Ismo Pänkäläinen

Spatial analysis of sound field for parametric sound reproduction with sparse microphone arrays

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 10.3.2018

Thesis supervisor:

Prof. Ville Pulkki

Thesis advisors:

D.Sc. (Tech.) Oliver Thiergart

Prof. Emanuël Habets

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Ismo Pänkäläinen

Title: Spatial analysis of sound field for parametric sound reproduction with sparse microphone arrays

Date: 10.3.2018 Language: English Number of pages: 7+69

Department of Signal Processing and Acoustics

Professorship: Acoustics and Audio Technology Code: ELEC3030

Supervisor: Prof. Ville Pulkki

Advisors: D.Sc. (Tech.) Oliver Thiergart, Prof. Emanuël Habets

In spatial audio capturing the aim is to store information about the sound field so that the sound field can be reproduced without a perceptual difference to the original. The need for this is in applications like virtual reality and teleconferencing. Traditionally the sound field has been captured with a B-format microphone, but it is not always a feasible solution due to size and cost constraints. Alternatively, also arrays of omnidirectional microphones can be utilized and they are often used in devices like mobile phones. If the microphone array is sparse, i.e., the microphone spacings are relatively large, the analysis of the sound Direction of Arrival (DoA) becomes ambiguous in higher frequencies. This is due to spatial aliasing, which is a common problem in narrowband DoA estimation.

In this thesis the spatial aliasing problem was examined and its effect on DoA estimation and spatial sound synthesis with Directional Audio Coding (DirAC) was studied. The aim was to find methods for unambiguous narrowband DoA estimation. The current State of the Art methods can remove aliased estimates but are not capable of estimating the DoA with the optimal Time-Frequency resolution. In this thesis similar results were obtained with parameter extrapolation when only a single broadband source exists. The main contribution of this thesis was the development of a correlation-based method. The developed method utilizes pre-known, array-specific information on aliasing in each DoA and frequency. The correlation-based method was tested and found to be the best option to overcome the problem of spatial aliasing. This method was able to resolve spatial aliasing even with multiple sources or when the source's frequency content is completely above the spatial aliasing frequency. In a listening test it was found that the correlation-based method could provide a major improvement to the DirAC synthesized spatial image quality when compared to an aliased estimator.

Keywords: DirAC, spatial aliasing, microphone array, direction of arrival

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Ismo Pänkäläinen

Työn nimi: Äänikentän tila-analyysi parametrista tilaäänentoistoa varten käyttäen harvoja mikrofoniasetelmia

Päivämäärä: 10.3.2018 Kieli: Englanti Sivumäärä: 7+69

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka ja audiotekniikka Koodi: ELEC3030

Valvoja: Prof. Ville Pulkki

Ohjaajat: TkT Oliver Thiergart, Prof. Emanuël Habets

Tilaäänen tallentamisessa tavoitteena on tallentaa äänikentän ominaisuudet siten, että äänikenttä pystytään jälkikäteen syntetisoimaan ilman kuuloaistilla havaittavaa eroa alkuperäiseen. Tarve tälle on löytyy erilaisista sovelluksista, kuten virtuaalitodellisuudesta ja telekonferensseista. Perinteisesti äänikentän ominaisuuksia on tallennettu B-formaatti mikrofonilla, jonka käyttö ei kuitenkaan aina ole koko- ja kustannussyistä mahdollista. Vaihtoehtoisesti voidaan käyttää myös pallokuvioisista mikrofoneista koostuvia mikrofoniasetelmia. Mikäli mikrofonien väliset etäisyydet ovat liian suuria, eli asetelma on harva, tulee äänen saapumissuunnan selvittämisestä epäselvää korkeammilla taajuuksilla. Tämä johtuu ilmiöstä nimeltä tilallinen laskostuminen.

Tämän diplomityön tarkoituksena oli tutkia tilallisen laskostumisen ilmiötä, sen vaikutusta saapumissuunnan arviointiin sekä tilaäänisynteesiin Directional Audio Coding (DirAC) -menetelmällä. Lisäksi tutkittiin menetelmiä, joiden avulla äänen saapumissuunta voitaisiin selvittää oikein myös tilallisen laskostumisen läsnäollessa. Työssä havaittiin, että nykyiset ratkaisut laskostumisongelmaan eivät kykene tuottamaan oikeita suunta-arvioita optimaalisella aika-taajuusresoluutiolla. Tässä työssä samantapaisia tuloksia saatiin laajakaistaisen äänilähteen tapauksessa ekstrapoloimalla suunta-arvioita laskostumisen rajataajuuden alapuolelta. Työn pääosuus oli kehittää korrelaatioon perustuva saapumissuunnan arviointimenetelmä, joka kykenee tuottamaan luotettavia arvioita rajataajuuden yläpuolella ja useamman äänilähteen ympäristöissä. Kyseinen menetelmä hyödyntää mikrofoniasetelmalle ominaista, saapumissuunnasta ja taajuudesta riippuvaista laskostumiskuviota. Kuuntelukokeessa havaittiin, että korrelaatioon perustuva menetelmä voi tuoda huomattavan parannuksen syntetisoidun tilaäänikuvan laatuun verrattuna synteesiin laskostuneilla suunta-arvioilla.

Avainsanat: DirAC, tilallinen laskostuminen, mikrofoniasetelma, äänen saapumissuunta

Preface

About six years ago, I told my friend that I'm not sure if I want to go to university. *"I don't have the mathematical skills to be an engineer."*, *"I'm not sure what I want to do when I grow up."*, *"I don't think I'm capable of doing that."* The list of these thoughts was long. Some of these thoughts that I had were correct, some were wrong. At this point there are still things I'm uncertain of, e.g., Fourier analysis, and that is the best part of life. However, there is one thing I know for sure. I could not have finished this degree without the awesome people around me and these words are for all of you. Not just the ones mentioned here, but all.

First of all, I want to thank Oliver Thiergart for his support, advices and patience throughout this thesis work. He showed me what it means to be precise and accurate. I also want to thank Emanuël Habets for the supportive and encouraging attitude he had everytime we met. From Aalto University I want to thank Ville Pulkki for introducing me to the wonderful world of acoustics and audio technology. It was always a pleasure to attend to his lectures. A major thankyou goes to all the fellow students that I had the pleasure to study, work and spend time with. Because I can't name you all in this limited space, I'll just say thank you Wappuryhmä and Joutomiehet. Thank you friends that have been around a lot longer. Thank you Aino and Eeva, I would not be the person I am now if I was not the little brother. Thank you Äiti and Isi for the roots and wings you gave me.

Thank you Laura for being there. I promise to do the same to You.

Otaniemi, 10.3.2018

Ismo Pänkäläinen

Contents

Symbols

c	Speed of sound
d	Interchannel distance
D	Number of analysis dimensions
E	Sound field energy density
f	Frequency
h	Microphone pair index
H	Number of microphone pairs
i	Imaginary unit
$\mathbf{I}$	Intensity vector
k	Frequency index
K	Number of frequency bins
m	Microphone index
M	Number of microphones
n	Time index
$\mathbf{n}$	Direction of arrival vector
p	Air pressure
r	Microphone array radius
$\mathbf{u}$	Particle velocity
δ	Elevation angle
κ	Wavenumber
λ	Wavelength
μ	Interchannel phase difference
τ	Time delay
ϕ	Phase
φ	Azimuth angle
Φ	Power
Ψ	Diffuseness

1 Introduction

Spatial audio has become an important part of current entertainment and communication systems. By including it in applications like virtual reality, teleconferencing or home theaters, it is possible to create even more immersive user experiences. In addition, by using information about the sound field it is possible to utilize various algorithms to achieve other improvements, like noise reduction, in the recordings. An example of a parametric spatial audio reproduction method is the Directional Audio Coding [1], [2], for which the sound field analysis is performed in the time-frequency domain. This way the direction of arrival and diffuseness parameters are estimated for each time-frequency bin separately. Using these parameters and the recorded sound, it is possible to reproduce a sound field that it is perceptually equivalent to the original. When capturing spatial sound, different microphone setups are used. A common example of a microphone setup for capturing the sound field is the B-format microphone [3] [4]. In practical consumer devices, like mobile phones, the use of the B-format microphone is not applicable due to size and cost considerations. In addition, the problem of spatial aliasing occurs in the higher frequencies. Spatial aliasing happens when the spacing between the microphone capsules is too large for the considered frequency. Alternatively to the B-format microphone, several omnidirectional microphones can be used and the interchannel phase difference (IPD) information is utilized for the parameter estimation.

The direction of arrival estimation with omnidirectional microphone arrays is the main focus in this thesis. For this, several methods have been proposed like the Estimation of Signal Parameters via Rotational Invariance Techniques [5] or the Weighted Least Squares estimator [6]. However, in these methods the placing of the omnidirectional microphones around the device causes some drawbacks in the direction of arrival estimation. Most importantly in this thesis, the problem of spatial aliasing in higher frequencies. When the microphones are relatively far away from each other, the spatial aliasing is even more of a problem as it occurs on even lower frequencies. In this case the use of interchannel phase difference information produces ambiguity on which direction the sound is actually arriving from [7].

There are current solutions to overcome the problem of spatial aliasing, but these methods are not capable of estimating the direction of arrival parameter in the desired frequency resolution. In addition, some methods like phase unwrapping [8], [9] assume only one source per time frame. Another approach is the envelope detection [10], which is capable of estimating the direction of arrival also in higher frequencies but requires the processing to be made in wider bands. This way one direction of arrival can be estimated reliably for each band. This allows multiple sources to be identified in each time frame, but because the processing is performed in frequency bands, the frequency resolution is decreased.

This thesis studies the spatial aliasing problem when performing narrowband direction of arrival estimation with sparse microphone arrays of omnidirectional

microphones. The reasons behind the spatial aliasing problem in the current state of the art direction of arrival estimators are explained. Some state of the art solutions for the problem are explained and implemented in Matlab so that their performance can be compared. The contributions of this thesis for the spatial aliasing problem can be summarized as:

- Study of the perceptual effects of spatial aliasing in Directional Audio Coding reproduced spatial sound.
- Study of the effects of reducing the time-frequency-resolution of the direction of arrival parameter for Directional Audio Coding reproduction. This is done using an objective measure and a listening test. The results of this study can be used in development of algorithms that aim to reduce and resolve the negative effects of spatial aliasing.
- Development of direction of arrival parameter extrapolation method, which aims to reduce the negative perceptual effects of spatially aliased direction of arrival estimates. This method utilizes only the direction of arrival estimates below the aliasing frequency, which are extrapolated to the higher frequencies. Matlab implementation of this method.
- Development of a correlation-based direction of arrival estimator, which aims to resolve the spatial aliasing problem. In this method the information on how the aliasing changes the interchannel phase difference values, is used. Because this happens differently for each frequency and direction of arrival, the correlation between measured interchannel phase differences and previously computed interchannel phase difference values for each frequency and direction of arrival can be used. Matlab implementation of this method.
- Comparison of the direction of arrival estimation accuracy under the effect of spatial aliasing for the extrapolation, correlation, phase unwrapping and envelope methods in plane wave and room impulse response simulations.
- Listening test to compare the spatial image accuracy in synthesized sound scenes. The used direction of arrival parameters are estimated using the extrapolation and correlation methods and they are compared to the Weighted Least Squares method which suffers from spatial aliasing.

1.1 Structure of the thesis

This thesis is organized as follows: Sec. 2 provides the background information that is needed for this thesis. The section includes basic information on physics of sound, time-frequency audio processing, human hearing and spatial audio capturing and reproduction. Also the signal model, that is used throughout the thesis, is explained. Sec. 3 presents some state of the art direction of arrival estimation methods. Sec. 4 explains the problem of spatial aliasing and the effects it causes in spatial audio reproduction. In this section there is also a brief study on the effects of decreasing

the time-frequency resolution. Sec. 5 explains some state of the art methods that aim to overcome the spatial aliasing problem. In Sec. 6 the proposed extrapolation and correlation approaches are presented. In Sec. 7 the results of explained methods are presented and they are also compared with with some state of the art methods. Also the listening test results for the proposed methods are presented here. Sec. 8 provides a look on the future work and concludes the thesis.

1.2 Notations

The following notations are used in this thesis; scalar variables are expressed as italic lower case and upper case letters for time and frequency domains respectively, e.g., x and X. Vectors and matrices are presented as bolded lower case and upper case letters, e.g., $\mathbf{x}$ and $\mathbf{X}$, respectively. The expectation operator is denoted as $\mathrm{E}\{\cdot\}$. The transpose operator is denoted as $(\cdot)^{\mathrm{T}}$. Complex conjugate is denoted as $(\cdot)^{*}$. Hermitian operator is denoted as $(\cdot)^{\mathrm{H}}$. Absolute value is denoted as $|\cdot|$ and vector norm as $||\cdot||$.

2 Background

This section provides the needed background for this thesis. It starts by introducing the properties of sound in time and frequency domains. After these, the signal model is presented. Then, relevant properties of human hearing, especially in spatial hearing, are presented. The section ends with an introduction to spatial sound and a brief explanation of Directional Audio Coding.

2.1 Physics of sound

Sound is produced when a sound source, like a vibrating loudspeaker cone or the human speech production system, moves the air particles. This causes pressure fluctuations around the static air pressure. The compression and rarefaction of particles transfer to the particles around and the sound propagates as longitudinal sound waves [11]. These pressure changes can be perceived as vibrations and in the inner ear they are interpreted as different frequencies and in the end as sound from the source. Because the particles transfer the sound, a medium is needed and there is no sound propagation in vacuum [12]. Sounds can be divided in three main categories which are tonal, non-tonal and transients. Examples of these are vibrating string, white noise, and a clap of hands, respectively.

In air the particles are in constant pressure which is about 101325 Pa (1 atm) at mean sea level [13]. The pressure changes caused by everyday sound sources are much smaller. Usually the sound pressure level L_p is expressed in decibels (dB) which is a logarithmic scale. The sound pressure level is defined as a relation of measured sound pressure p and the reference pressure p_0 of 20µPa as [14]

$$L_p = 20 \log_{10} \left(\frac{p}{p_0} \right). \tag{1}$$

As the pressure changes move the air particles back and forth, the phase ϕ of the signal reveals at which part of this cycle is the observation point currently at [11]. The phase is usually presented as a number in the range of $[0\ 2\pi]$ due to the cyclic nature. The distance in meters at which a periodic wave completes a full cycle is known as the wavelength λ. The number of these cycles within one second determines the frequency f of the signal. The unit of frequency is Herz [Hz] [14]. The speed of the propagation, i.e., the speed of sound, can be determined as the multiplication of f and λ as [11]

$$c = f\lambda. \tag{2}$$

The speed of sound is dependent on factors like temperature and an often used constant for air is 343m/s in room temperature [11]. In Fig. 1 two sine signals are presented to demonstrate ϕ, f and λ.

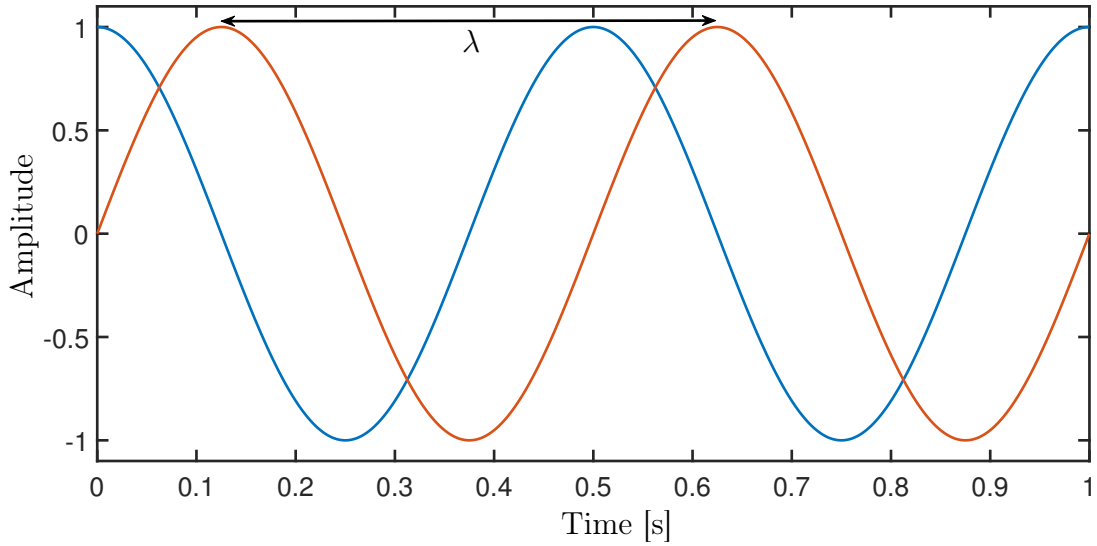

Figure 1: Two sine signals that complete two full cycles in one second, i.e, $f = 2$Hz. Red curve reaches the maximum $\lambda/4$ wavelengths later than the blue signal, i.e., there is a $\pi/2$ phase difference.

The electrical or mechanical power driving the sound source is transformed into acoustical energy. This energy is transferred and distributed to the space around the source. The local density and direction of this energy transfer is presented as the intensity $\mathbf{i}$ [14]. Intensity is defined as the time average $\langle\cdot\rangle$ of the product of $p(t)$ and particle velocity $\mathbf{u}(t)$ [15] [16]

$$\mathbf{i} = \langle p(t)\mathbf{u}(t)\rangle. \tag{3}$$

The particle velocity is the first time derivative of the particle displacement caused by the fluctuations [11]. The term intensity usually refers to the active intensity. Reactive component of intensity describes the part of energy flow that is not propagating. The reactive intensity is the imaginary part of $p\mathbf{u}^*$, where $(\cdot)*$ denotes the complex conjugate [15]. In the near field of the source, p and $\mathbf{u}$ are out of phase which makes the reactive part dominant. In the far field, p and $\mathbf{u}$ are in phase and there is propagation of energy [15]. As $\mathbf{i}$ is a vector measure that points to the direction of the energy flow, it is an important measure in sound field analysis [16].

2.2 Frequency domain

Often in audio signal processing the frequency content provides more information than just the time domain signal that is captured with a microphone. Fourier transfrom is used to transform between these two domains. The underlying principle is that any signal can be presented as a combination of sinusoidals of different

frequencies [17]. The coefficients for these are calculated with the Fourier transform for continuous signals or its discrete version, i.e.,

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-i2\pi kn/N}, \tag{4}$$

for frequency and time indexes k and n respectively. The time domain signal $x(n)$ is sampled at N points and the coefficients $X(k)$ are calculated for frequency indexes $k = 0, 1, 2, \ldots, N-1$ [17]. In practice, the used method is the fast Fourier transform that is much less computationally complex than the discrete Fourier transform in (4) [17]. To combine the information in time and frequency domains it is possible to use the time-frequency representations that can be obtained with for example the short-time Fourier transform [18]. In short-time Fourier transform the signal is assumed to be slowly varying so that it can be assumed stationary and ergodic within a time window [17]. For speech and audio signals the length of the time window is typically around 5-20ms [17]. These time frames are transformed into the frequency domain and the result can be presented as a spectrogram, which shows the signal power in each time-frequency instant. The time-frequency resolution of the short-time Fourier transform domain is dependent on the fast Fourier transform length [17]. If high frequency resolution is desired then a longer time frame is needed and correspondingly the frequency content can not be determined accurately if only a short time instant is examined [18]. This means that there is always a trade-off between time and frequency resolution.

