e^{-j\frac{2\pi n}{M}} \right\} , \tag{2.5}$$

where $\psi_{\text{diff}}(n)$ is the differential TKEO given by $\psi_{\text{diff}}(n) = \psi(x(n)) - \psi(x_{\text{opp}}(n))$. Here $x_{\text{opp}}(n)$ corresponds to a signal constructed via circular sampling which starts at the microphone opposite of the first microphone from $x(n)$.

To enable this approach to function with multiple, more complex sound sources the constructed signals are divided into a number of subbands and frames and a phase estimate is performed for each band and frame. All of the estimates are subsequently combined into a histogram, which shows clear peaks at the source positions. In [31] this approach was able to locate 5 speech sources with reasonable accuracy.

The advantages of this DOA estimation algorithm are its simplicity, low computational complexity and ability to locate multiple complex sources. As shown in [32], however, the TKEO is susceptible to noise and only suitable for frequencies up to $f_s/8$. The susceptibility to noise is particularly problematic when attempting to perform this DOA estimation approach using a moving microphone since the microphone movement will introduce wind and motor noise into the recorded signal. Additionally, the computation of $\psi_{\text{diff}}(n)$ would require two moving microphones placed opposite of each other. This in itself does not pose a problem, since our REM prototype contains two oppositely mounted microphones (see Section 5.1 and [22]), however, preferably we wish to perform DOA estimation using only one microphone.

An improvement to this approach was made in [32], where the TKEO was replaced by the Center-of-gravity (CoG) algorithm from [6] for the estimation of the instantaneous frequency $f_{\text{IF}}(n)$. As shown in [12], the CoG algorithm is given by

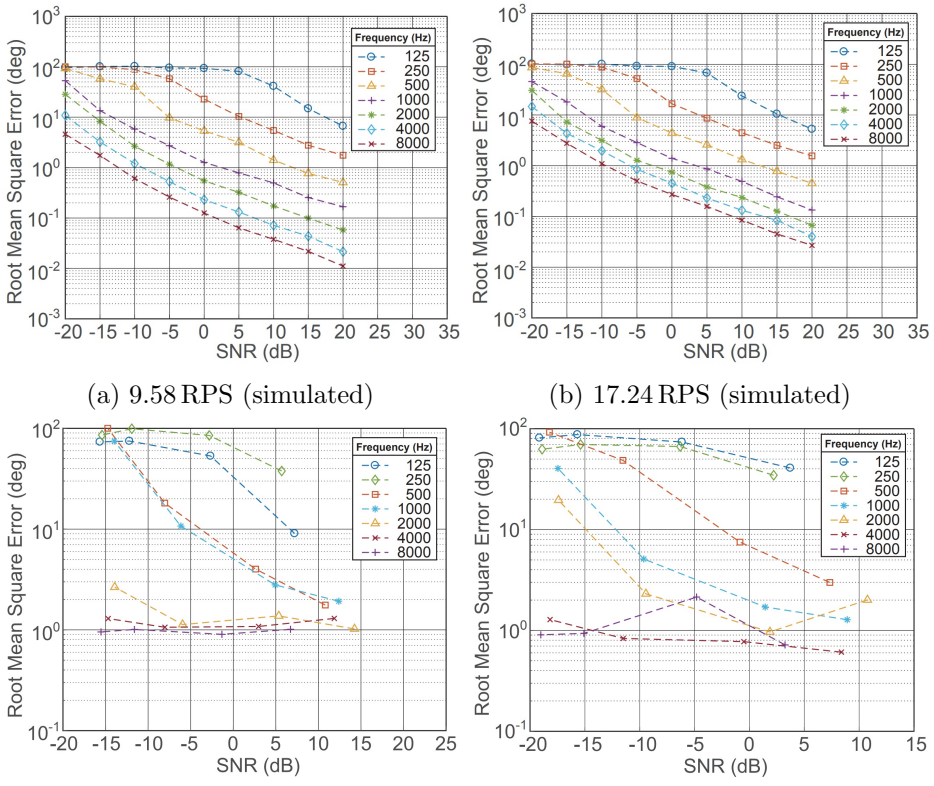
$$f_{\text{IF}}(n) = \frac{f_s}{N} \cdot \frac{\sum_{k=0}^{\frac{N}{2}-1} k \cdot |X(k,n)|^2}{\sum_{k=0}^{\frac{N}{2}-1} |X(k,n)|^2} \tag{2.6}$$

for discrete-time signals, where $X(k,n)$ is the spectrogram of $x(n)$ with k frequency bins and N is the length of the discrete Fourier transform (DFT). The subsequent DOA estimation is performed by replacing $\psi_{\text{diff}}(n)$ in Equation (2.5) with $f_{\text{IF}}(n)$. The estimate can be made even more robust using the differential instantaneous frequency, which analogous to $\psi_{\text{diff}}(n)$ is the difference of the instantaneous frequencies computed over two circularly sampled signals with opposite starting points. The paper showed that the CoG approach consistently outperformed the TKEO approach.

The paper also investigated the DOA estimation accuracy for different distances between the sound source and the microphone array. For distances above 20 cm and an array diameter of 8 cm the estimation accuracy is reasonably constant and only deteriorates at closer distances. This is due to the fact that the phase of the instantaneous frequency only gets noticeably distorted at close distances due to curved wavefronts.

In [12] the previously described approach was implemented in practice using a fast rotating microphone with a 25 cm diameter. To our knowledge it is also the only publication in which a fast rotating microphone was designed and evaluated in practice, aside from our REM prototype. The rotational speeds reached were between approximately 6 RPS and 17 RPS. At these speeds the individual microphone samples are taken at very close distances to each other, much closer than would be possible with the circular sampling approach, potentially leading to a higher DOA estimation accuracy.

The authors first simulated the DOA estimation accuracy of various single-frequency tones at different rotational speeds and signal-to-noise ratios (SNRs) using additive pink noise and subsequently tested these scenarios in practice in an anechoic chamber. The results can be found in Figure 2.6.



(a) 9.58 RPS (simulated)

(b) 17.24 RPS (simulated)

(c) 9.58 RPS (real)

(d) 17.24 RPS (real)

Figure 2.6: DOA estimation accuracy for different rotational speeds and frequencies at various SNRs (images from [12]).

As it can be seen, the DOA estimation accuracy increases for higher SNRs and frequencies both in theory and practice. This was to be expected, since the Doppler frequency shift is larger at high frequencies and thus the change in the instantaneous frequency can be estimated more accurately. Higher SNRs similarly allow for an improved instantaneous frequency estimation. Additionally, it was expected that higher rotational speeds would increase the DOA estimation accuracy due to larger Doppler shifts. However, simulations showed that the accuracy slightly deteriorated at larger rotational speeds, especially for higher frequencies. This may be due to a decrease in the estimation accuracy of the instantaneous frequency as it changes more quickly at higher speeds. In practice, this effect was also discernible to some extent, although it is challenging to determine the impact of the additional wind and motor noise as a result of higher rotational speeds on this observation.

The research showed that DOA estimation using a rotating microphone is feasible in practice and reasonably accurate for higher frequencies. A major drawback of the used approach, however, is the inability to locate low frequencies reliably. Additionally, the utilized implementation relies on the sound sources to be single, constant and known frequencies. Dividing the recorded signal into subbands could help circumvent this problem, but only as long as each subband can be assumed to only contain one frequency.

We will now elaborate on the introduced Doppler shift due to rotation in more detail and subsequently derive a new method of DOA estimation, which not only has the potential to locate complex and low frequency sources more reliably, but also has the ability to remove the distortion introduced due to the microphone rotation.

Chapter 3

Theoretical Foundations

In this chapter we elaborate on the theoretical background needed to perform DOA estimation with a rotating microphone and subsequently use this knowledge to implement a DOA detection algorithm. Section 3.1 shows a connection between the aforementioned movement-induced Doppler shift and frequency modulation and in Section 3.2 a simple DOA detection algorithm is derived. Section 3.3 introduces an alternative representation of frequency modulation using Bessel functions and Section 3.4 derives an algorithm to compensate frequency modulation. Finally, Section 3.5 shows an alternative to the previously derived frequency modulation compensation algorithm.

3.1 Frequency Modulation due to Rotation

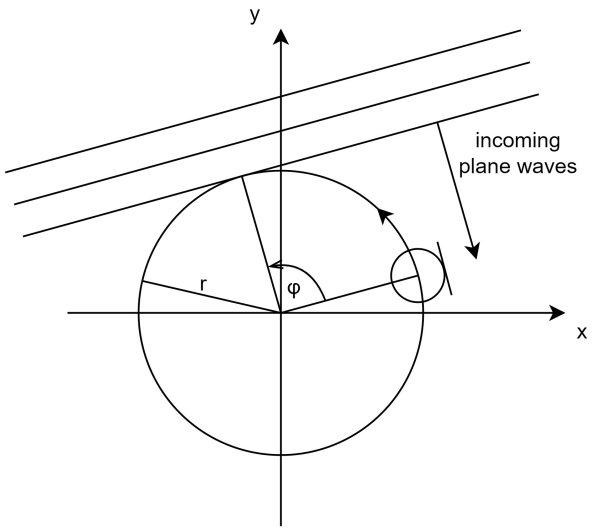


Figure 3.1: Rotating microphone in a sound field.

As we already briefly outlined in Section 2.3, a circular rotation in a sound field introduces a periodic Doppler shift into the recorded signal. In the following, a more detailed explanation of this effect will be given.

Consider a circularly moving microphone in the x-y plane with rotational radius r and rotational speed f_{rot} placed in a sound field composed of only a monochromatic wave of frequency f_{src} arriving at angle φ relative to the initial microphone position. The sound source is assumed to be in the x-y plane and sufficiently far away from the microphone such that we can assume all incoming waves to be plane waves. Furthermore, there are no acoustic reflections, i.e. the microphone is in a free field. This scenario is depicted in Figure 3.1.

If we further assume the microphone to be perfectly omnidirectional the circular microphone movement can be simplified to a sinusoidal movement along a line perpendicular to the incoming sound waves, as shown in Figure 3.2. This movement can be seen as a projection of the circular movement onto an axis perpendicular to the plane waves.

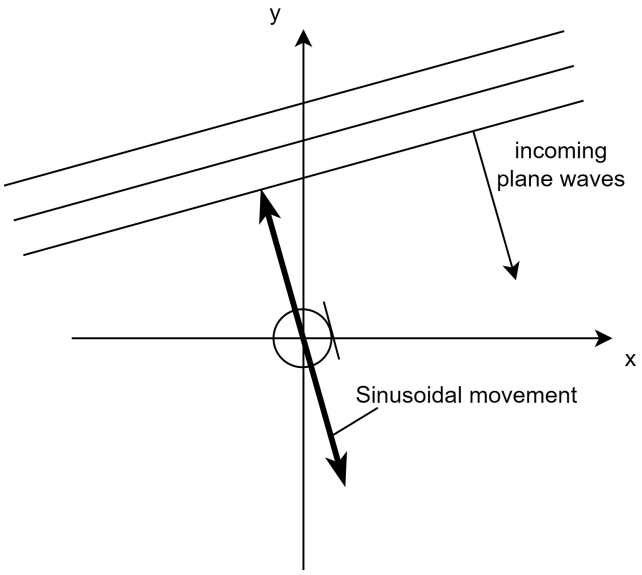


Figure 3.2: Equivalent linear sinusoidal movement of a microphone in a sound field.

From the depicted scenario it becomes clear that a periodic Doppler shift is introduced into the recorded sound. This shift reaches its maximum whenever the microphone is at the center of its linear movement, since this is the point at which the speed relative to the incoming sound waves is at its maximum. The maximum speed of the microphone $v_{m,max}$ relative to the plane waves corresponds to the angular velocity of the circularly rotating microphone:

$$v_{m,max} = 2\pi r \cdot f_{rot} \tag{3.1}$$

Since the speed of the linear movement changes in a sinusoidal fashion with respect to f_{rot}, the instantaneous speed of the microphone relative to the plane waves can be computed as

$$v_r(t) = v_{m,max} \cdot \cos(2\pi f_{rot}\, t) = 2\pi r \cdot f_{rot} \cdot \cos(2\pi f_{rot}\, t)\ , \tag{3.2}$$

if the initial microphone position is assumed to be at the center of the linear movement, i.e. $\varphi = 90°$. To compute the observed instantaneous frequency due to the movement, we make use of the well-known Doppler shift formula

$$f_{obs} = \left(\frac{c \pm v_r}{c \pm v_s}\right) \cdot f_{src}\ , \tag{3.3}$$

where c is the speed of sound and f_{obs} is the observed frequency of a receiver moving at speed v_r relative to a sound source moving at speed v_s, emitting frequency f_{src} [29]. In our case $v_s = 0$ and we use the positive sign in front of v_r since the microphone initially moves towards the sound source. Our instantaneous observed frequency therefore becomes:

$$f_{obs}(t) = \left(1 + \frac{v_r(t)}{c}\right) \cdot f_{src} \tag{3.4}$$

To find an expression for the signal $x(t)$ captured by the microphone, we make use of the fact that

$$x(t) = A \cdot \cos(\phi(t))\ , \tag{3.5}$$

where $\phi(t)$ is the instantaneous phase and A is the maximum amplitude of the recorded signal. Furthermore, the relationship $\frac{d\phi(t)}{dt} = 2\pi f_{obs}(t)$ holds, as shown in [33]. The instantaneous phase can therefore be computed as:

$$\phi(t) = \int_{-\infty}^{t} 2\pi f_{obs}(\tau) d\tau = \phi(0) + \int_{0}^{t} 2\pi f_{obs}(\tau)\, d\tau = \phi(0) + 2\pi f_{src}\, t + \frac{2\pi r \cdot f_{src}}{c} \cdot \sin(2\pi f_{rot}\, t) \tag{3.6}$$

For simplicity, we assume $\phi(0) = 0$. Therefore, our captured signal is

$$x(t) = A \cdot \cos\left(2\pi f_{src}\, t + \frac{2\pi r \cdot f_{src}}{c} \cdot \sin(2\pi f_{rot}\, t)\right) . \tag{3.7}$$

A keen eye may notice that this expression is very similar to the commonly used equation for a frequency modulated signal $x_{\mathrm{FM}}(t)$

$$x_{\mathrm{FM}}(t) = A_c \cdot \cos(2\pi f_c\, t + \beta \cdot \sin(2\pi f_m\, t)) , \tag{3.8}$$

where f_c is the carrier frequency with amplitude A_c, which in our case correspond to f_{src} and A, respectively, f_m is the frequency of the modulating wave, which corresponds to f_{rot}, and β is the so-called modulation index [5]. It is defined as $\beta = \frac{\Delta f}{f_m}$, where Δf describes the peak frequency deviation from carrier f_c. The modulation index can be seen as a measure of *how much* we modulate the carrier frequency. An example plot of $x_{\mathrm{FM}}(t)$ for $\beta = 0$ and $\beta = 1.5$ is shown in Figure 3.3.

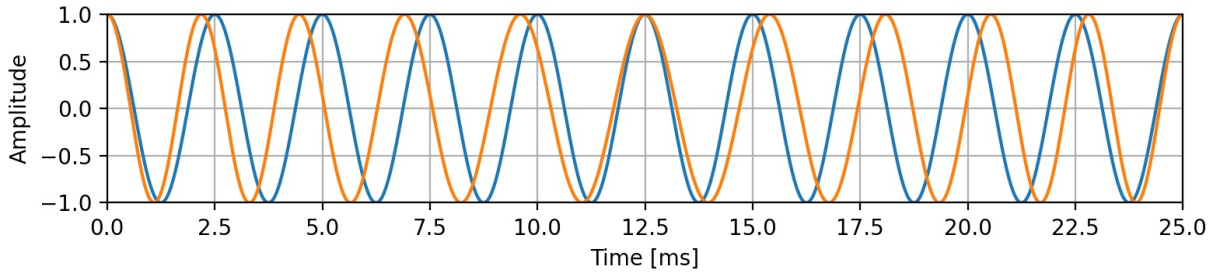


Figure 3.3: $x_{\mathrm{FM}}(t)$ for $f_c = 400\,\mathrm{Hz}$, $f_m = 40\,\mathrm{Hz}$, $\beta = 0$ (blue) and $\beta = 1.5$ (orange).

In the case of modulation introduced by the microphone movement the modulation index corresponds to

$$\beta = \frac{2\pi r \cdot f_{src}}{c} , \tag{3.9}$$

as it can be observed from Equation (3.7). Note that, contrary to what might be expected, it is independent of the rotational speed f_{rot}.

Our derived expression for the captured signal $x(t)$ can be easily extended to arbitrary angles of arrival $\varphi \in [0°, 360°]$ of plane waves in the x-y plane relative to the initial microphone position as

$$x(t, \varphi) = A \cdot \cos(2\pi f_{src}\, t + \beta \cdot \sin(2\pi f_{rot}\, t + \varphi')) \;, \tag{3.10}$$

where $\varphi' = \varphi - 90°$, since the angle of arrival only impacts the phase of the modulating wave ($90° = \frac{\pi}{2}$ rad). Finally, we can further extend this expression for arbitrary elevation angles of arrival $\theta \in [0°, 180°]$ of the plane waves relative to the rotational plane. Consider the scenario illustrated in Figure 3.4:

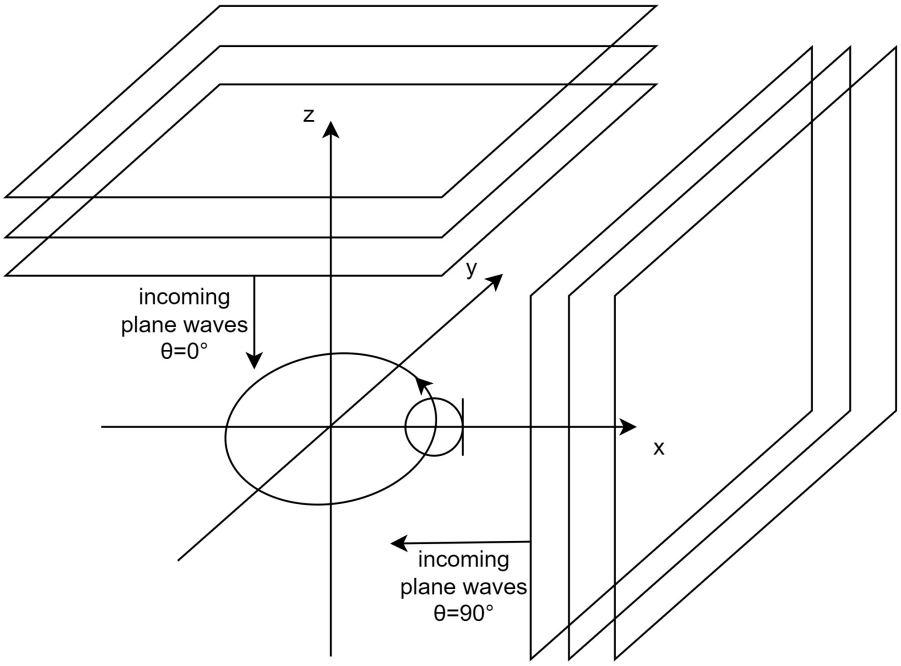


Figure 3.4: Plane waves arriving at different elevation angles.

The depicted scenario is identical to the 2D case we considered in Figure 3.2 for plane waves arriving at $\theta = 90°$. For plane waves are arriving at $\theta = 0°$, however, there is no movement of the microphone relative to the sound waves. Therefore no frequency modulation takes place and the modulation index is zero. It transpires that the relative displacement between the microphone and the sound source changes with the elevation of the incoming sound waves. This effect can be modeled by defining an elevation dependent modulation index $\beta(\theta)$ as follows:

$$\beta(\theta) = \sin(\theta) \cdot \beta = \sin(\theta) \cdot \frac{2\pi r \cdot f_{src}}{c} \tag{3.11}$$

With this modification of the modulation index we can express the recorded signal of a plane wave arriving from an arbitrary direction as

$$x(t, \varphi, \theta) = A \cdot \cos(2\pi f_{src}\, t + \beta(\theta) \cdot \sin(2\pi f_{rot}\, t + \varphi')) \ . \tag{3.12}$$

3.2 Single Frequency Direction of Arrival Estimation

To determine φ and θ from a given $x(t, \varphi, \theta)$, let us assume that we have knowledge of the instantaneous frequency $f_{obs}(t)$ of $x(t, \varphi, \theta)$, given by

$$f_{obs}(t) = \frac{d}{dt}\frac{\phi(t)}{2\pi} = \beta(\theta) \cdot f_{rot} \cdot \cos(2\pi f_{rot}\, t + \varphi') + f_{src} \ , \tag{3.13}$$

where $\phi(t)$ corresponds to the argument of the cosine function in Equation (3.12). Since $f_{obs}(t)$ is solely composed of a weighted frequency f_{rot} with initial phase φ' plus a DC offset, we can obtain φ by computing the argument of the Fourier transform of $f_{obs}(t)$ at frequency f_{rot} as

$$\varphi = \arg(\mathcal{F}\{f_{obs}(t)\}_{f=f_{rot}}) + 90^\circ \ , \tag{3.14}$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform.

To compute θ, recall that the modulation index can be expressed as

$$\beta(\theta) = \frac{\Delta f(\theta)}{f_m} = \sin(\theta) \cdot \frac{2\pi r \cdot f_{src}}{c} \ , \tag{3.15}$$

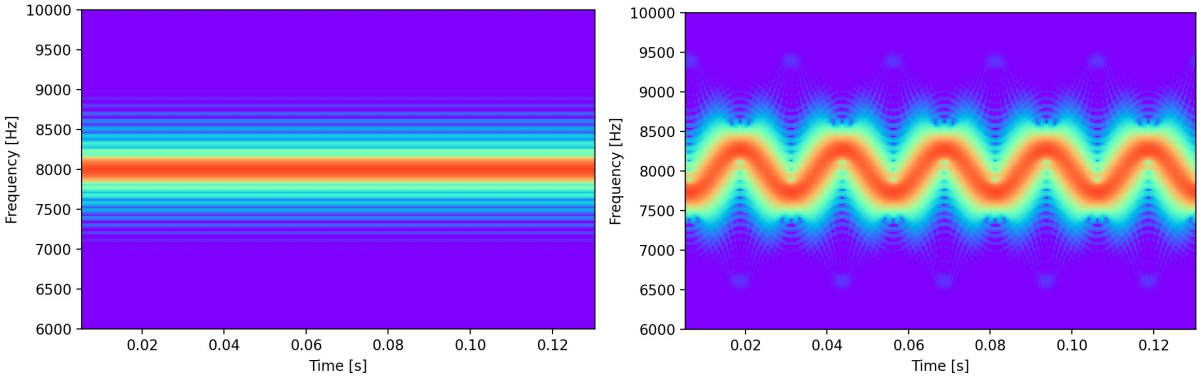
where f_m corresponds to f_{rot} in our case. Furthermore we can compute the peak frequency deviation as $\Delta f(\theta) = f_{obs,max} - f_{src}$, where $f_{obs,max}$ denotes the maximum of $f_{obs}(t)$, which allows us to express θ as

$$\theta = \arcsin\left(\frac{(f_{obs,max} - f_{src}) \cdot c}{f_{rot} \cdot 2\pi r \cdot f_{src}}\right) \ . \tag{3.16}$$

Note that $f_{obs,max} - f_{src} \geq 0$, therefore $\theta \in [0^\circ, 90^\circ]$. This may seem problematic at first, since in reality $\theta \in [0^\circ, 180^\circ]$, however, since $\beta(\theta)$ is the only component in $x(t, \varphi, \theta)$ dependent on θ

and $\beta(90° + \theta') = \beta(90° - \theta')$ for $\theta' \in [0°, 90°]$ due to its sine-dependency, plane waves arriving at a given azimuth φ and elevation $\theta = 90° \pm \theta'$ will produce identical audio signals, i.e. it is not possible to differentiate whether a signal is arriving above or below the rotational plane.

The question that remains is how we can obtain $f_{obs}(t)$ from $x(t, \varphi, \theta)$. To answer this question, consider the spectrogram of an 8 kHz sine wave and the corresponding modulated version depicted in Figure 3.5:



(a) Spectrogram of a sine wave

(b) Spectrogram of a modulated sine wave

Figure 3.5: Original and modulated sine wave for $f_{src} = 8\,\text{kHz}$, $f_{rot} = 40\,\text{Hz}$ and $r = 5\,\text{cm}$.

The center frequency $f_{ctr}[n]$ of Figure 3.5b over discrete time n, which can be computed using the CoG algorithm from [6] for example, can be seen as a discrete approximation of the instantaneous frequency $f_{obs}(t)$ of $x(t, \varphi, \theta)$. This enables us to compute φ by replacing the Fourier Transform and $f_{obs}(t)$ in Equation (3.14) with a DFT and $f_{ctr}[n]$, respectively. In essence, this is the approach used in [32] and [12] for DOA estimation. Similarly, θ can be computed by replacing $f_{obs,max}$ in Equation (3.16) by the maximum value of $f_{ctr}[n]$.

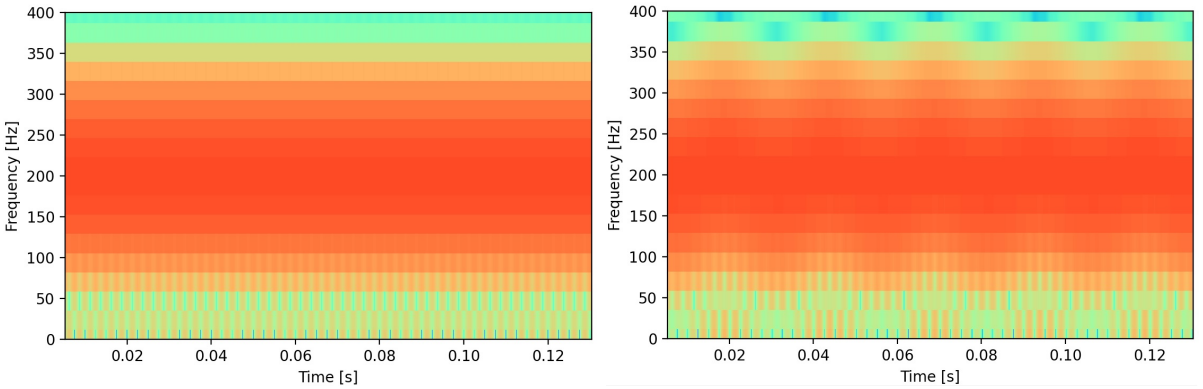
The main advantage of this DOA estimation approach is its simplicity, however, there are also multiple drawbacks:

1. To accurately approximate the instantaneous frequency $f_{obs}(t)$, we need to compute our DFT over a very short time interval, especially as we increase f_{rot}. As a consequence we have a low frequency resolution, which makes DOA detection challenging for low source frequencies f_{src}, as can be seen in Figure 3.6.

2. f_{src} needs to stay constant for at least one full rotation of the microphone to detect its DOA, time varying signals can therefore not be located. If f_{src} only varies slowly over time, increasing f_{rot} can help mitigate this problem.

3. Since this approach currently relies on the CoG algorithm, which only detects a single frequency [6], multiple tones arriving from different directions or more complex source

signals, e.g speech signals, cannot be located. Splitting the recorded signal into multiple subbands can potentially circumvent this problem, however, only as long as each subband contains only one frequency.



(a) Spectrogram of a sine wave

(b) Spectrogram of a modulated sine wave

Figure 3.6: Original and modulated sine wave for $f_{src} = 200\,\text{Hz}$, $f_{rot} = 40\,\text{Hz}$ and $r = 5\,\text{cm}$.

To obtain an alternative DOA estimation approach, the following chapter shows a relationship between frequency modulation and Bessel functions, which we can use to our advantage for deriving an improved DOA estimation algorithm.

3.3 Frequency Modulation and Bessel Functions

The spectrograms shown in Figure 3.5 and Figure 3.6 were computed using a DFT window length of $L = 512$ samples. This value has been chosen to provide a trade-off between the accuracy of the instantaneous frequency estimation and the frequency resolution (when using a sampling rate of $f_s = 48\,\text{kHz}$). If we substantially increase the value of L such that the DFT is computed over at least one full rotation of the microphone, we obtain the spectrograms shown in Figure 3.7.

Side note: These spectrograms and all following spectrograms were computed using a Hann-window if not specified otherwise.

As it can be observed, increasing L from 512 to 8192 not only provides a substantially higher frequency resolution, the spectrograms of the modulated signals also exclusively show multiple constant frequencies with a spacing of $40\,\text{Hz} = f_{rot}$. Furthermore, many more of these constant frequencies are visible for the 8 kHz signal as compared to the 200 Hz signal.

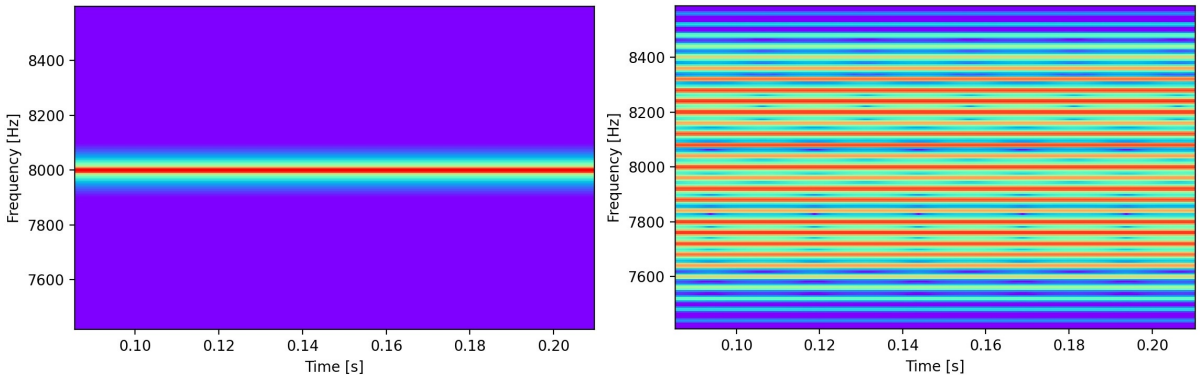
These unexpected findings can be explained by the following alternative representation of $x(t, \varphi, \theta)$ from Equation (3.12)

$$A \cdot \cos(2\pi f_{src}\, t + \beta(\theta) \cdot \sin(2\pi f_{rot}\, t + \varphi')) = A \cdot \sum_{n=-\infty}^{\infty} J_n(\beta(\theta)) \cdot \cos(2\pi (f_{src} + n f_{rot})\, t + n\varphi') \,, \tag{3.17}$$

where $J_n(\cdot)$ denotes the Bessel function of the first kind for integer order n. The derivation of this equation is omitted here and can be found in [7]. Note that the derivation does not include terms A and φ', but can be easily extended to include an amplitude and a starting phase.