Time-frequency representation can also be achieved with filterbanks where the signal is divided to subbands with a specific number of adjacent bandpass filters. An example of filterbank with perfect reconstruction property is the Quadrature Mirror Filter Bank (QMF) [17]. Perfect reconstruction means that the original signal can be achieved with the inverse transform without any errors [17]. This assumes that no processing is performed in the subbands.

2.3 Single-wave sound field model

Often in acoustic research the sound sources are assumed to be omnidirectional point sources. This means that the sound pressure caused by the source is equal in every direction. Therefore, the sound field caused by a point source is composed of spherical waves [11]. Another form of waves are plane waves where the wavefronts are planes [14]. Both of these are ideal versions of the real wavefronts, but these models can be assumed in some situations. Plane wave assumption means that locally in the far field the pressure of a curved wavefront is constant in all directions normal to the propagation direction [14]. This assumption is often used in microphone array processing. In practise sound fields can also be represented as a superposition of plane waves [19]. The plane wave assumption can be made with distances from the source that are large compared to the array size and wavelength of the signal, making the assumption frequency dependent [14]. According to literature [14] the

assumption can be used at distance r from the source when

$$r \gg \lambda/2\pi. \tag{5}$$

In the following, the sound field model, which is used throughout the thesis in the time-frequency domain, is presented. This model is commonly used in recent studies related to this topic, e.g., in [20]. The sound field is captured with M omnidirectional microphones at points $\mathbf{r}_{1\ldots M}$ in space. It is assumed that in the sound field there is only a single plane wave $P_{\mathrm{s}}(k, n, \mathbf{r})$ at each time-frequency bin. The assumption of a single plane wave is valid even in multi-source scenarios, when the source signals are sparse in time-frequency domain. This can be assumed, e.g., for speech signals [21]. In addition to the single plane wave assumption, also diffuse sound should be considered in the model. In an ideal diffuse sound field the instantaneous sound energy flow is uniformly distributed in all directions and the sound energy is equal in all points in space, i.e., the field is isotropic and homogenous [22] [23]. The total sound field becomes

$$P(k, n, \mathbf{r}) = P_{\mathrm{s}}(k, n, \mathbf{r}) + P_{\mathrm{d}}(k, n, \mathbf{r}), \tag{6}$$

which is a superposition of the direct sound component $P_{\mathrm{s}}(k, n, \mathbf{r})$ caused by a plane wave and the diffuse component $P_{\mathrm{d}}(k, n, \mathbf{r})$. The M microphone signals can be presented as

$$\mathbf{x}(k, n) = \mathbf{x}_{\mathrm{s}}(k, n) + \mathbf{x}_{\mathrm{d}}(k, n) + \mathbf{x}_{\mathrm{n}}(k, n), \tag{7}$$

in which the M microphone signals caused by the plane wave are denoted with $\mathbf{x}_{\mathrm{s}}(k, n) = [X_{\mathrm{s},1}(k, n), ..., X_{\mathrm{s},M}(k, n)]^{\mathrm{T}}$ and respectively for the diffuse part $\mathbf{x}_{\mathrm{d}}(k, n)$ and additive noise $\mathbf{x}_{\mathrm{n}}(k, n)$. Using the plane wave model, the plane wave at microphone m can be presented as

$$P_{\mathrm{s},m}(k, n, \mathbf{r}) = \sqrt{\Phi_{\mathrm{s}}(k, n)} a(k, \mathbf{n}, \mathbf{r}_m) e^{i\phi_{\mathrm{s}}(k,n)}, \tag{8}$$

in which $\Phi_{\mathrm{s}}(k, n)$ is the power of the wave, ϕ_{s} the phase at the origin of the coordinate system and i is the imaginary unit, i.e., $i = \sqrt{-1}$. The phase at point $\mathbf{r}$ is defined with $a(k, \mathbf{n}, \mathbf{r})$, which describes the phase shift of the plane wave along $\mathbf{r}$ as

$$a(k, \mathbf{n}, \mathbf{r}) = e^{i\kappa(k)\mathbf{r}^{\mathrm{T}}\mathbf{n}(k,n)}, \tag{9}$$

where the $\kappa(k) = \frac{2\pi}{\lambda(k)}$ is the wavenumber of the corresponding frequency. The vector $\mathbf{n}(k, n)$ describes the direction of arrival of the wave with azimuth $\varphi(k, n)$ and elevation $\delta(k, n)$ angles, which are illustrated in Fig. 2. The $\mathbf{n}(k, n)$ is described as

$$\mathbf{n}(k, n) = \begin{bmatrix} \cos(\varphi)\cos(\delta) \\ \sin(\varphi)\cos(\delta) \\ \sin(\delta) \end{bmatrix}. \tag{10}$$

In the case of 2-dimensional coordinate system the $\mathbf{n}(k, n)$ is described as

$$\mathbf{n}(k,n) = \begin{bmatrix} \cos(\varphi) \\ \sin(\varphi) \end{bmatrix}. \tag{11}$$

Based on (8) and (9), the direct sound vector can be obtained as

$$X_{\mathrm{s},m}(k,n) = a(k,\mathbf{n})P_{\mathrm{s}}(k,n,\mathbf{r}_1), \tag{12}$$

where $a(k,\mathbf{n})$ is the transfer functions between the reference position $\mathbf{r}_1$ and other microphones as defined in (9). For the ideal diffuse sound field the expected power is given by

$$\Phi_{\mathrm{d}}(k,n) = \mathrm{E}\left\{|P_{\mathrm{d}}(k,n,\mathbf{r})|^2\right\}, \tag{13}$$

where $\mathrm{E}\{\cdot\}$ is the expectation operator that can be approximated as a windowing function or time averaging. Operator $|\cdot|$ is the absolute value.

The ratio of the powers of the direct and diffuse components is known as the signal-to-diffuse ratio (SDR), given by

$$\mathrm{SDR}(k,n) = \frac{\Psi_{\mathrm{s}}(k,n)}{\Psi_{\mathrm{d}}(k,n)}. \tag{14}$$

As described in [24] the diffuseness of the sound field can be defined with (14) as

$$\Psi(k,n) = \frac{1}{1+\mathrm{SDR}(k,n)} \tag{15}$$

which by definition $\in [0\ 1]$ when $\Psi(k,n) = 1$ indicates fully diffuse sound field and $\Psi(k,n) = 0$ only direct sound.

2.4 Human hearing

This section focuses on the relevant principles of human hearing that are applied in the field of spatial audio. The audible frequency range for a healty individual is 16-20000 Hz [25]. However, because of the structure of the inner ear, there are limitations to the temporal and spectral resolution. The sound arrives through the ear canal to the middle ear and on to the oval window where inner ear starts [26]. The inner ear has spiral formed cochlea inside which there are two liquid filled canals separated by basilar membrane. The frequency resolution is due to the resonances of basilar membrane inside the cochlea [27].

Different sections of the membrane can be seen as overlapping bandpass filters for frequency bands [26]. Because one filter allows a range of frequencies to pass, the ability to distinguish nearby frequencies, also known as frequency selectivity, is limited. The widths of the bandwidths are roughly proportional to the center frequency and are known as critical bands [25]. Within one band the strongest tone is dominating the auditory perception but also exciting adjacent frequencies. This

causes frequency masking, i.e., the nearby frequencies with lower loudness are perceptually attenuated or even undistiquishable [26]. The masking effect takes place also in the time domain. Sounds arriving before or after a louder sound can be masked. These are known as backward and forward masking respectively [26].

Also other phenomena like short breaks or modulation in stimuli over time are related in the time domain acuity of hearing, known as the temporal resolution [26]. Informatic signals like speech and music deliver the information often in the changes of the signal. The hearing system is not able to distinguish these changes if they happen too rapidly, say, in the magnitude of microseconds [28]. If frequency resolution can be modeled with bandpass filters then temporal resolution can be presented as a lowpass filter. Instantaneous changes are attenuated and slower ones are passed and perceived. Temporal resolution is not an unambiguous time frame but it depends on the signal properties like bandwidth. In the field of spatial hearing the precedence effect is an important temporal aspect [29]. It explains the shift in localization when two nonperiodic and coherent signals are arriving from different directions and either one has a small lag. The direction delivering the first sound is dominating the localization. Due to the effect the two sound events are perceived as one auditory event with a shifted location.

2.4.1 Spatial hearing

The human hearing is highly developed and accurate system for sensing the environment. One of the most important functions of it is the sound source localization i.e. the ability to determine the sound direction of arrival and even distance to the source (to some extent) [28]. In literature the source location is described with three variables. The changes in lateral direction are measured with the azimuth angle φ, in vertical direction with elevation δ and in distance with r. The coordinate system is presented in Fig. 2.

Normally functioning hearing utilizes the binaural signals and can determine the direction of arrival from the differences in the signals between the ears. The most important cue for the localization is the phase difference between the ears. This is more commonly referred as the interaural time difference which is interpreted by the brain from the phase differences [28]. The difference comes directly from the path length inequality from the sound source to the ear canals. The time shifts can be from the carrier or envelope waveforms of the sound. These two function well in lower and higher frequencies respectively, when it is assumed that the sound is not a pure sinusoidal signal and there is an envelope [28]. The interaural time difference cue works well in the lateral direction but lacks performance in elevation, because the shortest delay is the one in lateral path difference [28]. In addition to time there is also a difference in sound level between the ears, the interaural level difference [28]. Because the low frequencies are more easily diffracted around the head, the interaural level difference cues are more important in higher frequencies, which have

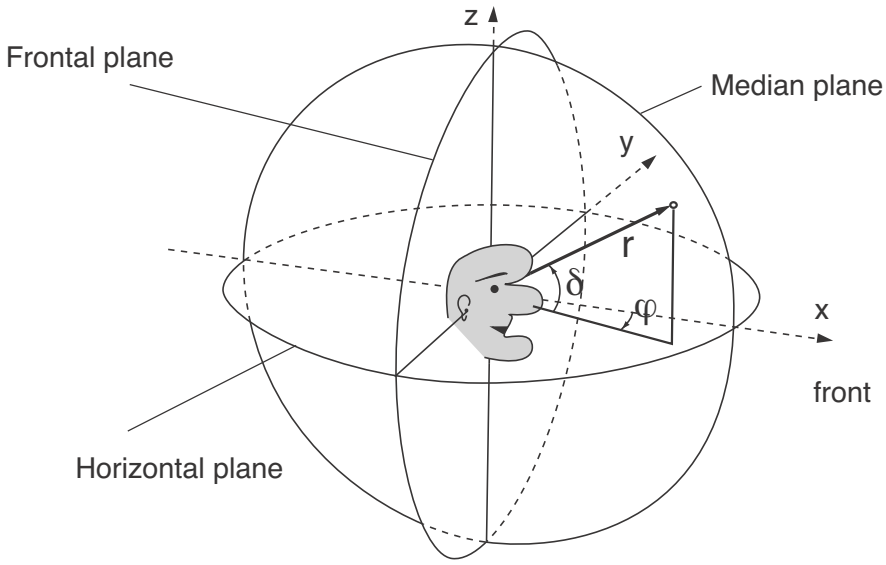

Figure 2: Coordinate system used in auditory research. φ is the azimuth, r is the distance and δ elevation. Figure adopted from [30].

higher attenuation from side to side [26]. Due to the phase ambiguity for sinusoidal signals, starting from around 800Hz, the interaural time difference is more important than interaural level difference in lower than higher frequencies and thus the interaural time and level difference cues can complement each other. However, both of these cues are missing proper elevation interpretation which is especially true in the median plane where there are no path differences and thus no interaural time or level difference. Therefore, more cues are needed for an accurate localization ability in every direction.

On both sides of the head, around the y-axis in Fig. 2, there are cone-shaped areas where the path differences are equal even with varying azimuth and elevation (when discarding the effect of the pinnae and unsymmetry of the head). These areas are called the cones of confusion [28]. If the sound source is inside this cone there are ambiguities in the direction of arrival. This problem can be solved by movements of the head which provides small changes in cues [26]. Also the shape of pinnae, head and upper body filter the spectral shape of the sound differently for different directions of arrival, which provides cues for the localization. Commonly this effect is known as head-related transfer function [28].

The similarity of the signals in the ears is described with the interaural coherence that is used as a cue for the diffuseness of the sound field. It can be treated as a separate cue but it is not clear whether if it is just perceived because of its effect on the other cues that were discussed above [30]. With a single source without any reflections the interaural coherence is high since the same signal is received in both ears. Instead, when there are many sources and a lot of reflections the coherence between the ears is low. The interaural coherence can also be seen as a measure of

the signal-to-diffuse ratio in (14), e.g., when the direct sound dominates the sound field, the signal-to-diffuse ratio is high, like shown in (14), and so is the coherence.

2.5 Spatial sound capturing and reproduction

In Sec. 2.4, the human ability to analyze the sound field was discussed. Nowadays, there are many applications in which more pleasant and immersive user experience can be provided when spatial audio is used, e.g., in teleconferencing, virtual reality and home theaters. To use spatial audio in these applications, there is a need for methods to analyse (capture) and synthesize (reproduce) the sound field. In the following, the sound field analysis and synthesis are discussed with the main focus on the analysis.

2.5.1 Sound field capturing

Acoustic signals can be captured with microphones of different types which transform the pressure signal into electrical signals. One of the main characteristics for a microphone is the directivity which is described with the polar pattern. Directivity can be for example described as omnidirectional or dipole but ideal monopole patterns are not available in practice [31]. Directivity is in practice proportional to the frequency [31].

Analysing the sound field means that the spatial parameters, i.e., the direction of arrival and the diffuseness, are captured. A common way is to use the B-format signals, proposed in [3], which consist of one omnidirectional and three orthogonal figure-of-eight signals that reveal the pressure gradient in x-, y- and z-directions. The omnidirectional signal $B_w(k,n)$ is used as the pressure signal and the other signals $B_x(k,n)$, $B_y(k,n)$, $B_z(k,n)$ are used to compute the corresponding components of the particle velocity. Usually, the B-format signals can be obtained for example from tetrahedron-shaped set of cardioid or subcardioid capsules, known as the sound field microphone [4]. These signals are referred to as the A-format and can be transformed to the B-format. The sound field microphone is relatively expensive and difficult to integrate for example in mobile phones due to size constraints. For these reasons, a more practical microphone setup like a planar array or an uniform linear array, presented in Fig. 3, can be applied in spatial sound capturing [32] [33]. Generally, the microphone array should cover as many dimensions as is used in the sound field analysis, e.g., when using an uniform linear array, the direction of arrival can be determined between [0 180] degrees, but there is an ambiguity on which side of the array axis the source is located. When using a planar array the problem is that it is not known whether the source is above or beyond the array. Arrays like these with a reasonably low number of cheap omnidirectional microhones, where the microphone distances are within the range from few centimeters to about 20 cm, are the main focus in this thesis. The number of microphone pairs in an array is denoted as H.

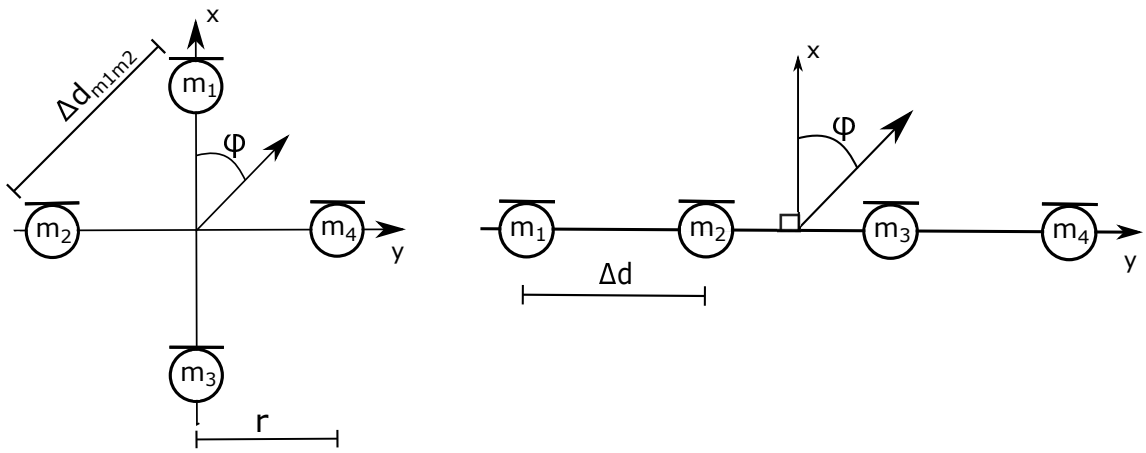

Figure 3: Examples of a planar microphone array (left) and a uniform linear array (right).

2.5.2 Directional Audio Coding

During the history of spatial audio different methods for capturing and reproduction of spatial sound have been proposed. These include for example stereo microphone techniques [34], Ambisonics [35] and Multichannel Audio Coding [36]. Many of these methods are not capable of reproducing the spatial image accurately or without distortions. For example the first-order ambisonics suffer from too high coherence between the loudspeaker channels [37]. In addition, these methods often lack the ability to be used with arbitrary loudspeaker setups. Directional Audio Coding [2] is an approach that utilizes information about the functioning of spatial hearing. This way it is possible to reproduce spatial image that is perceptually equivalent to the original sound field. In addition, the synthesis can be made for arbitrary loudspeaker setups.

The Directional Audio Coding method and some of its applications are presented in [2]. In the analysis phase the microphone signals are analysed in time and frequency steps that correspond to human hearing resolution. In the analysis phase two parameters of the sound field for each time-frequency bin are extracted: direction of arrival and diffuseness. The direction of arrival vector consists of azimuth and elevation angles as defined in (10) and depicted in Fig. 2. Diffuseness describes the proportion of sound that is coming from random directions and its connection to the signal-to-diffuse ratio is seen in (15). The direction of arrival corresponds to the interaural time/level difference and monaural localization cues whereas diffuseness corresponds to the interaural coherence. The fundamental idea is that by capturing these parameters together with the sound spectrum, it is possible to reproduce a spatial image that is perceptually equivalent to the original spatial image [38]. This is performed with temporal and spectral resolution that is as close as possible to the human hearing capabilities. It can be achieved with a filterbank composed of multiple narrow-band filters that match to human frequency resolution. With this filterbank also the temporal resolution can be determined for each

subband separately [2]. This way the best quality can be achieved, with the cost of higher complexity compared to the short-time Fourier transform approach [39]. In short-time Fourier transform the time and frequency resolutions are constant and defined by the fast Fourier transform length. Examples of Directional Audio Coding applications are teleconferencing [40] and stereo upmixing [1]. In the following, the processing pipeline is briefly explained after which the parameter estimation in short-time Fourier transform domain with the B-format input is described.

The processing flow diagram is presented in Fig. 4. The Directional Audio Coding assumes a single wave sound field model that was presented in Sec. 2.3. First, the microphone signals are transformed to time-frequency domain with short-time Fourier transform, resulting in the microphone signals $\mathbf{x}(k, n)$ in (7). Then the direction of arrival and diffuseness parameters, defined in (10) and (15) respectively, are estimated. This step will be discussed later in more detail. The obtained parameters are the metadata that are stored and/or transferred with the microphone signal(s) $\mathbf{x}(k, n)$ for the reproduction phase. The mono omnidirectional signal, like the $B_w(k, n)$ in B-format input, is the minimum requirement, but more can also be used in high quality applications. Based on the diffuseness $\Psi(k, n)$ values the transmitted audio is divided in two streams, one for direct sound and the other for diffuse, and these will be reproduced differently in the synthesis stage. This division of streams produces the direct and diffuse sound field components in (6), $P_{\mathrm{s}}(k, n, \mathbf{r})$ and $P_{\mathrm{d}}(k, n, \mathbf{r})$, respectively. These components compose the final output spectra $S(k, n)$. To reduce coherence between loudspeakers the diffuse signal $P_{\mathrm{d}}(k, n, \mathbf{r})$ is decorrelated before it is played by all the loudspeaker channels. The direct sound $P_{\mathrm{s}}(k, n, \mathbf{r})$ is reproduced as if it was coming from a point source. For this purpose the vector base amplitude panning method is a good solution for arbitrary loudspeaker configurations [41].

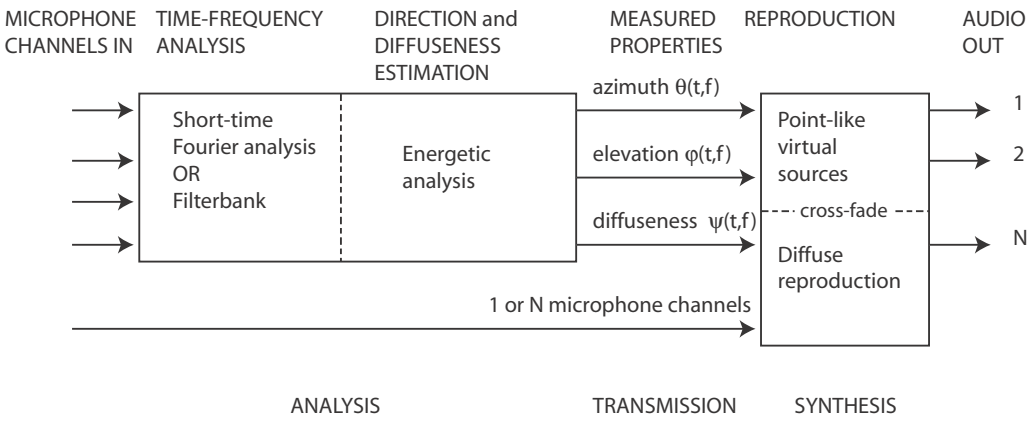

Figure 4: Simplified Directional Audio Coding flow diagram showing the input, analysis, transmission, synthesis and output phases. Figure adapted from [2]. Note that the notations here differ compared to the ones used in this thesis.

2.5.3 Parameter estimation with B-format input

The parameter estimation in Directional Audio Coding by using the B-format is presented in [37]. The analysis here is done using the single wave sound field model in Sec. 2.3 for a single location $\mathbf{r}$ so it is omitted in the notations. The omnidirectional signal $B_w(k,n)$ is an estimate of the total sound field pressure $P(k,n)$ in (6) as

$$B_w(k,n) = P(k,n). \tag{16}$$

With the plane wave assumption the particle velocity can be computed as

$$\mathbf{u}(k,n) = \frac{1}{\rho_0 c\sqrt{2}} \begin{bmatrix} B_x(k,n) \\ B_y(k,n) \\ B_z(k,n) \end{bmatrix}, \tag{17}$$

in which ρ_0 is the mean density of air. In short-time Fourier transform domain the intensity equation in (3) can be expressed as [42, p.24]

$$\mathbf{i}(k,n) = \mathrm{Re}[B_w^*(k,n)\mathbf{u}(k,n)]. \tag{18}$$

As the intensity points to the direction of energy flow, then the direction of arrival vector $\mathbf{n}(k,n)$ for the plane wave in (10) can be obtained from the negative intensity vector as

$$\mathbf{n}(k,n) = -\mathbf{i}(k,n). \tag{19}$$

The estimate of energy density $E(k,n)$ can be computed as

$$E(k,n) = \frac{\rho_0}{2}||\mathbf{u}(k,n)||^2 + \frac{|B_w(k,n)|^2}{2\rho_0 c^2}, \tag{20}$$

where $||\cdot||$ is the vector norm operator. With $\mathbf{i}(k,n)$ and $E(k,n)$, the diffuseness can be estimated as

$$\widehat{\Psi}(k,n) = 1 - \frac{||\mathrm{E}\left\{\mathbf{i}(k,n)\right\}||}{c\mathrm{E}\left\{E(k,n)\right\}}. \tag{21}$$

The diffuseness can also be estimated by using the statistics of the intensity vectors over time with a method called coefficient of variation [43]. The idea is that in diffuse field the length of averaged intensity vectors tends to zero because the direction varies constantly. In contrast, for a single plane wave the direction is constant so the length of averaged vector tends to a finite value and equals to the average of the vector length. Thus, the diffuseness can be estimated as

$$\widehat{\Psi}(k,n) = \sqrt{1 - \frac{||\mathrm{E}\{\mathbf{i}(k,n)\}||}{\mathrm{E}\{||\mathbf{i}(k,n)||\}}}. \tag{22}$$

This has advantages over (21), especially when using other microphone arrays than the B-format. In the next chapter the direction of arrival estimation for other microphone setups is presented.

3 State of the art direction of arrival estimation

The spatial parameters in Directional Audio Coding processing are the direction of arrival vector $\mathbf{n}(k,n)$ and diffuseness $\mathbf{\Psi}(k,n)$, defined in (10) and (15), respectively. The B-format input for the state of the art Directional Audio Coding processing, discussed in Sec. 2.5.3, is not always available or a practical option. In these cases also other estimators can be used that can be applied to more practical microphone arrays, for example [44] and [45]. In this chapter some state of the art direction of arrival estimators for arrays of omnidirectional microphones are presented. First, broadband estimation methods based on the time difference of arrival for a microphone pair is presented. Also some well known narrowband methods are explained as they are the main interest in time-frequency domain processing. In this thesis the direction of arrival estimation is limited to the azimuth plane and so only the azimuth $\varphi(k,n)$ will be considered in the direction of arrival from now on.

3.1 Time difference of arrival

A straight-forward method for direction of arrival estimation is to utilize the time difference of arrival value, also known as the delay τ, between microphones [46]. The scenario is presented in Fig. 5, where a plane wave arrives at the array. The basic idea is to take a set of samples from one sensor and observe the signal from the other microphone for the same part of the signal. One of the most fundamental ways to identify this similarity in signals is to use cross-correlation [47] [48]. The cross-correlation function is the average product of two signals when either one is being time shifted. The function can be presented for discrete time domain signals $x_1(n)$ and $x_2(n)$ as

$$R_{\mathrm{CC}}(\tau) = \mathrm{E}\{x_1(n)x_2(n-\tau)\}, \tag{23}$$

with the time shift $\tau \in [-\tau_{\max}, \tau_{\max}]$. The maximum delay $\tau_{\max}$ is the propagation time of sound for distance Δd between the microphones. By finding the maximum of (23) as

$$\hat{\tau}_{\mathrm{CC}} = \underset{\tau}{\mathrm{argmax}}(R_{\mathrm{CC}}(\tau)), \tag{24}$$

the delay between the microphones is found. Once the delay τ between the microphones has been estimated, the direction of arrival can be calculated as [46]

$$\hat{\varphi} = \arccos\frac{c\tau}{\Delta d}. \tag{25}$$

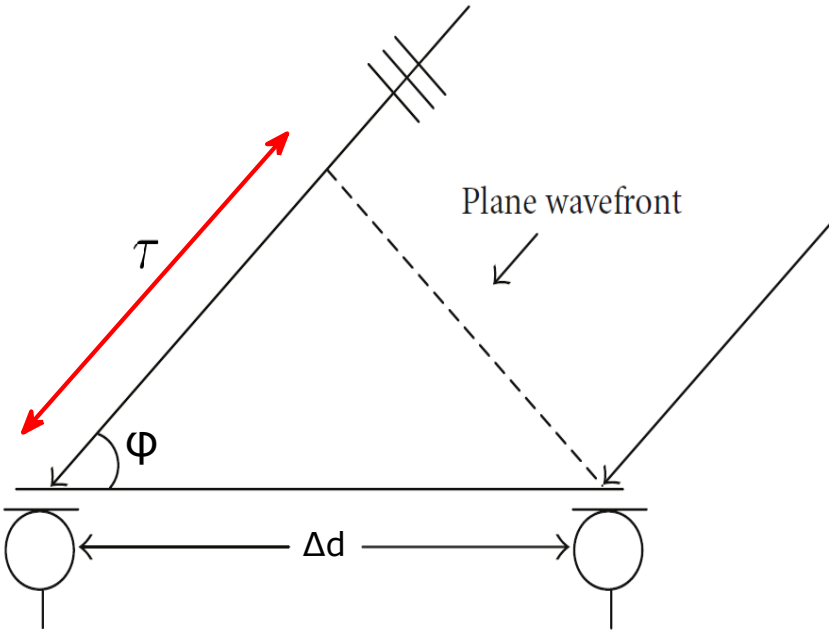

Figure 5: A plane wave arriving at a microphone pair. The delay τ between the sensors can be used to estimate the φ. Figure adapted from [46].

The problem of this method is that in reality even if $x_2(n)$ is shifted with $\hat{\tau}_{\mathrm{CC}}$ the signals are not the same. This is because they are also attenuated and spectrally distorted due to for example reverberation [46]. In a reverberant room the signal can be reflected multiple times before arriving at the sensors and this multipath propagation should be considered. In real life applications also ambient noise and moving sources increase the possibility for error [46]. The delay can also be found out by comparing the magnitude differences of the two signals with time shifts. The shift that produces minimal differences indicates the actual delay. This average magnitude difference function and other delay estimation techniques are presented in [49] and [50].

The Generalized Cross-Correlation is a correlation-based method for time difference of arrival estimation in frequency domain [51]. It allows the use of a priori knowledge through weighting function $\Lambda(k)$ that can be seen as a prefilter. The Generalized Cross-Correlation is formulated as

$$R_{\mathrm{GCC}}(\tau) = \sum_{k=0}^{N-1} \Lambda(k) \chi_{x_1 x_2}(k) e^{j2\pi\tau k/N}, \tag{26}$$

where

$$\chi_{x_1 x_2}(k) = \mathrm{E}\big\{X_1(k) X_2^*(k)\big\}, \tag{27}$$

is the cross-spectrum of the signals. The frequency domain is useful because the convolution of the weighting function and cross-spectrum is transformed into multiplication, making it more efficient [18]. The delay is found by replacing $R_{\mathrm{CC}}(\tau)$ by $R_{\mathrm{GCC}}(\tau)$ in (27).

If unit weights are used in $\Lambda(k)$, the Generalized Cross-Correlation corresponds to the classical cross-correlation. In addition, the weighting can be frequency dependent and give more emphasis on bands with higher signal-to-noise ratio (SNR) value. Proposed methods for weighting are for example Phase Transform (PHAT) [52] [51], maximum likelihood processor [51], Roth processor [53] or combination of these [54]. The PHAT algorithm is one of the most popular due to its performance in reverberant conditions and consistensy in performance with alternating source signal. Methods based on the time difference of arrival are popular in broadband estimation. In the following, some narrowband direction of arrival estimators are presented, as the Directional Audio Coding processing requires the narrowband estimates.

3.2 Weighted least squares

Least squares is an optimization method that provides an estimate of the parameter(s) that produced the observed values by minimizing the sum of squared errors between the estimate and the observations [55]. This can be applied in direction of arrival estimation and an example of a Weighted Least Squares (WLS) estimator is presented in [6]. In this method the direction of arrival is estimated based on the interchannel phase differences of the microphone pairs in the array. The problem is approached with the signal model presented in Sec. 2.3 but with only direct sound and additive noise. The interchannel phase difference information is contained in the microphone signals $\mathbf{x}(k,n)$ in (7). The power spectral density (PSD) matrix of the signals is

$$\boldsymbol{\Upsilon}_{\mathrm{x}}(k,n) = \mathrm{E}\big\{\mathbf{x}(k,n)\mathbf{x}^{\mathrm{H}}(k,n)\big\}, \tag{28}$$

where the $(\cdot)^{\mathrm{H}}$ is the Hermitian operator. Like the model in (7), also the (28) can be separated to the plane wave and noise parts, i.e., to power spectral density matrices $\boldsymbol{\Upsilon}_{\mathrm{s}}(k,n)$ and $\boldsymbol{\Upsilon}_{\mathrm{n}}(k,n)$, respectively. The cross power spectral density of the plane waves between microphones m and m' can be computed as

$$\Upsilon_{\mathrm{s},m'm}(k,n) = \mathrm{E}\big\{X_{\mathrm{s},m'}(k,n)X^{*}_{\mathrm{s},m}(k,n)\big\}, \tag{29}$$

and it can be represented with the phase shift in (9) as

$$\Upsilon_{\mathrm{s},m'm}(k,n) = \Psi_{\mathrm{s}}(k,n)a(k,\mathbf{n},\mathbf{r}_{mm'}). \tag{30}$$

The noise is assumed independent and identically distributed, i.e., it does not affect the phase shift between microphones. With this assumption the off-diagonal elements of $\boldsymbol{\Upsilon}_{\mathrm{s}}(k,n)$ and $\boldsymbol{\Upsilon}_{x}(k,n)$ are the same and can be used to compute the interchannel phase difference μ_{s} between the points $\mathbf{r}_{m'}$ and $\mathbf{r}_{m}$. The relevant values, i.e., corresponding to all pairs $m'm$ but not mm', from (28) are collected to a vector

$$\boldsymbol{\phi}_x(k,n) = [\Upsilon_{x,12}, \ldots, \Upsilon_{x,zj}, \ldots, \Upsilon_{x,ZM}]^{\mathrm{T}}, \tag{31}$$

where $Z = M - 1$, $1 \leq z \leq Z$ and $z \leq j \leq M$. The expectation operator is approximated with time averaging, resulting in the estimate $\hat{\boldsymbol{\phi}}_x(k,n)$ which is connected to the interchannel phase difference as

$$\hat{\boldsymbol{\mu}}_{\mathrm{s}} = \angle\hat{\boldsymbol{\phi}}_x(k,n), \tag{32}$$

i.e., the phase of the observed power spectral density values. The connection between the direction of arrival vector $\mathbf{n}(k,n)$ in (11) and the observed interchannel phase difference can be written as

$$\hat{\boldsymbol{\mu}}_{\mathrm{s}}(k,n) = \mathbf{Q}(k)\mathbf{n}(k,n) + \triangle(k,n). \tag{33}$$

The $\triangle(k,n)$ is used as an estimate of the errors in the interchannel phase difference estimation, which depend on the signal-to-noise ratio, f, microphone spacings and the direction of arrival. The $\mathbf{Q}(k)$ describes the phase shifts like in (9) by using the wavenumber $\kappa(k)$ and interchannel vectors $\mathbf{r}_{m'm}$ as

$$\mathbf{Q}(k) = \kappa(k)[\mathbf{r}_{12}, \ldots, \mathbf{r}_{zj}, \ldots, \mathbf{r}_{ZM}]^{\mathrm{T}}. \tag{34}$$

The errors are minimized by using a least squares solution as

$$\hat{\mathbf{n}}(k,n) = [\mathbf{Q}(k)^{\mathrm{T}}\mathbf{W}(k,n)\mathbf{Q}(k)]^{-1}\mathbf{Q}(k)^{\mathrm{T}}\mathbf{W}(k,n)\hat{\boldsymbol{\mu}}_s(k,n) \tag{35}$$

The solution assumes that when estimating the $\mathbf{n}(k,n)$, the matrix $\mathbf{Q}(k)^T\mathbf{W}(k,n)\mathbf{Q}(k)$ has full rank. This can be quaranteed if the array covers D-dimensions when estimating the direction of arrival vector in D-dimensions. As not all the pairs have equal distances and thus the spatial aliasing frequencies, a weighting matrix $\mathbf{W}(k,n)$ is used. This way the array pairs are discarded as they reach their spatial aliasing limit, i.e., if $\mathbf{r}_{zj}$ is too large for a certain frequency the weights for that pair are set to zero. As was mentioned, the array needs to cover D dimensions so the discarding of the microphone pairs is only done if it does not risk this condition. Spatial aliasing will be discussed more in a later section and this estimator will be used there. In the following, a simulation model is explained so that some simulation results with this method can be presented.

3.2.1 Plane wave simulation using the single plane wave signal model

A plane wave simulation according to the signal model in Sec. 2.3 is made to present an example of estimation results with the Weighted Least Squares estimator. Two source signals were used, 5 seconds of male and female speech. Both contained the same sentence with a small time shift. The time-frequency domain signals $X_{\mathrm{s}1}(k,n)$ and $X_{\mathrm{s}2}(k,n)$ are obtained from the corresponding time domain signals via short-time Fourier transform with fast Fourier transform length of 1024 and 50% overlap using a sine window. Sampling frequency $f_{\mathrm{s}} = 48\mathrm{kHz}$. A microphone array like in

the Fig. 3 with $r = 0.3$cm is used. The r was chosen so that no spatial aliasing occurs in the observable frequency range.

Using (8) in the signal model, the short-time Fourier transformed signals are phase shifted corresponding to the array structure and directions of arrival. This way the $P_{\mathrm{s},m}(k, n, \mathbf{r})$ for both plane waves and for each microphone m are obtained. The direction of arrival vectors shown in (11), $\mathbf{n}_1 = [\cos(52)\ \sin(52)]^{\mathrm{T}}$ and $\mathbf{n}_2 = [\cos(-92)\ \sin(-92)]^{\mathrm{T}}$, are used to indicate the true source locations. After the phase shifts are added, the $P_{\mathrm{s1},m}(k, n, \mathbf{r})$ and $P_{\mathrm{s2},m}(k, n, \mathbf{r})$ are added together and also additive noise $\mathbf{x}_{\mathrm{n}}$ in (7) was added to correspond to 60dB SNR. The noise vector was created by generating gaussian white noise individually for each microphone. These microphone signals were used as an input to the Weighted Least Squares estimator, where an averaging is performed within 10 time frames of power spectral density values in (29).

The simulated results are presented in Fig. 6 b) where the estimated directions of arrival are color coded as a function of time and frequency. In Fig. 6 a) and c) the first and second source are shown above and beyond the plotted estimates respectively. This way it can be seen how the activity and power of the source affects the direction of arrival estimates. When either one of the sources is active, the corresponding estimated direction of arrival is appointed to the time-frequency bin. If neither of the sources are active, like above 20kHz, the estimates are random. It can be seen that as the first source has larger power, it is dominating the estimates. When the first source is inactive, e.g., right before 2 seconds mark, the second source direction of arrival is appointed for the bins. The light green areas, corresponding to $\varphi \approx 0$ in the estimates, appear when the powers of the sources are approximately the same. This double talk scene will be simulated multiple times in following chapters, e.g., in Sec. 4.2.