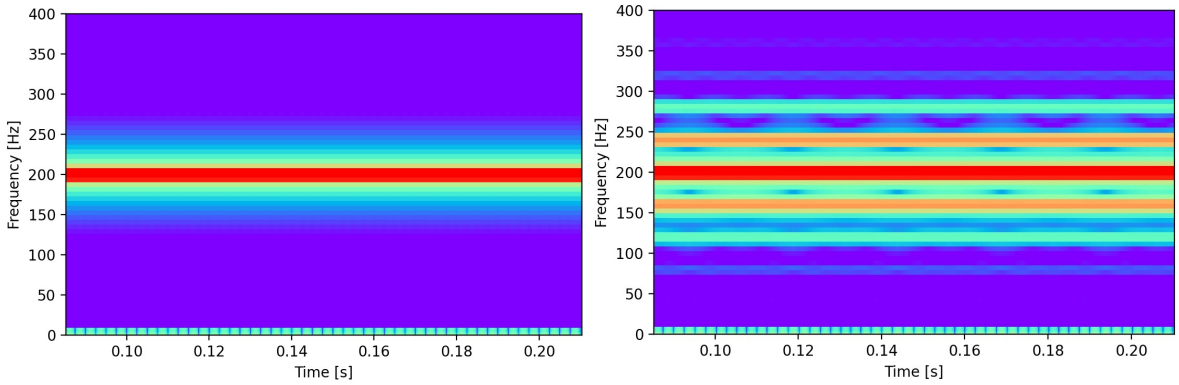
As can be seen from Equation (3.17), a frequency modulated signal can be represented as an infinite weighted sum of frequencies spaced integer multiples n of f_{rot} around the carrier frequency f_{src}. The weights are computed by evaluating n-th order Bessel functions of the first kind at the modulation index $\beta(\theta)$. An example plot of five of these functions is depicted in Figure 3.8.

If no frequency modulation is present, i.e. $\beta(\theta) = 0$, all Bessel functions except $J_0(\cdot)$, and therefore all the sidebands of f_{src}, are zero. As we increase the modulation index, the remaining Bessel



(a) Spectrogram of 8 kHz sine wave

(b) Spectrogram of modulated 8 kHz sine wave



(c) Spectrogram of 200 Hz sine wave

(d) Spectrogram of modulated 200 Hz sine wave

Figure 3.7: Spectrograms of the signals from Figure 3.5 and Figure 3.6 for DFT length $L = 8192$.

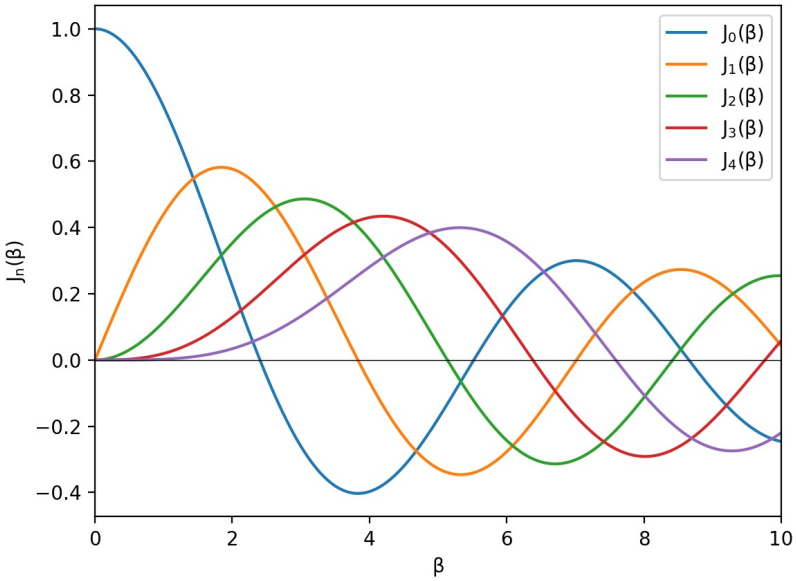


Figure 3.8: Bessel functions of the first kind for integer orders $n = 0, 1, 2, 3, 4$.

functions gradually take on significant values, resulting in more sidebands to become visible. Note that higher order Bessel functions stay close to zero longer, therefore the distant sidebands only start becoming visible at very high modulation indices. Since the modulation index is dependent on f_{src}, more sidebands will be visible for higher frequencies if all other parameters remain unchanged.

Another noteworthy property of Bessel functions of the first kind is given by

$$\sum_{n=-\infty}^{\infty} J_n^2(x) = 1\,,\ \forall x \geq 0\,, \tag{3.18}$$

as shown in [1]. This identity becomes useful when computing the energy E of the frequency modulated signal from Equation (3.17) as

$$E = \int_{-\infty}^{\infty} |x(t,\varphi,\theta)|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |X(2\pi f)|^2 df = \frac{A^2}{2\pi}\sum_{n=-\infty}^{\infty} J_n^2(\beta(\theta)) = \frac{A^2}{2\pi}\,, \tag{3.19}$$

where $X(2\pi f)$ denotes the Fourier transform of $x(t,\varphi,\theta)$ [4]. As it can be observed, the modulation index has no impact on the energy of the signal. This is advantageous since it allows us to interpret frequency modulation as an *energy-conserving redistribution* of the input energy, where the energy gets dispersed more as the modulation index increases. Likewise the energy

becomes more focused as the modulation index approaches zero. The degree to which the energy of the spectrogram is focused, which we will call *focusedness*, can therefore be used as a measure for the modulation index. This property will be used in Section 3.4 and Section 3.5, where an algorithm for unmodulating a given signal for multiple DOAs will be derived, effectively modifying the modulation index. The DOA will then be estimated based on which unmodulated signal has the highest focusedness. Since the focusedness acts on the entire spectrogram, this approach requires no knowledge or estimation of the source frequency.

To quantify the focusedness of a spectrogram, note that it can be shown that if we have a set of N numbers $\{x_1, x_2, \dots, x_N\}$, $x_n \geq 0 \; \forall \, n \in [1, 2, \dots, N]$, where $\sum_{n=1}^{N} x_n = 1$ holds, the squared sum $\sum_{n=1}^{N} x_n^2$ will reach its maximum if $\exists!\; i \in [1, 2, \dots, N]$ for which $x_i = 1$, $x_{n \neq i} = 0$ holds, i.e. if all except one x_n are zero. Furthermore the squared sum over all x_n will reach its minimum if $x_n = \frac{1}{N} \, \forall \, n \in [1, 2, \dots, N]$, i.e. if all x_n have the same value. Therefore, given a set of numbers which sum up to one, the squared sum can be used as a measure for how distributed the sum is over the numbers. Likewise we can use the squared sum of the energies of each frequency in a spectrogram frame as a way to quantify its focusedness.

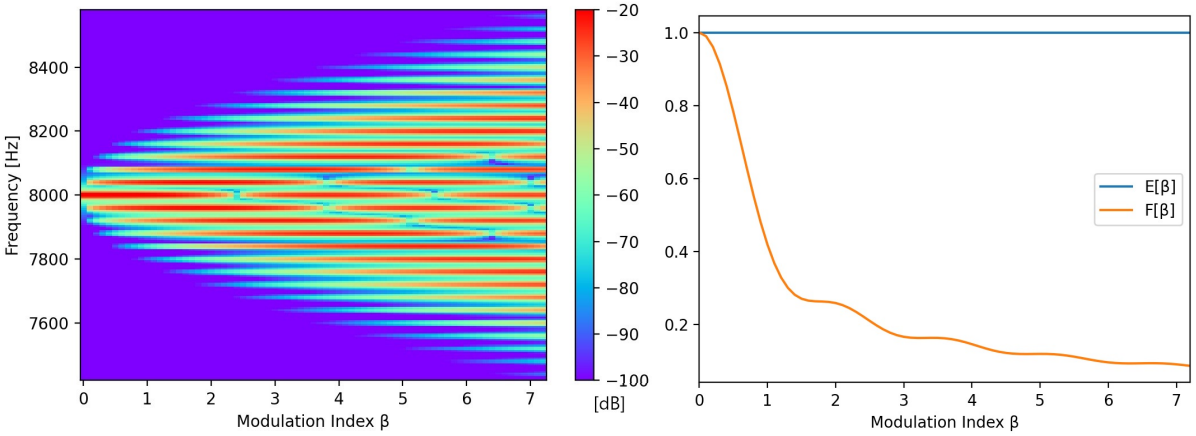
In practice, obtaining the energy of a given DFT frame n simply requires the computation of the squared sum of the absolute value of all the DFT bins $X[n, k]$ in frame n as

$$E[n] = \sum_{k=0}^{N/2} |X[n, k]|^2 \, , \tag{3.20}$$

where N denotes the used DFT length. Computing the focusedness $F[n]$ of DFT frame n can be achieved by squared summation of the energies of each bin as

$$F[n] = \sum_{k=0}^{N/2} |X[n, k]|^4 \, . \tag{3.21}$$

An example plot of the energy and focusedness of the spectrum of an 8 kHz sine wave for an increasing modulation index is depicted in Figure 3.9. The energy and the focusedness have been normalized to output a maximum value of 1. As it can be seen, the energy remains constant regardless of β, as it has been shown in Equation (3.19), whereas the focusedness decreases as the modulation index increases.



(a) β-dependent 8 kHz sine wave spectrum (b) β-dependent energy and focusedness

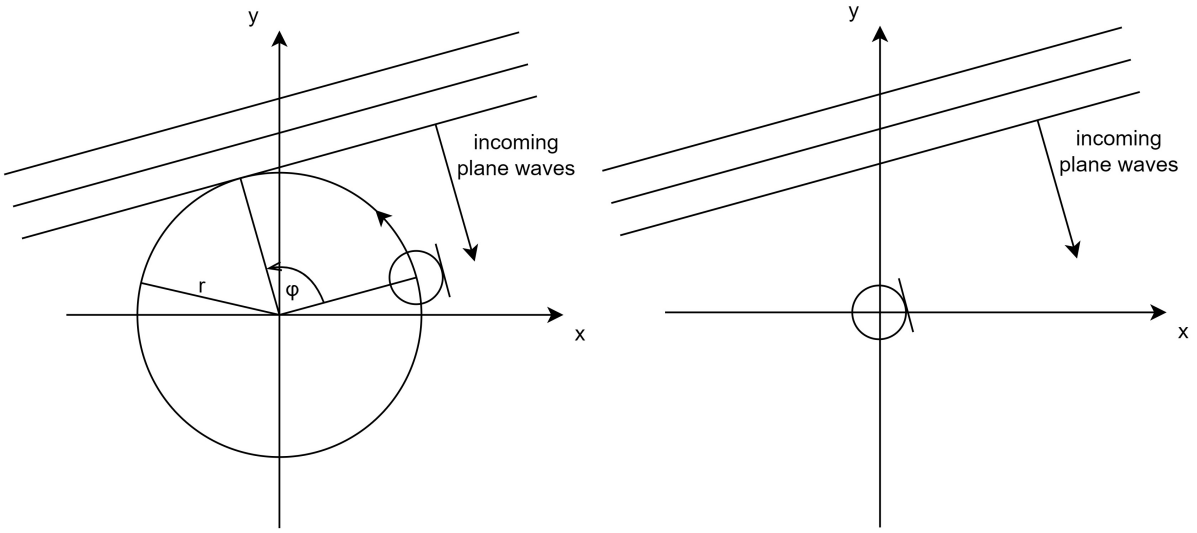
Figure 3.9: Normalized energy and focusedness of an 8 kHz sine wave for an increasing β.

3.4 Time Warping Algorithm

As it has been briefly mentioned in the previous section, the main idea behind the sound source localization approach in this thesis is to derive an algorithm which compensates the frequency modulation introduced into a recorded audio signal for various DOAs and subsequently compute the focusedness of the resulting signal spectrograms. The unmodulated signal with the highest focusedness should, in theory, provide information about the DOA of the audio source. In this section, a DOA-dependent algorithm for unmodulating a frequency modulated audio signal will be derived which in essence performs an accurate time stretching and compression of the modulated audio signal to undo the frequency modulation.

Consider again the scenario from Figure 3.1 and its simplified form in Figure 3.2. We now assume the incoming plane waves to be an arbitrary, unknown audio signal rather than monochromatic sound waves of a known frequency. Additionally, as before, we assume that the microphone is in a free field, perfectly omnidirectional and that its rotational speed f_{rot} as well as the starting phase relative to the sound source φ are known. To unmodulate the audio signal captured by the moving microphone M_{mov}, we need to determine what signal would have been captured by a stationary microphone M_{stat} sampling at the same sampling rate f_s. Figure 3.10 shows a comparison of the scenario from Figure 3.1 and a stationary microphone placed at the center of the rotational movement.

To derive an algorithm for the unmodulation it is helpful to change our reference frame from Figure 3.10: Rather than fixing our coordinate system in space we fix it to a wavefront of the plane waves, resulting in the waves to appear stationary. In this reference frame microphone M_{stat} moves at the speed of sound c towards the sound waves. Similarly, Microphone M_{mov} moves in the same direction as M_{stat} at variable speed $v(t) = c + v_r(t)$, where $v_r(t)$ is given by



(a) Moving Microphone (M_{mov}) (b) Stationary microphone (M_{stat})

Figure 3.10: Moving microphone compared to stationary case at the center of the rotation.

Equation(3.2), i.e. the speed periodically fluctuates around c. The movement of M_{mov} has an additional component parallel to the sound waves, however, we neglect this component since it has no influence on the recorded sound.

As M_{mov} is moving faster than M_{stat} it is capturing fewer data points of the sound field as compared to M_{stat}, since the sampling rates of both microphones are identical. Likewise, as M_{mov} is moving slower than M_{stat}, it is capturing more data points of the sound field. If we now want to capture the same data as M_{mov} with M_{stat} we can achieve this by precisely modifying the sampling rate of M_{stat} such that it is sampling more slowly as M_{mov} is moving faster and sampling more quickly as M_{mov} is moving slower. Correspondingly, if we choose to modify the sampling rate of M_{mov} such that it samples more quickly as it is moving faster and more slowly as it is moving slower, we can capture the same data points as M_{stat}.

To illustrate this concept in a more comprehensive manner we will now demonstrate how to perform frequency modulation by varying the sampling rate of a stationary microphone. Consider the sampling points shown in Figure 3.11:

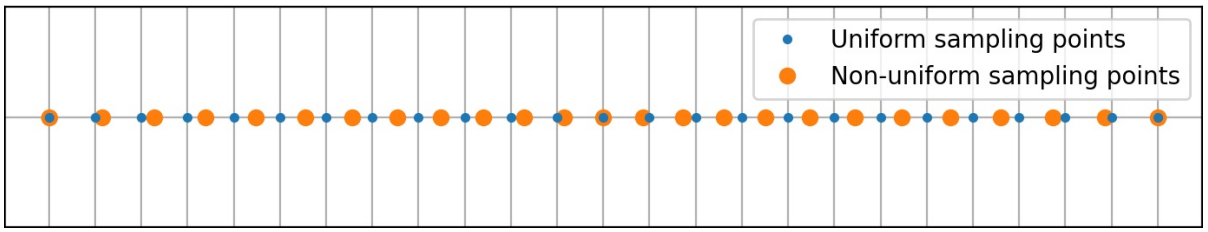


Figure 3.11: Uniform and non-uniform sampling points.

The blue points represent uniform sampling positions at sampling rate f_s, whereas the orange points are non-uniformly spaced. The computation of the exact locations of the non-uniform sampling points will follow later. Let us now use these sampling points to sample, for example, a sinusoidal wave, as illustrated in Figure 3.12.

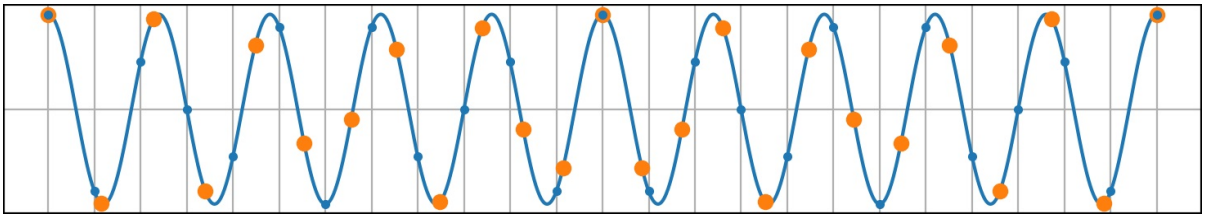


Figure 3.12: Uniform and non-uniform sampling of a sinusoidal wave.

Now, if we time-shift the non-uniform orange sampling points to match the sampling grid of the blue points, we obtain the red points shown in Figure 3.13.

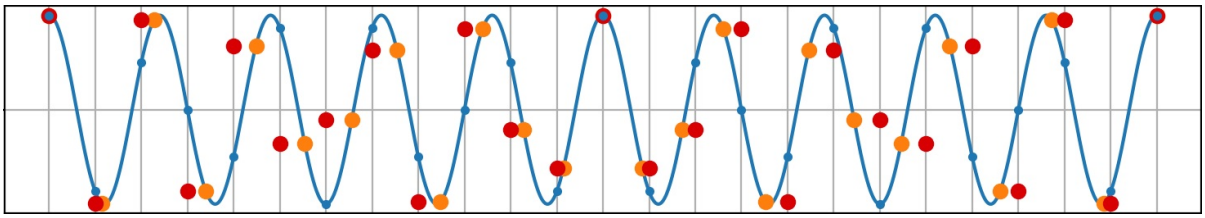


Figure 3.13: Non-uniform sampling points time-shifted to the sampling grid.

The resulting signal described by the red points is depicted in Figure 3.14. As it can be observed the red graph represents a frequency modulated version of the original signal. This does not only hold for a sinusoidal wave but for any arbitrary audio signal.

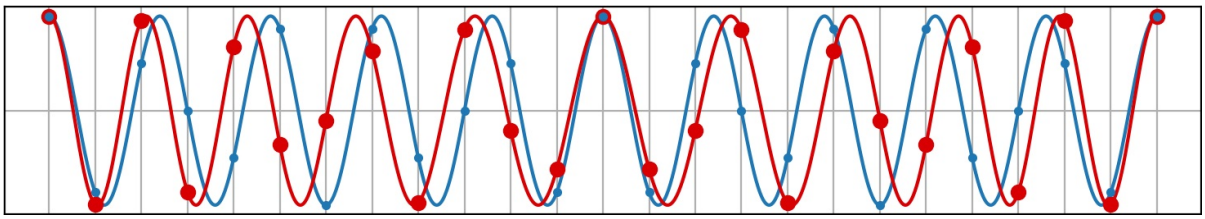


Figure 3.14: Original and frequency modulated signal.

The demonstrated concept can now be made use of for unmodulation: The red sampling points can be seen as the samples taken by the rotating microphone and the blue points as their unmodulated counterpart. To obtain the blue points from the red points we first need to time-shift the red samples to the correct positions in time, as shown in Figure 3.15. Then we can interpolate the blue points from the orange points using a high quality interpolation method.

Some interesting notes on interpolation: Generally speaking, perfect interpolation is only possible if the sampling rate is above the Nyquist rate, i.e. the sampling rate is at least twice the bandwidth

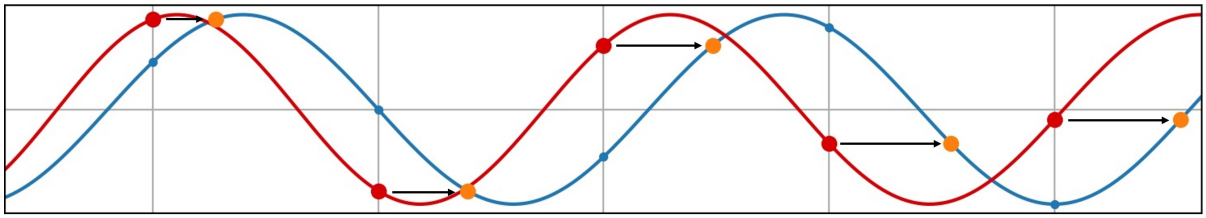


Figure 3.15: Frequency modulation compensation - first method.

of the sampled signal [21]. Here we have the unusual case that the sampling rate of the orange points varies over time. Interestingly, this does not pose a problem for interpolation, since perfect signal reconstruction is theoretically possible as long as the average sampling rate is above the Nyquist rate [8]. In practice, however, interpolation of non-uniform data is more challenging and often less accurate than that of uniform data, as the authors of [23] show.

The necessary time-shifts are obtained by computing the time at which the wavefronts that were sampled by the moving microphone arrived at the stationary microphone. The new sampling positions $t_{\text{new}}(t, \varphi)$ for sound waves arriving at angle φ and sampled at time t can therefore be computed by:

$$t_{\text{new}}(t, \varphi) = t - \frac{\Delta d(t, \varphi)}{c} = t - \frac{r \cdot \cos(2\pi f_{rot}\, t + \varphi)}{c} \,, \tag{3.22}$$

where $\Delta d(t, \varphi)$ is the time-dependent difference in distance of the moving microphone to the sound source and the stationary microphone to the sound source.

To interpolate the blue data points from our time-shifted samples, it is desirable to use sinc interpolation, given by

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT) \cdot \text{sinc}\left(\frac{\pi(t - nT)}{T}\right) \,, \quad \text{sinc}(x) = \begin{cases} \frac{\sin(x)}{x}, & \text{if } x \neq 0. \\ 1 \quad , & \text{if } x = 0. \end{cases} \,, \tag{3.23}$$

where $T = \frac{1}{f_s}$ refers to the sampling period. The reason we would like to use sinc interpolation is that it is exact if we assume to have an infinite number of discrete samples, no quantization errors and a sampling rate above the Nyquist rate. Although we will always have quantization errors and a finite number of discrete samples in practice, sinc interpolation still provides very accurate results [30].

Unfortunately, we cannot directly implement Equation (3.23) since it requires uniform spacing of our input samples, which is not the case for the orange samples in Figure 3.15. However, as the

authors of [23] show, it is possible to use modified versions of sinc interpolation for non-uniform samples, given that certain constraints are met. The downside of these approaches is that they represent an approximation of sinc interpolation and are more difficult to implement.

Fortunately, we can avoid using non-uniform sinc interpolation altogether by slightly changing our unmodulation approach from Figure 3.15. Consider Figure 3.16:

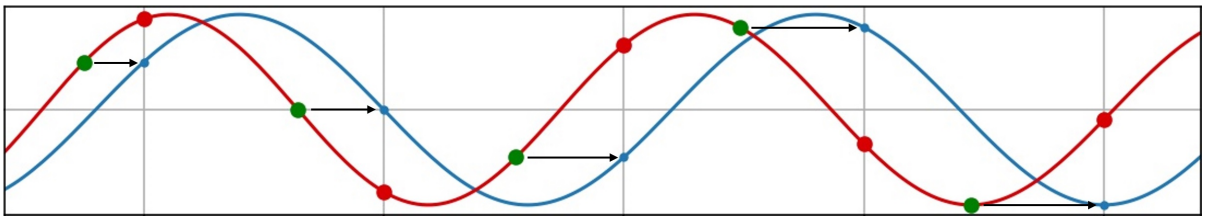


Figure 3.16: Frequency modulation compensation - second method.

Instead of first time-shifting the samples and then interpolating, we could instead choose to first interpolate and then time-shift. This is advantageous since we can interpolate the green samples from the red samples using Equation (3.23), as the red data points are uniformly spaced. The disadvantage of this approach, however, is that determining the equivalent positions of the blue samples on the red graph, i.e. the timestamps of the green points, is more difficult than the time-shift computation from Equation (3.22) as we will now show.

The timestamps of the green samples can be determined by computing the time at which the wavefronts that were sampled by the stationary microphone arrived at the moving microphone. Let us assume that a given wavefront arrived at the stationary microphone at time t_0. Then, the time-dependent difference in distance $\Delta d(t, t_0, \varphi)$ of the moving microphone to the sound source and the stationary microphone to the sound source reads:

$$\Delta d(t, t_0, \varphi) = r \cdot \cos(2\pi f_{rot}\,(t + t_0) + \varphi)\ , \tag{3.24}$$

The wavefront travels along an axis parallel to the linear sinusoidal movement of the microphone at the speed of sound. Furthermore it reaches the stationary microphone at time t_0, which is placed at the origin of the coordinate system. Therefore, along this axis, we can model the time-dependent position of the wavefront as $c \cdot t$. To determine at what time $\hat{t}(t_0, \varphi)$ the wavefront meets the moving microphone, we need to find the intersection between these two equations:

$$r \cdot \cos(2\pi f_{rot}\,(t + t_0) + \varphi) = c \cdot t \tag{3.25}$$

Unfortunately, this expression cannot be solved analytically for t. Hence, we need to numerically approximate the root of function

$$f(t, t_0, \varphi) = r \cdot \cos(2\pi f_{rot}\,(t + t_0) + \varphi) - c \cdot t \overset{!}{=} 0\ , \tag{3.26}$$

for a given t_0 and φ, which can be achieved by minimizing the absolute value of $f(t, t_0, \varphi)$ as

$$\hat{t}(t_0, \varphi) = \underset{t}{\operatorname{argmin}}(|f(t, t_0, \varphi)|)\ . \tag{3.27}$$

Note that $f(t)$ has a unique root and $|f(t)|$ is convex only as long as $c \geq 2\pi r f_{rot}$ holds, i.e. as long as the microphone speed does not exceed the speed of sound, which we can safely assume to be the case.

To summarize, computing the unmodulated version $\hat{x}[n]$ from a given sampled audio signal $x[n]$ can be achieved by the following algorithm, which we will refer to as *time warping algorithm* or TWA for short:

1. Compute $\hat{t}(t_0, \varphi)$ using Equation (3.27), where t_0 are the uniformly spaced sampling times and φ is the DOA of the sound waves
2. Use Equation (3.23) to interpolate $x[n]$ at timestamps $\hat{t}(t_0, \varphi)$
3. Uniformly space the interpolated samples from 2. to obtain $\hat{x}[n]$

Side note: For simulation purposes it may be helpful to have an inverse TWA, i.e. an algorithm that modulates a given audio signal. Fortunately, the inverse TWA follows the exact same steps as the TWA, with the only difference being that $\hat{t}(t_0, \varphi)$ is replaced by $t_{\text{new}}(t, \varphi)$ from Equation (3.22).

This algorithm can be easily extended to work for arbitrary incoming elevation angles θ of the plane waves by following the same argumentation as for the derivation of $\beta(\theta)$ from Equation (3.11): We simply need to modify Equation (3.26) to

$$f(t, t_0, \varphi, \theta) = \sin(\theta) \cdot r \cdot \cos(2\pi f_{rot}\,(t + t_0) + \varphi) - c \cdot t \overset{!}{=} 0\ , \quad c \overset{!}{\geq} \sin(\theta) \cdot 2\pi r f_{rot}\ , \tag{3.28}$$

and instead of estimating $\hat{t}(t_0, \varphi)$ we estimate $\hat{t}(t_0, \varphi, \theta)$ by

$$\hat{t}(t_0, \varphi, \theta) = \underset{t}{\operatorname{argmin}}(|f(t, t_0, \varphi, \theta)|) \ . \tag{3.29}$$

One final note on the TWA: To perfectly unmodulate a signal it is necessary to have perfect knowledge of its DOA. This, of course, completely defies the purpose of DOA estimation. However, as the next section will show, if we guess the DOA for a signal and the guess is close to the true DOA, the TWA will increase the focusedness of the signal spectrogram. Likewise, if our DOA guess is poor, the algorithm will decrease the focusedness. The true DOA can therefore be approximated by either making a large number of DOA guesses and selecting the DOA with the maximum associated focusedness or by making a few strategical guesses and iteratively refining our DOA search. The latter approach is more efficient from a computational standpoint, however, we will use the former method for the sake of simplicity, since efficiency or real-time functionality are not of a concern in this thesis and left for future research. Unfortunately, the computational complexity of our approach will most likely be significantly greater as compared to the localization technique from [12] regardless of the implementation. However, our proposed method has the great advantage of removing or at the very least reducing the distortions introduced by the microphone rotation. This allows us to not only locate sound sources, but also to reconstruct the incoming audio signals, potentially enabling applications such as blind source separation or beamforming.

3.5 Matrix-based Time Warping

To enable applications such as blind source separation or beamforming it would be beneficial to encapsulate the functionality of the TWA into a modulation matrix $\mathbf{Z}_\varphi$ such that

$$\mathbf{y}^{(1)} = \mathbf{x}^{(1)} \cdot \mathbf{Z}_\varphi \ , \tag{3.30}$$

where $\mathbf{x}^{(1)}$ is the first spectrogram frame of the source signal and $\mathbf{y}^{(1)}$ is the corresponding modulated frame with DOA φ (we assume $\theta = 90°$). To be more specific,

$$\mathbf{x}^{(1)} = [x_1^{(1)} \ x_2^{(1)} \ ... \ x_n^{(1)}] \in \mathbb{C}^n \ , \quad \mathbf{y}^{(1)} = [y_1^{(1)} \ y_2^{(1)} \ ... \ y_n^{(1)}] \in \mathbb{C}^n \ , \quad Z_\varphi \in \mathbb{C}^{n \times n} \ , \tag{3.31}$$

where $n = \frac{L}{2} + 1$ for a given (even) DFT length L and $x_i^{(1)}$, $y_i^{(1)}$ correspond to the i-th DFT bins

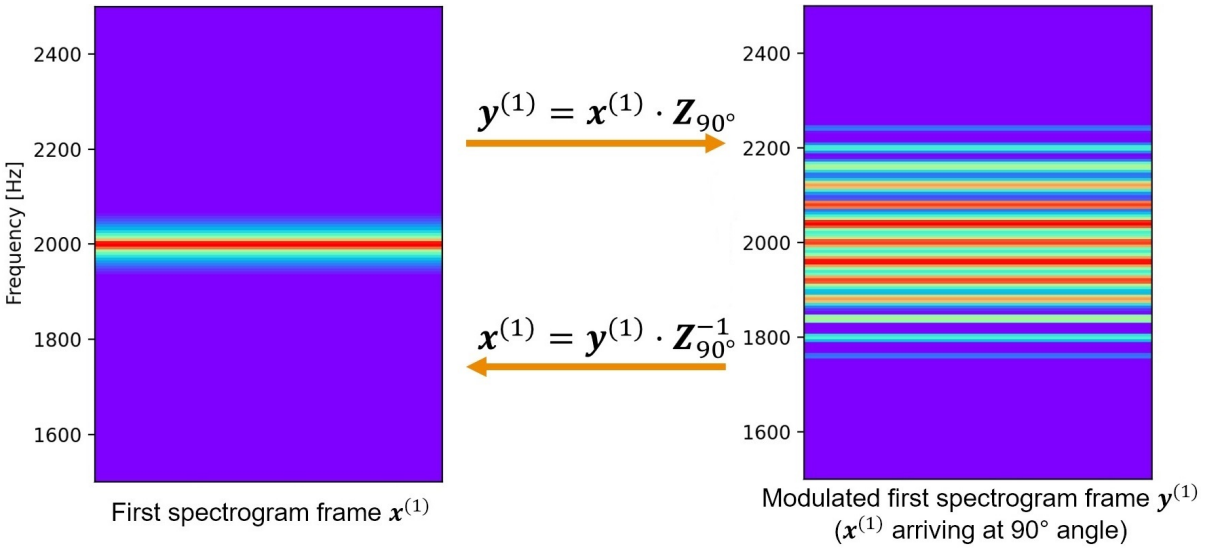


Figure 3.17: Matrix-based time warping - idea.

of the first spectrogram frame of the source signal and the modulated signal, respectively.

Inverting $\mathbf{Z}_\varphi$ provides us with the unmodulation matrix $\mathbf{Z}_\varphi^{-1}$ which allows us to compute $\mathbf{x}^{(1)}$ from $\mathbf{y}^{(1)}$ as

$$\mathbf{x}^{(1)} = \mathbf{y}^{(1)} \cdot \mathbf{Z}_\varphi^{-1} \,. \tag{3.32}$$

This idea is visualized in Figure 3.17 for a 2 kHz source signal. The advantage of this representation becomes apparent once we have two audio sources arriving from two different directions. Let us denote the first spectrogram frame of an audio signal with a DOA of 90° as $\mathbf{a}^{(1)}$ and the first spectrogram frame of another audio signal with a DOA of 180° as $\mathbf{b}^{(1)}$. The first spectrogram frame captured by a stationary microphone can therefore be expressed as $\mathbf{x}^{(1)} = \mathbf{a}^{(1)} + \mathbf{b}^{(1)}$. Separating these two signals is most likely impossible if no additional constraints are given, however, separation becomes possible with some prior knowledge of the audio sources. As an example, [20] separates a male and a female speaker from a single microphone recording given the characteristics of male and female speech.