3.3 Subspace methods

Multiple Signal Classification [56] and Estimation of Signal Parameters via Rotational Invariance Techniques [5] are direction of arrival estimation methods that allow multiple source directions to be detected simultaneously. These methods provide high resolution when the source signals are uncorrelated. They are known as subspace methods because they are based on analyzing signal and noise subspaces. The Multiple Signal Classification is computationally very complex as it requires eigenvalue decomposition and the search over all directions of arrival. It is also sensitive to errors in the array modeling, i.e., the positions, gains and phases of the microphones [57]. Root Multiple Signal Classification was developed to decrease the complexity by using polynomial rooting but it can only be applied to an uniform linear array [58]. In [59] the algorithm was modified so that it could be used also with a nonuniform linear array. In addition to direction of arrival estimation, Multiple Signal Classification can be used also for other applications like estimating the number of sources [56].

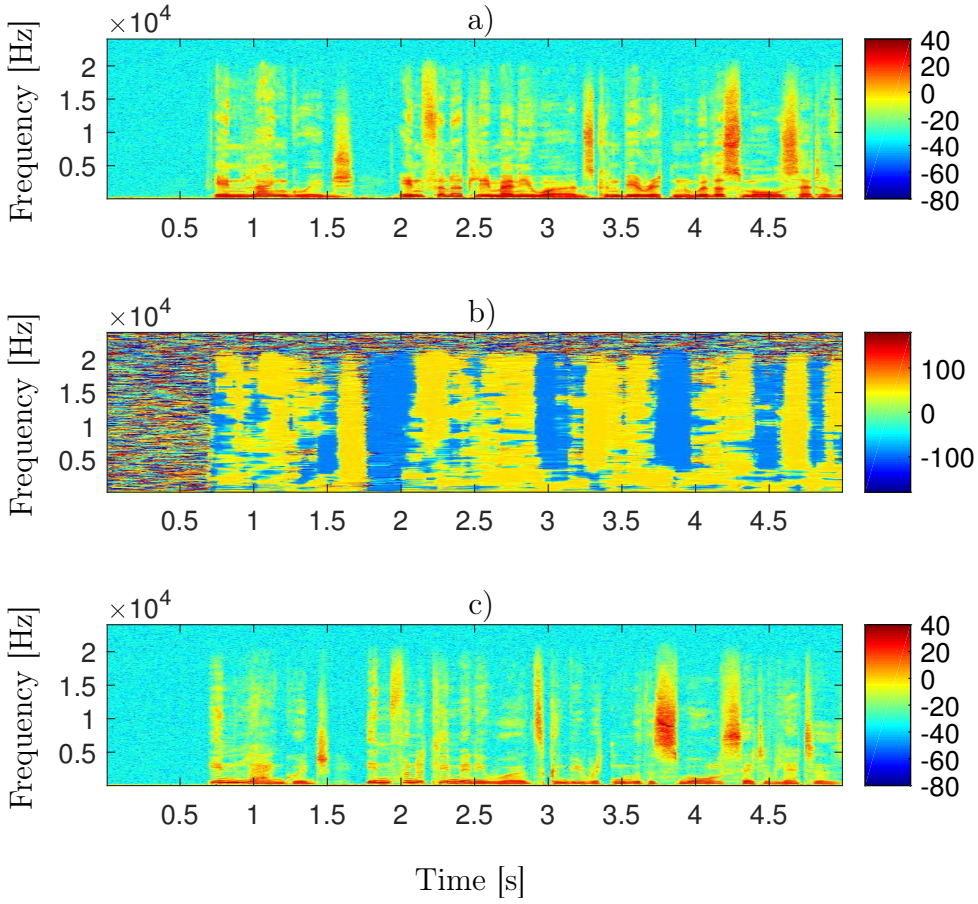

Figure 6: a) Spectrogram of the male speech on dB-scale, source located at $\varphi_1 = 52°$. b) Color coded direction of arrival estimates with the Weighted Least Squares method [6]. c) Spectrogram of the female speech on dB-scale, source located at $\varphi_2 = -92°$. When sources are active the directions of arrival are mostly estimated to either one of the correct directions of arrival.

The Estimation of Signal Parameters via Rotational Invariance Techniques works with similar eigenvalue decomposition approach and can overcome some of the problems in the Multiple Signal Classification. Mainly, the computational complexity is decreased because the search over all directions of arrival is not needed and it is also less sensitive to errors in the positions of the microphones and their gain/phase [5]. The main drawback is that the array is required to consist of two identical subarrays. The microphones in the first subarray are paired with a corresponding microphone in the second subarray. This way the interchannel phase differences between each pair are equal. In the original Estimation of Signal Parameters via Rotational Invariance Techniques publication [5] it was found to perform better than Multiple Signal Classification but also contrary results have been obtained [60].

4 Spatial aliasing

The state of the art direction of arrival estimators presented in the Sec. 3 can be used in different applications where omnidirectional microphones are used. In time-frequency processing, like Directional Audio Coding in Sec. 2.5.2, the estimator should be capable of processing the data in narrow bands, so that each frequency bin can be appointed with separate direction of arrival parameters. These estimators are usually based on the interchannel phase difference information, which provides reliable results when certain assumptions about the signal bandwidth are met. In low frequencies the phase difference is so small that the noise might dominate the samples, making estimation difficult. In higher frequencies a phenomenon, known as spatial aliasing, occurs and it is discussed in this section. The introduction to the spatial aliasing problem is started by explaining the temporal aliasing, which is a common problem in digital signal processing. Then, the connection between temporal and spatial aliasing is explained. After this, the spatial aliasing is explained in more detail and with examples. Finally, also some perceptual aspects of spatial aliasing are discussed. This section also includes a brief study on how decreasing the time-frequency resolution affects perceptually to the spatial image. This information could be used when designing aliasing free direction of arrival estimators.

When capturing any continuous time signal like the air pressure to digital form, it is sampled, i.e, its amplitude is captured at discrete time instants. This sampling is performed with a finite sampling frequency f_s. The Nyqist sampling theorem states that the maximum achievable frequency that can be reconstructed from the sampled values is equal to half of f_s [18]. Inversely, the minimum f_s needed to capture a signal with bandwidth β is

$$f_\mathrm{s} \geq 2\beta. \tag{36}$$

If this condition is not satisfied then the phenomena known as temporal aliasing occurs. In aliasing, frequencies higher than $f_\mathrm{s}/2$ are sampled identically to frequencies within the bandwidth [18]. When spatially sampling the sound field with a microphone array, the same principle applies. The pressure signal is sampled at discrete positions in space and only a certain spatial bandwidth can be sampled without aliasing [7]. The sampling problem is presented in Fig. 7 with three signals of different frequencies. The x-axis can be considered time, when the sampling points are separated by $1/f_\mathrm{s}$, or distance, in which case the sampling points describe the position of microphones in space. If the sampling is done at the points of the black markers, only the blue signal can be reconstructed and the other two will be temporally aliased, i.e., reconstructed as the blue one. Spatial aliasing causes the phase differences to appear as the same for each of the signals, which makes the use of interchannel phase difference problematic in direction of arrival estimation. From

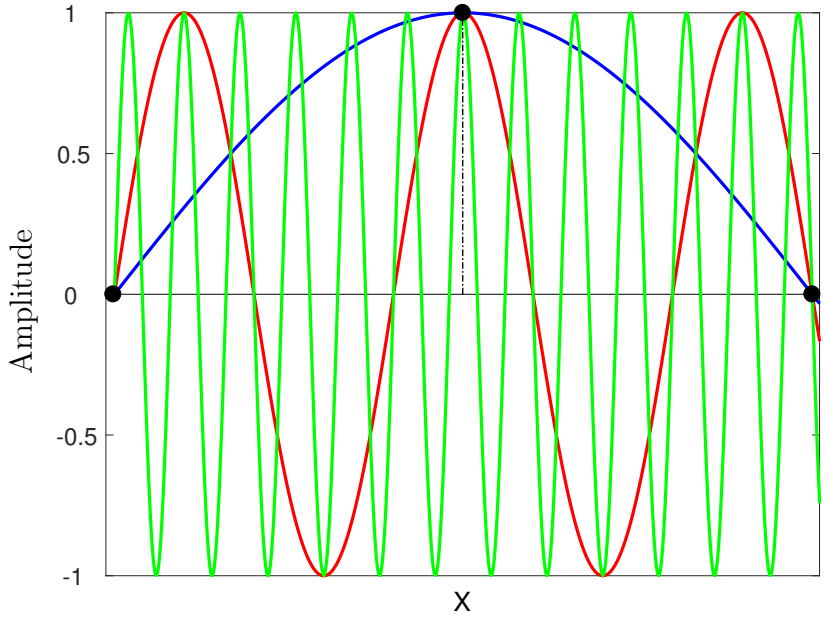

Figure 7: Three signals of different frequencies. From samples at the black markers, only the blue signal can be reconstructed and the other signals will be aliased. X can be either time or distance to illustrate temporal and spatial aliasing respectively.

now on the term aliasing will be solely used to describe the spatial aliasing.

4.1 Effects of spatial aliasing on parameter estimation

Spatial aliasing appears particularly in systems that apply interchannel phase difference in the direction of arrival estimation, such as [6], where aliasing causes wrong direction of arrival estimates. In addition, aliasing is also a problem in beamforming applications where aliasing causes the sidelobes to appear too large [7]. The focus on this work is on direction of arrival estimation but it should be mentioned that the diffuseness estimator based on intensity-energy ratio in (21) is affected by spatial aliasing [45]. The narrowband diffuseness estimation with (22) is not affected because it only considers the lengths of the intensity vectors and their variation but the direction does not matter [43].

The sampling interval of a microphone pair is determined by the microphone spacing Δd. Together with the speed of sound c and the incident angle φ (relative to the normal of the microphone pair), the aliasing frequency f_a for a microphone pair can be computed as [61]

$$f_\mathrm{a} = \frac{c}{2|\sin(\varphi)|\Delta d}. \tag{37}$$

As can be seen the lowest aliasing frequency can be found at $|\varphi| = 90°$. When $\varphi = 0$ there is no interchannel phase difference and also no aliasing from this direction. However, some other direction can indicate this interchannel phase difference when aliased, hence aliasing affects all directions. For a microphone pair or linear array the lowest aliasing frequency is found at [44]

$$f_{\mathrm{a}} = \frac{c}{2\Delta d}. \tag{38}$$

This means that the Δd must be at maximum half the size of the wavelength. Otherwise the interchannel phase difference information becomes ambiguous due to multiple wavelength periods between the microphones. For a planar array (like in Fig. 3) the f_{a} is found at [44]

$$f_{\mathrm{a}} = \sqrt{\frac{1}{2}}\frac{c}{\Delta d}. \tag{39}$$

The maximum microphone spacing preventing aliasing at a certain frequency f can be found as

$$\Delta d = \frac{c}{2f}. \tag{40}$$

For example to prevent aliasing in teleconferencing with frequency range of 3400Hz, the microphone spacing can maximally be around 5cm.

A noiseless plane wave simulation was made to estimate azimuth angles when aliasing is present. The simulation was carried out the same way as was described in Sec. 3.2.1. The difference to the earlier simulation is that here it was run for each direction of arrival with 1° steps while using a white noise source. One time frame of the result for each direction of arrival was added to the Fig. 8. A 4-mic array like in Fig. 3 with microphone distances $r = 9$cm was used. In this kind of microphone array the smallest microphone spacings are $\Delta d \approx 13$cm. The method uses weighting to prioritize smaller spacing microphone pairs, as was discussed in Sec. 3.2. This way the lowest aliasing frequency is set around 1.3kHz. As is seen in Fig. 8, in frequencies below the f_{a} the true φ in x-axis corresponds to the estimated $\hat{\varphi}$ that is coded in colors. In addition, for each frequency each estimate is only presented once in the x-axis, i.e, the estimates are unambiguous. Above the f_{a} the estimate is incorrect and they are ambiguous. For example, the red color indicating estimates above 150° appears multiple times for different true direction of arrival. By the way the estimates change, it can be seen that the aliasing is dependent on the frequency and the direction of arrival. The higher the frequency the more there are ambiguities. When looking at a single direction of arrival the aliased estimates may vary a lot over the frequency range. However they might also only change between two values, like happens at for example $\varphi = 0°$.

When estimating the interchannel phase difference from the power spectral density matrix in (29), only the interchannel phase difference $\hat{\mu}_{\mathrm{w}}$ wrapped between

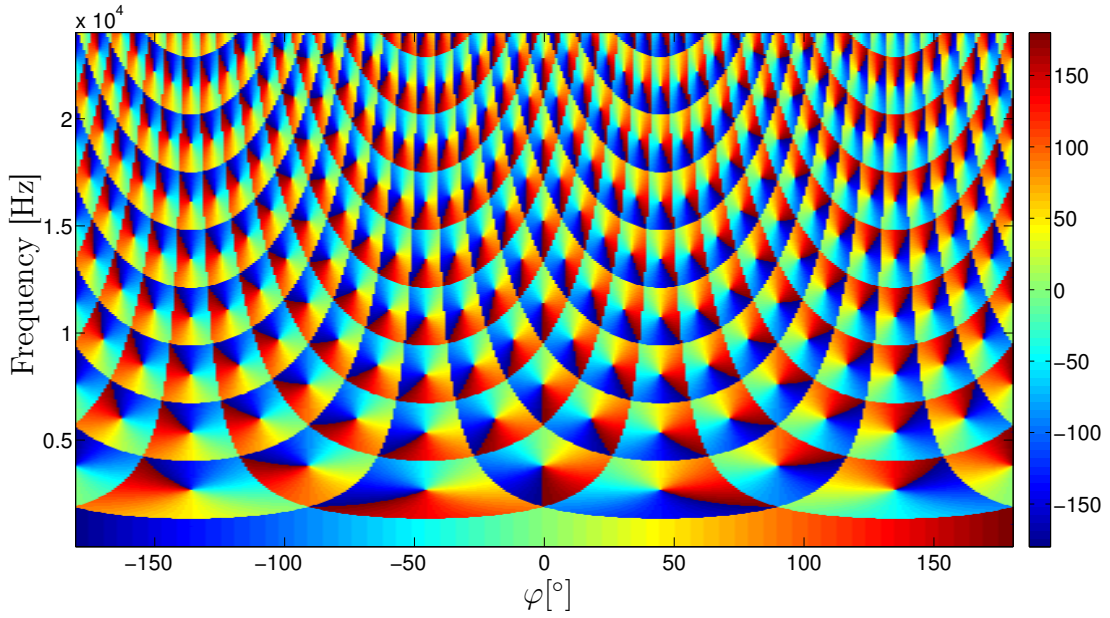

Figure 8: Aliasing effect with a 4 microphone planar array with a radius of 9cm. Directions of arrival are calculated for a single white noise source in a plane wave simulation using the Weighted Least Squares method [6]. Below $f_\mathrm{a} = 1.3$kHz the directions of arrival are correct but above the aliasing limit the ambiguities appear.

$[-\pi\ \pi]$, can be calculated [62]. When there are multiple wavelengths between the microphone positions, i.e., above the f_a, this wrapped phase difference can not be used in the direction of arrival estimation. This is because the interchannel phase difference does not anymore describe the correct delay between the microphones. For this, the unwrapped phase difference $\hat{\mu}_\mathrm{uw}$ is needed which is not bound between the $[-\pi\ \pi]$ [62]. The relation between the unwrapped and wrapped phase can be found as

$$\hat{\mu}_\mathrm{uw}(k,n) = \hat{\mu}_\mathrm{w}(k,n) + 2\pi l(k,n), \tag{41}$$

in which l is an integer value indicating how many periods of $\pm 2\pi$ are there between the sampling points. When the $\hat{\mu}_\mathrm{w}$ is used in the direction of arrival estimation to describe the phase shift presented in the signal model (9), it will lead to wrong estimates.

In Fig. 9 there are the interchannel phase differences as a function of frequency, calculated for a micophone pair when white noise sound arrives from three different directions ($\varphi = 0/10/90°$). There is a small amount of noise and the simulation procedure is same as was explained in Sec. 3.2.1. It can be seen how the values are in the range of $[-\pi\ \pi]$. At the f_a around 1.3kHz, the interchannel phase difference for $\varphi = 90°$ reaches the limit of this range and it is estimated as the wrapped value. The interchannel phase difference for $\varphi = 10°$ reaches the wrapping limit at higher frequency and the interchannel phase difference for $\varphi = 0°$ is never wrapped.

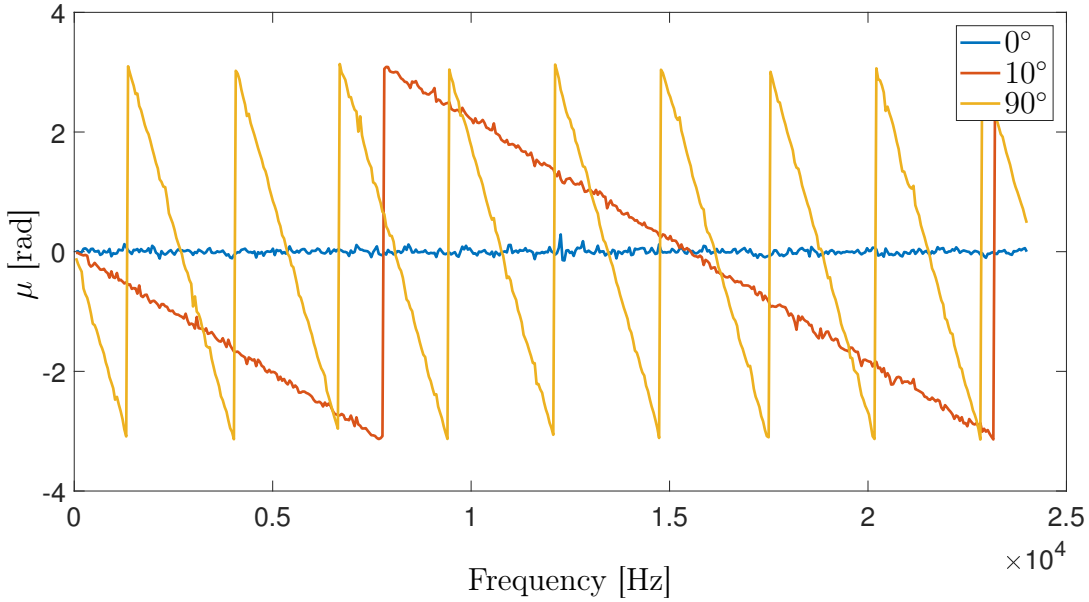

Figure 9: Observed phase differences with a 4 microphone planar array with $r = 9$cm and lowest f_a around 1.3kHz. Three different directions of arrival are shown here to show the dependency on the direction of arrival. Lowest f_a is with $\varphi = 90°$ and with $\varphi = 0°$ the interchannel phase difference is always zero, i.e., never wrapped.