On the other hand, the first spectrogram frame captured by the rotating microphone can be formulated as $\mathbf{y}^{(1)} = \mathbf{a}^{(1)} \cdot \mathbf{Z}_{90°} + \mathbf{b}^{(1)} \cdot \mathbf{Z}_{180°}$. This expression is visualized in Figure 3.18, where $\mathbf{a}^{(1)}$ and $\mathbf{b}^{(1)}$ correspond to single frequencies.

It can be observed that deriving $\mathbf{a}^{(1)}$ and $\mathbf{b}^{(1)}$ from $\mathbf{y}^{(1)}$ is not a straightforward task despite the simplicity of the source signals. However, the matrix-based representation allows us to define the

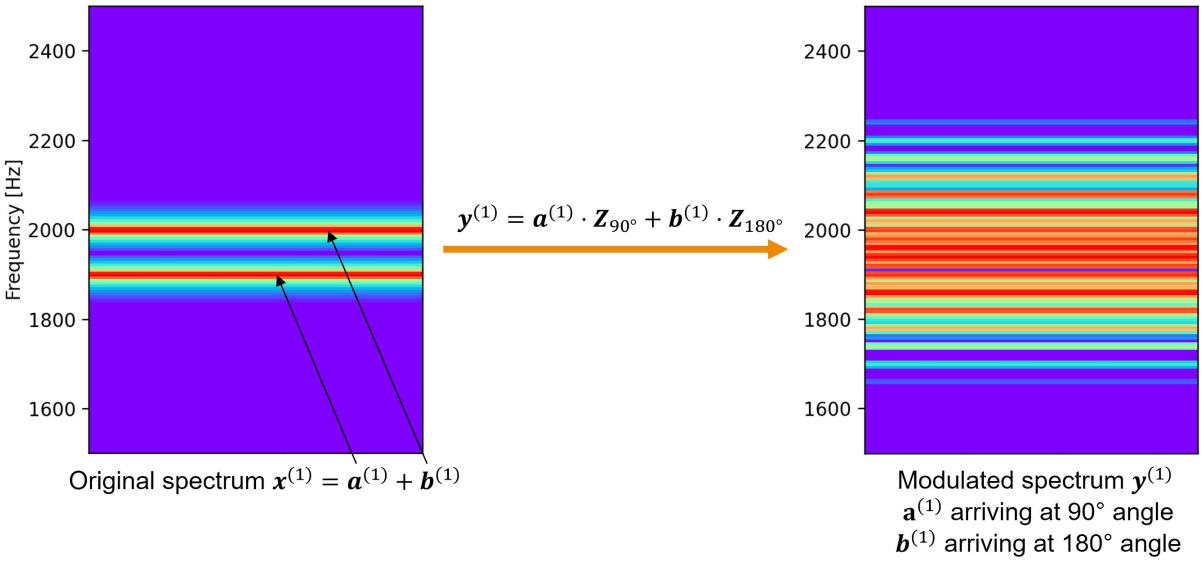


Figure 3.18: Matrix-based time warping of two source signals with different DOAs.

following function:

$$f(\mathbf{a}, \mathbf{b}) = \mathbf{y}^{(1)} - \mathbf{a} \cdot \mathbf{Z}_{90°} + \mathbf{b} \cdot \mathbf{Z}_{180°} \tag{3.33}$$

This function should allow us to approximate $\mathbf{a}$ and $\mathbf{b}$, which correspond to the first spectrogram frames of the audio signals to be determined, by minimizing its absolute value as

$$\mathbf{a}, \mathbf{b} = \underset{\mathbf{a},\mathbf{b}}{\operatorname{argmin}}(|f(\mathbf{a}, \mathbf{b})|) \ . \tag{3.34}$$

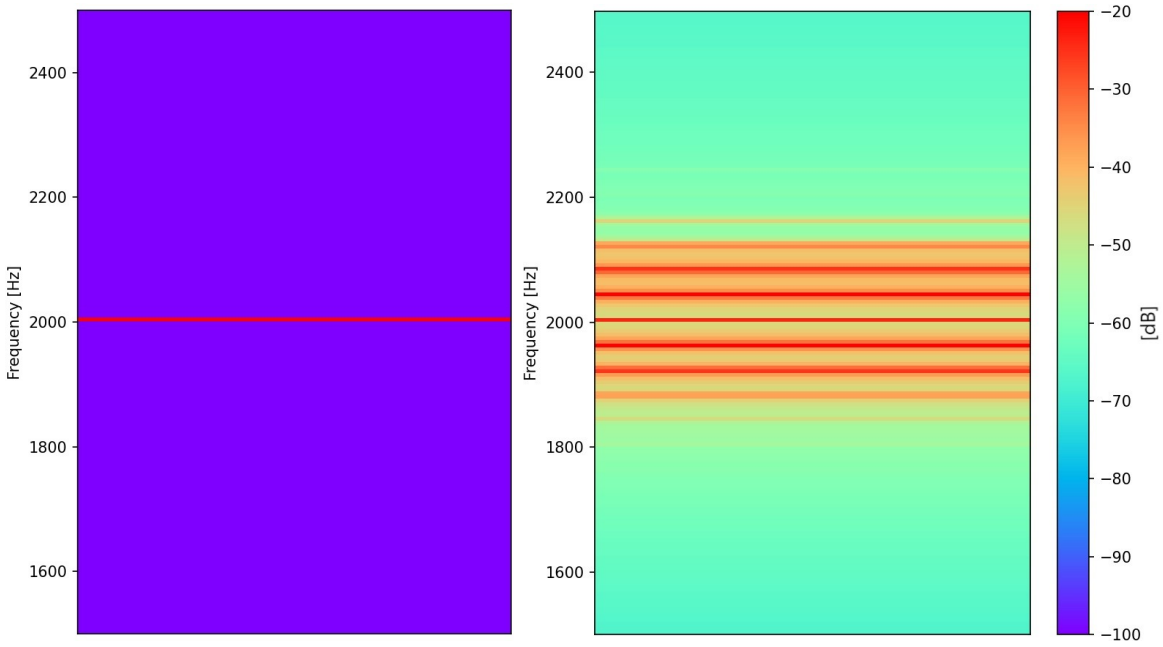
This, in theory, enables us to perform source separation of arbitrary audio signals as long as we have perfect knowledge of the location of our source signals and we can find a way to determine the corresponding modulation matrices. We will not investigate source separation specifically as it goes beyond the scope of this thesis, however, we will nonetheless explore how the modulation and unmodulation matrices can be derived. For this we define a set $\mathrm{F} = \{f_1, f_2, \dots, f_n\}$ which contains the center frequencies of each DFT bin, where $n = \frac{L}{2} + 1$ and L is the utilized DFT length. The center frequencies can be determined as

$$f_i = (i-1) \cdot \frac{f_s}{L} \ , \quad i \in [1, 2, \dots, n] \ , \tag{3.35}$$

where f_s corresponds to the sampling rate. If we compute the Fourier transform of $x_i(t) = \cos(2\pi f_i\, t)$ for any $f_i \in \mathrm{F}$ using a rectangular window we obtain a 'perfect' spectrum. An example is depicted in Figure 3.19a for $f_s = 48\,\mathrm{kHz}$, $L = 8192$ and $i = 343$, i.e. $f_{343} \approx 2004\,\mathrm{Hz}$. The depicted spectrum can be considered as being 'perfect' since it shows precisely one peak at the source frequency bin and is very close to zero everywhere else, i.e. there is almost zero spectral leakage. The modulated counterpart of $x_i(t)$, given by

$$z_i(t, \varphi, \theta) = \cos(2\pi f_i\, t + \beta(\theta) \cdot \sin(2\pi f_{rot}\, t + \varphi - 90°)) \;, \tag{3.36}$$

is shown in Figure 3.19b for $\beta(\theta) \approx 1.84$, $\varphi = 90°$ and $f_{rot} = 40\,\mathrm{Hz}$.



(a) Magnitude spectrum of $x_{343}(t)$ (b) Magnitude spectrum of $z_{343}(t, \varphi, \theta)$

Figure 3.19: Modulated and original spectrum using a frequency in F and a rectangular window.

Side note: Recall that $\beta(\theta) = \sin(\theta) \cdot \frac{2\pi r \cdot f_i}{c}$. The values we have chosen to obtain $\beta(\theta) \approx 1.84$ are $r = 5\,\mathrm{cm}$, $c = 343\,\frac{\mathrm{m}}{\mathrm{s}}$ and $\theta = 90°$.

The great advantage of the spectrum and therefore energy of $x_i(t)$ being perfectly concentrated into the i-th DFT bin is that the spectrum of $z_i(t, \varphi, \theta)$ indicates precisely how the energy and phase of the i-th DFT bin gets redistributed as we perform modulation for a given $\beta(\theta)$, f_{rot} and φ. If we now formulate the spectrum of $z_i(t, \varphi, \theta)$ as a vector $\mathbf{z}_{i_\varphi}$ as

$$\mathbf{z}_{i_\varphi} = [z_{i1_\varphi}\ z_{i2_\varphi}\ ...\ z_{in_\varphi}] \in \mathbb{C}^n \tag{3.37}$$

and normalize this vector to have unit energy, we can compute the modulated version of individual DFT bins of arbitrary signals by simple multiplication of their i-th bin with $\mathbf{z}_{i_\varphi}$. As an example, let us assume we have a signal whose spectrum $\mathbf{x}$ is zero at every DFT bin except the $i = 100$-th position, which we denote as x_{100}. The modulated spectrum $\mathbf{y}$ for DOA φ can then be computed by

$$\mathbf{y} = x_{100} \cdot \mathbf{z}_{100_\varphi}\ . \tag{3.38}$$

An arbitrary signal, such as $\mathbf{x}^{(1)}$ from Equation (3.31), can be modulated by applying the same principle to all DFT bins as

$$\mathbf{y}^{(1)} = \sum_{i=1}^{n} x_i^{(1)} \cdot\ \mathbf{z}_{i_\varphi}\ . \tag{3.39}$$

We can formulate this expression more compactly by stacking all $\mathbf{z}_{i_\varphi}$ to form matrix $\mathbf{Z}_\varphi$ as

$$\mathbf{Z}_\varphi = \begin{bmatrix} \mathbf{z}_{1_\varphi} \\ \mathbf{z}_{2_\varphi} \\ \vdots \\ \mathbf{z}_{n_\varphi} \end{bmatrix}\ . \tag{3.40}$$

This matrix allows us to modulate $\mathbf{x}^{(1)}$ by computing

$$\mathbf{y}^{(1)} = \mathbf{x}^{(1)} \cdot \mathbf{Z}_\varphi\ , \tag{3.41}$$

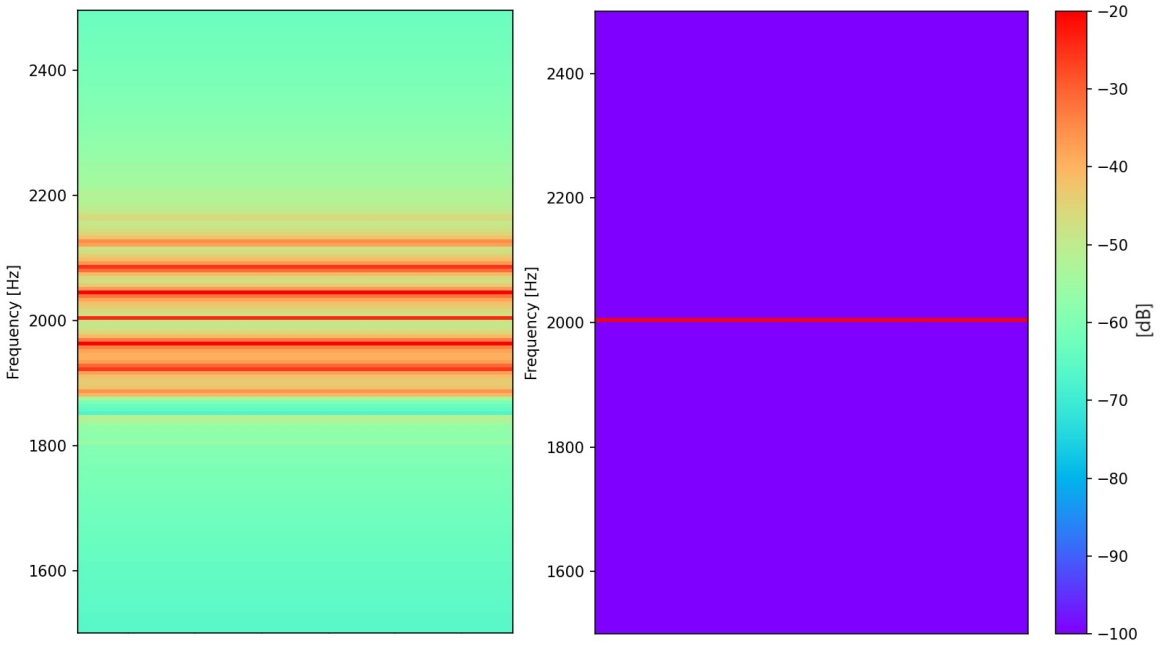
which is identical to the expression from Equation (3.30). $\mathbf{Z}_\varphi$ can be easily extended to consider arbitrary elevations θ as well by modifying θ in the computation of $z_i(t, \varphi, \theta)$, however, we will restrict ourselves to $\theta = 90°$ in this section. To summarize, the modulation matrix $\mathbf{Z}_\varphi$ can be obtained using the following algorithm:

1. Compute the spectra of $z_i(t, \varphi, \theta) = \cos(2\pi f_i\, t + \beta(\theta) \cdot \sin(2\pi f_{rot}\, t + \varphi - 90°))$ with a given DFT length L, rotational radius r, rotational speed f_{rot}, sampling rate f_s and DOA (φ, θ) for all $f_i \in \mathrm{F}$
2. Stack the resulting spectra as shown in Equation (3.40) to form matrix $\mathbf{Z}_\varphi$
3. Normalize the energy of $\mathbf{Z}_\varphi$ by multiplication with factor $2 \cdot \sqrt{\frac{f_s}{L}}$

Side note: Normalization of the energy of $\mathbf{Z}_\varphi$ is not strictly necessary for sound source localization, however, we choose to perform normalization nonetheless since this results in transformations using $\mathbf{Z}_\varphi$ to be energy conserving.

As we have mentioned earlier, the inverse of $\mathbf{Z}_\varphi$ can, in theory, be used to unmodulate $\mathbf{y}^{(1)}$ by computing $\mathbf{x}^{(1)} = \mathbf{y}^{(1)} \cdot \mathbf{Z}_\varphi^{-1}$. In practice, however, $\mathbf{Z}_\varphi$ has a very large condition number, leading to sizeable errors when computing its inverse. As the condition number is a topic beyond the scope of this thesis we will not elaborate further on this metric and refer to [26] for more details regarding its computation and influence on matrix inversion.

A workaround that allows for a direct computation of $\mathbf{Z}_\varphi^{-1}$ without the need of inverting $\mathbf{Z}_\varphi$ is to apply the same approach used to derive $\mathbf{Z}_\varphi$ to a signal that has a 'perfect' spectrum in the modulated domain, i.e. we search for a signal $\hat{z}_i(t)$ that produces the spectra shown in Figure 3.20 (in the case of $i = 343$) when using a rectangular window. In other words, we require a signal



(a) Magnitude spectrum of $\hat{z}_{343}(t)$ (b) Modulated magnitude spectrum of $\hat{z}_{343}(t)$

Figure 3.20: Ideal (modulated) magnitude spectrum of a signal $\hat{z}_i(t)$ using a rectangular window.

which is distorted in a way such that the modulation introduced when recording the signal with a rotating microphone perfectly cancels out these distortions to produce a perfect sinusoidal tone in F. Such a signal can be constructed by applying the TWA to a frequency in F. The downside of this approach is that any inaccuracies introduced by the interpolation step in the TWA will also be introduced into the spectrum used to construct the unmodulation matrix. To circumvent this we can choose to only compute $\hat{t}(t, \varphi)$ from Equation (3.27) (or $\hat{t}(t, \varphi, \theta)$ from Equation (3.29)) and use these time stamps to approximate $\hat{z}_i(t)$ as

$$\hat{z}_i(t) \approx x_i(\hat{t}(t, \varphi)) = \cos(2\pi f_i \, \hat{t}(t, \varphi)) \; . \tag{3.42}$$

The modulated counterpart of $x_i(\hat{t}(t, \varphi))$, given by computing $z_i(\hat{t}(t, \varphi), \varphi, \theta)$ using Equation (3.36), produces a close to perfect spectrum. The spectra of $x_i(\hat{t}(t, \varphi))$ and $z_i(\hat{t}(t, \varphi), \varphi, \theta)$ for $i = 343$ look identical to the spectra from Figure 3.20.

This means that $x_i(\hat{t}(t, \varphi))$ gives us a good indication of how the energy and phase of the i-th modulated frequency bin gets a redistributed as we unmodulate for a given $\beta(\theta)$, f_{rot} and φ. Therefore we can formulate the spectrum of $x_i(\hat{t}(t, \varphi))$ as a vector $\hat{\mathbf{z}}_{i_\varphi}$

$$\hat{\mathbf{z}}_{i_\varphi} = [\hat{z}_{i1_\varphi} \; \hat{z}_{i2_\varphi} \; ... \; \hat{z}_{in_\varphi}] \in \mathbb{C}^n \tag{3.43}$$

and then stack these vectors as

$$\hat{\mathbf{Z}}_\varphi^{-1} = \begin{bmatrix} \hat{\mathbf{z}}_{1_\varphi} \\ \hat{\mathbf{z}}_{2_\varphi} \\ \vdots \\ \hat{\mathbf{z}}_{n_\varphi} \end{bmatrix} \; . \tag{3.44}$$

to obtain $\hat{\mathbf{Z}}_\varphi^{-1}$. Note that we use the hat on $\hat{\mathbf{Z}}_\varphi^{-1}$ since it does not perfectly represent the inverse of $\mathbf{Z}_\varphi$.

It may at first seem as though performing unmodulation using $\hat{\mathbf{Z}}_\varphi^{-1}$ is vastly more efficient than using the TWA, since the unmodulation matrix needs to only be computed once for each DOA an then the unmodulation of arbitrary spectrogram frames simplifies to a vector-matrix multiplication. However, there are multiple caveats to the matrix-based approach:

1. So far, we have only shown unmodulation with respect to the first spectrogram frame as $\mathbf{x}^{(1)} = \mathbf{y}^{(1)} \cdot \mathbf{Z}_{\varphi}^{-1}$. If we want to unmodulate the second spectrogram frame we cannot simply compute $\mathbf{x}^{(2)} = \mathbf{y}^{(2)} \cdot \mathbf{Z}_{\varphi}^{-1}$, since φ refers to the DOA at the start of the spectrogram frame. If the microphone is not at exactly the same position at the start of our next spectrogram frame we need to compute $\mathbf{Z}_{\varphi}^{-1}$ again for a different DOA to compensate for the different microphone starting position. As an example, for $f_{rot} = 40\,\text{Hz}$, $f_s = 48\,\text{kHz}$ and $L = 8192$ the microphone completes $f_{rot} \cdot \frac{L}{f_s} \approx 6.83$ rotations in one DFT frame. This means that at the start of each spectrogram frame the DOA changes by $-61.2°$ with respect to the previous frame, assuming there is no frame overlap. This requires the computation of a large number of unmodulation matrices, which poses a problem since determining $\mathbf{Z}_{\varphi}^{-1}$ is expensive from a computational standpoint, especially for large L.

2. It may seem as though we could counteract this problem by making sure that f_{rot} is an integer multiple of $\frac{L}{f_s}$ or using appropriate spectrogram frame overlap, however, in practice f_{rot} is not perfectly constant, requiring us to compute multiple unmodulation matrices for the varying rotational speed in addition to the DOA shift.

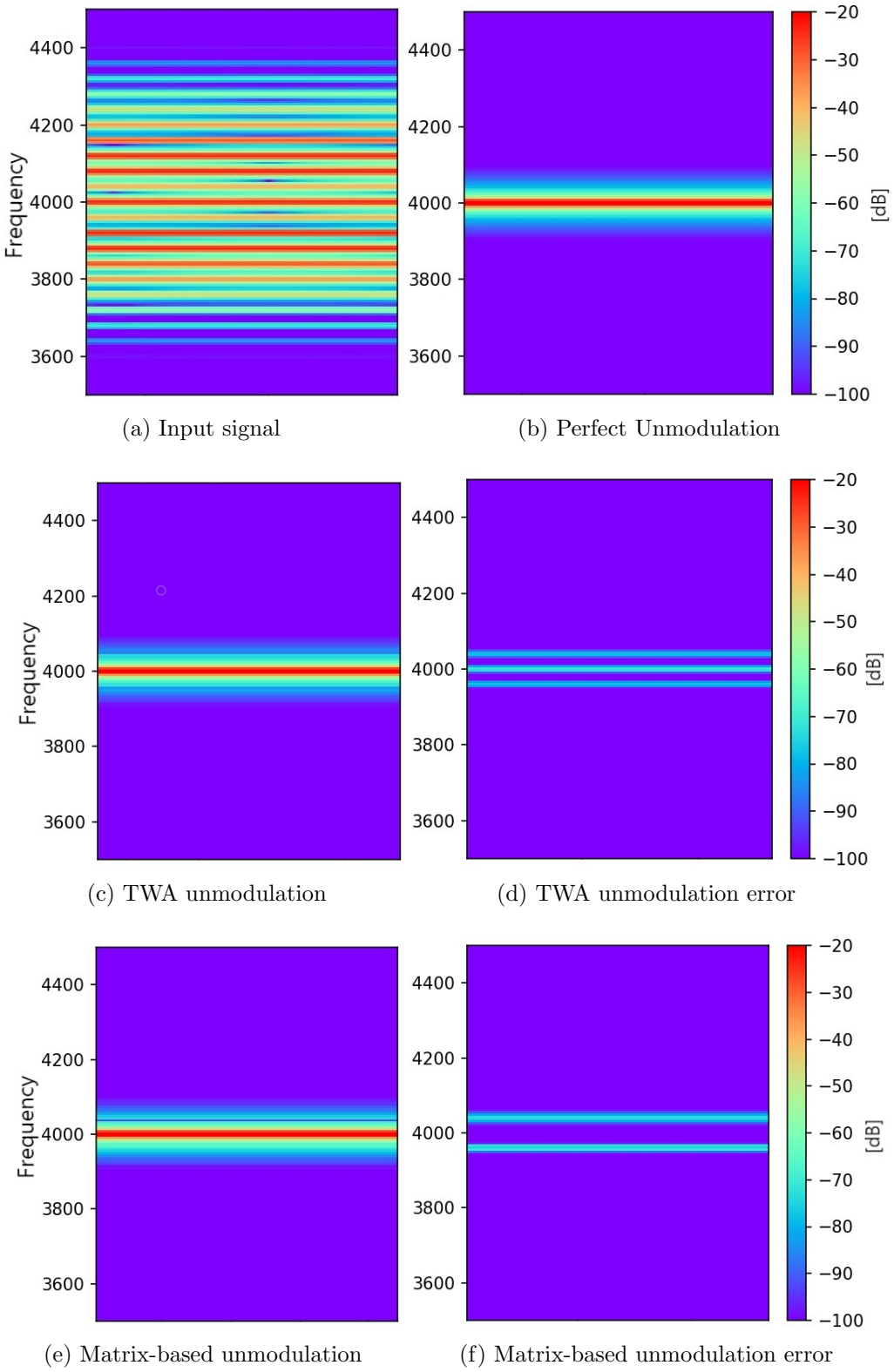
3. The unmodulation matrices have a large memory requirement. Using the `numpy.complex128` datatype and $L = 8192$ requires $8192^2 \cdot 128\,\text{bit} \cdot 0.125\frac{\text{byte}}{\text{bit}} \approx 1\,\text{GB}$ of memory for a single unmodulation matrix.

We could mitigate the first two problems if the rotational speed of the rotating microphone could be set more precisely. Additionally this would make the third concern less problematic since fewer unmodulation matrices need to be stored. This incentivises the design of a more accurate successor to the REM prototype such that matrix-based time warping can be utilized without the need for computing a separate unmodulation matrix for each frame. Therefore, a microphone with sufficiently precise speed control could allow for all the matrices to be computed beforehand, potentially enabling real time audio processing. This would allow for some interesting applications, e.g. real time tracking of multiple audio sources.

To conclude this chapter, let us give a quick comparison between the two unmodulation approaches. Figure 3.21 shows a comparison between the unmodulation performance of the TWA and the matrix based unmodulation for a 4 kHz source signal. It can be observed that both approaches perform similarly well, however, they appear to produce slightly different artifacts. The matrix-based approach reconstructs the main frequency almost perfectly but there is a noticeable error at the first sidebands. The TWA, on the other hand, does not reconstruct the main frequency as well as the matrix-based approach, however, the error at the first sidebands is reduced. To evaluate the accuracy of both approaches more precisely we compute the total error between the true unmodulated signal spectrum $\mathbf{x}$ and the computed unmodulated signal spectrum $\hat{\mathbf{x}}$ as

$$Err = \sum_{i=1}^{n} |x_i - \hat{x}_i|^2 \,. \tag{3.45}$$

For the example shown in Figure 3.21 we find that the errors of the TWA and matrix-based time warping are approximately $2.12 \cdot 10^{-7}$ and $2.72 \cdot 10^{-7}$, respectively. For other input signals the performance of both algorithms is also similar, however the TWA always outperformed the matrix-based approach for all the signals we tested. For performance reasons we will be exclusively utilizing the TWA in the following chapters.



(a) Input signal (b) Perfect Unmodulation

(c) TWA unmodulation (d) TWA unmodulation error

(e) Matrix-based unmodulation (f) Matrix-based unmodulation error

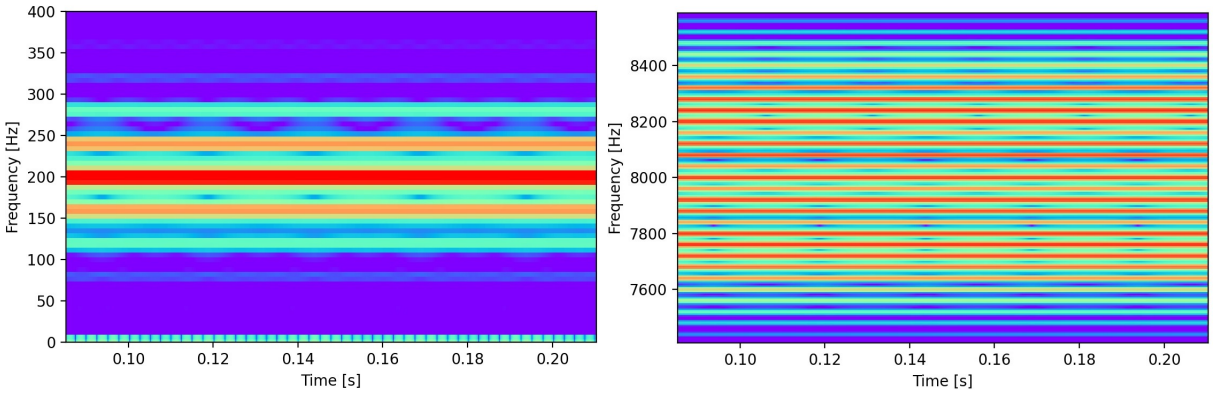
Figure 3.21: Unmodulation performance using the TWA and matrix-based time warping.

Chapter 4

Direction of Arrival Estimation - Theoretical Verification

We will now use the TWA to estimate the location of one or multiple sound sources. Section 4.1 presents the DOA estimation approach and determines the localization accuracy of one single frequency sound source in two-dimensional space, i.e. it is assumed that $\theta = 90°$. In Section 4.2 this localization method will be refined to allow for separate processing of different frequency bands. Subsequently, Section 4.3 and Section 4.4 explore the localization accuracy of one and multiple complex sound sources in two-dimensional space, respectively. Finally, Section 4.5 investigates the DOA estimation accuracy in three-dimensional space.

4.1 Localization of a Single Frequency Source in 2D Space

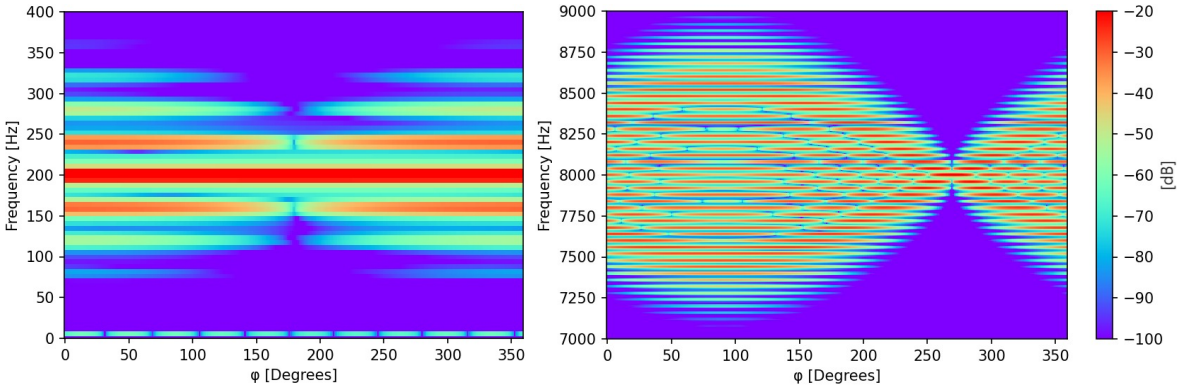


(a) Modulated 200 Hz sine wave

(b) Modulated 8 kHz sine wave

Figure 4.1: Spectrograms for $f_s = 48\,\text{kHz}$, $f_{rot} = 40\,\text{Hz}$, $r = 5\,\text{cm}$ and DFT length $L = 8192$.

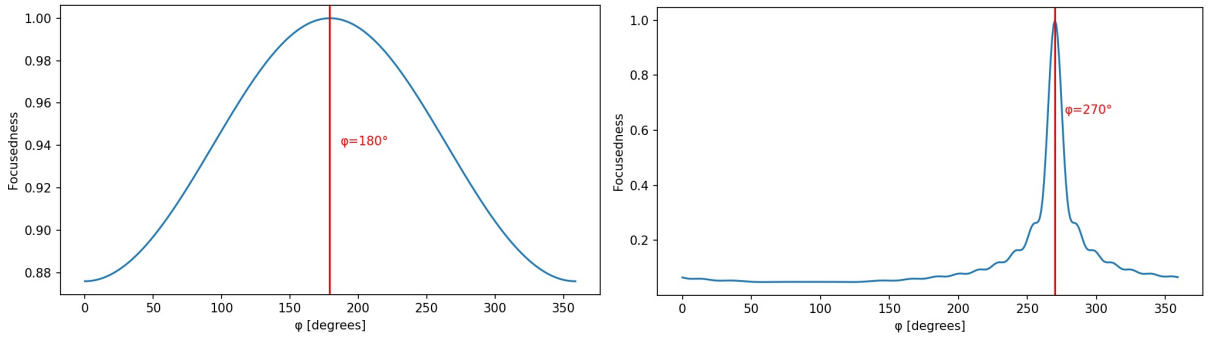
Consider again the spectrograms of the modulated 200 Hz and 8 kHz sine waves from Section 3.3. They are depicted again in Figure 4.1 for convenience. We can now use our derived TWA and apply it to the first frame of these spectrograms for $\varphi \in [0°, 1°, ..., 358°, 359°]$, for example. Placing all the resulting spectrogram frames side-by-side results in a φ-dependent spectrogram, which we will refer to as *azimuth-spectrogram*, shown in Figure 4.2.