The figure illustrates why the relation between the μ_uw and μ_w is not as simple as suggested in (41) because the unknown direction of arrival affects the aliasing frequency and the slope of the interchannel phase difference curve [61]. This is why there is not enough information to perform a trivial unwrapping like in [8]. In addition, noise and multiple sources can make the phase information difficult to unwrap [62].

4.2 Perceptual aspects of spatial aliasing

Spatial aliasing causes large deviations to the direction of arrival estimate because the aliased estimate might be located on a completely wrong direction, like is shown in Fig. 8. Unlike small deviations around the correct estimate, these errors are perceptually significant especially in spatial sound reproduction like Directional Audio Coding processing. As for human the smallest just noticeable difference in source localization is around 1° [28], it is obvious that spatial aliasing is severly reducing the perceived quality of spatial image. On the other hand, the acuity for localization is degraded at higher frequencies where the aliasing is a problem [26].

To test the degradation of spatial image quality due to aliased estimates, the double talk scene in Fig. 6 was simulated again with the same procedure as explained in Sec. 3.2.1. The estimation was done with an array radius $r = 9$cm to include spatial aliasing. The simulation result is presented in Fig. 10, where the direction of arrival estimates are coded in color and presented as a function of time and fre-

quency. Comparison with the non-aliased estimates in Fig. 6 shows that for most of the frequencies the direction of arrival estimates are aliased. When the sources are active, there should be only two sources at $\varphi_1 = 52°$ and $\varphi_2 = -92°$ indicated by yellow and blue respectively. Below the $f_a \approx 1.3$kHz the estimation result is the same as in the non-aliased version but in higher frequencies the estimates are mostly completely wrong. However, as there are still relevant frequencies localized correctly, the perceived spatial image is not as distorted as the figure would imply.

By using Directional Audio Coding synthesis, described briefly in Sec. 2.5.2, with vector base amplitude panning for 5.0 loudspeaker setup the perceptual effects can be examined. In informal listening test, the talkers were localized somewhat in the correct positions. The aliasing could be noticed especially when speech contains fricatives like /s/, this was noticed as a sudden, individual, sound in the wrong direction. This was audible for the fricatives because they contain a lot of energy in the higher frequencies [63]. When using Directional Audio Coding processing in multiple sources scene, the diffuseness values are higher when the sources overlap. For this reason, the wrong localization might not be perceived so easily as the diffuse stream distributes the sources in all channels. When comparing this to the non-aliased version the difference is clear in how strong the localization is, i.e., the spatial image sounds less diffuse. Because this is a strongly source dependent issue it is best not to assume that the aliasing might not be perceived. Due to this, solutions to overcome the aliasing are needed.

In parametric time-frequency domain processing the temporal and spectral resolutions for the reproduced sound and the estimated parameters are limited due to

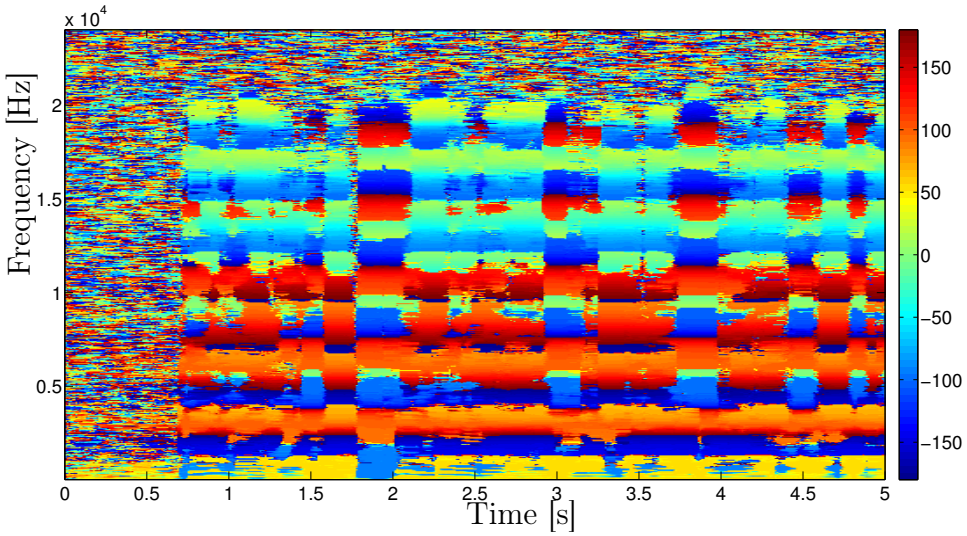

Figure 10: Same simulation as in Fig. 6 with array radius $r = 9$cm which causes aliasing to appear above the f_a. When the sources are active the estimates should be either coded with yellow or blue, but aliasing causes completely wrong estimates.

the tradeoff between temporal and spectral resolution explained in Sec. 2.2. Because Directional Audio Coding is perceptually motivated, the human hearing resolution, discussed in Sec. 2.4, should be considered when determing the actual required time-frequency accuracy for the direction of arrival parameter. In the following, the aim is to gather some insight on whether it is possible to use lower resolution by averaging time-frequency bins together to larger blocks. This could allow different, non-aliasing, direction of arrival estimators to be used. The quality degradation with decreased resolution is examined both with an objective and a subjective measure.

4.2.1 Objective quality degradation of time-frequency averaging

Log-spectral distortion is the difference between two spectra in dB-scale [64]. The value for each time frame n is calculated as a mean over the K frequency bins. The value is calculated as follows:

$$\mathrm{LSD}(n) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} |20\log_{10}\{|S(k,n)|\} - 20\log_{10}\{|\hat{S}(k,n)|\}|^2}, \tag{42}$$

where $S(k,n)$ is the reference spectra and $\hat{S}(k,n)$ is the distorted one. This was used to measure how much the source signal is distorted when the direction of arrival parameter is averaged for the Directional Audio Coding synthesis. The simulation was made as a room impulse response simulation. The simulation computes the impulse responses at the microphone location when the location of the impulse is at the source location. These impulse responses are then convolved with the source signals to obtain the signals with reflections and reverberation. The room size was $7.5 \times 6.8 \times 3.5$ [L $\times$ W $\times$ H] meters. The microphone position was at [3.3 3.5 1.6] and 1.6m from the source positions. The $T_{60} = 0.25$s and SNR was 60dB. The used fast Fourier transform length was 512 with 50% overlap with a sine window. To neglect the direction of arrival estimation errors, the direction of arrival parameters were computed by calculating a weighted average of the true directions of arrival for each time-frequency bin. As a weighting, the powers of each source were used. The diffuseness was computed with the coefficient of variation method in (22), where the intensity vector length was the sound power and direction the computed direction of arrival.

Three different scenes were used. The first scene had two talkers, a male and female at directions of arrival of $\varphi_{1,1/2} = 30/-135°$. The second scene had three sources, a male talker, a bass guitar and a barking dog at directions of arrival of $\varphi_{2,1/2/3} = 30/-135/135°$. The third scene had five music instruments located at $\varphi_{3,1/2/3/4/5} = 30/-30/0/135/-135°$. The Directional Audio Coding processing was performed for all scenes and the synthesis was made for a 5.0 loudspeaker setup.

The reference spectra for the log-spectral distortion calculation was in all scenes the source signal that was located at $\varphi = 30°$. The distorted spectra was the signal that was synthesized to the loudspeaker at the same location. The direction of

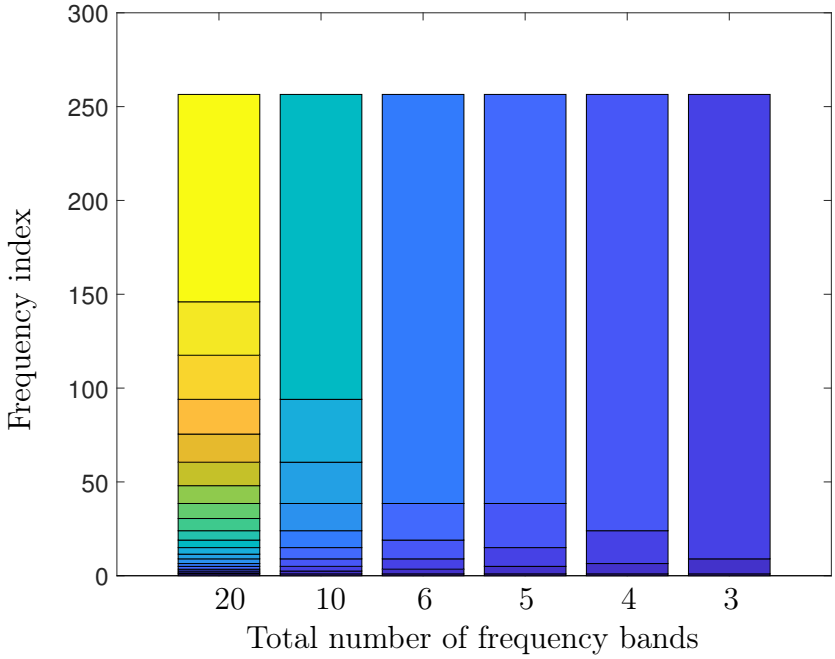

Figure 11: Frequency bands that were used for the direction of arrival averaging. The whole frequency scale was divided to 20, 10, 6, 5, 4 and 3 frequency bands which are visualized as bars covering the frequency bins.

arrival parameter resolution was decreased by averaging the estimated directions of arrival in 2, 4, 6, 8, 10 and 12 adjacent time frames of approximately 11ms. The averaging was done without a sliding window. In frequency scale the averaging was performed so that the whole frequency scale was divided to 20, 10, 6, 5, 4 and 3 frequency bands inside which the averaging was performed. The division of frequency bands is visualized in Fig. 11, which shows the frequency bins that were included in each averaging band. The averaging was a simple mean of the computed directions of arrival.

The calculated log-spectral distortion values that were averaged over the whole 5 second scenes are presented in Fig. 12. The figure shows the color coded log-spectral distortion values. X-axis shows the size of the averaging block in time frames and y-axis shows the number of frequency bands inside which the averaging was made. In the double talk scene in plot a) there is a quite clear increase in distortion with averaging, especially in time. The mixed scene (talk, bass and dog) in b) shows less clear pattern and the difference between largest averaging and no averaging is only about 2dB. In the music scene results in c) there is an increase in distortion when averaged in frequency but temporal averaging does not show that large increase in the log-spectral distortion values. In double talk scene the sources say the same sentence but with a small time shift. This is why time averaging makes a greater difference in the log-spectral distortion values than averaging in frequency. In mixed scene the source signals are completely different and that is why the averaging does

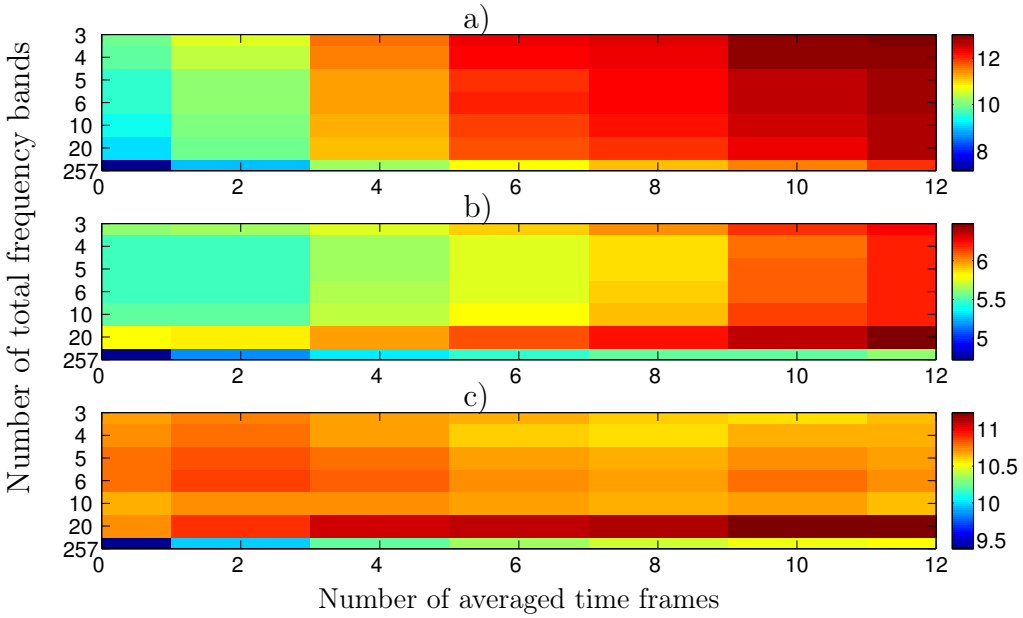

Figure 12: Color coded log-spectral distortion values of a) double talk, b) Speech-bass-dog (mixed) and c) Music scene. Note the different scaling between figures. The y-axis shows the number of frequency bands inside which the direction of arrival was averaged and x-axis the size of time-averaging window.

not affect as much as in the double talk scene. In music scene the sources are overlapping in time anyway but are located in different frequency bands. This is why averaging in time does not cause as much distortion as averaging in frequency.

4.2.2 Subjective quality degradation of time-frequency averaging

The log-spectral distortion does not provide information on how much the averaging affects perceptually. To obtain information on this, a listening test was organized. In the listening test the same signals as in the log-spectral distortion calculations were used. The test was organized as a MUSHRA test [65] where the anchor was a signal for which the direction of arrival information was randomized in 30×40 ($k \times n$) blocks. The reference was the signal without averaging. Although the log-spectral distortion values were calculated also for samples that were averaged in both time and frequency, only either one was applied to the listening test samples at once. This was done because the main interest was in the effect of these averagings individually and also to keep the listening test length within reasonable limits. The participants were asked to evaluate how stable and sharp the spatial image is. The used scale was from 0-100. There were 9 participants in the listening test.

The results of the listening test are presented in Figs. 13. The vertical lines present the result for each sample. Middle point of the line is the mean of the answers and the height of the line presents the 95% confidence interval. In time averaging in plot a) it should be noticed that the double talk and mixed scene have

quite clear decrease in evaluated spatial image sharpness and stability. In music scene the quality drop is clear compared to the reference but it is not significantly decreasing with larger time averaging. In frequency averaging the overall drop for all scenes is larger and especially now the music scene suffered the most. These results are somewhat in line with the decrease in quality that was noticed from the log-spectral distortion values. Based on these results the spatial image quality decreases when averaging is done so that the computed direction of arrival parameter covers bins where the sources would otherwise be non-overlapping. Because the signals overlap partly even without the averaging, the sources were not perfectly panned in the correct locations even in the reference signal. This is why the reference was not always graded at 100.

It seems clear that averaging decreases the perceived spatial image quality. However, in some cases it could be possible to use estimators utilizing a wider band at least above the aliasing limit in which case they should remove completely wrong directions of arrival due to aliasing, which are probably a bigger issue than reduced time-frequency-resolution.

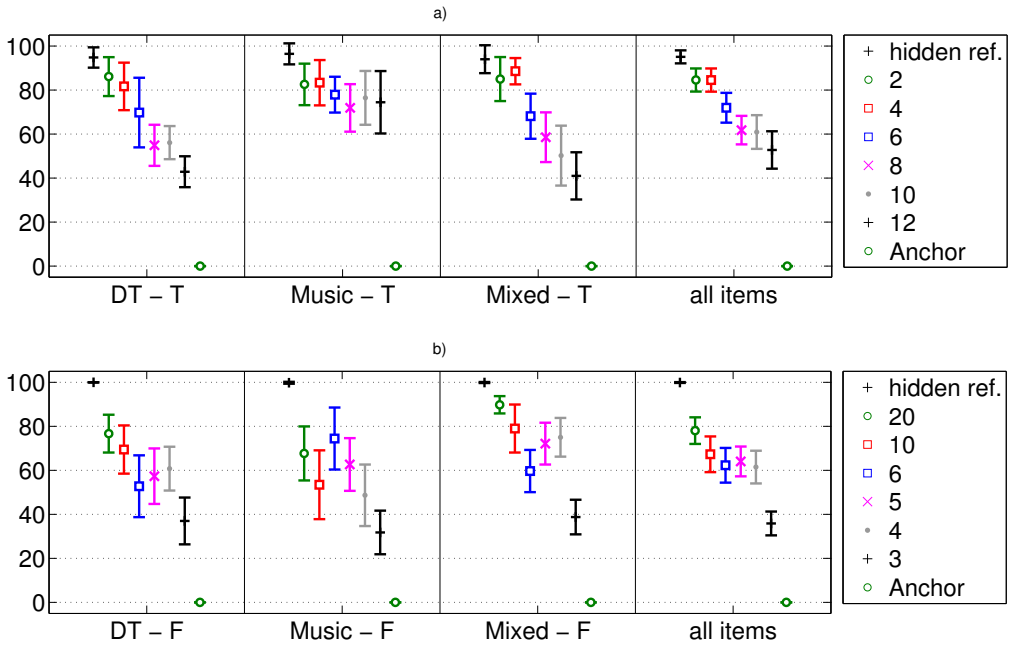

Figure 13: Results of spatial image sharpness listening test. Direction of arrival parameters averaged a) within 2,4,6,8,10,12 time frames of 11ms and b) within 20,10,6,5,4,3 frequency bands. In reference there was no averaging and in the anchor the direction of arrival was random. The spatial image quality decreases more when the averaged direction of arrival covers non-overlapping parts of the sources.

5 State of the art approaches for the spatial aliasing problem

The state of the art direction of arrival estimators in Sec. 3 suffer from spatial aliasing, which causes ambiguities in the estimated source direction. Spatial aliasing was explained in Sec. 4.1 and in this section some state of the art methods are presented to overcome the problem of spatial aliasing. Also the drawbacks of these methods are explained. The section starts by discussing some possibilities that are based on the modification of the microphone array. Then, some state of the art methods, that are based on unwrapping the observed interchannel phase difference information, are discussed. After these, an estimator based on the signal envelope is explained.

5.1 Physical changes to microphone arrays

This thesis focuses on ways to solve spatial aliasing with algorithms when using microphone arrays of omnidirectional microphones. However, it should be remembered that there are also ways to prevent aliasing by optimizing the microphone spacing or by utilizing microphone directivity. These approaches are briefly discussed in the following.

Spatial aliasing is caused by non-optimal array dimensions and thus, the simplest way to prevent it is to optimize the microphone spacing. In the direction of arrival estimation literature this is often made by choosing small microphone spacing or by limiting the frequency bandwidth, e.g., [66] and [67]. On the other hand, too small microphone spacing causes noise domination and decreased spatial resolution in the lower frequencies. To prevent this, it is possible to combine microphone arrays for different frequency bands. These setups are called nested arrays [68]. The basic idea is to use the array optimal in the frequency band under inspection. The problem is that adding a new array for each band is impractical due to number of microphones and computational complexity. To overcome this, one option is to nest the arrays with harmonically increasing distances [69]. Different sized uniform linear arrays are combined in a way that the overlapping microphones can be shared by different arrays and the total number of sensors is decreased. This method requires careful choosing of the subbands for each subarray but can be effective in preventing the spatial sampling problems.