(a) Azimuth-spectrogram of a 200 Hz sinusoid (b) Azimuth-spectrogram of an 8 kHz sinusoid

Figure 4.2: Azimuth-spectrograms of the signals from Figure 4.1.

The true DOAs of the 200 Hz and 8 kHz sine waves were $\varphi = 180°$ and $\varphi = 270°$, respectively. We can clearly see that the azimuth-spectrograms focus into one point as φ approaches the correct values. This becomes even more clear when computing the φ-dependent normalized focusedness of the plots, which is depicted in Figure 4.3.

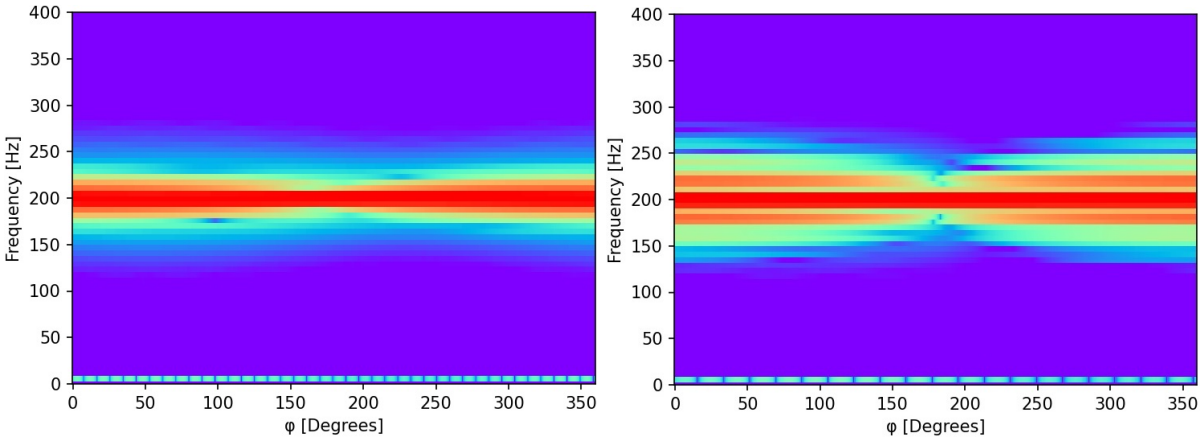


(a) Focusedness of 200 Hz azimuth-spectrogram (b) Focusedness of 8 kHz azimuth-spectrogram

Figure 4.3: Focusedness of the azimuth-spectrograms from Figure 4.2 with marked peaks.

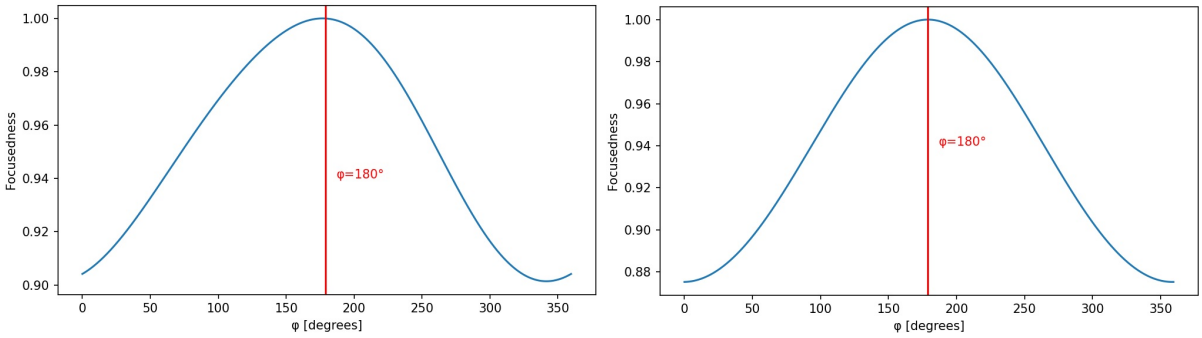
It can be observed that the peaks of the focusedness plots precisely correspond to the true DOAs of the sine waves. Furthermore it is clearly visible that the focusedness has a substantially sharper peak and a lower minimum value for the 8 kHz signal as compared to the 200 Hz signal. This is due to the modulation index being larger for higher frequencies (see Equation (3.11)), which also leads us to presume that DOA estimation of high frequencies is more accurate and robust to noise in theory.

Let us now investigate the impact of f_{rot} on the localization accuracy. Figure 4.4 and Figure 4.5 depict a comparison between 10 RPS and 20 RPS for the 200 Hz signal.



(a) 'Unmodulated' 200 Hz sinusoid (10 RPS) (b) 'Unmodulated' 200 Hz sinusoid (20 RPS)

Figure 4.4: Azimuth-spectrograms of a 200 Hz sinusoid for 10 RPS and 20 RPS.



(a) Focusedness for 10 RPS (b) Focusedness for 20 RPS

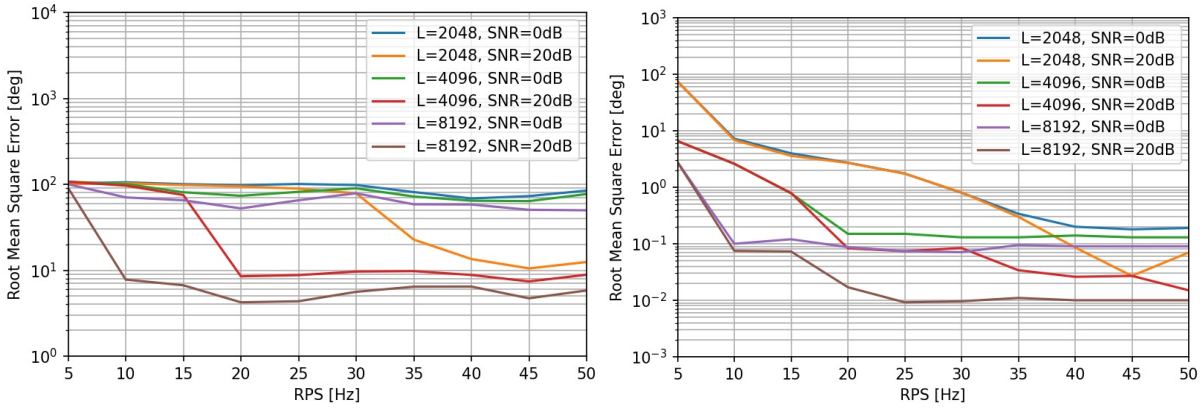
Figure 4.5: Focusednesses of the azimuth-spectrograms from Figure 4.4 with marked peaks.

It can be observed that the focusedness is less precise in the 10 RPS case, which is due to the sidebands overlapping. This leads us to presume that once the sidebands fully separate there should be no increase in DOA estimation accuracy as we further increase f_{rot}. Thus, in practice, the lowest rotational speed should be selected which causes the sidebands to separate, since unnecessarily high rotational speeds result in additional motor and wind noise. We expect this minimum required rotational speed to be dependent on the utilized DFT length L, since smaller transformation lengths result in wider sidebands (and mainbands).

To verify these presumptions we simulate 100 DOA estimation trials for a 125 Hz and 8 kHz sound source at microphone rotational speeds varying between 5 RPS and 50 RPS. Randomly generated pink noise is added to the signals at 0 dB and 20 dB SNR and three DFT lengths $L = 2048/4096/8192$ are used. To quantify the DOA estimation error we use the root mean square error (RMSE), which according to [13] is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\varphi - \hat{\varphi}_i)^2}\,, \tag{4.1}$$

where N is the number of estimation trials, φ is the true DOA and $\hat{\varphi}_i$ is the estimated DOA of the i-th trial (in degrees). The diameter of the rotational motion is chosen to be 10 cm and the microphone sampling rate as $f_s = 48\,\text{kHz}$, since this corresponds to the values used for our REM prototype. The results of the simulations are shown in Figure 4.6. Note that RMSE $\approx 104°$ if we randomly guess the DOA. This value was derived by performing 10^8 random guesses and subsequently computing the RMSE.



(a) 125 Hz source frequency

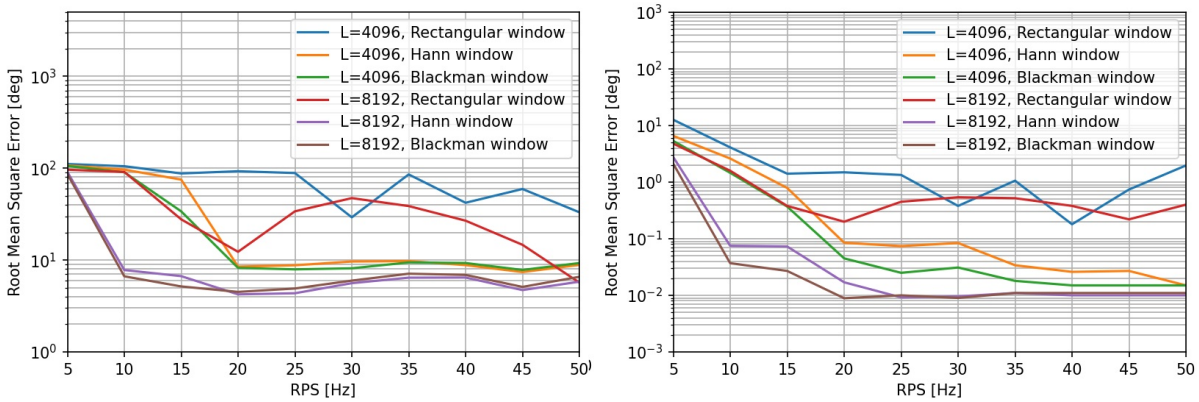
(b) 8 kHz source frequency

Figure 4.6: RMSE of two frequencies for various DFT lengths L, RPS and SNR.

From the depicted results it can be observed that for each L there is a clear rotational speed at which the RMSE substantially improves. The minimum required rotational speeds RPS_{min} for $L = 2048/4096/8192$ appear to be $\text{RPS}_{min} \approx 40/20/10$, i.e. doubling the DFT length approximately halves the required rotational speed. Interestingly, the SNR has virtually no impact on the RMSE before this minimum rotational speed is reached, however, after RPS_{min} has been reached the RMSE is consistently lower for the 20 dB SNR case. Another noteworthy finding is that a larger L results in a lower RMSE, which is most likely due to an improved frequency resolution and noise suppression inherent with the increased DFT length. A final notable observation is that the localization accuracy of the 8 kHz signal is almost three orders of magnitude better than that of the 200 Hz signal. This agrees with our hypothesis from earlier that localization of higher frequencies is more accurate and robust to noise.

Up until this point we have exclusively utilized Hann windows for the computation of all shown spectrograms, however, different windows may perform better in regards to the localization accuracy. In theory, windows that have a narrow main lobe should require a lower minimum

rotational speed since the sidebands will separate sooner. On the other hand, narrower main lobes result in higher side lobe levels, increasing the spectral leakage and therefore negatively affecting the accuracy of the focusedness. To see which of these factors affects the localization accuracy more we perform 100 localization trials of a 125 Hz and 8 kHz source signal using three windows: A rectangular window, a Hann window and a Blackman window. The rectangular window provides us with the narrowest main lobe at 1 bin (in regards to the equivalent noise bandwidth) and the highest maximum side lobe level at -13 dB. The Blackman window has the widest main lobe at 1.73 bins and the smallest maximum side lobe level at -58 dB. The Hann window lies approximately in the middle at 1.5 bins main lobe width and -32 dB maximum side lobe level [28][11]. A rotational diameter of 10 cm is chosen and additive pink noise is added to the signals at 20 dB SNR. Furthermore, we test two DFT lengths $L = 4096$ and $L = 8192$. The results are plotted in Figure 4.7.



(a) 125 Hz source frequency

(b) 8 kHz source frequency

Figure 4.7: RMSE of two frequencies for various windows, DFT lengths L and RPS.

It can be observed that the rectangular window provides comparably unreliable and inconsistent results. The difference between the Hann and the Blackman window are much smaller, with the Blackman window showing a slight improvement in consistency and localization accuracy. Therefore, every following spectrogram will be computed using a Blackman window from now on.

We will now investigate the impact of the rotation diameter on the DOA estimation accuracy. Similarly to before we perform 100 estimation trials for a 125 Hz and 8 kHz sound source at rotational diameters ranging from 2 cm to 20 cm. The rotational speed is chosen to be 40 RPS and the DFT length as $L = 8192$. Additionally, we add randomly generated pink noise to the signals at 0 dB and 20 dB SNR. The results of the simulations are depicted in Figure 4.8.

It can be observed that the RMSE decreases as the diameter increases. This was to be expected since the modulation index is dependent on the rotational radius (see Equation (3.11)), increasing the range of values for the focusedness for higher radii. It appears that we can increase the diameter indefinitely to achieve a higher localization accuracy, however, an increase in diameter leads to an

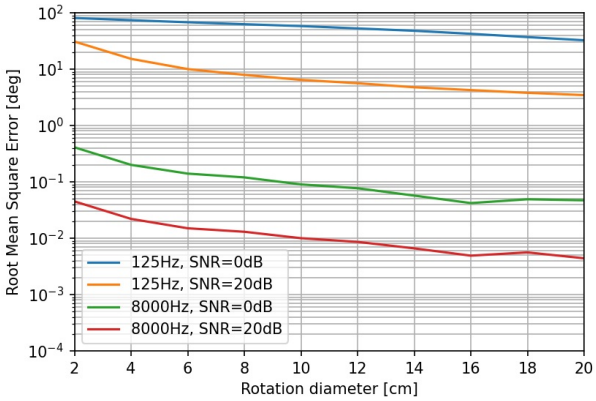


Figure 4.8: RMSE of two frequencies for various rotation diameters and SNR.

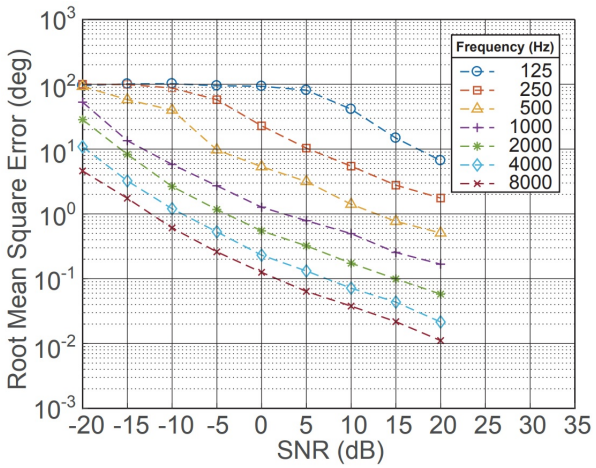
increase in microphone speed, which in turn results in increased wind noise. Additionally, most real sound sources do not emit plane waves, resulting in a certain wave curvature depending on the distance between the microphone and the sound source, as well as the geometry of the sound source. This curvature is not taken into account by the TWA, leading to an error that becomes larger as we increase the diameter.

Side note: If we extend the capabilities of the TWA such that it can take into account different wavefront curvatures we could potentially not only estimate the DOA of sound sources but also their distance to the microphone. In such a case it would be beneficial to choose the maximum possible rotational diameter to allow for more accurate distance estimation.

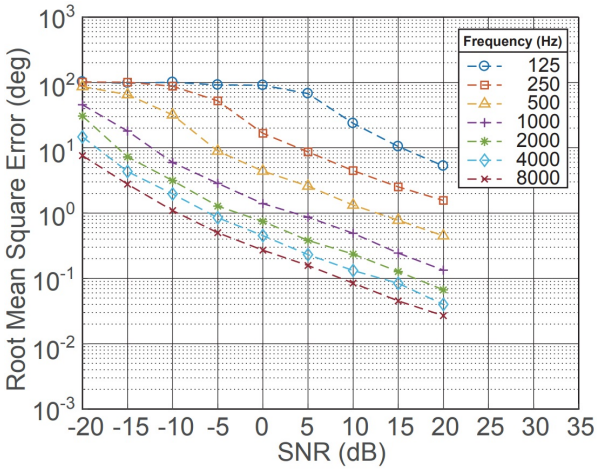
We will now directly compare the DOA estimation accuracy of our approach with the method from [12]. Similarly to the paper we perform 100 DOA estimation trials of source frequencies 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz and 8 kHz mixed with randomly generated pink noise at various SNR ranging from -20 dB to 20 dB. The rotational diameter is chosen to be 25 cm since this is the value used in [12]. Additionally, we use a sampling rate of $f_s = 48\,\text{kHz}$, a DFT length of $L = 8192$ and rotational speeds of 20 RPS and 40 RPS. We choose these speeds since 20 RPS correspond to the minimum speed of our REM prototype and 40 RPS represent the maximum speed the REM can reach without corrupting the audio signal too much due to self noise. The results of our simulations are depicted in Figure 4.9.

It can be observed that all the plotted curves follow a very similar characteristic pattern, however, there are some noteworthy findings: Unlike the approach from [12] our DOA estimation method does not work regardless of the source frequency below an SNR of -10 dB at 20 RPS and -5 dB at 40 RPS. Beyond this point, however, our method consistently outperforms [12], especially for low source frequencies. Interestingly, the localization accuracy of low frequencies is worse and that of high frequencies is better at 40 RPS as compared to 20 RPS.

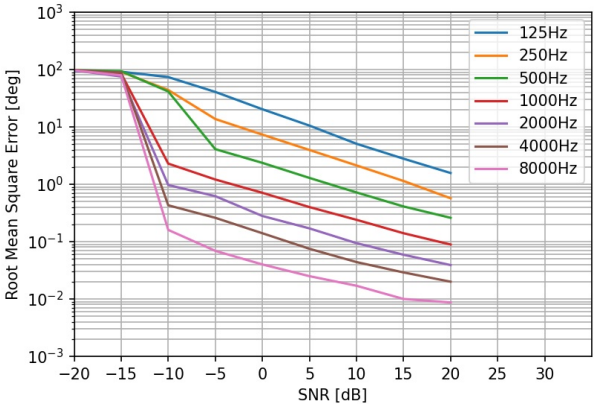
At this point it must be noted that the approach from [12] assumed knowledge of the source



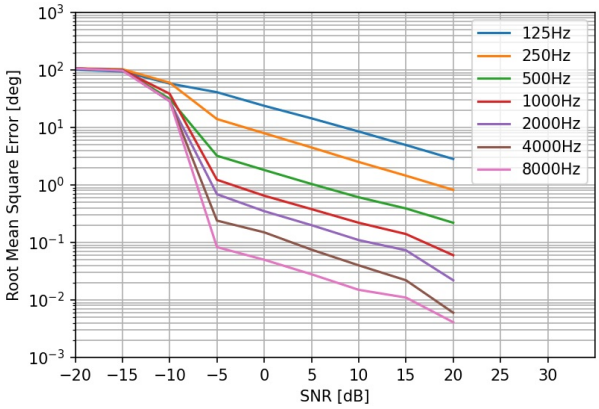
(a) 9.58 RPS, approach from [12]



(b) 17.24 RPS, approach from [12]



(c) 20 RPS, our approach



(d) 40 RPS, our approach

Figure 4.9: Comparison of the DOA estimation accuracy for various RPS using the approach from [12] and the TWA.

frequencies, therefore the CoG algorithm was reduced to a small interval when computing the instantaneous frequency. Our method does not make any assumptions regarding the source frequency, which may explain the poor performance at low SNR, despite the true DOA being visible by visual inspection of the -20 dB SNR case for the 8 kHz signal at 40 RPS, for example. Its azimuth-spectrogram is depicted in Figure 4.10.

It is clearly visible that the true DOA lies at approximately 180°, however, the focusedness fails to detect the peak since it is being overpowered by the pink noise which has more energy in lower frequency regions. If we assume knowledge of the source frequency we could choose to compute the focusedness only for the frequency bin most closely associated with 8 kHz and therefore easily approximate its DOA. However, since we would like to localize arbitrary sound sources we require a more general approach, which we will elaborate on in the following section.

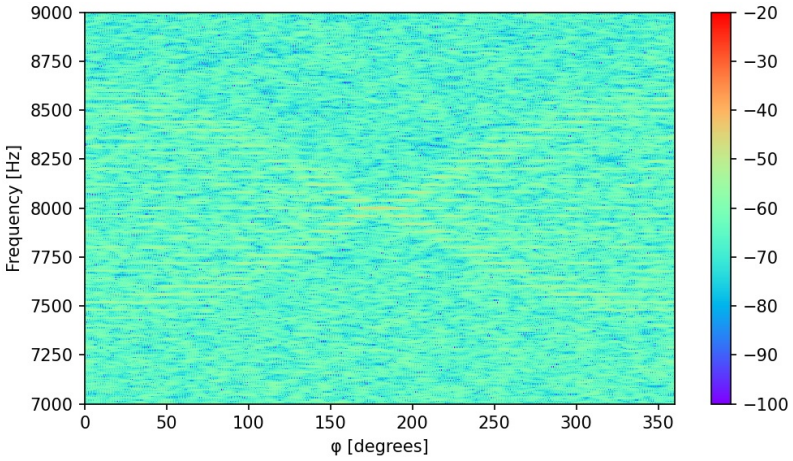


Figure 4.10: Noisy azimuth-spectrogram of an 8 kHz sine wave at 40 RPS and -20 dB SNR.

4.2 Localization Using Subband Processing

On important note regarding this section: The extensive testing in the previous section was performed to give us a general idea of the impact of various parameters and a baseline performance of the focusedness-based localization. We will now perform one final modification of the localization approach to improve performance in low SNR situations as well as allow for multiple source localization. This modification opens the door for a very large number of additional signal processing ideas which go beyond the scope of this thesis. Therefore we choose one possible implementation of this modification and investigate its accuracy, as well as point out what other possibilities could be explored in the future. All numerical parameters in this section were chosen 'by eye' and will not be optimized as vigorously as the parameters from the previous section, since the number of possible parameter combinations to evaluate are too great to be included in this thesis.

Up until this point the focusedness has been evaluated across the entire bandwidth of the spectrogram. In practical settings, however, this may lead to inaccuracies if the source signal is restricted to a small bandwidth under noisy conditions, as the previous section has shown. Additionally, localization of multiple sound sources in different frequency bands is not possible if there are large level differences between the signals. This motivates separating the spectrogram into multiple subbands and performing focusedness-based localization for each band individually. A suitable filter bank for separating the spectrogram is illustrated in Figure 4.11. We choose an overlapping logarithmic filter bank since for each frequency there should be at least one filter that fully encapsulates a frequency and all of its sidebands. Since the sidebands spread more as the main frequency increases we require wider filters at higher frequencies. We use 20 filters since rough tests showed that dividing the signal into 20 subbands produced good results.

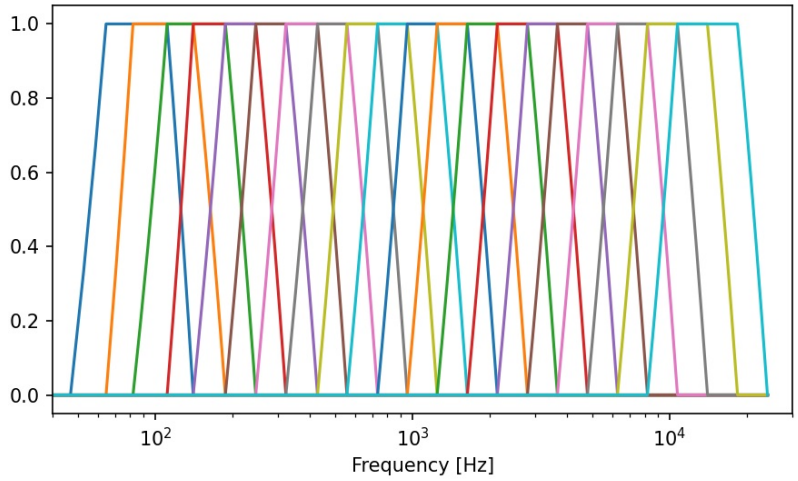
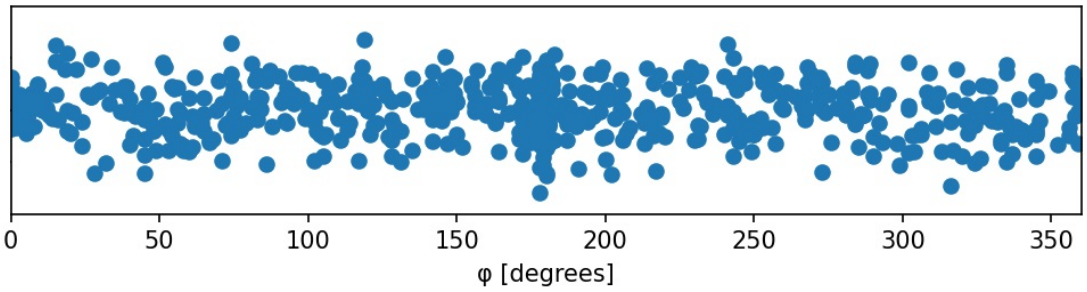


Figure 4.11: Proposed logarithmic filter bank with 20 filters.

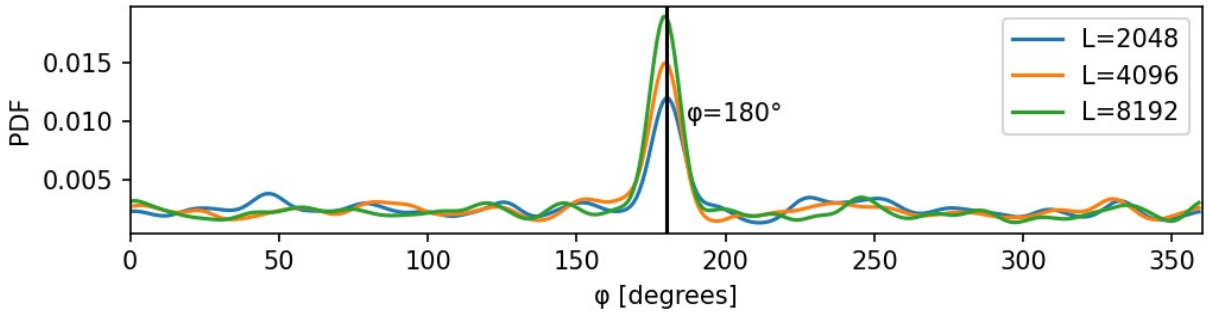
The main difficulty associated with subband processing is combining the DOA estimations from the individual subbands into one or multiple DOA predictions. One possibility is to perform a weighted average of the individual predictions based on the energy associated with each subband, however, this approach does not allow for the detection of multiple sound sources. Additionally, it would be beneficial to assign a higher credibility to DOA guesses from a filter associated with larger frequencies, since high frequencies can be located with a higher precision. Furthermore, the energy within a subband does not necessarily indicate DOA estimation reliability, since the energy could stem from noise rather than the signal to be located. A method for filtering out bad DOA predictions would be to devise an algorithm which detects frequencies within the signal that remain reasonably constant for short time periods and subsequently uses only the subbands containing these frequencies for DOA estimation. This would be beneficial since constant frequencies are needed to induce the Doppler shift required for the focusedness to work and thus the DOA predictions of subbands containing noise will be discarded. However, devising such an algorithm that works reliably on complex and noisy spectrograms is a very challenging task in itself and will therefore be left for future research.

For the sake of simplicity we use a straightforward combination method: We obtain the single angle with maximum focusedness for each subband and spectrogram frame and use these values as our DOA predictions. Subsequently we estimate the probability density function (PDF) of the obtained predictions using Parzen window density estimation with a Gaussian kernel. For more information regarding this density estimation method see [25]. Our final DOA prediction corresponds to the angle at which our estimated PDF has its maximum. Applying this modified approach to the -20 dB SNR 8 kHz signal from Figure 4.10 gives us the DOA estimation points in Figure 4.12a when observing the signal for 1 s and using $L = 4096$. Note that the y-axis has been added for visualization purposes only. The PDFs of these estimation points and for points

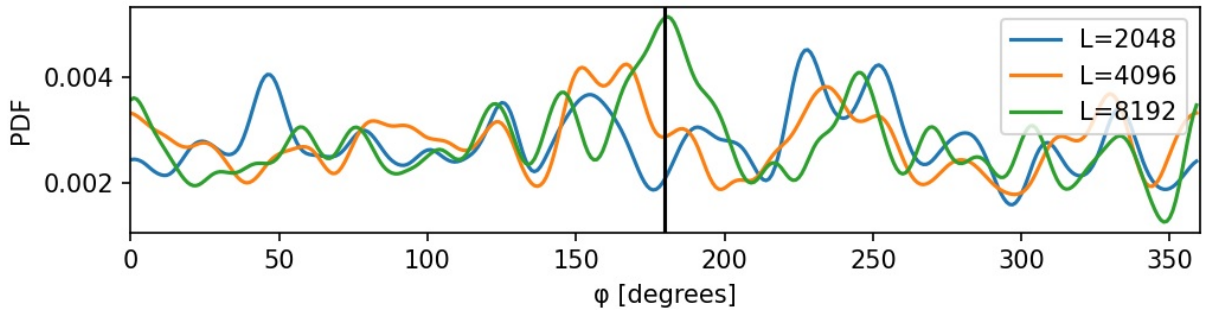
obtained using other DFT lengths are depicted in Figure 4.12b, which show clear peaks at the true DOA of the source signal. Localization is even possible down to a -30 dB SNR for $L = 8192$ as can be observed from Figure 4.12c.



(a) DOA estimation points for $L = 4096$ and -20 dB SNR



(b) PDF of DOA estimation points with peak at $\varphi = 180°$ for various DFT lengths at -20 dB SNR



(c) PDF of DOA estimation points with peak at $\varphi = 180°$ for various DFT lengths at -30 dB SNR

Figure 4.12: Localization using Parzen window density estimation of the DOA estimation points.

Some final remarks regarding this new approach: From a practical standpoint we perceive an accuracy of 1° to be sufficient for our DOA estimation. Therefore, from now on, we compute the TWA for all angles at 1° increments for performance reasons. Furthermore, we assume that all sound sources are separated by at least 22.5°. Following this assumption we found that a Gaussian kernel width of 31° produced reliable PDFs. Additionally, we use a spectrogram frame shift of 1024 samples to provide a trade-off between performance and the number of DOA predictions. Finally, it must be noted that although this modified approach increases the

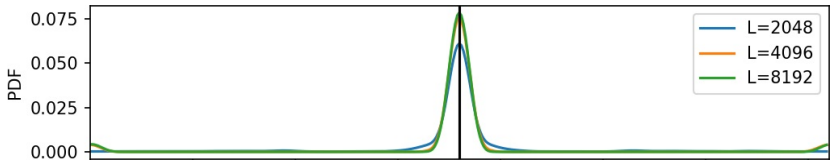
reliability of DOA estimation, the absolute localization accuracy decreases due to the smoothing we apply when using the Gaussian kernel. For this reason we will not provide a direct comparison with the DOA estimation plots from Figure 4.9 and instead use the shape of the obtained PDFs as an indication of the localization accuracy.