In addition to nesting microphone arrays, it should be noted that microphone directivity can also be used as an advantage against spatial aliasing. This advantage can be obtained by placing omnidirectional microphones on the surface of a rigid baffle like a cylinder or a sphere. The structure causes shadowing especially on higher frequencies where the aliasing occurs. This method is used in [45] where microphones are placed on the surface of a sphere and direction of arrival estimation is based on the magnitude sensor response.

5.2 Phase unwrapping

As was discussed in Sec. 4.1, the failure of estimators based on the interchannel phase difference is caused by the wrapped phase differences. The simplest method for phase unwrapping is the Itoh's algorithm [8]. In this method the estimated interchannel phase difference values $\hat{\mu}(k)$ are examined along the frequency axis. The method compares the estimated interchannel phase differences between adjacent frequencies. If the difference

$$\Delta\hat{\mu}(k) = \hat{\mu}(k-1) - \hat{\mu}(k), \tag{43}$$

between two adjacent samples is out of the range of [-π π], it is considered as a wrapping point. The first wrapping point, i.e., the aliasing frequency bin, is denoted as K_a. Depending on whether the difference was positive or negative, then 2π is added or subtracted from all following values respectively. This method unwraps the values and in theory the phase difference slope for a single broadband source is then monotonic. This naïve method is effective when there is only one source but it is not very robust when noise is present. An improved version, presented in [70], applies Kalman filter which combines the noise reduction with phase unwrapping. The signal can also be divided to subbands so that the direction of arrival can first be estimated unambiguously in the first band and then proceed to upper bands, see for example [67] and [71]. In a recent paper [9] a method is proposed that provides both accurate direction of arrival estimate in wide frequency range and robustness to noise. In the following explanation the time indexes n are omitted for clarity as only one time frame is processed at once. The method has 7 stages which are explained in the following. At each stage the interchannel phase difference values are analyzed or modified and these stages are also plotted in Fig. 14 b).

1. *Narrow-band signal subspace estimation.* Frequency bins that do not contain enough energy from the source are neglected as noise. These bins are replaced by interpolating them from the surrounding values.

2. *Aliasing frequency f_{a} estimation.* Assuming the phase is wrapped in periods of K_a, applying autocorrelation to the wrapped interchannel phase differences $\hat{\mu}_{\mathrm{w}}(k)$ results a symmetrical vector. After taking the latter half of this vector the first maxima is at $k = 0$ and the first minima points the frequency bin K_a, where the first phase wrapping occurs and aliasing starts. The latter half of the autocorrelation vector is presented in Fig. 14 a) as a function of frequency.

3. *Wrapping direction estimation.* At the first wrapping point the phase difference is either $-\pi$ or π. From there the next value jumps to the other end of that range. In [9], the wrapping direction is estimated by fitting a line to the blue curve in 14 b). The sign of the slope of this line determines the wrapping direction w_d.

4. *Phase unwrapping.* As stated before, the unwrapping means adding or subtracting 2π to/from the wrapped value. This can be done using the following

equation:

$$\hat{\mu}_{\mathrm{uw}}(k) = \hat{\mu}_{\mathrm{w}}(k) + 2w_d \left\lfloor \frac{K_a + k}{2K_a} \right\rfloor \pi, \tag{44}$$

in which the wrapped $\hat{\mu}$ is added with $\pm 2\pi$ multiplied by the factor inside the floor function $\lfloor \cdot \rfloor$. Wrapping direction w_d determines the sign of the addition. In Fig. 14 b) the red curve shows the unwrapped values, i.e., values that are not anymore bound between $[-\pi\ \pi]$

5. *Failed unwrapped points correction.* In this step the unwrapped values are checked in case some values were missed by the algorithm. In Fig. 14 b) it can be seen how not all values on the red curve were unwrapped, because they are still the same as before unwrapping.

6. *Outlier removal for denoising.* The points which have distance higher than a certain threshold are removed to have a steady slope of the interchannel phase difference values.

7. *Final direction of arrival estimation.* Finally the unwrapped interchannel phase difference vector can be used in the direction of arrival estimation the same way as the wrapped interchannel phase difference would have been used.

A more detailed explanation can be found in [9]. Similarly to the Itoh's algorithm, this method also assumes a single source for all k, which makes it inadequate for accurate time-frequency processing. In [61], an interchannel phase difference replication method is presented for a linear array. This method considers all the possible aliasing periods for each frequency bin. In the end a matrix is created which will contain all the unwrapped phase values for both wrapping directions and all aliasing periods. These unwrapped values will be transformed to corresponding directions of arrival. As explained in [61], a histogram analysis will reveal the true directions of arrival as peaks in the histogram. This way it is possible to find more than one source within one time frame. Still, this only means that these sources are active in the time frame but their distribution along the frequencies remains unknown.

Because the above discussed phase unwrapping algorithms require information from a wider frequency region and may assume only a single source, these methods are not capable of resolving the true directions of arrival with the desired time-frequency resolution. However, because they are applicable in certain situations the phase unwrapping method of [9] was implemented in Matlab and it was tested in simulations. In Sec. 7.2 some estimation results of this method are shown.

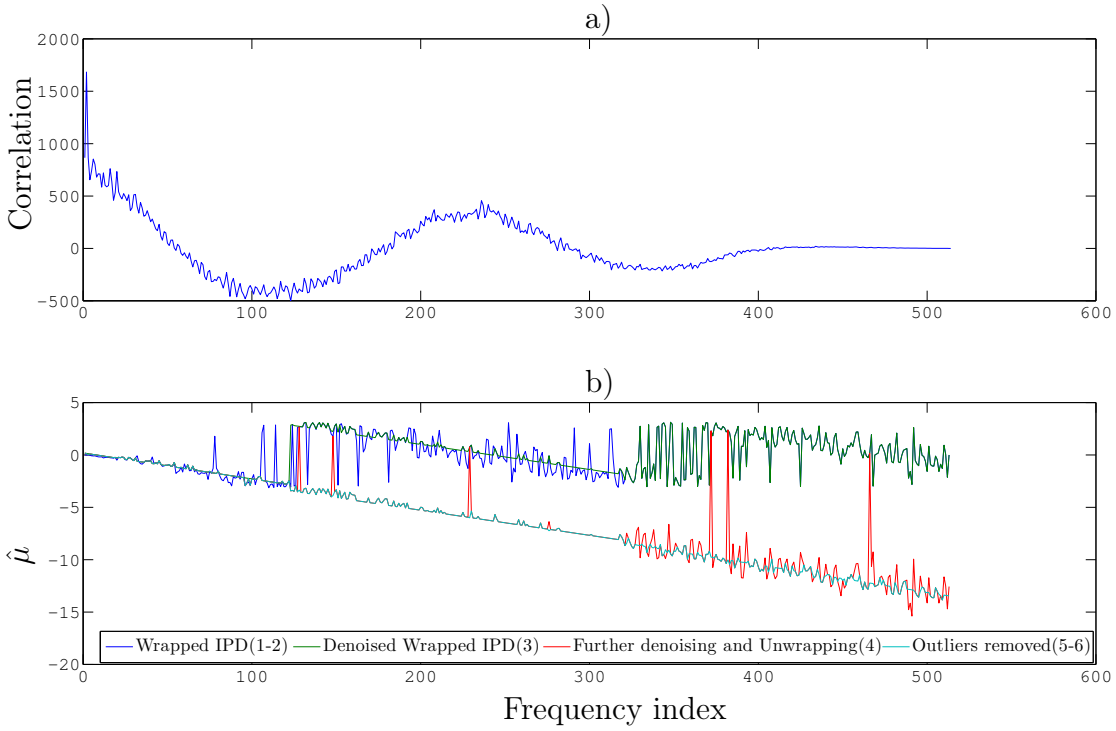

Figure 14: a) Shows the autocorrelation vector as a function of frequency which is used for finding the first aliasing frequency bin. b) shows the phase difference as a function of frequency. The different stages of phase unwrapping in [9] are color coded and named in the legend.

5.3 Envelope-based direction of arrival estimator

To overcome the limitations of spatial sampling with large microphone spacing, a parameter estimator based on signal envelope is presented in [10]. This method discards the fine structure of the signal and focuses on the amplitude modulations present in the common signals like speech and music. Because this envelope is relatively slowly varying, the spatial sampling problems have no affect in the frequency range of interest. Obtaining the time delay $\triangle n$, and furthermore the direction of arrival, from the envelope is quite straight-forward. The sound is divided into frequency bands from which the envelope is extracted. These bands can be for example 1/3 octave bands or equivalent rectangular bandwidth bands. The sound field model notations in Sec. 2.3 are used and only direct sound is assumed. As presented in (7) and (8), without the diffuse sound component the signal of the m-th microphone is

$$X_m(k,n) = P_{s,m}(k,n,\mathbf{r}_m) + X_{n,m}(k,n), \tag{45}$$

where $P_{s,m}(k,n,\mathbf{r}_m)$ is the direct sound plane wave and $X_{n,m}(k,n)$ is the uncorrelated microphone noise. In time-frequency domain the jth bandpass signal $X_{m,j}(k,n)$ is obtained as the multiplication of the bandpass filter transfer function $T_j(k)$ and the broadband signal

$$X_{m,j}(k,n) = T_j(k)X_m(k,n). \tag{46}$$

To obtain the envelope of the signal, the analytical signal spectrum $\check{X}_{m,j}(k,n)$ is first constructed. It is constructed with the following statements depending on frequency bin k and fast Fourier transform length N:

$$\check{X}_{m,j}(k,n) = \begin{cases} X_{m,j}(0,n), & \text{for } k = 0 \\ 2X_{m,j}(k,n), & \text{for } 1 \leq k \leq \frac{N}{2} - 1 \\ X_{m,j}(\frac{N}{2},n), & \text{for } k = \frac{N}{2} \\ 0, & \text{else.} \end{cases} \tag{47}$$

These statements form the discrete time analytic signal transform, as presented in [72]. This sets the spectrum's negative frequency half to zero to obtain the one-sided spectrum. The inverse Fourier transform of $\check{X}_{m,j}(k,n)$ gives the analytic signal $\check{s}_j(n)$. The magnitude of this is the signal envelope as [73]

$$\varepsilon_j(n,\mathbf{r}_m) = |\check{s}_j(n)|. \tag{48}$$

After the envelopes for at least the two points $\mathbf{r}_1$ and $\mathbf{r}_2$ are calculated, the time delays can be estimated by maximizing the cross-correlation over different delays τ, similarly to Sec. 3.1 as

$$\widehat{\triangle n_j} = \underset{\tau}{\operatorname{argmax}} \Big[\mathrm{E}\big\{ \varepsilon_j(n,\mathbf{r}_1)\varepsilon_j(n+\tau,\mathbf{r}_2) \big\} \Big]. \tag{49}$$

The obtained delays can now be used in the direction of arrival estimation that is based on the time difference of arrival. Because the delay can not be larger than travel the time between microphones, the range of τ can be limited to $\tau_{\max} = \pm\frac{\|\mathbf{r}_2-\mathbf{r}_1\|}{c}$. The achievable angular resolution is limited by number of time samples in this range. This in turn is directly dependent on the microphone spacing together with the sampling rate f_s [10]. This dependency is favourable because the larger the array is the more accurate this method is. Usually a large array causes the problem of spatial aliasing but this method takes the advantage of the large microphone spacing. The choice of bandpass filters determines the direction of arrival resolution in frequency as only one direction of arrival is estimated for each passband. Compared to the full time-frequency resolution this leads to decreased accuracy. Due to this, the envelope method should only be applied above the aliasing frequency where the methods based on the interchannel phase difference fail. In [10] the method was used together with the Estimation of Signal Parameters via Rotational Invariance Techniques estimator [5] above and beyond the f_a respectively. This combination produced improved perceptual spatial image quality when compared to using only the Estimation of Signal Parameters via Rotational Invariance Techniques estimation that produces aliased estimates. However, it did not reach the perceptual quality of the reference signal which had source locations rendered individually to correct locations. This means that reducing the frequency resolution to avoid spatial aliasing can lead to improved spatial image quality, as was assumed in Sec. 4.2.2.

6 Proposed approaches to overcome the spatial aliasing problem

With the state of the art direction of arrival estimators in Sec. 3 the direction of arrival can be estimated reliably below the spatial aliasing frequency f_a [5], [6]. Using the aliased direction of arrival estimates above the f_a does not produce perceptually desired results in the reproduction with Directional Audio Coding, as was discussed in Sec. 4.2. The methods presented in Sec. 5 can help reducing and preventing the wrong direction of arrival estimates caused by aliasing. However, the presented methods are not capable of providing the correct estimates in each frequency band individually for the time-frequency resolution used in Directional Audio Coding processing. In this section two methods are presented that could also be used to overcome the aliasing problem. The first one is a simple extrapolation that aims to provide similar results as phase unwrapping with low complexity. The second method is a correlation-based method that is uses the information on how the aliasing happens in different frequencies and directions of arrival.

6.1 Reducing spatial aliasing effects with parameter extrapolation

In Sec. 5.2 a method was presented that exploits the information below the aliasing frequency to unwrap the interchannel phase difference information. It could be better to use a low-complexity direction of arrival estimator that is accurate when aliasing is not present and extrapolate the non-aliased values from below the f_a to the higher bands. To perform this it is assumed that there exists only one source in each time frame. In addition, this source obviously needs to cover the frequencies below the f_a. However, the aliased estimates should first be calculated for all frequencies to capture the variation of the direction of arrival vector. These estimates can be used in the diffuseness estimation in (22). The value to be extrapolated can be calculated as a simple mean of the non-aliased estimates or then as a diffuseness-weighted mean. After using an aliased estimator to obtain the direction of arrival vector $\hat{\mathbf{n}}(k,n)$ in (11) and diffuseness $\hat{\Psi}(k,n)$ in (22), the diffuseness-weighted mean can be computed as

$$\tilde{\mathbf{n}}(n) = \frac{\sum_{k=2}^{K_a-1}[\hat{\mathbf{n}}(k,n)(1-\hat{\Psi}(k,n))]}{\sum_{k=2}^{K_a-1}[1-\hat{\Psi}(k,n)]}, \tag{50}$$

where K_a is the lowest frequency bin with aliasing and starting from bin $k = 2$ leaves out the DC component. The range of the averaging can be narrowed, e.g., if it is known that the low frequencies have low signal-to-noise ratio.

The extrapolation was first tested with a single speech source plane wave simulation with $\varphi = 52°$ and SNR = 30dB. The simulation follows the same procedure explained in Sec. 3.2.1. Obtained results are presented in Fig. 15, which shows the color coded direction of arrival estimates as a function of time and frequency. Plot a)

presents the aliased estimates with the Weighted Least Squares estimator and b) the extrapolation in use. In this scenario the correct direction of arrival is clear below the f_a which leads to correct extrapolation values. The time-frequency bins where the source is inactive are also denoted with the extrapolated value. In Directional Audio Coding reproduction this is not a major problem because the signal in these bins will be reproduced in the diffuse stream.

Problems arise when the assumption of a single source per time frame is violated or when the source direction of arrival can not be determined below the aliasing frequency. For example, the deficiency of this extrapolation method appears when trying to localize two sources in different directions. To test this, the simulation was run with the double talk scene, which has been simulated before in Sec. 4.2. In addition, a single source scene was simulated where the source is a bird sound at $\varphi = 52°$, which has majority of its frequency content above the f_a. The direction of arrival estimates for these scenes are color coded in Figs. 16 a) and b) as a function of time and frequency.

The Fig. 16 a) shows the extrapolation results in the double talk scenario. It can be seen that the male speaker at $\varphi = 52°$ is dominating the extrapolated estimates. This is due to its outstanding power in the lower frequencies, which can be seen by comparing the source spectrograms in Figs. 6 a) and c). Only when source 1 is inactive and the source 2 is active, the direction of arrival is extrapolated correctly for source 2, which can be seen for example right before 2 seconds. When neither of the samples is dominating in lower frequencies the extrapolated direction of arrival is placed in between the true directions, which causes spreading of the spatial image. This can be seen as light green areas indicating $\hat{\varphi} \approx 0$ which is approximately in the middle of these two sources. The spreading could be avoided by doing a histogram analysis of the non-aliased samples and picking the highest peak. However, this would present sudden changes of direction of arrival estimates in time which causes disturbing jumping of the sources.

In Fig. 16 b) it can be seen how the extrapolated values for a high pitch bird sound are distorted because there is no clear direction of arrival below the f_a to extrapolate. In general, it seems that the extrapolation works when a single source has dominant direction of arrival below f_a. In addition, multi-source scenes could benefit from this method if the sources are located nearby each other, e.g., at $\pm 30°$. In this case the sources would be localized in the frontal section whereas aliasing would cause parts of the signal being reproduced in completely wrong directions. This means that the spreading of the spatial image might be less of an issue than the spatial aliasing.

In informal listening test it was noticed that the single speech source scene with direction of arrival estimates in Fig. 15 b) did not have the negative perceptual effects of spatial aliasing. With the double talk and bird scenes in Fig. 16 there was a clear reduction in spatial image quality when compared to a scene without the

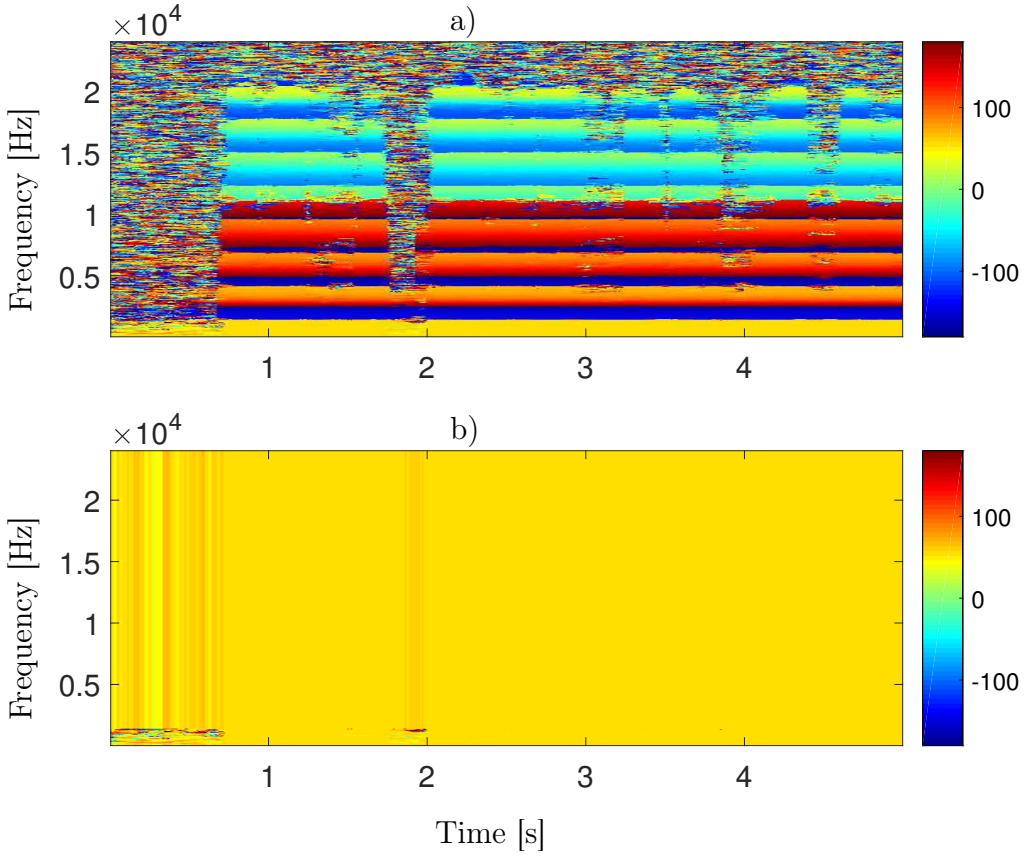

Figure 15: Speech source simulated at $\varphi = 52$. Results for a) aliased Weighted Least Squares estimator b) extrapolation by averaging the values below f_a. When the source is correctly estimated below the f_a, the extrapolation works as expected and appoints the correct direction of arrival estimate for all time-frequency bins.

aliasing effects. More experimental results for the extrapolation will be provided in Sec. 7, but based on these the extrapolation could be an efficient and robust way of reducing the negative effects of spatial aliasing when the broadband single source assumption is satisfied.