4.3 Localization of a Single Complex Source in 2D Space

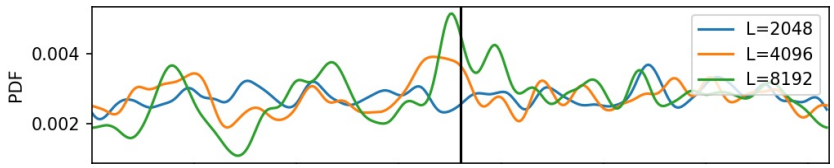
Localization of complex audio signals using our proposed approach is generally only possible if there are frequencies in the source signal which remain reasonably constant during one DFT frame. This encourages the use of shorter DFT lengths as it increases the chances that frequencies in the source signal will be constant for an entire frame. The expense of the shorter DFT length is a slight decrease in localization accuracy and an increased required rotational speed, which results in higher wind and motor noise, further decreasing the localization accuracy. Additionally, we expect that our approach does not allow for the localization of percussive and noise-like sources, since there will be no observable Doppler frequency shift. This, however, may be the case in general for any DOA estimation algorithm using a single moving microphone if no prior information regarding the source signal is given. We will experiment with $L = 2048$, $L = 4096$ and $L = 8192$, which for a sampling rate of 48 kHz corresponds to frame lengths of approximately 43 ms, 85 ms and 171 ms, respectively. Additionally, we explore localization accuracy for $f_{rot} = 20\,\text{RPS}$ and $f_{rot} = 40\,\text{RPS}$, since these are the minimum and maximum reasonable speeds of our REM prototype.

We test 7 different complex signals: All the frequencies 125 Hz, 250 Hz, ..., 4 kHz, 8 kHz combined, randomly generated pink noise, a male voice sample, a female voice sample, a simple drum groove, a short excerpt from a piano concerto and an exponential sine sweep. All signals are roughly 2 s long and modulated for a DOA of $\varphi = 180°$ and a rotational radius $r = 5\,\text{cm}$. To obtain the baseline performance we first investigate the localization accuracy without any additive pink noise. The results for various DFT lengths and $f_{rot} = 40\,\text{RPS}$ are depicted in Figure 4.13, where the true DOAs are indicated with a black line.

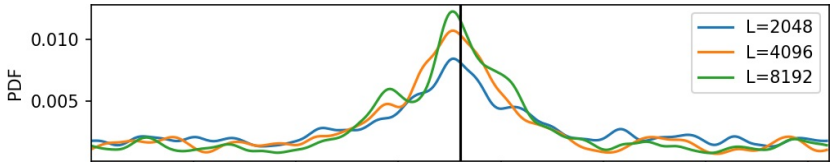
It can be observed that a larger DFT length leads to an improved, more decisive DOA estimation in all cases. Therefore it seems that the gain in localization accuracy from using shorter DFT lengths due to frequencies being more constant within a frame is minimal. The best results were achieved with the sine tones, since this signal features a large number of constant frequencies. An interesting observation is the small peak at $\varphi = 0°$. Further investigation showed that this peak is caused by subbands that encapsulate only the sidebands of a frequency and not the main frequency, since the energy in the sidebands is maximized when the modulation index is at its peak, i.e. when the TWA is performed in the opposite direction as the true DOA.



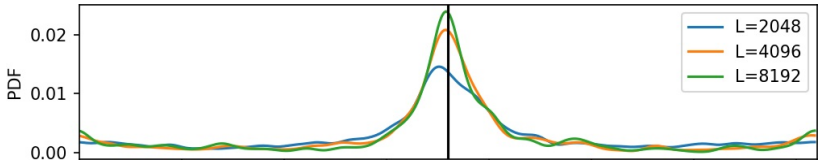
(a) 125 Hz, 250 Hz, ..., 4 kHz, 8 kHz sine tones combined



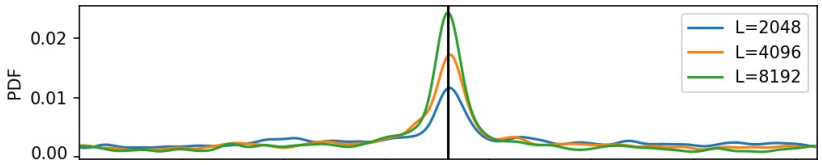
(b) Randomly generated pink noise



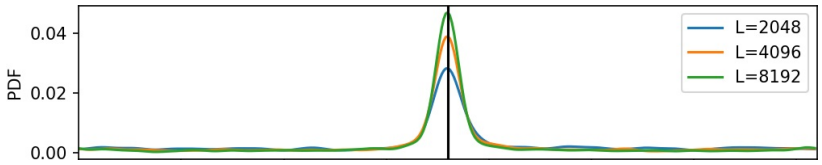
(c) Male voice sample



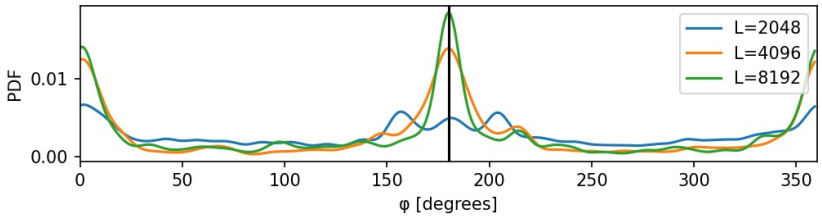
(d) Female voice sample



(e) Simple drum groove



(f) Piano concerto excerpt

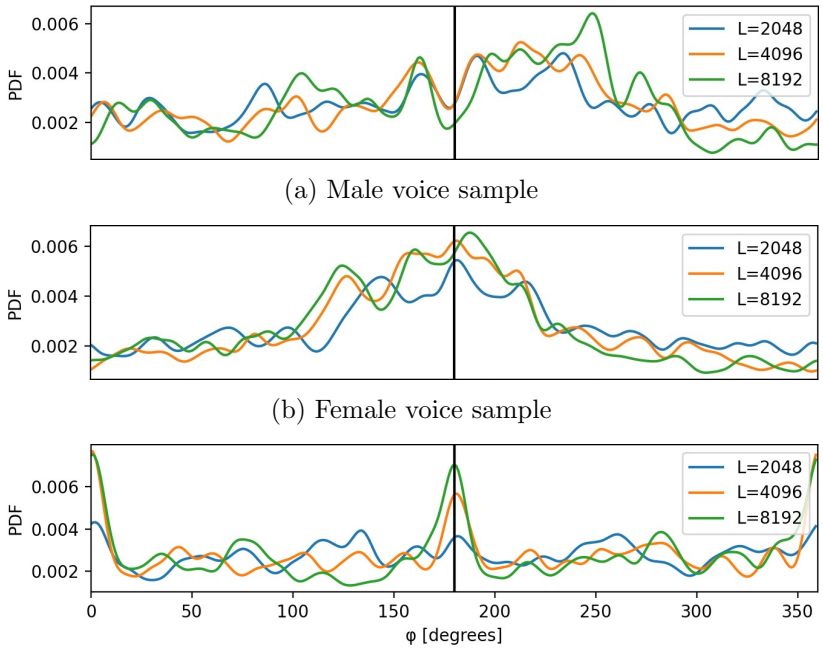


(g) Exponential sine sweep

Figure 4.13: DOA estimation accuracy for various sound sources at $f_{rot} = 40$ RPS.

As we expected, the classical music signal also showed very good performance, since it features many constant frequencies. A surprising observation is the localization accuracy of the drum groove. We believe this is due to the constant tonal components inherent to the bass and snare drum within the signal. The speech samples were detected with reasonable accuracy with the female voice being detected more reliably due to its higher fundamental frequency. The exponential sine sweep was detected very well for $L = 4096$ and $L = 8192$, which was unexpected since it lacks a constant frequency. However, a second clear peak is visible at $\varphi = 0°$ the cause of which we believe is the same as the small peak visible for the sine tones signal. Finally, the DOA of the pink noise signal could not be detected, which matched our expectations. Therefore we will not be including this signal in any further tests.

We performed the same tests for $f_{rot} = 20\,\text{RPS}$ and expected the localization accuracy to only deteriorate for $L = 2048$. Contrary to our expectations the localization accuracy decreased for every signal and every DFT length. The DOA estimation accuracy was only marginally worse for the sine tones, the classical music signal and the drum groove, however, in the case of the speech signals and the exponential sine sweep the detection was very unreliable or failed entirely. The results for these signals are depicted in Figure 4.14.



(a) Male voice sample

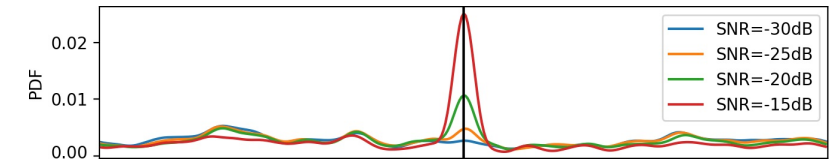
(b) Female voice sample

(c) Exponential sine sweep

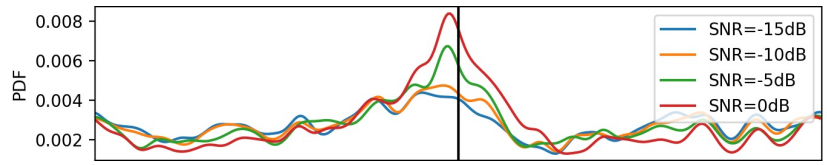
Figure 4.14: DOA estimation accuracy for various sound sources at $f_{rot} = 20\,\text{RPS}$.

Our expectation was that, in the case of the exponential sine sweep, the slower rotational speed would lead to a decrease in height of the peak at $\varphi = 0°$ since there is less separation of the sidebands, decreasing the chance of subbands only containing the sidebands of a frequency and not its main frequency. In our simulation, however, the height of the incorrect peak increased

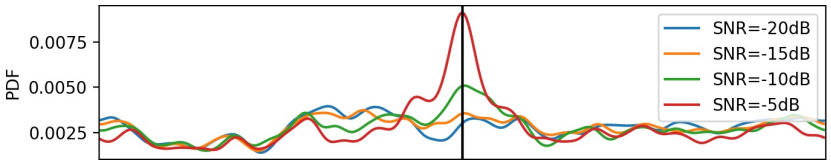
relative to the correct peak and even surpassed its amplitude. Further investigation is needed to explain this phenomenon as well as why the slower rotational speed produces overall worse results despite Figure 4.9 showing very similar performance for both $f_{rot} = 20\,\text{RPS}$ and $f_{rot} = 40\,\text{RPS}$.



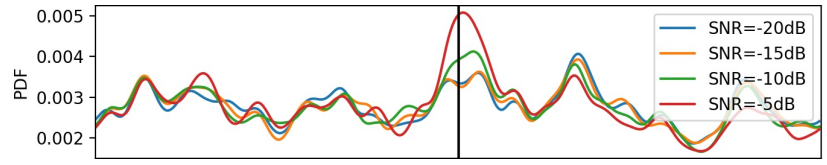
(a) 125 Hz, 250 Hz, ..., 4 kHz, 8 kHz sine tones combined



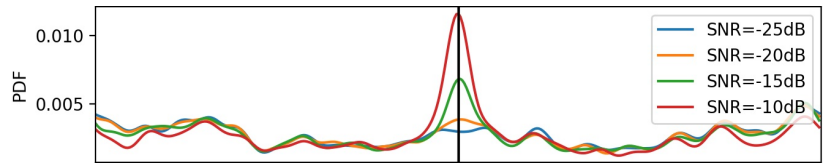
(b) Male voice sample



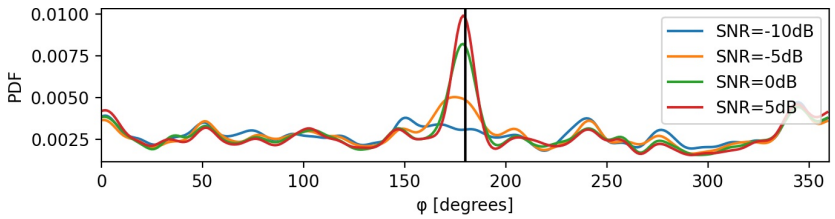
(c) Female voice sample



(d) Simple drum groove



(e) Piano concerto excerpt



(f) Exponential sine sweep

Figure 4.15: DOA estimation accuracy for various sound sources and SNRs at $f_{rot} = 40\,\text{RPS}$.

Since $f_{rot} = 40\,\text{RPS}$ and $L = 8192$ consistently produced the best results we will now be using these parameters to investigate the DOA estimation accuracy of the previously used signals at

various SNR. The results are depicted in Figure 4.15. Note the different SNR values.

It can be observed that all signals can be localized with reasonable accuracy once SNR $> -5\,\mathrm{dB}$. Additionally, the signals whose DOA estimation was most accurate in Figure 4.13 can be located at lower SNRs and with higher accuracy. An interesting finding is the decrease in the peak at $\varphi = 0°$ in the case of the combined sine tones and exponential sine sweep signals. We believe this is due to the low energy sidebands being overpowered by the pink noise, leading to fewer $\varphi = 0°$ detections in subbands that only contain the sidebands of a frequency.

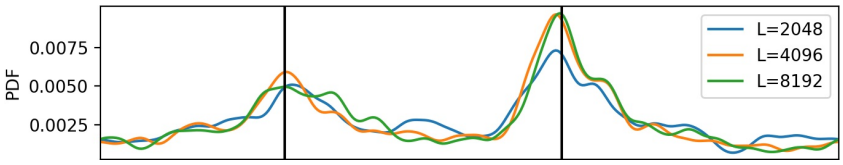
4.4 Localization of Multiple Sources in 2D Space

Let us now investigate DOA estimation accuracy when multiple sources are active at various spatial points simultaneously. We start with two active sources and investigate two signal combinations: the male speech + the female speech and the drum groove + the piano concerto excerpt. We fix the DOA of the first signal to $\varphi_1 = 90°$ and test three different DOAs for the second signal: $\varphi_2 = 225°/150°/113°$. Each signal and DOA combination is modulated at $f_{rot} = 40\,\mathrm{RPS}$, $r = 5\,\mathrm{cm}$ and the three previously used DFT lengths are utilized during localization. Additionally, all signals are normalized such that they have approximately the same total energy. The results are depicted in Figure 4.16 and Figure 4.17, where the true DOAs are indicated with black lines.

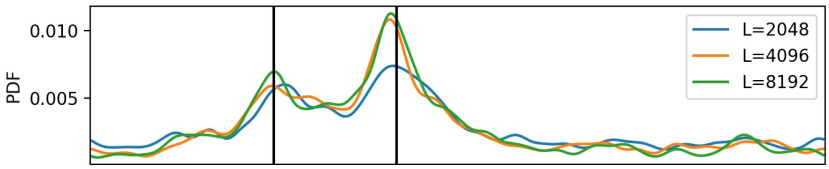
It can be observed that separate localization of two signals is possible, especially in the case of the music signals. The speech signals are localized less precisely, but their DOAs can still be obtained with reasonable precision as long as they are sufficiently spaced. When placed 23° apart the two peaks merge into one, making it challenging to determine the individual DOAs. An interesting observation can be made in Figure 4.16a as it appears that localization for $L = 4096$ outperforms the larger DFT length. This may, however, simply be a coincidence. Another notable observation is the large difference in the peak heights in the case of the music signals. We believe this is due to the energy of the piano concerto excerpt being more concentrated onto constant frequencies over a larger range of subbands and the energy of the drum groove being more concentrated onto the transients.

The depicted simulations were repeated using $f_{rot} = 20\,\mathrm{RPS}$. The results showed that the localization accuracy of the music signals was very comparable to the $f_{rot} = 40\,\mathrm{RPS}$ case. Localization of the speech signals, however, failed entirely, since neither of the signals could be detected.

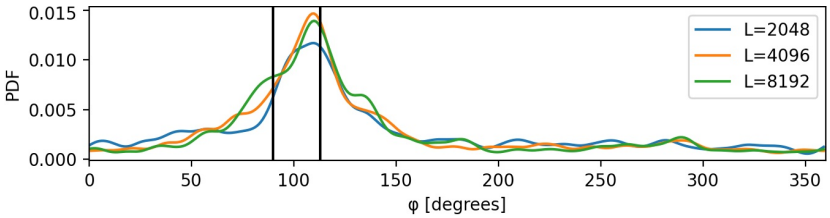
We will now explore the sensitivity of the localization accuracy to additive pink noise at various SNRs in the case of $L = 8192$, $f_{rot} = 40\,\mathrm{RPS}$, $\varphi_1 = 90°$ and $\varphi_2 = 150°$. The results of the simulations are shown in Figure 4.18, where the SNR was computed relative to the combined



(a) Male voice ($\varphi_1 = 90°$) + female voice ($\varphi_2 = 225°$)

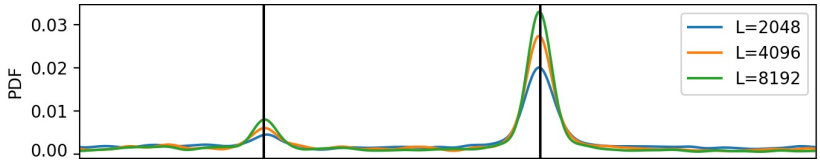


(b) Male voice ($\varphi_1 = 90°$) + female voice ($\varphi_2 = 150°$)

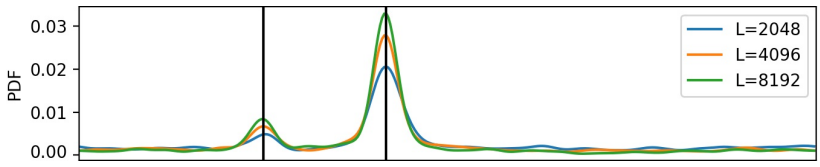


(c) Male voice ($\varphi_1 = 90°$) + female voice ($\varphi_2 = 113°$)

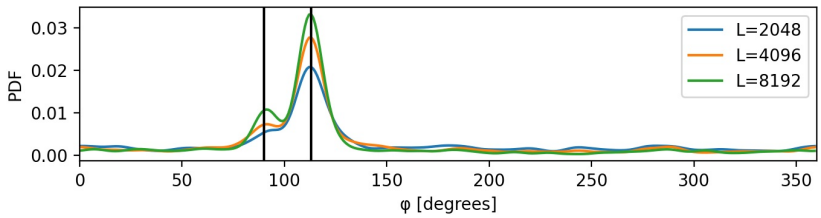
Figure 4.16: DOA estimation accuracy for two speech sources at $f_{rot} = 40\,\text{RPS}$.



(a) Drum groove ($\varphi_1 = 90°$) + piano concerto excerpt ($\varphi_2 = 225°$)



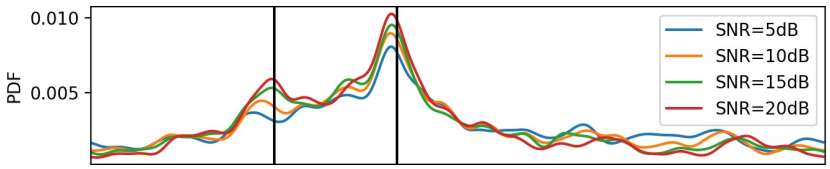
(b) Drum groove ($\varphi_1 = 90°$) + piano concerto excerpt ($\varphi_2 = 150°$)



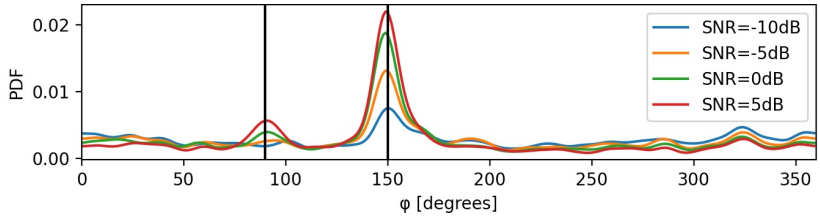
(c) Drum groove ($\varphi_1 = 90°$) + piano concerto excerpt ($\varphi_2 = 113°$)

Figure 4.17: DOA estimation accuracy for two music sources at $f_{rot} = 40\,\text{RPS}$.

source signals. The plots show that detection of the male voice only becomes possible at an SNR $> 10\,\text{dB}$. On the other hand, the drum groove can already be located at an SNR $> 0\,\text{dB}$. We therefore conclude that DOA estimation of signals which can be detected reliably by themselves are less sensitive to noise.



(a) Male voice ($\varphi_1 = 90°$) + female voice ($\varphi_2 = 150°$)



(b) Drum groove ($\varphi_1 = 90°$) + piano concerto excerpt ($\varphi_2 = 150°$)

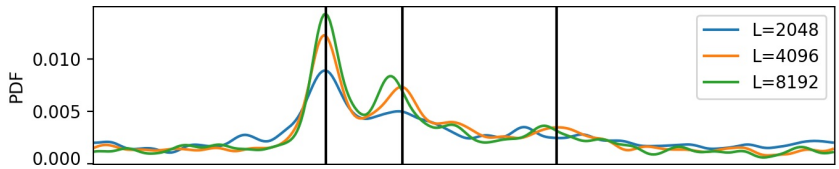
Figure 4.18: DOA estimation accuracy for two sources at various SNR and $f_{rot} = 40\,\text{RPS}$.

Finally, let us explore localization accuracy for three and four simultaneous audio sources. We choose the same four signals as before and place the drum groove at $\varphi_1 = 90°$, the piano concerto excerpt at $\varphi_2 = 113°$, the female voice at $\varphi_3 = 150°$ and the male voice at $\varphi_4 = 225°$. One of these sources is deactivated at a time resulting in simultaneous playback of three signals. The results of the DOA estimation using these signal combinations are depicted in Figure 4.19. Subsequently we tested simultaneous playback of all four audio sources. These results are shown in Figure 4.20a.

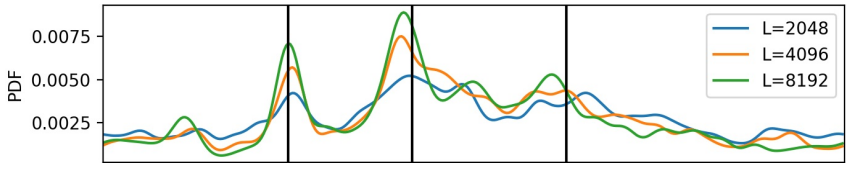
It can be observed that the music signals are localized accurately for all signal combinations. The female voice is detected with reasonable accuracy in all cases, whereas localization of the male voice is less reliable. In Figure 4.19a localization fails entirely, except for $L = 4096$, where a subtle peak is visible at the correct DOA of the male speech sample.

In the case of all four sources being active at the same time the localization is reasonably accurate, however, it fails to detect the male speech signal. Interestingly, $L = 4096$ once again slightly outperforms $L = 8192$. Since we believed the male voice signal was overpowered by the piano concerto excerpt we doubled the amplitude of the male speech sample and halved the amplitude of the piano concerto signal. The results of the DOA estimation using these modifications are displayed in Figure 4.20b. It can be observed that all four signals are now detected with reasonable accuracy.

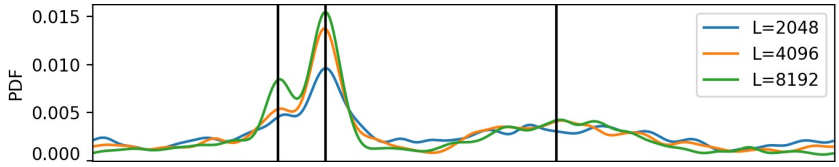
We conclude that localization of multiple audio sources is possible, especially if the source signals



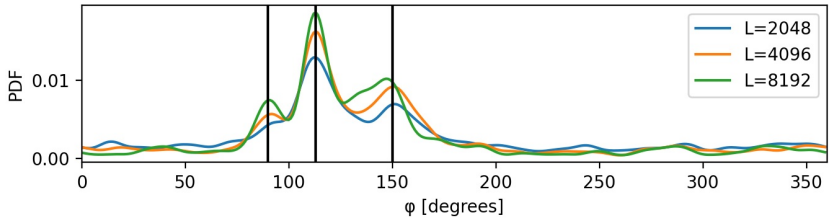
(a) Piano concerto excerpt + female voice + male voice



(b) Drum groove + female voice + male voice

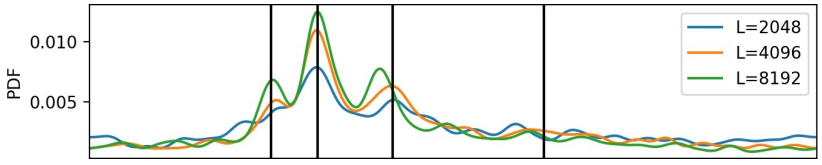


(c) Drum groove + piano concerto excerpt + male voice

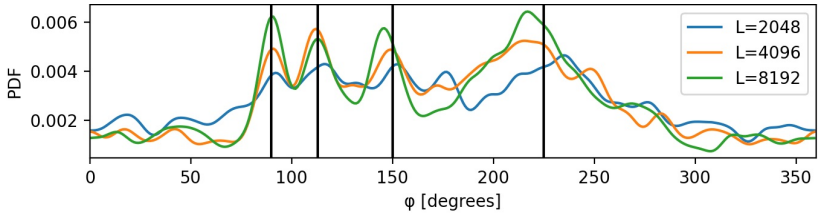


(d) Drum groove + piano concerto excerpt + female voice

Figure 4.19: DOA estimation accuracy for three sound sources and $f_{rot} = 40\,\mathrm{RPS}$.



(a) All four signals combined



(b) All four signals combined, modified amplitudes

Figure 4.20: DOA estimation accuracy for four sound sources and $f_{rot} = 40\,\mathrm{RPS}$.

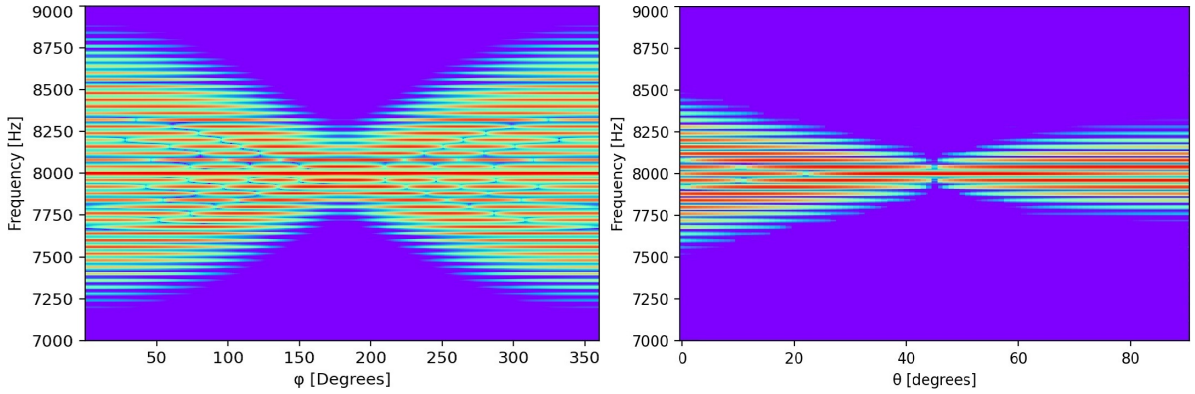
consist of many constant, higher frequencies. Generally speaking, signals which can be detected reliably by themselves are more likely to be detected accurately in the presence of other sources.

Signals consisting mainly of low frequencies which do not remain constant for very long, such as the male speech signal, are prone to being overpowered by other sources and detection may only be possible if the amplitudes of all signals are adjusted.

4.5 Localization in 3D Space

To conclude the theoretical investigation of our DOA estimation algorithm we will now explore 3D sound source localization. The main idea is to follow the same steps as before, that is computing the TWA for all angles $\varphi = [0°, 1°, ..., 358°, 359°]$, $\theta = 90°$ and finding peak angles $\hat{\varphi}$ in the PDF of the focusedness-based predictions for each subband and spectrogram frame. The 3D localization is then performed by taking each peak angle $\hat{\varphi}$ we found and applying the TWA for $\varphi = \hat{\varphi}$ and $\theta = [0°, 1°, ..., 89°, 90°]$. Subsequently we search for peak elevation angles $\hat{\theta}$ in a similar fashion to the search for $\hat{\varphi}$. Note that we only search the interval $[0°, 90°]$ and not $[0°, 180°]$ since the modulation index $\beta(\theta)$ is symmetric around 90°. A reminder as to why this is the case can be found under Equation (3.16). Since we focus more on 2D localization in this thesis our investigation of 3D localization will not be as vigorous as the 2D case and therefore we use only $f_{rot} = 40\,\text{RPS}$, $L = 8192$ and $r = 5\,\text{cm}$ for all simulations.

An example of the 3D localization approach is shown in Figure 4.21 for an 8 kHz source signal with a DOA of $\varphi = 180°$ and $\theta = 45°$. First, the azimuth-spectrogram is computed while fixing $\theta = 90°$, leading to a peak being detected at $\varphi = 180°$. Subsequently the elevation-spectrogram is computed while fixing $\varphi = 180°$, leading to a peak being detected at $\theta = 45°$.



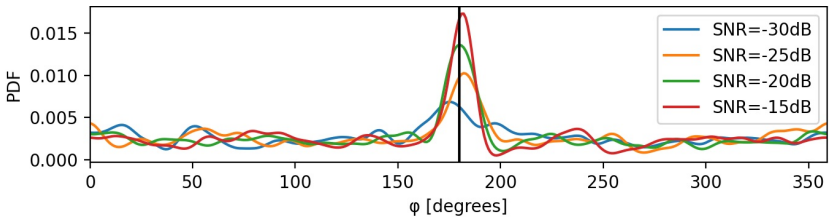
(a) Azimuth-spectrogram for $\theta = 90°$ (b) Elevation-spectrogram for $\varphi = 180°$

Figure 4.21: 3D localization for an 8 kHz source signal with DOA $\varphi = 180°$ and $\theta = 45°$.

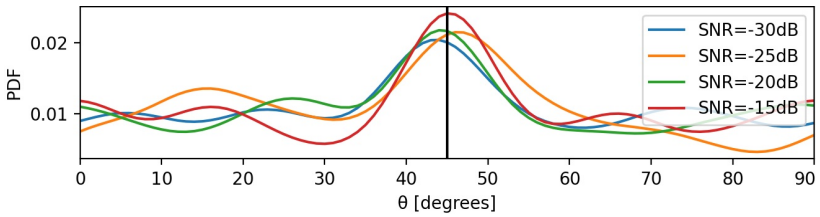
Side note: An alternative, perhaps more accurate 3D localization method would be to perform the TWA for all φ and θ simultaneously and subsequently compute the 2D focusedness over all azimuth and elevation angles for each subband and spectrogram frame. The DOA estimates would then be 2D points, allowing us to obtain a 2D PDF. The main problem with this approach

is that it is very expensive from a computational standpoint. However, a more efficient future implementation of the TWA could enable its use.

Using the signal from Figure 4.21 and performing DOA estimation with subband processing at various SNR leads to the results shown in Figure 4.22, where the true azimuth and elevation angles have been indicated with a black line. A clear peak is visible at or close to both the correct azimuth and elevation angles.



(a) PDF of azimuth estimation for $\theta = 90°$



(b) PDF of elevation estimation for $\varphi = 180°$

Figure 4.22: PDFs of the azimuth- and elevation-spectrograms from Figure 4.21 for various SNR.

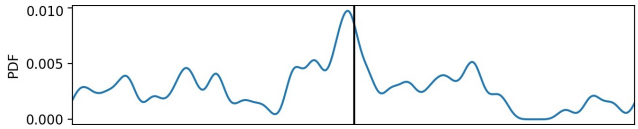
We will now explore azimuth as well as elevation detection for the previously used drum groove, female speech, piano concerto excerpt and male speech signals. The incoming azimuth angle is $\varphi = 180°$ in all cases and the incoming elevation is $\theta_{1/2/3/4} = 23°/45°/67°/90°$, respectively. First, detection of the azimuth angle is carried out and subsequently elevation detection is performed using the previously derived azimuth estimate. The results are displayed in Figure 4.23, where the angles in brackets indicate the fixed azimuth and elevation angles used for the TWA. It can be observed that azimuth estimation deteriorates for low values of θ, since the modulation index is reduced. Regardless, all the azimuth angles are detected reliably. The elevation estimation PDFs show two peaks at $\theta = 0°$ and $\theta = 90°$. Further investigation is needed to explain these phenomena. Ignoring the peaks at the sides in all cases except for the male speech signal allows for determination of the elevation angles with reasonable precision.

Subsequently, we performed 3D DOA estimation for two simultaneous sources. We placed the simple drum groove and the piano concerto excerpt at the same azimuth $\varphi = 180°$ and at different elevations $\theta_{1/2} = 23°/67°$. The results are depicted in Figure 4.24. It can be observed that two sources with the same azimuth and different elevations can be localized individually.

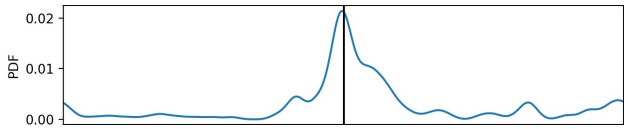
Finally, we performed 3D DOA estimation for two sources at completely different positions.

We placed the female speech signal at $\varphi_1 = 45°$ and $\theta_1 = 45°$ and the male speech signal at $\varphi_2 = 180°$ and $\theta_2 = 67°$. The results of this simulation can be found in Figure 4.25. Localization of the female speech signal is very precise, however, the elevation of the male speech signal is barely detectable and has a large error. We believe this is due to the male speech signal being difficult to localize overall, as our previous tests have shown.

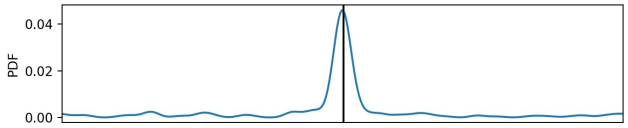
The simulations show that 3D localization is possible in theory, however, it is less precise than azimuth estimation. Additionally, azimuth estimation becomes less accurate as the elevation angle approaches 0°, since the modulation index is reduced. Simulations with additive noise furthermore revealed that elevation estimation is very sensitive to noise and therefore difficult to reliably implement in practice. Increasing the rotational diameter of the microphone may help decrease the inaccuracies and noise sensitivities of 3D localization, assuming the resulting increase in wind noise does not outweigh the benefits of the larger modulation index. Additional improvements in the combination method of the individual subband and spectrogram frame DOA estimations may also contribute to improvements of localization in 3D space.



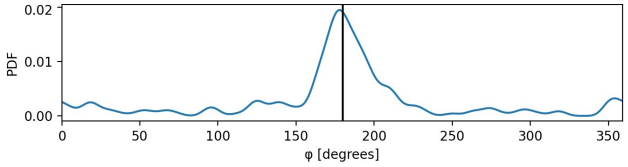
(a) Drum groove, detection at $\hat{\varphi} = 176°$ ($\theta = 90°$)



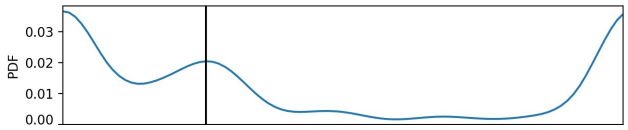
(b) Female speech, detection at $\hat{\varphi} = 179°$ ($\theta = 90°$)



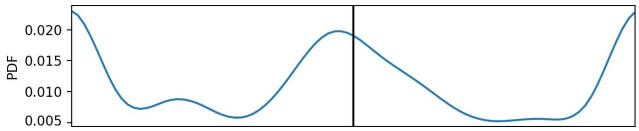
(c) Concerto excerpt, detection at $\hat{\varphi} = 179°$ ($\theta = 90°$)



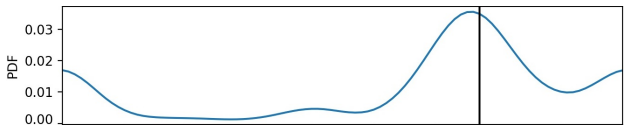
(d) Male speech, detection at $\hat{\varphi} = 178°$ ($\theta = 90°$)



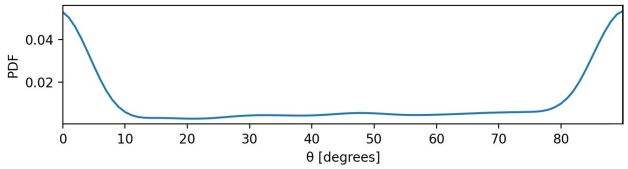
(e) Drum groove, detection at $\hat{\theta}_1 = 23°$ ($\varphi = 176°$)



(f) Female speech, detection at $\hat{\theta}_2 = 43°$ ($\varphi = 179°$)

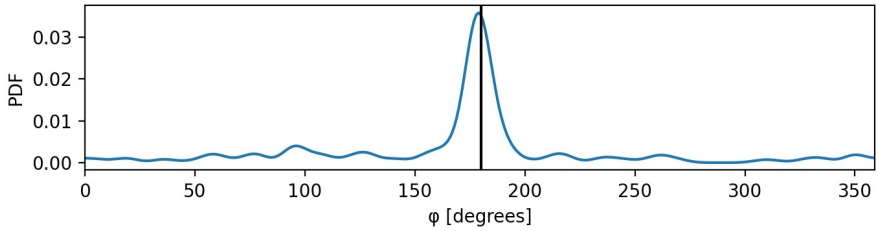


(g) Concerto excerpt, detection at $\hat{\theta}_3 = 66°$ ($\varphi = 179°$)

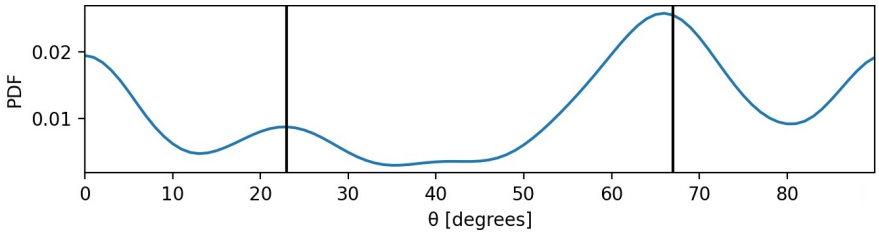


(h) Male speech, detection at $\hat{\theta}_4 = 90°$ ($\varphi = 178°$)

Figure 4.23: 3D localization for various source signals with varying elevation.

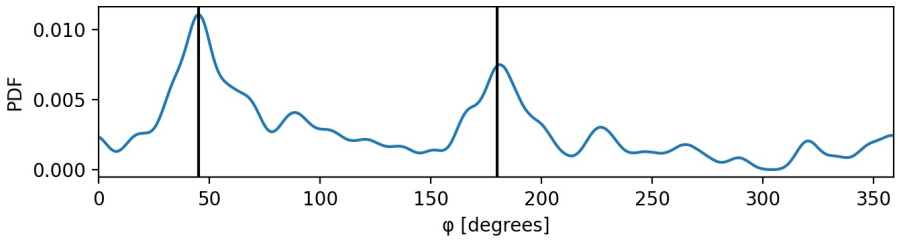


(a) Drum groove + concerto excerpt, detection at $\hat{\varphi} = 179°$ ($\theta = 90°$)

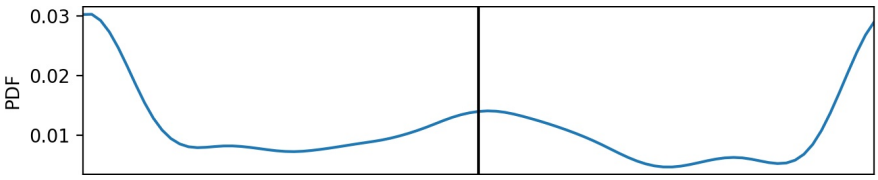


(b) Drum groove + concerto excerpt, detections at $\hat{\theta}_{1/2} = 23°/66°$ ($\varphi = 179°$)

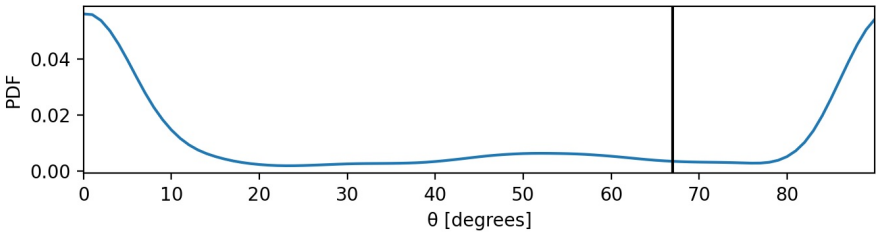
Figure 4.24: 3D localization of two sources at the same azimuth and different elevations.



(a) Female speech + male speech, detections at $\hat{\varphi}_{1/2} = 45°/181°$ ($\theta = 90°$)



(b) Female speech + male speech, detection at $\hat{\theta}_1 = 46°$ ($\varphi = 45°$)



(c) Female speech + male speech, detection at $\hat{\varphi}_2 = 52°$ ($\varphi = 181°$)

Figure 4.25: 3D localization of two sources at different azimuth and elevations.

Chapter 5

Direction of Arrival Estimation - Practical Verification

We will now attempt to verify the proposed DOA estimation approach in practice using our REM prototype. Section 5.1 provides details about the REM and subsequently Section 5.2 addresses the additional difficulties we are expected to face when the idealizing assumptions we made during the derivation of our approach are not met. Section 5.3 elaborates on the utilized practical setup and finally Section 5.4 and Section 5.5 show the results of our experiments.

5.1 REM Prototype

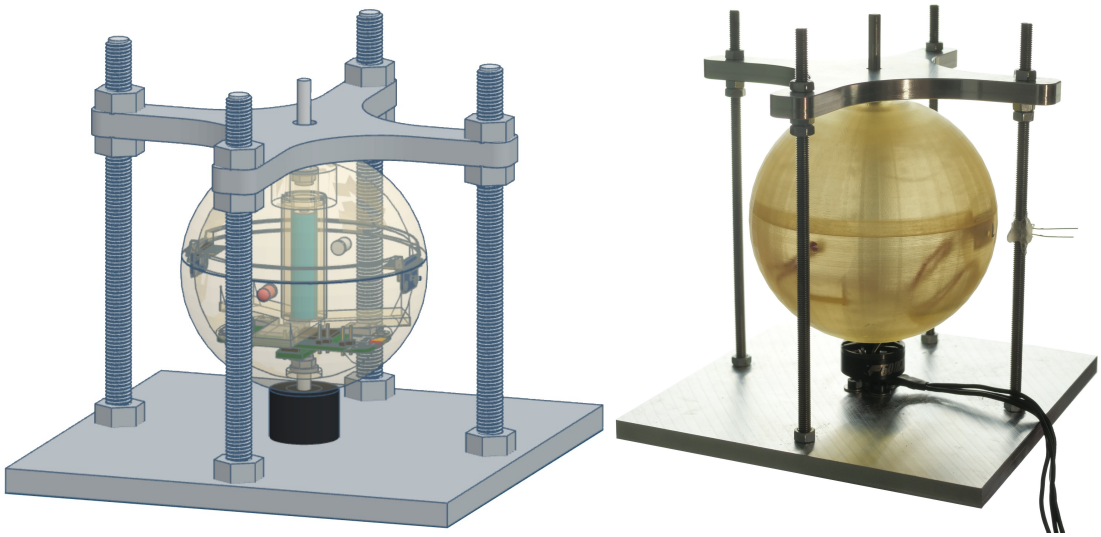


Figure 5.1: 3D model (left) and photograph (right) of the REM prototype.

The REM prototype is depicted in Figure 5.1 and a schematic showing the connections and components of the REM can be found in Figure 5.2.

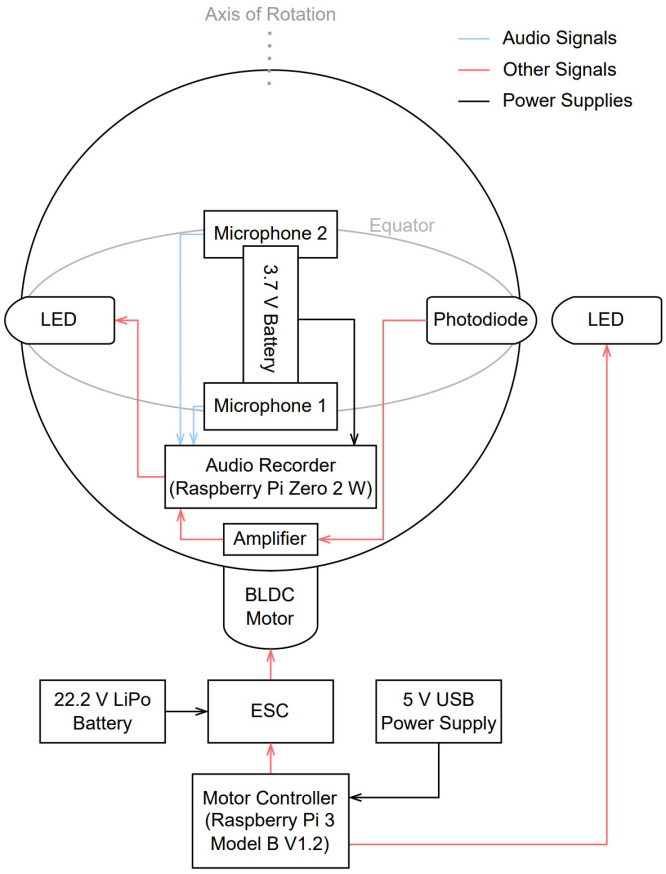


Figure 5.2: Electronic components and signal flow diagram of the REM prototype.

The equator of the REM is equipped with an LED, a photodiode and two omnidirectional SPH0645LM4H MEMS microphones which each record at a sampling rate of $f_s = 48\,\text{kHz}$ and have the frequency response shown in Figure 5.3 (3.072 MHz graph). Although the REM features two microphones we will only be utilizing one audio signal in this thesis since we wish to perform

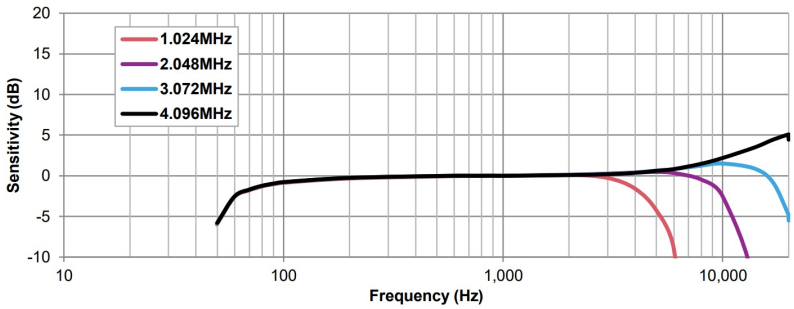


Figure 5.3: Frequency response of the SPH0645LM4H MEMS microphone (image from [19]).

DOA estimation using a single microphone.

The internal photodiode passes an external LED during rotation and the recorded light peaks allow the internal Raspberry Pi to determine the rotational speed as well as the starting phase with high precision. The rotation is driven by a brushless DC (BLDC) motor which is controlled by an electronic speed controller (ESC). The ESC does not allow for direct adjustment of the rotational speed but instead is set to a certain power level which causes the microphone to rotate at a certain speed depending on air resistance, friction and battery voltage. This causes minor fluctuations in rotational speed over time which are recorded and taken into account by the TWA. The minimum stable operating speed of the REM is approximately 20 RPS. Although the maximum speed we have achieved is roughly 200 RPS, we will limit the maximum speed to 40 RPS for the sake of audio quality. More details regarding the REM can be found in [22].

5.2 Real World Considerations

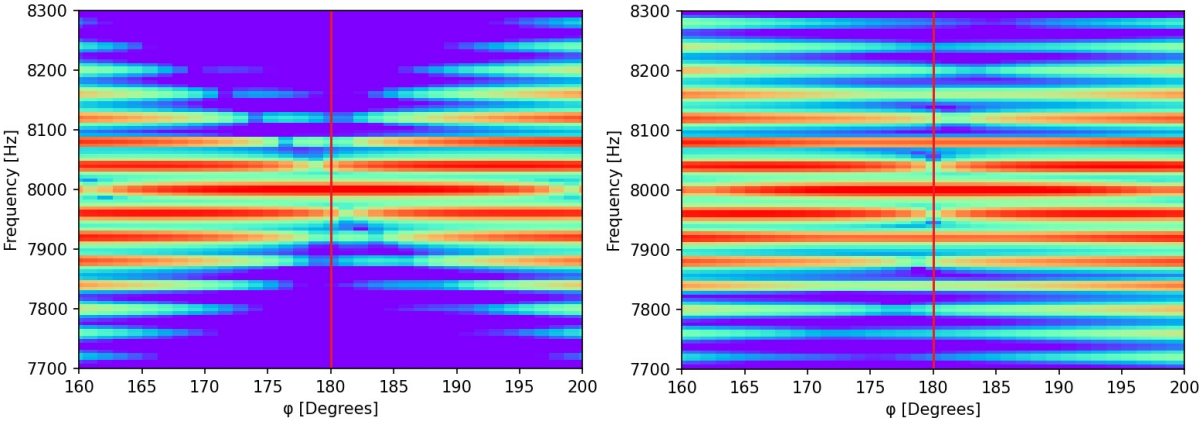
In the previous chapter we made many simplifying assumptions to facilitate the derivation of our DOA estimation algorithm. We will now give a brief overview of these assumptions and point out in which way they may not hold in the real world:

1. The amplitude A in Equation (3.12) was assumed to be constant since plane waves have the same amplitude regardless of where in space we place the microphone. In reality, however, the amplitude decreases with increasing distance to the sound source, which in our case would result in a periodic modulation of A.

2. We assumed all sound waves to be plane waves. In practical settings, however, sound waves will always have a slight curvature regardless of how distant the sound source is. As a consequence the TWA will not perfectly compensate the frequency modulation for a given direction unless the distance to the sound source is also known and taken into account.

3. The microphone was assumed to be perfectly omnidirectional, however, real world microphones are never perfectly omnidirectional. In addition, the sphere of our REM prototype acts as an acoustic barrier as the microphone is facing away from a sound source. Both of these effects result in an additional frequency-dependent modulation of the amplitude A.

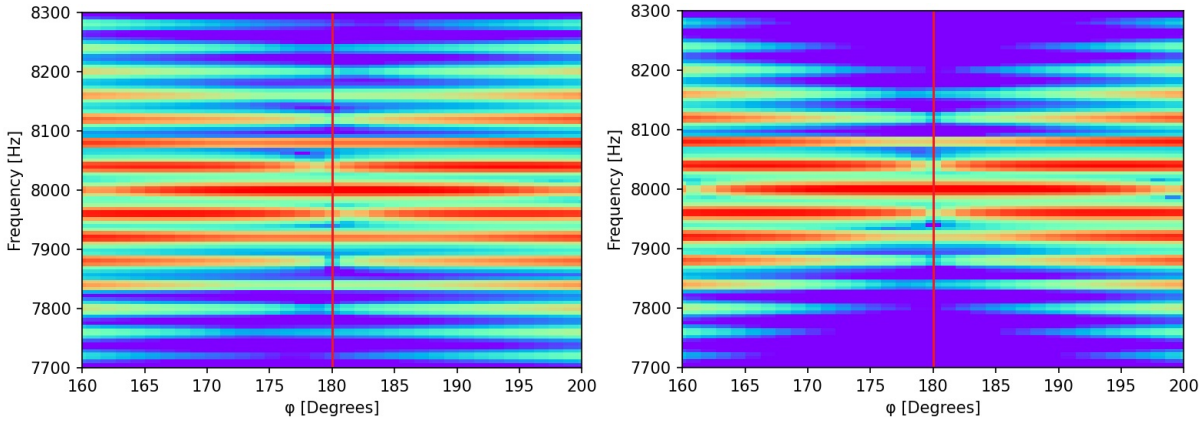
4. The assumption was made that the microphone is in a free field. In realistic environments, however, we will always have frequency-dependent acoustic reflections and acoustic scattering. As a consequence, any sound source will never arrive from only one direction but multiple frequencies within the sound source will arrive from multiple, different directions with different levels of distortion.



(a) Effects of amplitude modulation

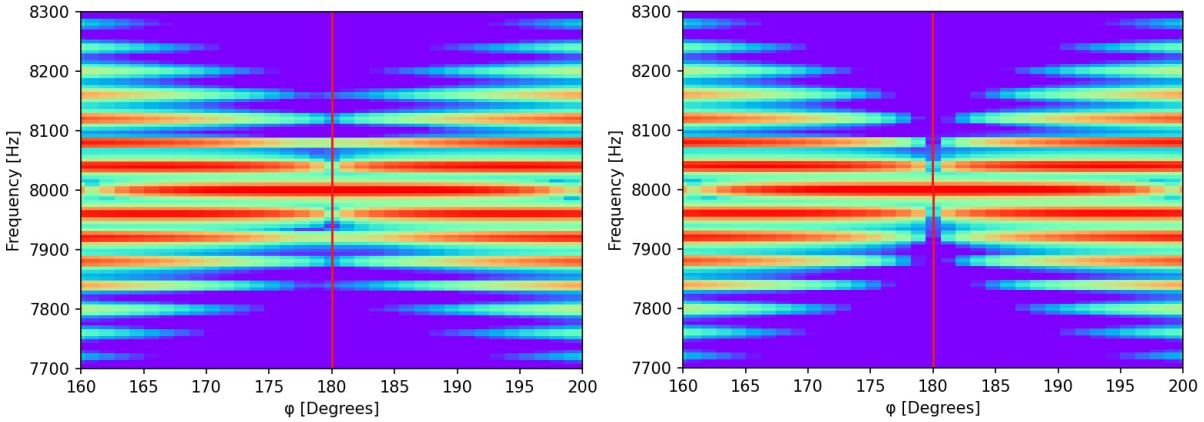
(b) Effects of wavefront curvature

Figure 5.4: Azimuth spectrograms for an 8 kHz sound source placed 20 cm from the microphone.



(a) 0.2 m distance

(b) 1 m distance



(c) 2 m distance

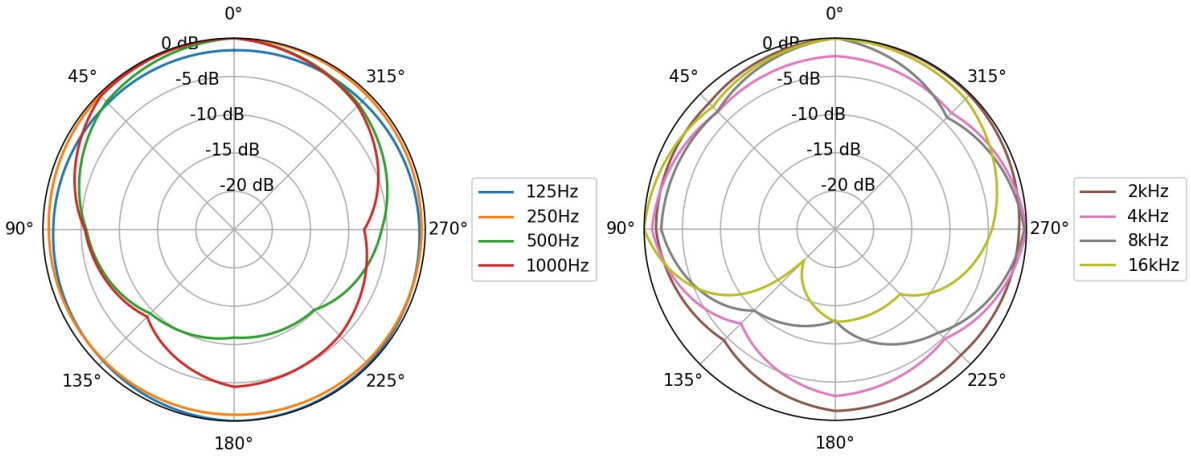
(d) 1000 m distance

Figure 5.5: Azimuth spectrograms for an 8 kHz sound source placed at multiple distances.

We will now investigate the influence of violations of these assumptions when performing DOA estimation in practice. Figure 5.4a and Figure 5.4b show the impacts of 1. and 2. on the azimuth spectrogram of an 8 kHz sound source placed 20 cm from the microphone with a 10 cm rotational diameter. The true DOA is indicated with a red vertical line.

It can be observed that amplitude modulation results in a minor deviation of the sidebands above the main frequency towards the left and below it towards the right. The wavefront curvature has a much stronger impact on the unmodulation accuracy, making it significantly more challenging to determine the exact DOA when observing the azimuth-spectrogram.

Figure 5.5 depicts the azimuth-spectrograms for multiple microphone-source distances when both 1. and 2. are taken into account. It can be observed that at a distance of 1 m the distortions have already drastically reduced and the DOA is clearly visible for the 1 m, 2 m and 1000 m case. Therefore these effects can already be neglected at relatively low microphone-source distances



(a) First microphone - low frequencies

(b) First microphone - high frequencies

(c) Second microphone - low frequencies

(d) Second microphone - high frequencies

Figure 5.6: Directivity of both microphones for various frequencies.

when performing DOA estimation. However, a more accurate implementation which takes both amplitude modulation and curved wavefronts into account is required if we wish to accurately reconstruct sound sources in addition to their localization.

Let us now address concern 3.: We measured the directional characteristics of our REM prototype for frequencies 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz and 8 kHz in an anechoic chamber. All these frequencies were played by a speaker placed at approximately 1 m distance from the microphone and the microphone was rotated by 45° after each measurement has been taken. The results of these measurements can be found in Figure 5.6.

As it can be observed, both microphones within the REM exhibit very similar directional characteristics. Additionally, higher frequencies are attenuated more as the microphone is facing away from a sound source. An unanticipated discovery is the strong directivity concerning the 1000 Hz and particularly the 500 Hz frequency as it remains unclear what induces the pronounced attenuation towards the rear of the microphone. One hypothesis is that this may have been caused by standing waves within the anechoic chamber, which may also explain the measured asymmetric directivity. Further investigation using an omnidirectional reference microphone is needed to definitively verify or disprove this hypothesis.

Nevertheless, we can conclude that the directivity of the microphone causes additional frequency-dependent amplitude modulation when recording an audio source during rotation, which will increase the effects shown in Figure 5.4a. To counteract these artifacts it would be necessary to take the precise directional characteristics of the microphone into account, however, the TWA currently lacks the ability to perform amplitude modulation compensation. Hence, we set aside the incorporation of the REM directivity into the TWA for future research.

Finally, let us briefly address concern 4.: Acoustic reflections and scattering are perhaps the most difficult distortions to compensate. Perfect compensation would require an intractable amount of knowledge regarding the precise acoustic properties of the space surrounding the microphone. Therefore, to reduce the impacts of acoustic scattering and reflections, we chose to perform all of our practical measurements in an anechoic chamber. This will give us a baseline performance which we can compare against when performing DOA estimation in more realistic, reverberant environments in the future.

5.3 Experiment Setup

Our practical measurements were carried out in an anechoic chamber which the FAU LMS chair kindly provided us access to. It features a reverb time of < 30 ms and a reflection coefficient < 0.1 for frequencies above 200 Hz. Additionally, the approximate dimensions of the room were measured as W×L×H = 2.75 m×2.5 m×2.4 m and the temperature inside of the room as

19° C. The REM prototype was placed at the center of the room at a height of 1.2 m and two Genelec 1029A loudspeakers were positioned at the same height at various spatial points to simulate one or two audio sources. The initial setup is depicted in Figure 5.7a. Additionally, we tested localization accuracy of an audio signal placed at 45° elevation. This setup is shown in Figure 5.7b.

(a) Setup in the horizontal plane (b) Setup with 45° speaker elevation

Figure 5.7: Microphone and speaker placement in the utilized anechoic chamber.

The loudspeakers were connected to a computer via a Steinberg UR44C audio interface which was placed in the neighbouring room, from which the microphone and the microphone motor were also controlled wirelessly. Each test consisted of placing the speakers at certain predefined points, spinning the microphone to the desired speed and subsequently playing a test file. For each speaker arrangement three microphone speeds were tested: 24 RPS, 32 RPS and 42 RPS. The self noise at these speeds was measured at a distance of 20 cm as 45 dB, 51 dB and 56 dB, respectively.

The loudspeakers were set to a volume such that the sound pressure level at the microphone center was roughly 95 dB for a distance of 139 cm between the microphone center and the speaker membrane. The utilized test file consisted of a concatenation of the audio samples listed in Table 5.1. Each sample was approximately 2 s long and normalized to -0.1 dB. The samples '125 Hz' - '8 kHz' are sinusoidal tones of the given frequency, 'Combined' is a combination of all the previous tones, 'Pink noise' is randomly generated pink noise, 'Male speech' and 'Female speech' are speech samples of a male and female speaker reading the phrase, "The audiovisual sector is very important", 'Drum groove' is a simple drum groove, 'Classical music' is a short excerpt from the 1st movement of Rachmaninoffs 2nd piano concerto where both the piano and the orchestra are active and 'Sine sweep' is an exponential sine sweep in the range 1 kHz - 5 kHz. In the case of simultaneous playback of the male and female speech samples the male speech sample was replaced by a different sample where a speaker reads the phrase, "Did we uhm... did

One source	Two sources Channel 1	Two sources Channel 2
125 Hz	125 Hz	1 kHz
250 Hz	1 kHz	250 Hz
500 Hz	500 Hz	1 kHz
1 kHz	1 kHz	1 kHz
2 kHz	2 kHz	1 kHz
4 kHz	1 kHz	4 kHz
8 kHz	8 kHz	1 kHz
Combined	Pink noise	Male speech
Pink noise	Female speech	Male speech
Male speech	Drum groove	Classical music
Female speech	Classical music	Female speech
Drum groove		
Classical music		
Sine sweep		

Table 5.1: Used audio samples for single source and two source localization.

we hire interns?".

The utilized speaker placements are listed in Table 5.2. The given angles refer to the azimuth relative to the position of the blue speaker in Figure 5.7a plus 90°. Furthermore, positive angles refer to anti-clockwise movement with respect to the 90° reference point, i.e. the black speaker in Figure 5.7a is placed at a 180° azimuth angle. We attempted to maximize the distance between the speakers and the microphone which lead to minor variations in the microphone-speaker distance in certain configurations due to the small room space. Note that the microphone-speaker distance was measured with respect to the microphone center and the speaker membrane.