6.2 Resolving spatial aliasing with correlation-based approach

As explained in Sec. 4.1, the spatial aliasing is only affected by the frequency, direction of arrival and microphone array setup. An example of this, meaning how the aliasing occurs over different frequencies and directions of arrival, i.e., the aliasing pattern, was presented in Fig. 8. This pattern is specific to the used microphone array and the use of the Weighted Least Squares direction of arrival estimator, explained in Sec. 3.2. In the situation in Fig. 8, spatial aliasing causes the wrong estimates above $f_a \approx 1.3$kHz. If the wrong estimates would appear only once across all directions of arrival in a specific frequency band, the mapping from aliased to correct values would be trivial. However, as there are multiple appearances of the same

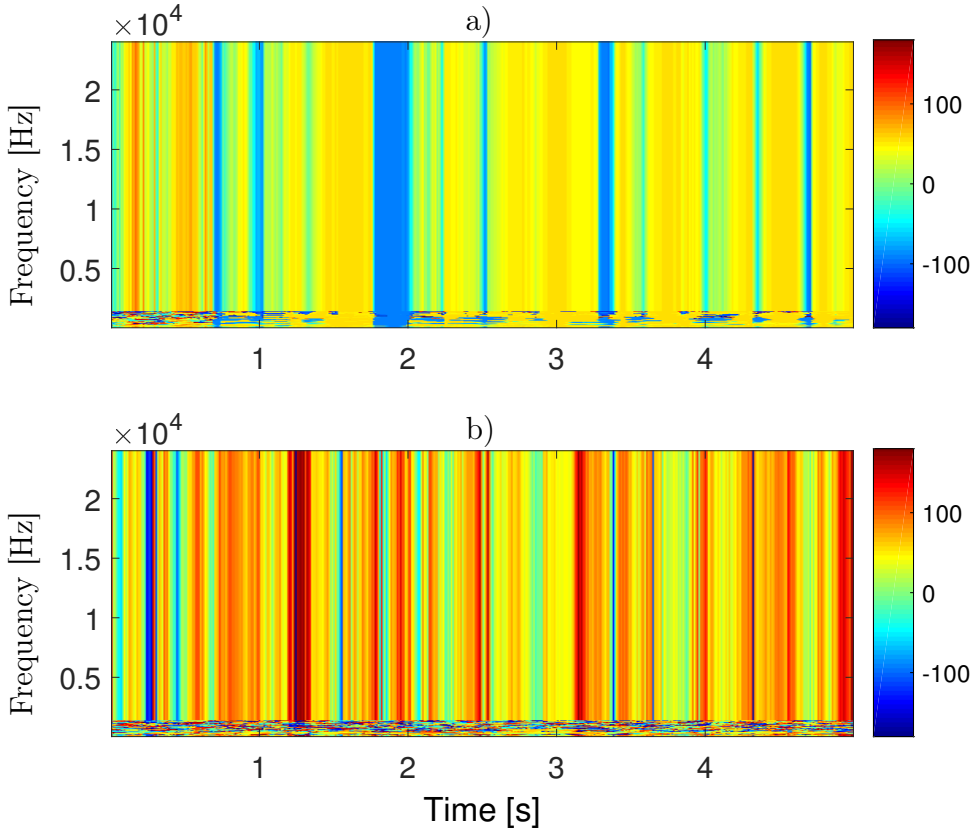

Figure 16: Problems of extrapolation method pointed out in a) double talk b) bird sound scenes. In a) the existence of two sources below f_a causes extrapolated value to be incorrect (light green values). In b) the extrapolated value is almost random because there is no clear direction of arrival to extrapolate.

aliased direction of arrival across the direction of arrival axis, i.e., ambiguities, it is not possible to reliably determine to which true direction of arrival the aliased value corresponds to. However, there are other ways than direct mapping to utilize this aliasing pattern and its properties. This is based on correlation between the aliased direction of arrival estimates and this aliasing pattern. This correlation is possible because the pattern is different for each direction of arrival when all frequencies are considered, which will be shown in the following. The highest correlation is found when the aliased values are correlated with the direction of arrival vector of the pattern that corresponds to the true direction of arrival. In fact, it is also possible to use directly the wrapped interchannel phase difference information in the correlation without using the Weighted Least Squares estimator. The following sections present the use of aliased Weighted Least Squares and wrapped interchannel phase difference data in the correlation approach to resolve spatial aliasing.

6.2.1 Correlation approach with Weighted Least Squares data

The aliasing pattern for a 4-microphone rectangular array in Fig. 3 with $r = 9$cm was shown in Fig. 8. The computing of these direction of arrival estimates was made by using (28)-(35) that use the signal model presented in Sec. 2.3. This pattern can be used to estimate the direction of arrival by finding the direction of arrival vector in the pattern that produces the highest correlation with the aliased direction of arrival vector. This approach is possible because the aliasing pattern has two properties; it is constant when the microphone array is kept constant and the patterns for each true direction of arrival are unique when considering the whole frequency range. In the following, the uniqueness property is shown and the correlation equations for wideband and narrowband estimation are presented. Then, an example of using the correlation for wideband direction of arrival estimation is presented. After that, it is explained why this approach becomes unreliable when attempting to perform narrowband direction of arrival estimation. However, it is possible to obtain correct narrowband estimates with this correlation method when the estimation is done in a 2-step manner, which will be explained last.

The uniqueness property of the aliasing pattern can be verified by calculating the correlation coefficients between each frame and all the other frames [74]. The result of this correlation is shown in Fig. 17 a) which shows the normalized color coded correlation coefficients as a function of the directions of arrival. These values were calculated using the Matlab corrcoef-function [75]. The figure shows that the highest (positive) correlation appears always on the diagonal, meaning that correlation is highest only when a single direction of arrival vector is compared to itself. It can also be seen that at $\varphi = \pm 45°$ and $\pm 135°$ there are higher correlation values also around the diagonal but the highest correlation is still at the diagonal.

Using the uniqueness property, the correlation can be performed by first using the Weighted Least Squares estimator, as explained in Sec. 3.2, to obtain aliased direction of arrival vector $\hat{\mathbf{n}}(k, n)$ for all frequencies and then finding the $\mathbf{n}(k, \varphi)$ along the aliasing pattern that best matches to the obtained aliased values. The direction of arrival vector of the pattern which produces the maximum correlation when considering all frequencies can be found as

$$\hat{\varphi}_{\mathrm{CC}} = \underset{\varphi}{\operatorname{argmax}} \left[\sum_{k=1}^{k=K} \left[\hat{\mathbf{n}}(k) \mathbf{n}(k, \varphi) \right] \right]. \tag{51}$$

The maximum correlation for specific frequency bin k, i.e., narrowband estimate is found using

$$\hat{\varphi}_{\mathrm{CC}}(k) = \underset{\varphi}{\operatorname{argmax}} \left[\hat{\mathbf{n}}(k) \mathbf{n}(k, \varphi) \right]. \tag{52}$$

For now, the wideband correlation in (51) is used and the problem of the narrowband version in (52) is explained after that.

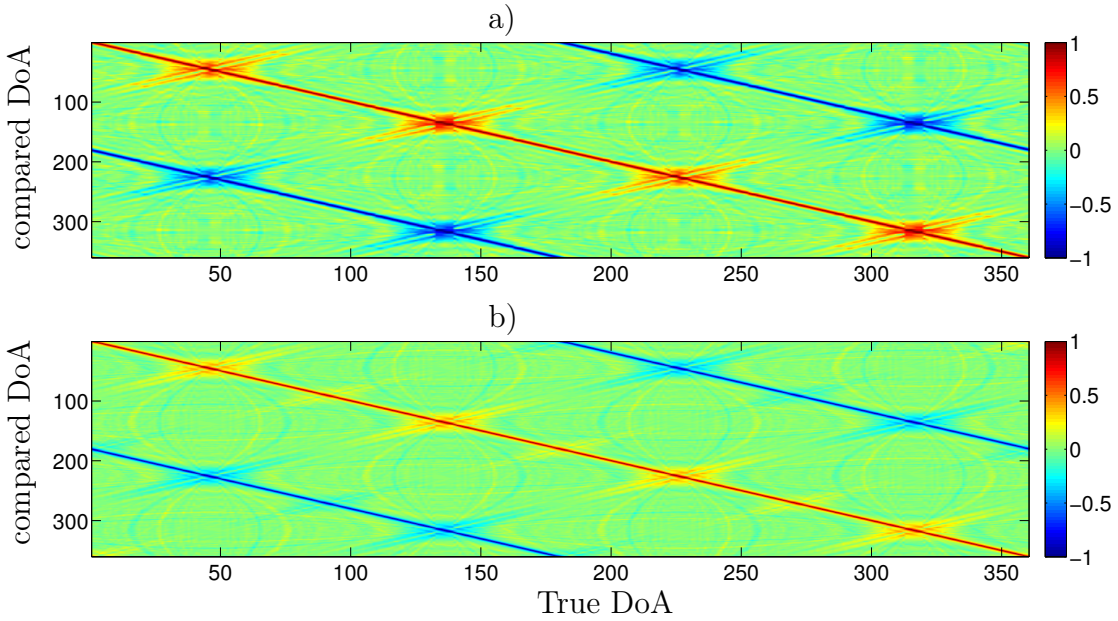

Figure 17: Color coded correlation coefficients as a function of the direction of arrival for all frames in a) aliasing pattern b) wrapping pattern. Frame here refers to values with single DoA, i.e., $K \times D$ bins in the aliasing and $K \times H$ bins in the wrapping pattern. The correlation of 1 is only found at the diagonal, i.e., all the pattern frames are unique. Narrowing down the frame size to single frequency makes also off-diagonal values approach to 1.

The correlation process for wideband direction of arrival estimation is illustrated in Fig. 18, which shows an aliased frame on the left and the aliasing pattern on the right. The result of the correlation with (51), as a function of the direction of arrival, is shown at the bottom. The aliased frame was obtained from a plane wave simulation similar to what was performed in Sec. 3.2.1, but with $r = 9$cm. The correct direction of arrival for most of the frequency bins in the frame is $\varphi = 52°$ and in the correlation result the highest peak appears on this location.

This way the correlation is able to reveal the true direction of arrival even when the aliased frame contains noise like is the case in Fig. 18. However, this result is a broadband solution, whereas in Directional Audio Coding processing narrowband estimates are needed. This broadband solution could be used when only one source is expected within one time frame and its direction can not be estimated below the f_a. If reliable estimation is possible below the aliasing frequency, then the extrapolation method, presented in Sec. 6.1, would be a more efficient way of resolving the aliased frequencies.

By looking at the aliasing pattern in Fig. 18, it can be seen that the difference between the frames of different directions of arrival decreases when narrowing the inspected frequency range. For example, the direction of arrival coded with

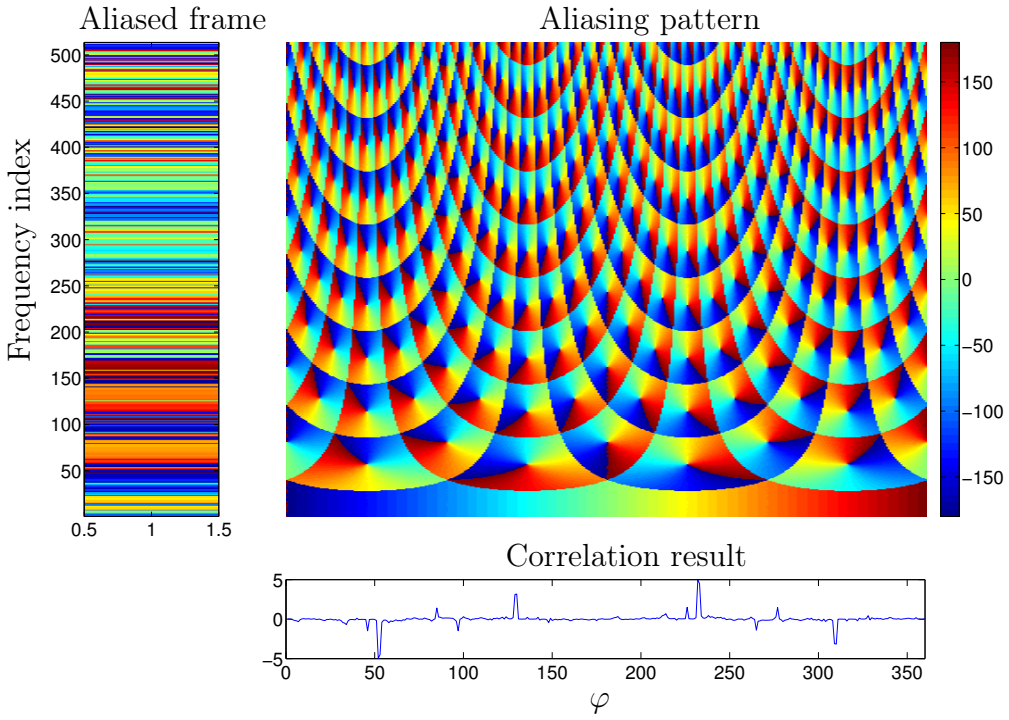

Figure 18: Correlation with the aliased frame and a frame on the aliasing pattern results a peak when the best match along the direction of arrival dimension of the pattern is used. The active source is located at $\varphi = 52°$. Notice that the pattern directions of arrival (x-axis) are between $[-180\ 180]°$ but the correlation x-axis is between $[0\ 360]°$.

orange color around frequency bin 140 appears multiple times on the direction of arrival axis. Narrowing the inspected frequency range leads to losing the uniqueness property and inevitably to unreliable correlation results when trying to perform narrowband estimation. However, the broadband correlation can be considered as the first step in resolving the ambiguities. In the second step the broadband estimates are used as a priori information to obtain the narrowband estimates. A priori information means that other source of information of the correct direction of arrival is used to reduce the ambiguity. For example the a priori information could come from a video camera system that recognizes the potential sources and their directions.

The 2-step correlation approach for a single time-frequency bin in the doubletalk scene (Fig. 10) is illustrated in Fig. 19. First the wideband correlation in (51) is used to obtain direction of arrival estimates above the aliasing frequency for each time frame. Plot a) shows these color coded estimates as a function of time and frequency for the whole scene. These broadband estimates are then collected to a histogram in plot b), which shows two peaks at the correct locations of the sources, i.e., $\varphi_1 = 52°$ and $\varphi_2 = -92°$. The plot c) shows the narrowband correlation values

for a single time-frequency bin as a function of the direction of arrival. These values were computed using (52) for one time frame and $k = 100$. At the inspected time instance, only the source at $\varphi_2 = -92°$ was active. It can be seen that there is high correlation at this source direction but also almost equal peaks for false directions of arrival, i.e., ambiguities. By first picking the peaks in c) that indicate a high correlation, e.g., above 0.9, and then choosing the direction of arrival that has highest peak in the histogram in plot b), it is possible to reveal the correct direction of arrival.

The problem with the 2-step correlation approach is the forming of the histogram of the direction of arrival candidates. In the example in Fig. 19, a 5 second section of the scene was analyzed for the histogram. Using shorter time sections for the analysis would also allow only occasionally appearing source directions to be represented in the histogram. However, too short time section might make the histogram peaks less clear. To avoid this trade-off, the aim is to perform the correlation only directly for individual time-frequency bins. In the next section the correlation approach for individual time-frequency bins is explained by utilizing the interchannel phase difference information.

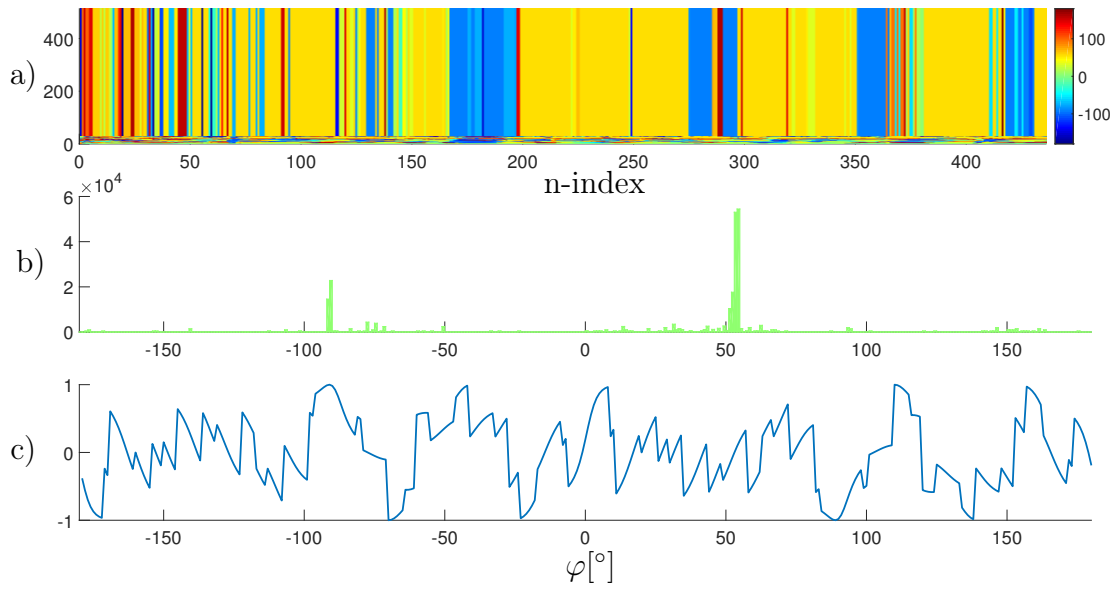

Figure 19: The 2 step correlation approach illustrated. a) Color coded broadband estimates as a function of time and frequency. b) Histogram of the broadband direction of arrival estimates that is used as a priori information, peaks appear at the correct $\varphi = 52°/-92°$. c) Shows the correlation of a single time-frequency bin where the source at $\varphi = -92°$ is dominant. When combining these correlation values with the histogram in b), it is possible to find the narrowband estimate.