	Speaker 1	Speaker 2	Microphone-speaker distance
Single source	90°	-	139 cm
	180°	-	139 cm
	90° + 45° elevation	-	139 cm
Two sources	90°	180°	139 cm
	90°	270°	130 cm
	90°	112.5°	111 cm
	90° + 45° elevation	180°	139 cm

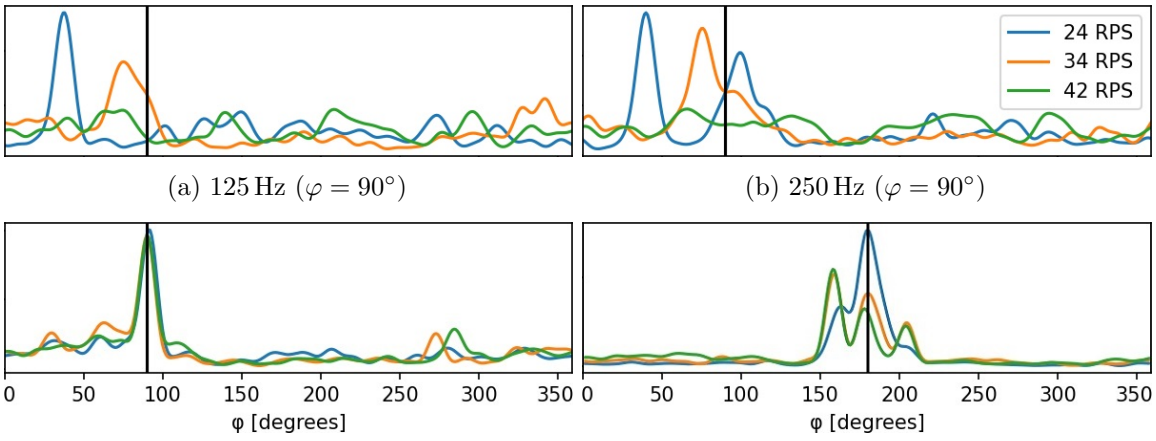
Table 5.2: Utilized speaker placements for single source and two sources localization

The following two sections will show the accuracy for our single source and subsequently two sources localization, as well as compare these findings to the simulations performed in Chapter 4.

5.4 Localization of a Single Source

The self noise created by the REM was found to be a combination of wind noise, which has pink noise characteristics, and motor noise, which can be described as a low frequency hum at the rotational frequency and its overtones. The SNRs at the rotational speeds 24 RPS/34 RPS/42 RPS were estimated to be approximately SNR = −3.8 dB/ − 9.2 dB/ − 12.6 dB. Due to these low SNRs we expect that localization of complex signals will prove difficult.

Let us first focus on the localization in the horizontal plane for $\varphi = 90°$ and $\varphi = 180°$: The 125 Hz signal was not detectable under any circumstances, which we believe is due to it being overpowered by the REM self noise. Localization of the 250 Hz signal is not very accurate and detection was only possible when using slower rotational speeds. The localization accuracy gradually improves as the source frequency increases up to 2 kHz, where we see the best results. At 4 kHz, however, we start seeing the detection peaks either split in two parts or getting shifted by approximately 20°. This phenomenon becomes worse for the 8 kHz signal, which makes localization impossible. Similar splitting of the peaks can be seen in the combined signal. Figure 5.8 shows the localization performance of four signals at three rotational speeds each. The full DOA estimation results for every signal can be found in Appendix A. Note that we omitted the numerical values of the y-axis in all plots, since we only find the shape of the PDF to be of importance.



(a) 125 Hz ($\varphi = 90°$)

(b) 250 Hz ($\varphi = 90°$)

(c) 2 kHz ($\varphi = 90°$)

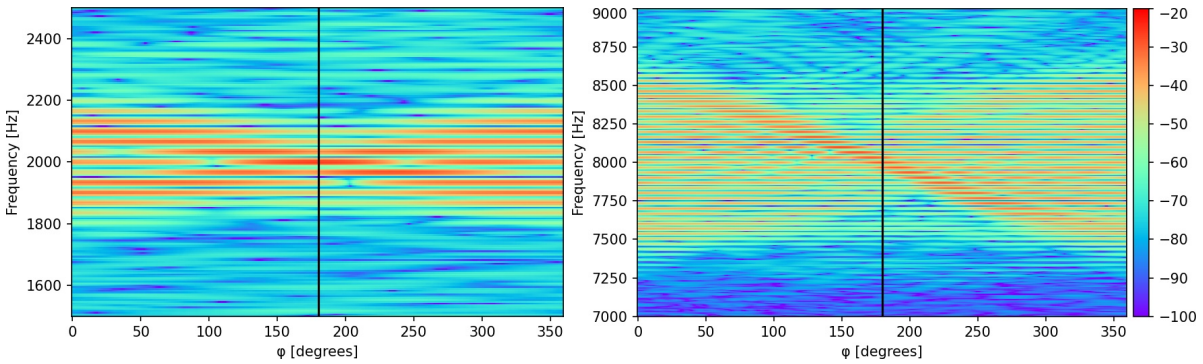
(d) Combined ($\varphi = 180°$)

Figure 5.8: Azimuth estimation for various sources.

Another interesting finding is the 'phantom peak' at the left of the 24 RPS plot in Figure 5.12a and Figure 5.12b. Numerous phantom peaks could be found in other plots as well and it is so far unclear whether these are due to acoustic reflections being localized, acoustic resonance of objects within the anechoic chamber or the REM prototype itself, standing waves, coincidence or other phenomena. Numerous overtones could be observed when investigating the spectra of the

recorded frequencies, suggesting that some sort of resonance took place. More research is needed to find and prevent the cause of these phantom peaks.

A possible explanation for the poor performance of high frequencies can be found when observing the azimuth-spectrograms for the well localized 2 kHz signal and the poorly localized 8 kHz signal. Figure 5.9 depicts these azimuth-spectrograms for $\varphi = 180°$ and 34 RPS, where the true DOA has been indicated with a black line. The characteristic shape from Figure 4.2 is clearly visible in these spectrograms, however, the sidebands do not perfectly disappear when approaching the correct DOA. This was to be expected, since the TWA does not compensate for wavefront curvature, amplitude modulation and acoustic scattering as well as reflections. Regardless, in the 2 kHz case we clearly see that the main frequency approaches a maximum at the correct DOA. In the 8 kHz case we also observe that the main frequency has a maximum at $\varphi = 180°$, however, numerous sidebands also show clearly defined maxima in an asymmetric fashion, making it difficult to determine the DOA without knowledge of the source frequency. We believe that these artifacts and the asymmetry are caused by the larger directivity of the REM for higher frequencies. This increases the amplitude modulation when recording high frequencies, magnifying the effects shown in Figure 5.4a. Therefore we conclude that the directivity of the REM must be taken into account for the localization of high frequencies.



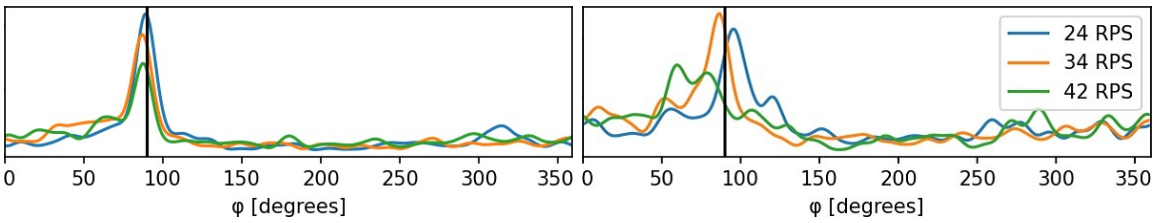
(a) 2 kHz ($\varphi = 180°$) (b) 8 kHz ($\varphi = 180°$)

Figure 5.9: Real azimuth-spectrogram for 2 kHz and 8 kHz source signals at 34 RPS.

Detection of the 8 kHz signal would be easy if we had knowledge of the source frequency, as we could simply search for the maximum within the frequency bin most closely associated with 8 kHz. This motivates the usage of a stationary reference microphone placed at or close to the center of the REM, since it would allow us to determine all incoming frequencies without any Doppler shifts. Therefore we would not have to rely on the focusedness for DOA estimation and could instead choose to, for example, correlate the azimuth spectrogram with one spectrogram frame of the reference microphone. We set aside this idea for future research.

Localization of the pink noise signal was not possible, which matched our expectations. However, the male voice, female voice and drum groove could also not be localized. Occasionally there

are peaks at approximately the right positions, but we believe these are coincidences. As far as we are concerned the inability to locate these complex signals is due to an insufficient SNR, especially since previous simulations showed that speech localization was only possible for higher rotational speeds where there is an even lower SNR. On the other hand, the classical music signal was localized very accurately, making it the only real world signal that could be detected. The exponential sine sweep was also localized with reasonable precision for lower rotational speeds. The results of the last two signals matched our expectations and are depicted in Figure 5.10 for $\varphi = 90°$.

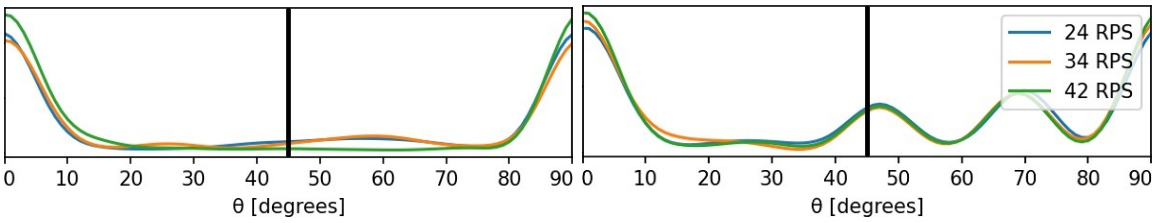


(a) Classical music ($\varphi = 90°$) (b) Sine sweep ($\varphi = 90°$)

Figure 5.10: Azimuth estimation for the classical music and sine sweep sources.

Let us now address the third single source localization with DOA $\varphi = 90°$ and $\theta = 45°$. The azimuth estimation was very comparable to the previous results with two small differences: Splitting of the peaks was less pronounced, therefore localization of the single frequencies was more accurate except in the case of the 2 kHz source signal, and localization of the exponential sine sweep was less accurate. It is unclear what causes the improvement in localization accuracy, since we expected azimuth localization to become worse due to the decreased overall modulation index as a result of the elevation. A possible explanation is that changing the speaker placement reduced some form of resonance in the anechoic chamber.

In a subsequent step we performed elevation estimation for all signals which had reliable and reasonably accurate azimuth estimation results. We chose the detected azimuth angles when applying the TWA. We found that elevation estimation did not produce any conclusive results regardless of the source signals. Figure 5.11 shows the results for two source signals, where the φ-values in brackets indicate the used azimuth angles when applying the TWA. Two clear



(a) Classical music ($\varphi = 89°/89°/88°$) (b) 4 kHz ($\varphi = 93°/92°/92°$)

Figure 5.11: Elevation estimation accuracy for the 4 kHz and classical music sources.

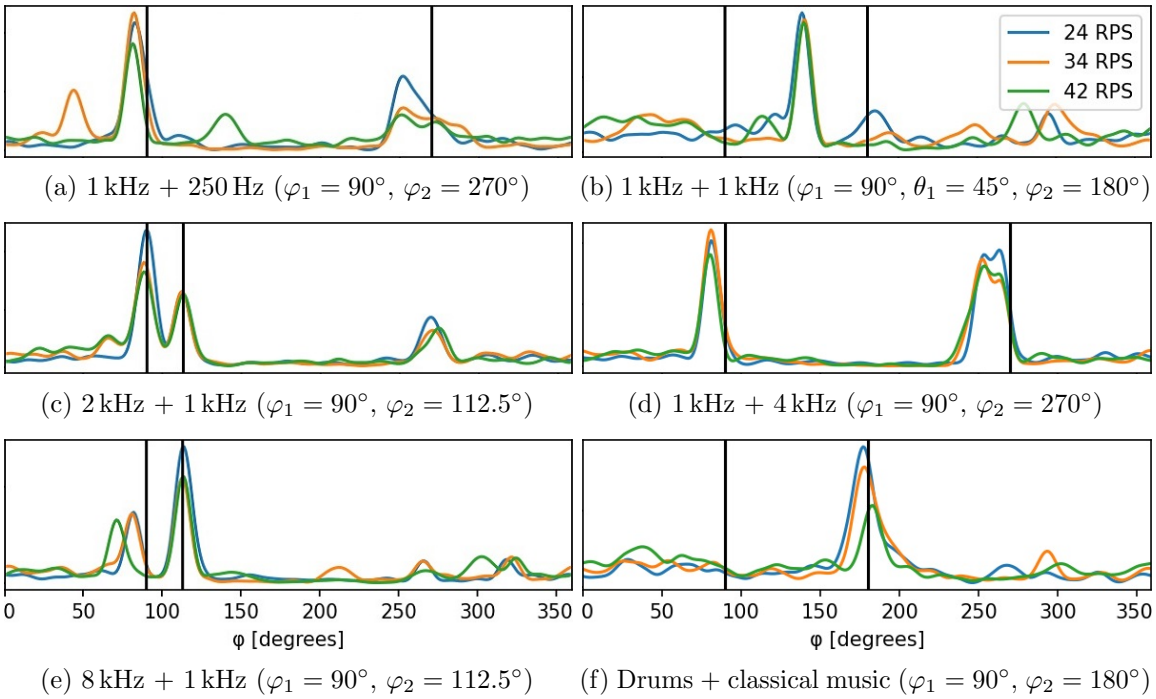
peaks are visible in the case of the 4 kHz signal, however, their origin could not be determined. Interestingly, the general shape of the elevation estimation curves matches the simulations in that there are two peaks at $\theta = 0°$ and $\theta = 90°$. More research is needed to explain these phenomena.

We conclude that single source localization in the horizontal plane is accurate most of the time for sound sources which mainly consist of constant tones in the mid frequency range. The 24 RPS and 34 RPS speeds clearly outperformed the 42 RPS case, however, it is difficult to ascertain which of the slower speeds performs better, as there are multiple instances in which one outperforms the other and vice versa. Performance may improve in general if the combination method of the individual subband predictions is adjusted such that bad DOA guesses are not included in the PDF, e.g. by only considering the subbands containing constant frequencies or having a certain SNR, and the TWA is extended to take into account amplitude modulation. This may also reduce the occurrence of phantom and split peaks as well as open the door for elevation estimation.

5.5 Localization of Two Sources

We will now examine the azimuth estimation results for two simultaneous audio sources. Note that we will omit the elevation estimation as it produced similarly inconclusive results as for the single source localization. The full results for all speaker placements, rotational speeds and source signal combinations are once again shown in Appendix A.

Generally speaking, signals that could be localized in the previous section could also be detected when a second source was present. Likewise, localization was not possible for signals that previously could not be detected. Additionally, as before, we see occasional phantom peaks, which may be caused by a multitude of different reasons. Interestingly, the presence of a second source appears to slightly reduce the occurrence of split peaks for 4 kHz and 8 kHz source signals. The proposed DOA estimation approach easily detects both sound sources when they are spaced at least 90° apart in the horizontal plane. In the case of them being spaced at 22.5°, however, the peaks merged in the 1 kHz + 250 Hz and 1 kHz + 4 kHz cases. Another interesting finding is that the peak appeared between the two true DOA angles when both speakers played the same 1 kHz tone. Finally, in some cases both peaks were shifted, but the relative angle between the peaks corresponded to the relative angle between the speakers. This suggests that there is a possibility that the zero point, which is detected by the REM, is not perfectly stable. However, these shifts are mostly observable at all three rotational speeds, indicating that these shifts have another origin. The most noteworthy results are depicted in Figure 5.12, where the left lines in the plots correspond to the first signal mentioned in their respective caption. Furthermore, $\theta = 90°$ applies to all signals if not stated otherwise.



(a) 1 kHz + 250 Hz ($\varphi_1 = 90°$, $\varphi_2 = 270°$)

(b) 1 kHz + 1 kHz ($\varphi_1 = 90°$, $\theta_1 = 45°$, $\varphi_2 = 180°$)

(c) 2 kHz + 1 kHz ($\varphi_1 = 90°$, $\varphi_2 = 112.5°$)

(d) 1 kHz + 4 kHz ($\varphi_1 = 90°$, $\varphi_2 = 270°$)

(e) 8 kHz + 1 kHz ($\varphi_1 = 90°$, $\varphi_2 = 112.5°$)

(f) Drums + classical music ($\varphi_1 = 90°$, $\varphi_2 = 180°$)

Figure 5.12: Azimuth estimation for two sources placed at various positions.

We conclude that sources that can be localized individually can also be located when other sources are present. Localization is possible even when the speakers are placed 22.5° from each other in most cases. Improvements in the utilized DOA estimation approach may increase the accuracy and reliability of localization, especially if these improvements lead to fewer peak shifts, as this reduces the possibility of peaks merging.

Chapter 6

Conclusions

The aim of this thesis was to perform sound source localization using the previously developed REM prototype. We saw this application as a first step towards allowing more complex spatial sound field analysis, since the detection of the DOA of sound sources may subsequently allow for their separation. To date, the research that has been conducted on the topic of DOA estimation using a single moving microphone is scarce and the only tested practical verification we are aware of is limited to localizing single, known frequencies.

We found that the circular rotation of a microphone introduces a periodic Doppler frequency shift into recorded signals. The phase of this frequency shift depends on the azimuth angle of an incoming sound source, while the degree to which the frequency is shifted is dependent on the elevation angle. Previous research estimates the phase of this periodic frequency shift to perform azimuth estimation. The limitations of this approach are the difficulty of localizing low frequencies and more complex sources, as well as not allowing for the removal of the distortions introduced by the microphone movement. To overcome these issues we first showed that a frequency modulated signal can be expressed as a weighted sum of Bessel functions spaced at the rotational frequency of the REM around any source frequency. The higher the modulation index of our frequency modulation, the more the energy of the source frequency is spread into these sidebands. In this context we introduced the focusedness as a measure of how concentrated or distributed the energy of a frequency is.

In a next step, we showed that frequency modulation can not only be introduced by a periodic circular microphone movement, but also by modifying the sampling rate of a stationary microphone. This knowledge allowed us to derive the TWA, which enables the compensation of the periodic Doppler frequency shift introduced into a recorded sound source given its azimuth and elevation angle by precisely time-shifting and interpolation the individual audio samples. An unknown sound source with an unknown DOA can then be localized by applying this algorithm for multiple azimuth and elevation angle guesses and finding the combination of angles at which

the focusedness is maximized. Additionally, this approach allows for the reconstruction of the source signal.

A second method of removing the distortions introduced by the microphone movement was presented, which acts on the spectrogram of the audio signal rather than the individual audio samples. Here, a modulation matrix and unmodulation matrix was derived by investigating how individual frequencies are modified when modulated, which allows for transformation of individual spectrogram frames into the modulated and unmodulated domain by vector-matrix multiplication. This approach has the potential of compensating the Doppler shifts much faster from a computational standpoint and potentially opens the door for defining an optimization problem which separates multiple audio sources given knowledge of their DOA. Unfortunately, the matrix-based approach was not applicable using our prototype since we found that more precise motor speed control is needed to use this method efficiently.

Using the TWA and focusedness-based localization we performed simulations which showed that our method of localization outperformed the accuracy of previous research when a large SNR is present. To overcome the lack of accuracy in low SNR situations we proposed splitting the spectrogram of the recorded audio into multiple frequency bands and performing focusedness-based DOA estimation on each band and spectrogram frame separately. In a subsequent step the PDF of all the DOA estimation points is estimated and the final DOA prediction is obtained from the peaks of this PDF. This modification significantly improved DOA estimation accuracy in low SNR situations and enabled the localization of multiple sources. Our simulations showed that this approach could detect up to four sources reliably in 2D space and two sources in 3D space as long as the source signals feature sufficiently constant frequencies.

Finally, we verified the proposed localization approach in practice using the REM placed in an anechoic chamber with one and two audio sources. We found that localization of one or two signals was possible and accurate for signals with many constant tones in the mid frequency range and speakers placed no closer than 22.5° apart. However, there were also many unexpected artifacts: Phantom peaks occasionally occurred in the PDFs, DOA detection peaks sometimes shifted by a few degrees and the peaks detected for higher frequencies split into multiple parts.

The discovery of these artifacts suggests further research is necessary. Improvements in the accuracy of the motor speed as well as wind and motor noise suppression may help overcome the DOA detection peak shifts and the phantom peaks. This will also open the door for the utilization of matrix-based unmodulation, greatly improving the speed of the localization. Additionally, modifications of the TWA to compensate not only the Doppler shifts, but also the frequency-dependent amplitude modulation may prevent the peaks detected for higher frequencies to split into multiple parts. The presented approach could be further improved by proposing a more intelligent method of combining the individual subband DOA predictions into a final DOA prediction by taking into consideration the REM self noise and the presence of constant

frequencies as well as the SNR within each subband.

Another avenue for future research is combining the REM with a stationary reference microphone. The current method relies on the focusedness to determine the DOA since the REM has no knowledge of the true source frequencies, however, the azimuth and elevation angles with maximum focusedness do not necessarily correspond to the true DOA for complex source signals. A stationary reference microphone would provide the REM with knowledge of the true source frequencies, allowing us to only search for the DOA of specific frequencies rather than the entire spectrum. This approach has the potential of drastically improving DOA estimation accuracy, since a large quantity of random DOA guesses due to noisy subbands containing no DOA information are removed. Additionally, if the stationary microphone is placed, for example, above the REM, it would allow for the differentiation of sounds arriving above and below the rotational plane, enabling localization in full 3D space.

Finally, the topic of source separation using the REM also warrants further exploration. We showed that the matrix-based unmodulation approach allows us to define an optimization problem which could potentially enable the separation of two or more audio sources, assuming we have perfect knowledge of their DOA. If this method is successful it would be a major breakthrough in spatial sound field analysis, as we would be able to localize and reconstruct individual audio sources using only a single microphone. This would allow the REM to perform similar audio processing applications to that of a spherical microphone array at a fraction of the hardware cost.

List of Abbreviations

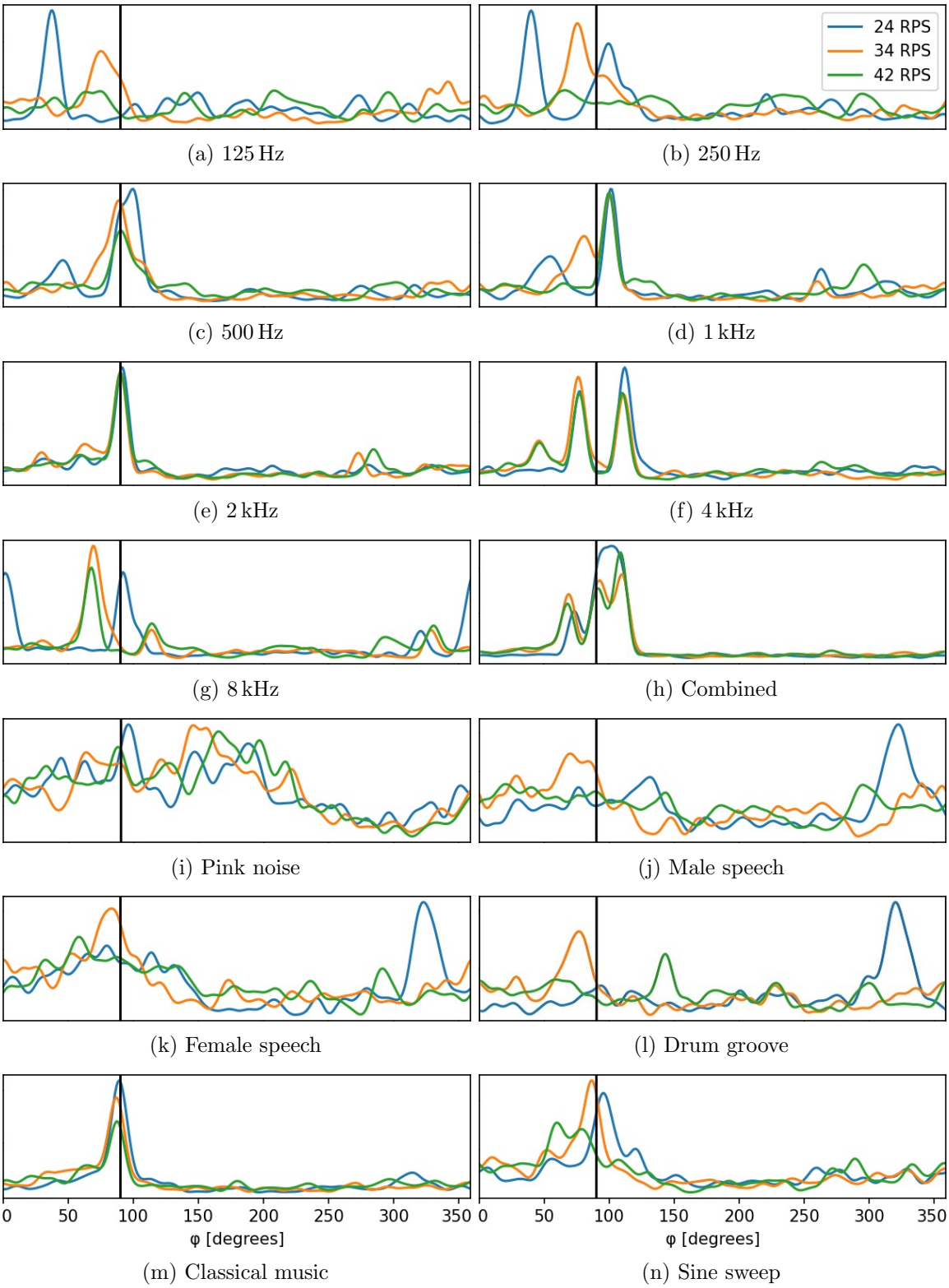
BLDC	BrushLess DC
CoG	Center of Gravity
DFT	Discrete Fourier Transform
DOA	Direction Of Arrival
EMA	Equatorial Microphone Array
ESC	Electronic Speed Controller
PDF	Probability Density Function
REM	Rotating Equatorial Microphone
RIR	Room Impulse Response
RMSE	Root Mean Square Error
RPS	Rotations Per Second
SNR	Signal-to-Noise Ratio
TKEO	Teager-Kaiser Energy Operator
TWA	Time Warping Algorithm

Appendix A

Direction of Arrival Estimation - Full Results of the Practical Verification

Here the full results of the experiments from Section 5.3 can be found. Figure A.1, Figure A.2 and Figure A.3 show the results of the azimuth estimation for a single speaker placed at respective positions $\varphi = 90°/180°/90°$ and $\theta = 90°/90°/45°$. Figure A.4 depicts the elevation estimations for the signals that were detected reliably from Figure A.3. The used φ-values for the elevation estimation are shown in brackets under each figure for the 24 RPS, 34 RPS and 42 RPS cases, respectively.

Figure A.5, Figure A.6, Figure A.7 and Figure A.8 show the results of the azimuth estimation for two speakers placed at respective positions $\varphi_1 = 90°$, $\theta_1 = 90°/90°/90°/45°$, $\varphi_2 = 180°/270°/112.5°/180°$ and $\theta_2 = 90°$. The elevation estimations have been omitted due to their inconclusive nature similarly to the results from Figure A.4.



(a) 125 Hz (b) 250 Hz

(c) 500 Hz (d) 1 kHz

(e) 2 kHz (f) 4 kHz

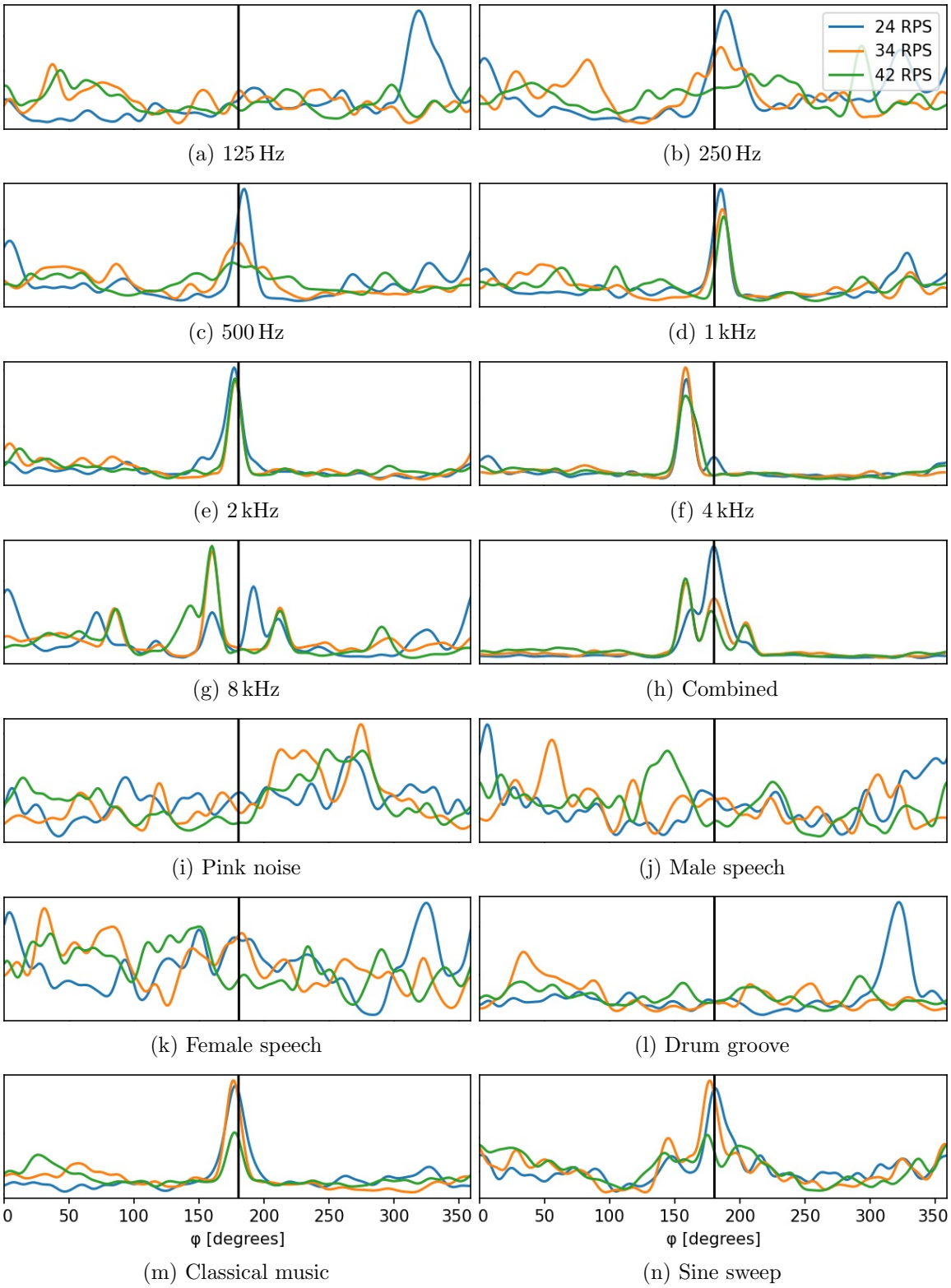
(g) 8 kHz (h) Combined

(i) Pink noise (j) Male speech

(k) Female speech (l) Drum groove

(m) Classical music (n) Sine sweep

Figure A.1: Azimuth estimation for a single source placed at $\varphi = 90°$, $\theta = 90°$.



(a) 125 Hz (b) 250 Hz

(c) 500 Hz (d) 1 kHz

(e) 2 kHz (f) 4 kHz

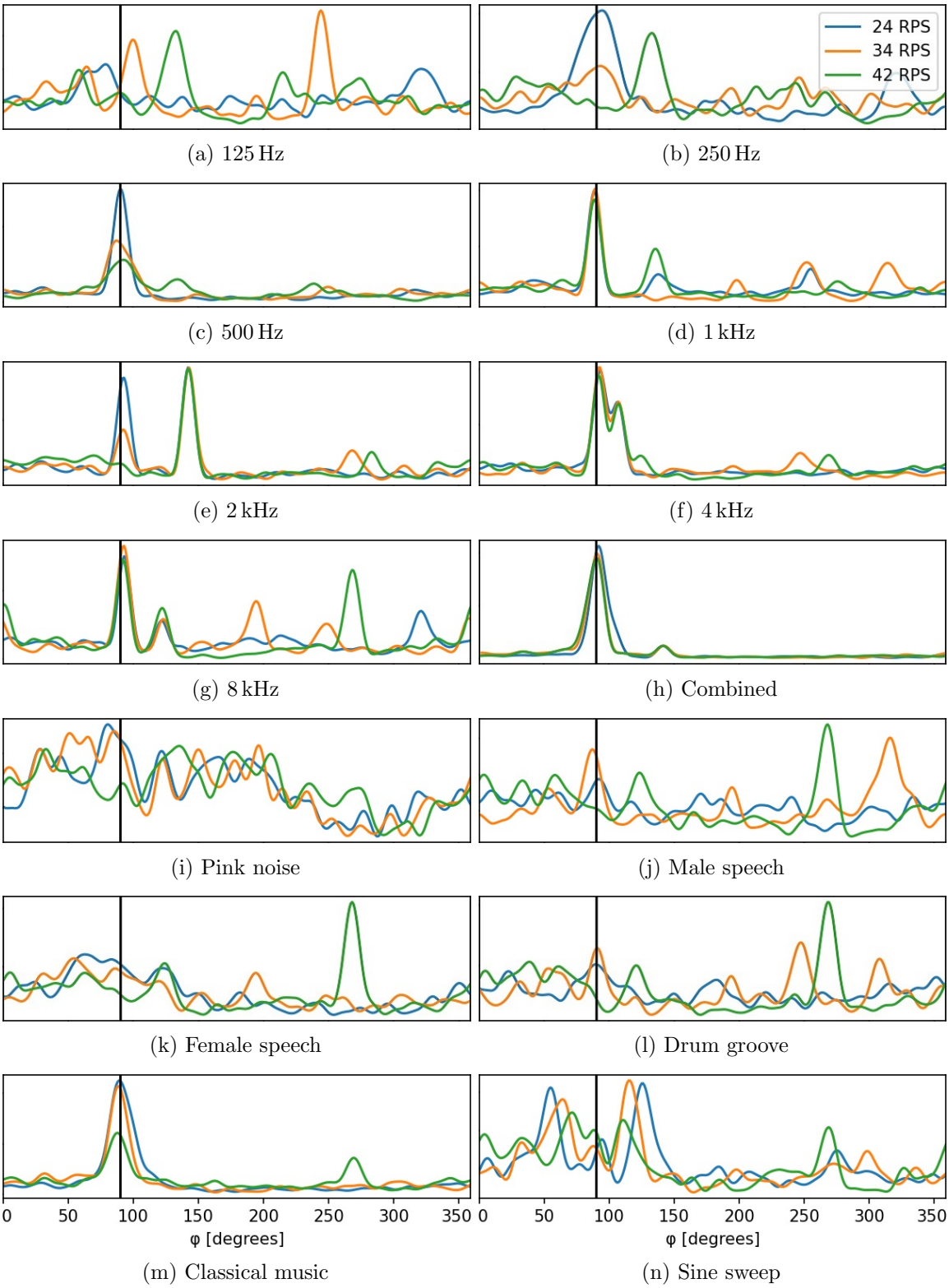
(g) 8 kHz (h) Combined

(i) Pink noise (j) Male speech

(k) Female speech (l) Drum groove

(m) Classical music (n) Sine sweep

Figure A.2: Azimuth estimation for a single source placed at $\varphi = 180°$, $\theta = 90°$.



(a) 125 Hz (b) 250 Hz

(c) 500 Hz (d) 1 kHz

(e) 2 kHz (f) 4 kHz

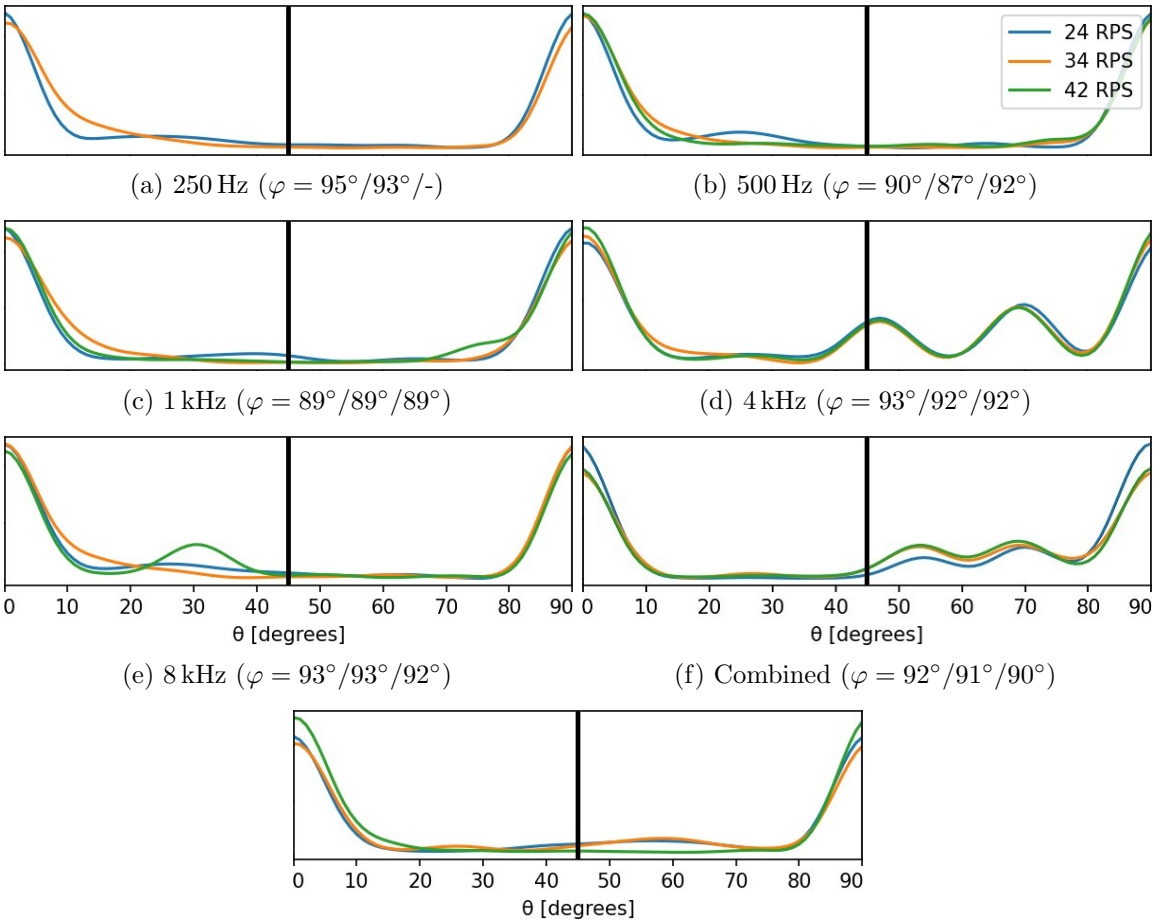
(g) 8 kHz (h) Combined

(i) Pink noise (j) Male speech

(k) Female speech (l) Drum groove

(m) Classical music (n) Sine sweep

Figure A.3: Azimuth estimation for a single source placed at $\varphi = 90°$, $\theta = 45°$.



(a) 250 Hz ($\varphi = 95°/93°/\text{-}$)

(b) 500 Hz ($\varphi = 90°/87°/92°$)

(c) 1 kHz ($\varphi = 89°/89°/89°$)

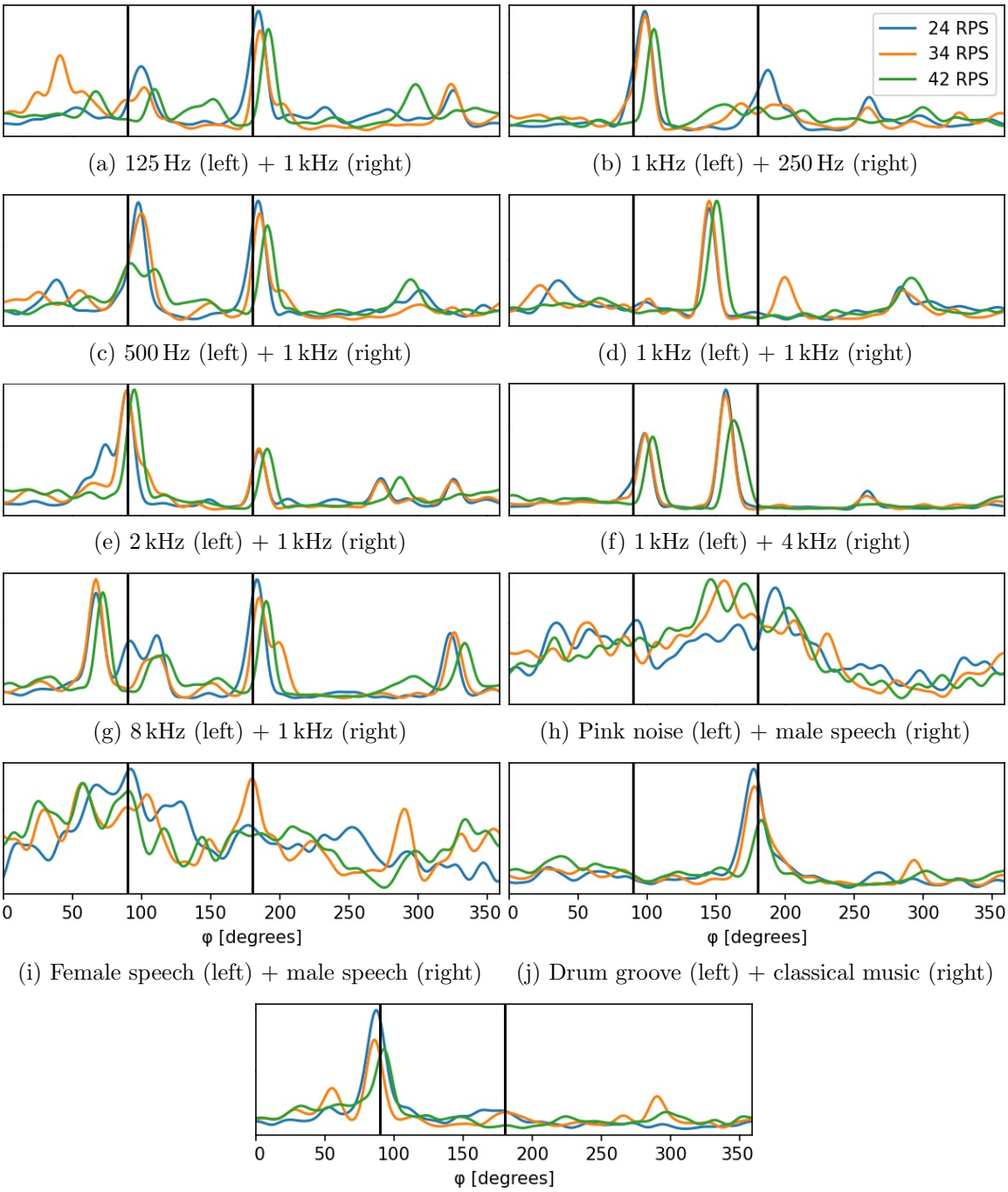
(d) 4 kHz ($\varphi = 93°/92°/92°$)

(e) 8 kHz ($\varphi = 93°/93°/92°$)

(f) Combined ($\varphi = 92°/91°/90°$)

(g) Classical music ($\varphi = 89°/89°/88°$)

Figure A.4: Elevation estimation for a single source placed at $\varphi = 90°$, $\theta = 45°$.



(a) 125 Hz (left) + 1 kHz (right)

(b) 1 kHz (left) + 250 Hz (right)

(c) 500 Hz (left) + 1 kHz (right)

(d) 1 kHz (left) + 1 kHz (right)

(e) 2 kHz (left) + 1 kHz (right)

(f) 1 kHz (left) + 4 kHz (right)

(g) 8 kHz (left) + 1 kHz (right)

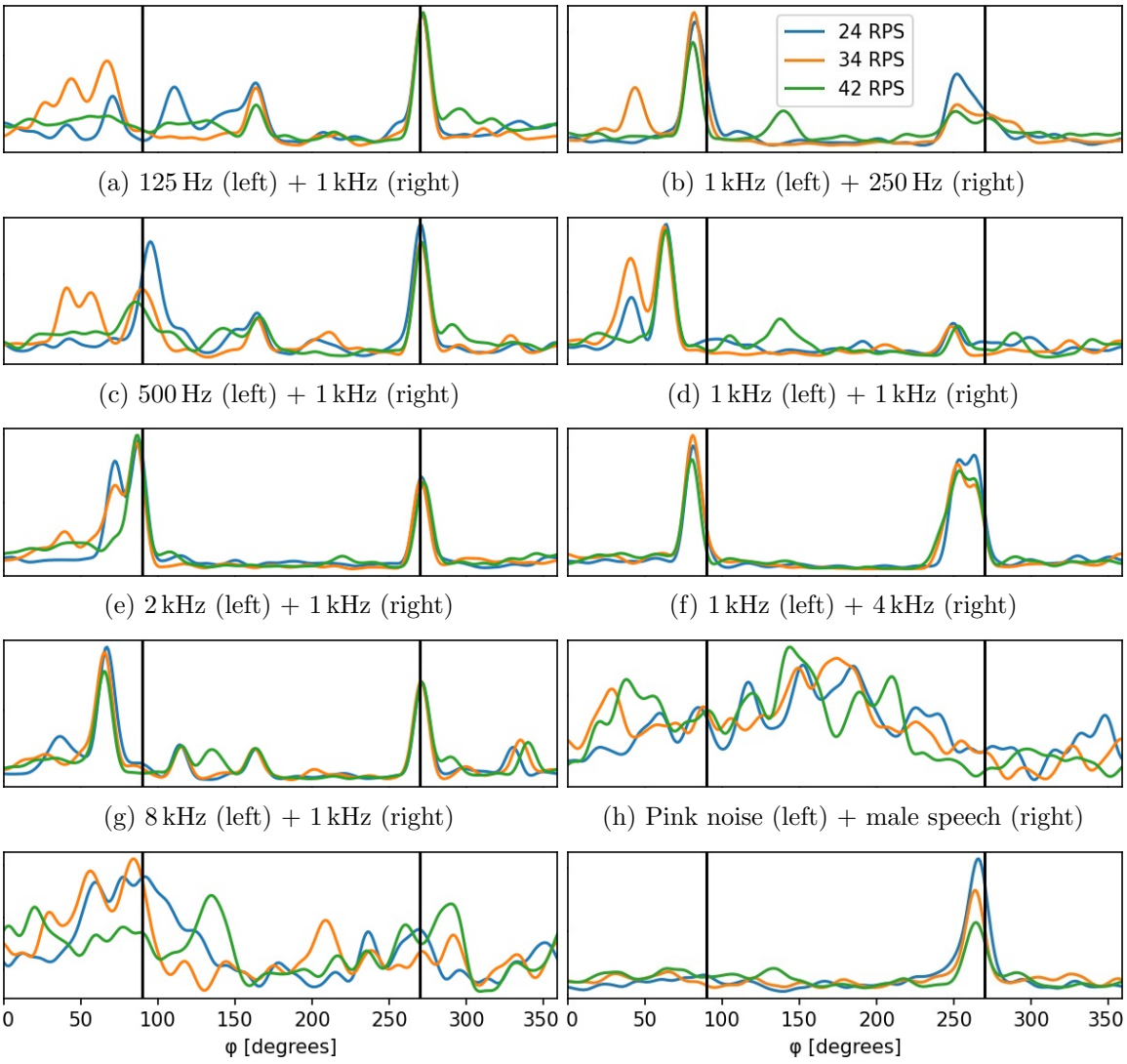
(h) Pink noise (left) + male speech (right)

(i) Female speech (left) + male speech (right)

(j) Drum groove (left) + classical music (right)

(k) Classical music (left) + female speech (right)

Figure A.5: Azimuth estimation for two sources placed at $\varphi_1 = 90°$, $\theta_1 = 90°$ (left) and $\varphi_2 = 180°$, $\theta_2 = 90°$ (right).



(a) 125 Hz (left) + 1 kHz (right)

(b) 1 kHz (left) + 250 Hz (right)

(c) 500 Hz (left) + 1 kHz (right)

(d) 1 kHz (left) + 1 kHz (right)

(e) 2 kHz (left) + 1 kHz (right)

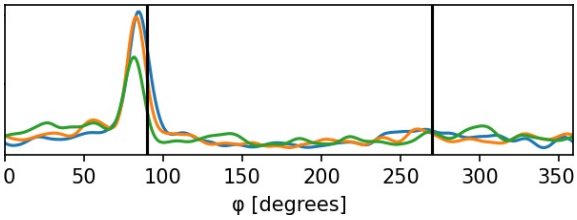
(f) 1 kHz (left) + 4 kHz (right)

(g) 8 kHz (left) + 1 kHz (right)

(h) Pink noise (left) + male speech (right)

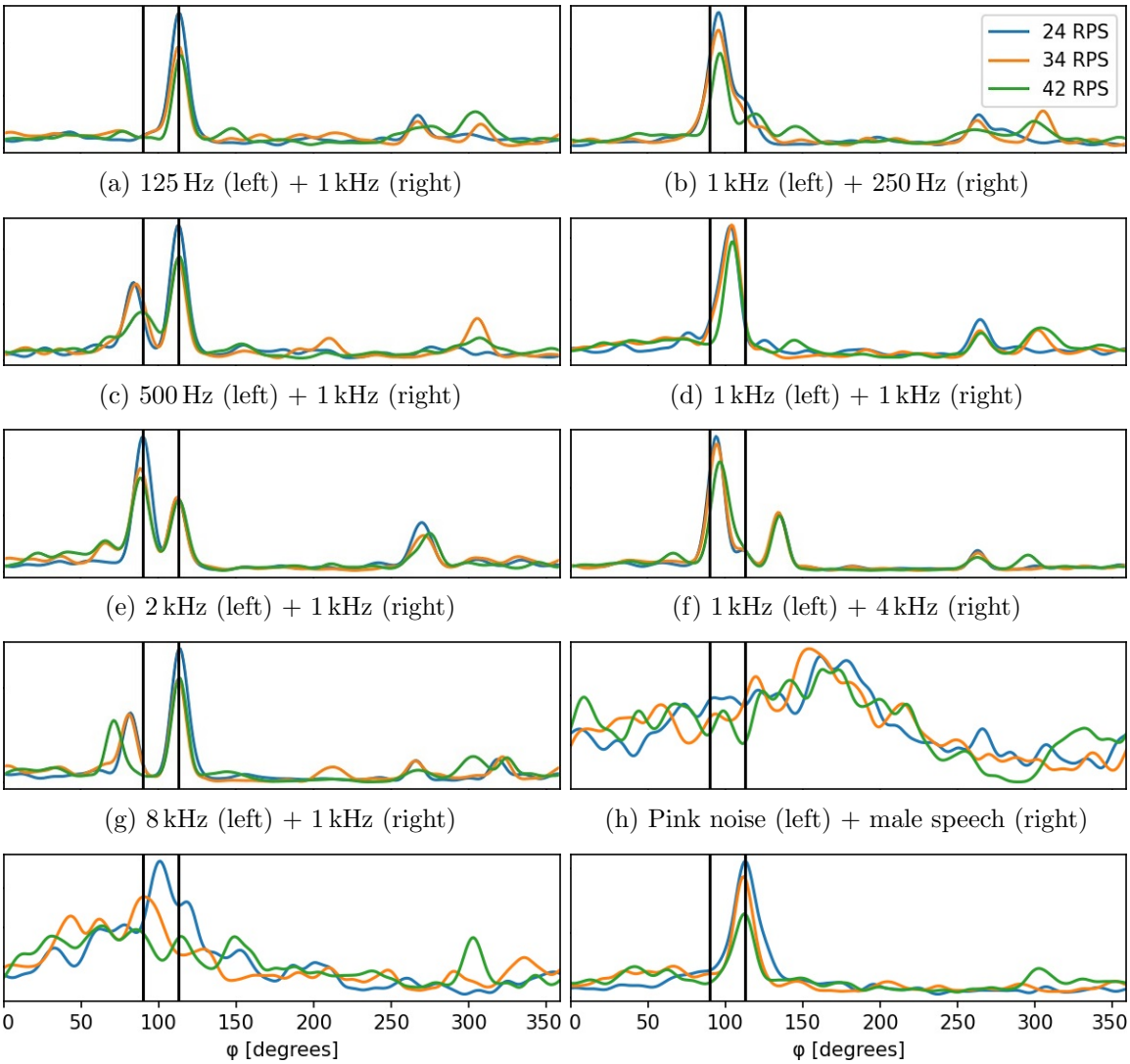
(i) Female speech (left) + male speech (right)

(j) Drum groove (left) + classical music (right)



(k) Classical music (left) + female speech (right)

Figure A.6: Azimuth estimation for two sources placed at $\varphi_1 = 90°$, $\theta_1 = 90°$ (left) and $\varphi_2 = 270°$, $\theta_2 = 90°$ (right).



(a) 125 Hz (left) + 1 kHz (right)

(b) 1 kHz (left) + 250 Hz (right)

(c) 500 Hz (left) + 1 kHz (right)

(d) 1 kHz (left) + 1 kHz (right)

(e) 2 kHz (left) + 1 kHz (right)

(f) 1 kHz (left) + 4 kHz (right)

(g) 8 kHz (left) + 1 kHz (right)

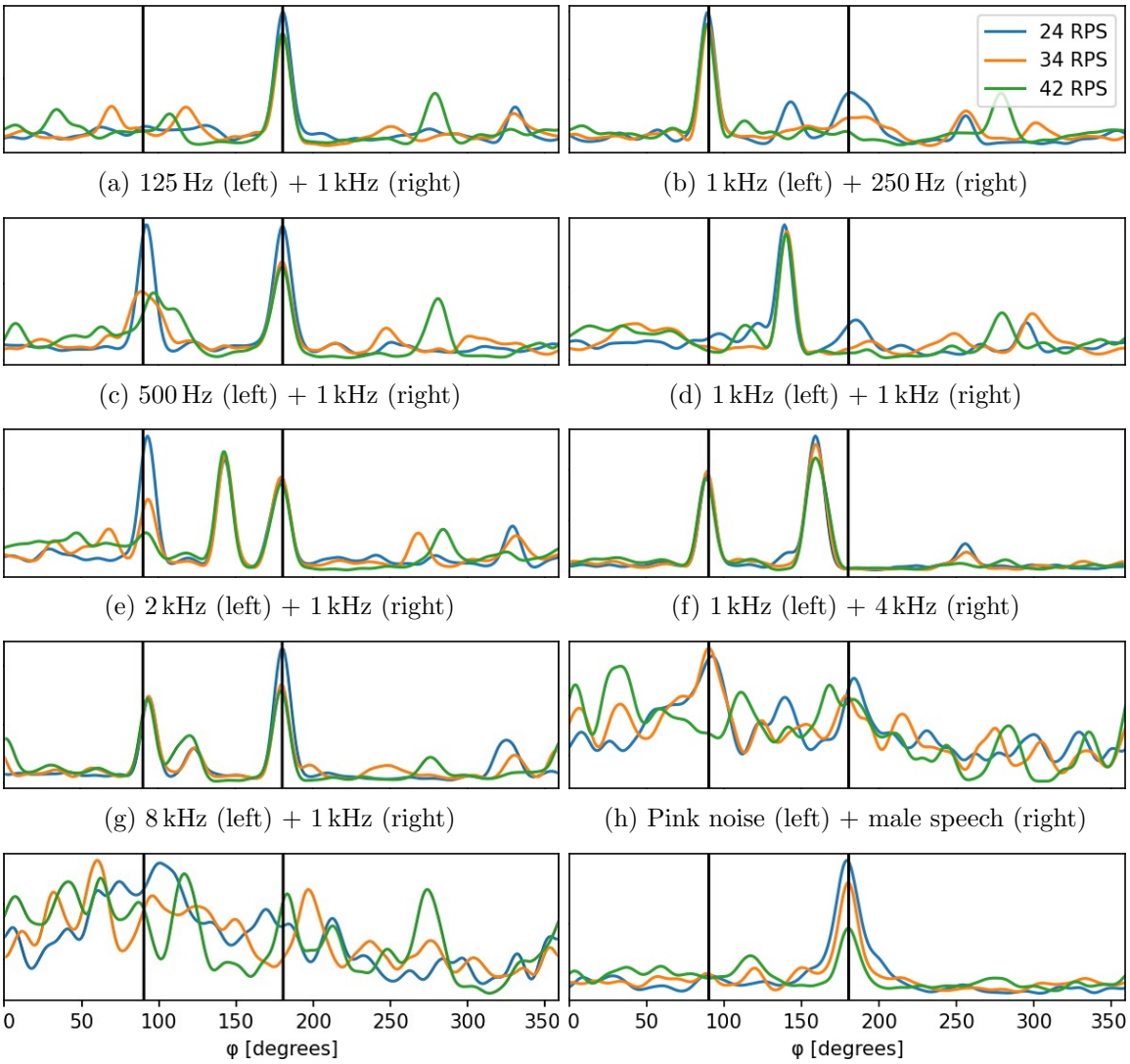
(h) Pink noise (left) + male speech (right)

(i) Female speech (left) + male speech (right)

(j) Drum groove (left) + classical music (right)

(k) Classical music (left) + female speech (right)

Figure A.7: Azimuth estimation for two sources placed at $\varphi_1 = 90°$, $\theta_1 = 90°$ (left) and $\varphi_2 = 112.5°$, $\theta_2 = 90°$ (right).



(a) 125 Hz (left) + 1 kHz (right)

(b) 1 kHz (left) + 250 Hz (right)

(c) 500 Hz (left) + 1 kHz (right)

(d) 1 kHz (left) + 1 kHz (right)

(e) 2 kHz (left) + 1 kHz (right)

(f) 1 kHz (left) + 4 kHz (right)

(g) 8 kHz (left) + 1 kHz (right)

(h) Pink noise (left) + male speech (right)

(i) Female speech (left) + male speech (right)

(j) Drum groove (left) + classical music (right)

(k) Classical music (left) + female speech (right)

Figure A.8: Azimuth estimation for two sources placed at $\varphi_1 = 90°$, $\theta_1 = 45°$ (left) and $\varphi_2 = 180°$, $\theta_2 = 90°$ (right).

Bibliography

[1] *§10.23 sums*. https://dlmf.nist.gov/10.23, 2023. Retrieved 16.03.2023.

[2] J. Ahrens, H. Helmholz, D. L. Alon, and S. V. A. Garí, *Spherical harmonic decomposition of a sound field based on observations along the equator of a rigid spherical scatterer*, in The Journal of the Acoustical Society of America 150, 2021, pp. 805–815.

[3] T. Ajdler, L. Sbaiz, and M. Vetterli, *Dynamic measurement of room impulse responses using a moving microphone*, in The Journal of the Acoustical Society of America 122, 2007, pp. 1636–1645.

[4] R. J. Beerends, H. G. ter Morsche, J. C. van de Berg, and E. M. van de Vrie, *Fourier and Laplace Transforms*, Cambridge University Press, 2003, pp. 15, 179, 229–232.

[5] W. R. Calvert, *Transmission systems: An introduction to frequency modulation*, in International Journal of Satellite Communications, 1984.

[6] P. Clark and L. Atlas, *Time-frequency coherent modulation filtering of nonstationary signals*, IEEE Transactions on Signal Processing, 57 (2009), pp. 4323–4332.

[7] J. D. Cook, *Analyzing an fm signal*. https://www.johndcook.com/blog/2016/02/17/analyzing-an-fm-signal/, 2016. Retrieved 16.03.2023.

[8] F. M. (Ed.), *Nonuniform Sampling: Theory and Practice*, Springer New York, 2001, p. 170.

[9] N. Hahn and S. Spors, *Continuous measurement of impulse responses on a circle using a uniformly moving microphone*, in 2015 23rd European Signal Processing Conference (EUSIPCO), 2015, pp. 2536–2540.

[10] ——, *Continuous measurement of spatial room impulse responses using a non-uniformly moving microphone*, in 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017, pp. 205–208.

[11] F. Harris, *On the use of windows for harmonic analysis with the discrete fourier transform*, Proceedings of the IEEE, 66 (1978), pp. 51–83.

[12] Y. Hioka, R. Drage, T. Boag, and E. Everall, *Direction of arrival estimation using a circularly moving microphone*, in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), 2018, pp. 91–95.

[13] T. O. Hodson, *Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not*. https://gmd.copernicus.org/articles/15/5481/2022/, 2022. Retrieved 18.04.2023.

[14] J. Kaiser, *On a simple algorithm to calculate the 'energy' of a signal*, in International Conference on Acoustics, Speech, and Signal Processing, 1990, pp. 381–384 vol.1.

[15] F. Katzberg, M. Maass, and A. Mertins, *Spherical harmonic representation for dynamic sound-field measurements*, in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 426–430.

[16] F. Katzberg, R. Mazur, M. Maass, P. Koch, and A. Mertins, *Measurement of sound fields using moving microphones*, in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 3231–3235.

[17] ——, *Multigrid reconstruction of sound fields using moving microphones*, in 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017, pp. 191–195.

[18] ——, *Sound-field measurement with moving microphones*, in The Journal of the Acoustical Society of America 141, 2017, p. 3220–3235.

[19] Knowles, *Sph0645lm4h-1 rev a datasheet.* `https://www.knowles.com/docs/default-source/default-document-library/sph0645lm4h-1-datasheet.pdf`, 2017. Retrieved 19.04.2023.

[20] T. Kristjansson, H. Attias, and J. Hershey, *Single microphone source separation using high resolution signal reconstruction*, in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 2004, pp. ii–817.

[21] H. Landau, *Sampling, data transmission, and the nyquist rate*, Proceedings of the IEEE, 55 (1967), pp. 1701–1706.

[22] J. Lawrence, J. Ahrens, and N. Peters, *Comparison of position estimation methods for the rotating equatorial microphone*, in 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), 2022, pp. 1–5.

[23] S. Maymon and A. V. Oppenheim, *Sinc interpolation of nonuniform samples*, IEEE Transactions on Signal Processing, 59 (2011), pp. 4745–4758.

[24] mh acoustics, *Eigenmike® microphone.* `https://mhacoustics.com/products`. Retrieved 07.04.2023.

[25] E. Parzen, *On estimation of a probability density function and mode*, The Annals of Mathematical Statistics, 33 (1962), p. 1065–1076.

[26] A. Pyzara, B. Bylina, and J. Bylina, *The influence of a matrix condition number on iterative methods' convergence*, in 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), 2011, pp. 459–464.

[27] I. Quilez, *File:spherical harmonics.png.* `https://commons.wikimedia.org/wiki/File:Spherical_Harmonics.png`, 2023. Retrieved 11.04.2023.

[28] S. Rapuano and F. J. Harris, *An introduction to fft and time domain windows*, IEEE Instrumentation & Measurement Magazine, 10 (2007), pp. 32–44.

[29] J. Rosen and L. Q. Gothard, *Encyclopedia of Physical Science*, Infobase Publishing, 2009, p. 155.

[30] T. Schanze, *Sinc interpolation of discrete periodic signals*, IEEE Transactions on Signal Processing, 43 (1995), pp. 1502–1503.

[31] A. Schasse and R. Martin, *Localization of acoustic sources based on the teager-kaiser energy operator*, in 2010 18th European Signal Processing Conference, 2010, pp. 2191–2195.

[32] A. Schasse, C. Tendyck, and R. Martin, *Source localization based on the doppler effect*, in IWAENC 2012; International Workshop on Acoustic Signal Enhancement, 2012, pp. 1–4.

[33] M. Tohyama, *Acoustic Signals and Hearing: A Time-Envelope and Phase Spectral Approach*, Elsevier Science, 2020, pp. 1–2.

List of Figures