6.2.2 Correlation approach with interchannel phase difference data

The aliasing pattern and its use for the correlation approach was presented in Sec. 6.2.1. For that approach, the aliasing pattern was created using the Weighted Least Squares estimator which combines the information from all microphone pairs by using a least squares solution, as explained in Sec. 3.2. The use of the least squares solution might lead to loss of information. For this reason, in this section the correlation approach for direction of arrival estimation is developed to utilize directly the interchannel phase difference information from each microphone pair. In this case the interchannel phase difference information for all the frequencies, directions of arrival and microphone pairs is referred to as the wrapping pattern. The problem with the aliasing pattern was that the uniqueness of the pattern for different directions of arrival is lost when considering only a single frequency. For the wrapping pattern this is not the case, which is why it could yield more reliable results. However, the basic idea is the same; utilize the uniqueness property of the wrapping pattern for each direction of arrival and find the best match along the wrapping pattern and the measured interchannel phase difference values. In the following, more details of this approach are explained.

The wrapping pattern can be obtained by simulating a white noise source in every direction of arrival (with 1° steps) and use the equations (28)-(32) to compute the interchannel phase difference values $\mu_h(k, \varphi)$. An example of this wrapping pattern for a single direction of arrival is shown in Fig. 20. This figure shows the color coded wrapped interchannel phase differences for all six microphone pairs and frequency bins when only $\varphi = 0°$ is considered. The used microphone array is the 4-microphone rectangular array shown in Fig. 3 with $r = 9$cm. In this case the pattern for a single direction of arrival is called a layer. It can be seen that in this layer the wrapping pattern is the same for microphone pairs 1, 3 and 4. This is not the case with different directions of arrival, because the microphone pairs are oriented differently. Changing the direction of arrival makes the interchannel phase difference values change and so an unique layer can be achieved for each direction of arrival.

The uniqueness is shown similarly to the aliasing pattern by calculating the correlation between each of the layers and the result is shown in Fig. 17 b). For the use of Matlab corrcoef-function [75], the K number of interchannel phase difference values for each microphone pair and direction of arrival are collected to a single array of values so that there are 360 arrays with $K \times H$ values. Similarly to the aliasing pattern, the highest correlation is found at the diagonal but the correlation around $\varphi = \pm 45°$ and $\pm 135°$ is not as high as it is for the aliasing pattern. This means that using the wrapping pattern in the correlation approach for direction of arrival estimation could be more accurate, especially around these angles. Calculating the correlation coefficients for a single frequency bin makes the off-diagonal values approach to 1. With the aliasing pattern this happens for more direction of arrival combinations and already with wider frequency range than with the wrap-

ping pattern. This means that the uniqueness property is stronger, i.e., the use of correlation approach with the wrapping pattern could be more reliable even for narrowband estimation. The narrowband correlation coefficients are not shown in the figure but it was tested separately.

The improved uniqueness means that the correlation approach using the wrapping pattern could yield more reliable results. For the correlation approach, first the interchannel phase differences $\hat{\mu}_h(k)$ for each microphone pair h, are computed using (28)-(32). Using these values and the wrapping pattern values $\mu_h(k, \varphi)$, the narrowband correlation can be calculated as,

$$\hat{\varphi}_{\mathrm{PCC}}(k) = \underset{\varphi}{\operatorname{argmax}} \left[\sum_{h=1}^{H} \left[\hat{\mu}_h(k) \mu_h(k, \varphi) \right] \right]. \tag{53}$$

When calculating the correlation for a single time-frequency bin with (53), a more reliable estimate can be achieved when compared to using the Weighted Least Squares data in (51). This is due to the improved uniqueness and more specifically because the ambiguity peaks in the correlation are likely to be distributed differently for each microphone pair that is oriented differently or has a different spacing. Only the correct correlation peak is the same for all microphone pairs. To show an example, a plane wave simulation was made at a single frequency of 6.5kHz and $\varphi = 52°$. The planar microphone array was the one in Fig. 3 with $r = 9$cm. The

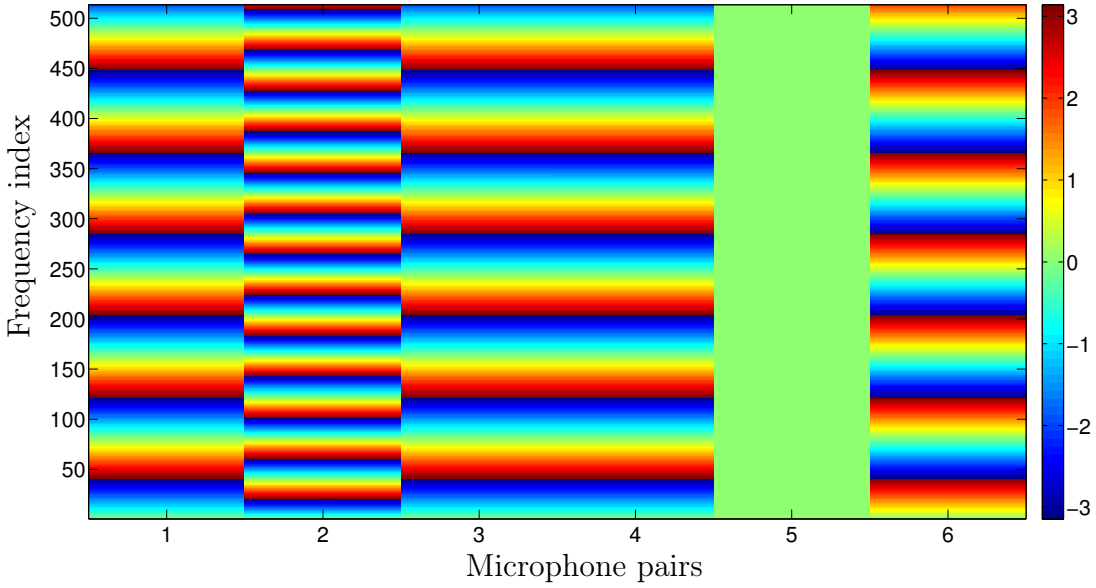

Figure 20: Example of wrapped interchannel phase differences with $\varphi = 0$ for the planar array in Fig. 3 with $r = 9$cm. Microphone pairs that are symmetrically positioned have the same phase differences. Pair number 5 has axis perpendicular to the direction of arrival so the interchannel phase differences are zero at all frequencies.

noise level was 30dB and the results were obtained by running the simulation 2000 times so that a distributions for the interchannel phase differences were obtained. The correlation was made separately for each microphone pair and the correlation results were summed to obtain the total correlation. Fig. 21 a)-f) shows the correlations for each microphone pair as a function of the direction of arrival. It can be seen that there are multiple high correlation peaks that are practically equal in height. Among the peaks there is one peak that points the correct direction of arrival and the rest of the peaks are the ambiguities. The distributions vary between the microphone pairs, but there are also pairs that are oriented the same way and thus the distribution of the peaks is the same for these pairs, e.g., pairs 3 and 4 in plots c) and d). Plot g) shows the summation of the correlation values for individual microphone pairs. The summation shows the highest peak at the correct direction of arrival, which means that in this case, the narrowband estimation would be possible. However, depending on the direction of arrival and frequency, it is possible that also the aliased peaks line up and the summation does not point out only the correct direction of arrival. This becomes a problem especially when there are multiple sources and reflections/reverberation included in the simulation. This means that the wrapping pattern uniqueness along the direction of arrival dimension is not a valid assumption in every case.

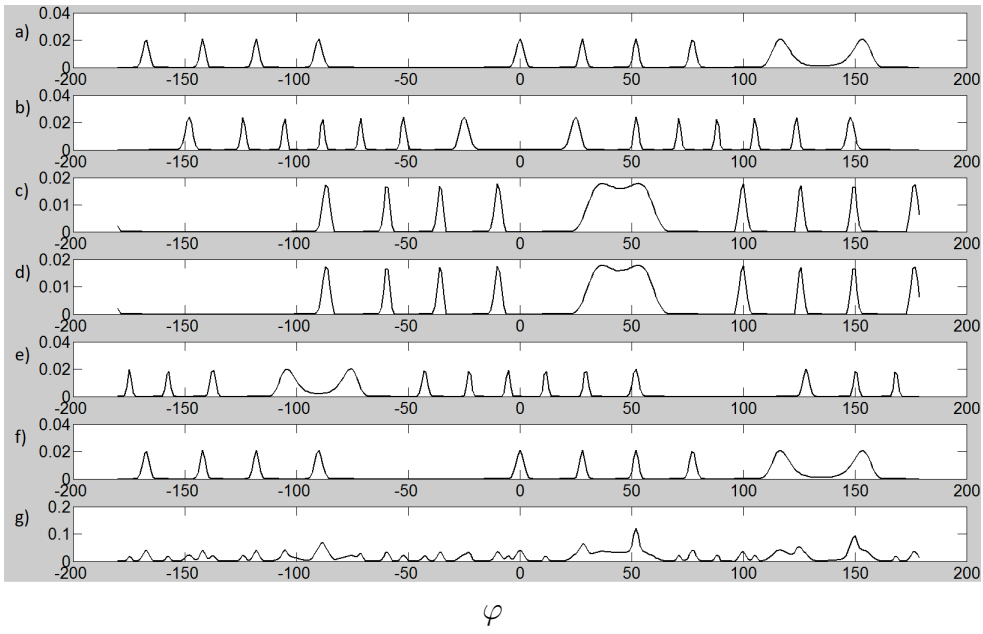

Figure 21: Plane wave simulation with $f = 6.5$kHz and correlation results calculated for each microphone pair individually. a)-f) shows the correlation peaks, including one correct and several wrong peaks. The changing orientation and microphone spacing changes the distribution of the wrong peaks. Summation in g) of all the correlation values results a single high peak at the correct direction of arrival $\varphi = 52°$.

To increase the uniqueness of the wrapping pattern and thus the reliability of this approach, the correlation can be performed over a broader frequency range, i.e., use a frequency buffer around the frequency bin under inspection. This means that the dominating source within this buffer is computed with the correlation and the result is used as a direction of arrival estimate for examined bin. When using the frequency buffer, the correlation in (53) becomes

$$\hat{\varphi}_{\mathrm{PCCB}}(k) = \underset{\varphi}{\operatorname{argmax}} \left[\sum_{\omega=-\Omega}^{\Omega} \sum_{h=1}^{H} \left[\hat{\mu}_h(k+\omega) \mu_h(k+\omega, \varphi) \right] \right], \tag{54}$$

where Ω denotes the size of the frequency buffer. This is a trade-off between frequency resolution and direction of arrival estimation accuracy. The frequency buffer can be used also for the Weighted Least Squares data correlation in (51). As was discussed in Secs. 4.2.1 and 4.2.2, decreasing the frequency resolution decreases the spatial image quality. However, it was noticed that this decrease of resolution is also a source dependent issue. For example, the scene with musical instruments covering different frequencies suffered a lot from the decreased resolution, whereas the double talk scene did not suffer so much. For now it will be assumed that reducing the frequency resolution is acceptable to achieve more reliable direction of arrival estimation above the aliasing limit. In Sec. 7.3, some listening test results are shown that prove that the reduced frequency resolution is perceptually less of an issue than using the aliased estimates.

The Fig. 22 shows the flow diagram of the narrowband correlation approach using the aliasing and wrapping patterns. The steps are:

1. Simulate a plane wave arriving from every direction. Here a 1° resolution was used.

2. Save either the obtained interchannel phase difference information directly to the wrapping pattern or use the Weighted Least Squares method to obtain the aliasing pattern.

3. Measure the interchannel phase differences of the microphone pairs when a source is located at φ_1.

4. Use the Weighted Least Squares method to obtain an aliased direction of arrival vector or keep the interchannel phase difference layer as is.

5. Compute the correlation for all φ between the aliased direction of arrival vector and the aliasing pattern, or between the wrapped layer and the wrapping pattern, using (52) or (53) respectively.

6. Search for the maximum correlation along the direction of arrival dimension to obtain the direction of arrival estimate $\hat{\varphi}$.

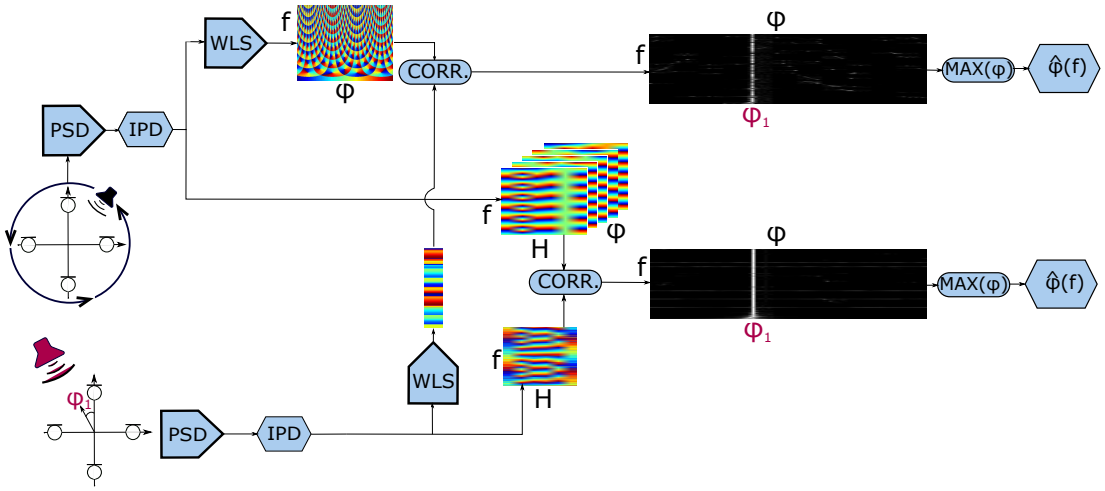

Figure 22: The flow diagram of the correlation-based direction of arrival estimation. First, the aliasing or wrapping pattern is created by simulating a white noise source in all directions of arrival. Second, the aliased or wrapped values are measured and the correlation with the corresponding pattern is calculated. The maximum correlation along the direction of arrival dimension points out the direction of arrival estimate $\hat{\varphi}$.

The correlation approach is based on similarity comparison between the measured values and the pattern, i.e., the direction of arrival estimation becomes a template matching problem. Another way of measuring the similarity would be to calculate the least squares error between the measured value and the pattern values. The direction of arrival estimate yield by this measure can be computed as

$$\hat{\varphi}_{\mathrm{PLS}}(k) = \underset{\varphi}{\operatorname{argmax}} \left[\frac{1}{\sum_{h=1}^{H} \left[\hat{\mu}_h(k) - \mu_h(k, \varphi) \right]^2} \right]. \tag{55}$$

This can be used in the direction of arrival estimation by replacing the correlation calculation in the flow diagram in Fig. 22. Also the use of the frequency buffer works similarly. In the following section these two similarity measures, used with both the Weighted Least Squares and interchannel phase difference data, are applied and their results are compared to find the best combination.

7 Experimental results

In this section the correlation-based method and extrapolation method, which were proposed in Sec. 6, are tested for different scenarios with plane wave and room impulse response simulations. Also the phase unwrapping and envelope-based direction of arrival estimation methods, explained in Sec. 5, are used for comparison. Before comparing all these methods, a single white noise source test is made only for the correlation method. This is because the correlation method is of main interest and the point is to compare the use of Weighted Least Squares and interchannel phase difference data and the correlation and least squares similarity measures discussed in Sec. 6.2. Then, the correlation, extrapolation, phase unwrapping and envelope detection methods are tested with speech signals and in multiple source scenario. In the end also some listening test results are presented. The listening test shows the perceptual spatial image quality that is achieved with Directional Audio Coding reproduction by using the extrapolation and correlation-based methods in the direction of arrival estimation.

7.1 Results of the correlation approach with a single white noise source

In this section the correlation-based method is tested with a single white gaussian noise source. This source provides equal power in all frequency bins so the true direction of arrival φ can be used as a reference in root-mean-square error calculation. The noise signal is simulated as a plane wave arriving at the planar microphone array of 4 microphones (see Fig. 3) with $r = 9$cm. The plane wave simulation procedure was explained in more detail in Sec. 3.2.1. With this array the spatial aliasing frequency is $f_a \approx 1.3$kHz. The noise source is moved from $\varphi = 0°$ to $\varphi = 180°$ with $5°$ steps. The used source signal is the same in all directions of arrival. Microphone noise is added to observe root-mean-square error values with different signal-to-noise ratios. The aliased estimates are calculated using the Weighted Least Squares method, discussed in Sec. 3.2, which provides valid estimated up to the aliasing limit. The correlation method is only applied above the aliasing limit. The root-mean-square error calculation is made for estimates between $f_a - 20$kHz.

The difference $\triangle\varphi(k,n)$ between true direction of arrival φ and the estimate $\hat{\varphi}(k,n)$ is defined as

$$\triangle\varphi(k,n) = \begin{cases} |\hat{\varphi}(k,n) - \varphi|, & \text{for } |\hat{\varphi}(k,n) - \varphi| \leq 180 \\ 360 - |\hat{\varphi}(k,n) - \varphi|, & \text{else.} \end{cases} \tag{56}$$

The root-mean-square error is calculated as

$$\text{RMSE}_{\hat{\varphi}} = \sqrt{\frac{1}{K}\frac{1}{L}\sum_{k=1}^{K}\sum_{l=1}^{L}\left[\triangle\varphi(k,n)\right]^2}, \tag{57}$$

where k and n are the frequency and time bins respectively and K and L are the number of frequency and time bins respectively. The signal length is 5s and sampling frequency $f_s = 48$kHz. The correlation method is applied using the narrowband correlation presented in (52) and (53) for the Weighted Least Squares and interchannel phase difference data respectively. The results using the least squares distance measure are computed using (55) for the interchannel phase difference data and its corresponding modification to the Weighted Least Squares data.

The root-mean-square error values for each direction of arrival are presented in Fig. 23. In plot a), where the SNR = 60dB, it can be seen that the root-mean-square error values are low for all four versions of the correlation approach. The errors for the least squares measure with interchannel phase difference data is close to 0 at all different directions of arrival. The correlation measure with interchannel phase difference data shows slightly higher error values. When using the Weighted Least Squares data the least squares and correlation produces identical results, which are higher than for the interchannel phase difference data. In plot b) the noise level was increased so that the SNR = 20dB. Now the differences between the different approaches is more clear, but their order is the same; least squares with interchannel phase difference data produces the lowest root-mean-square error value and correlation with interchannel phase difference data is more prone to errors. Using the Weighted Least Squares data produces still the worst results. For comparison, with SNR = 20dB the use of the aliased Weighted Least Squares estimates directly led to an average RMSE = 107. This means that the correlation approach was still improving the estimation accuracy because the maximum error with the Weighted Least Squares data in correlation was RMSE = 94. With SNR = 5dB in plot c) it can be seen that the differences between the methods are small because the microphone noise is causing a lot of failures for all the methods.

To show a visual example how much wrong estimates there are with each approach, a plane wave simulation was run for a single direction of arrival $\varphi = 50°$ and with SNR = 20dB. Fig. 24 shows the color coded direction of arrival estimates of this simulation as a function of time and frequency. The correct estimates (yellow color) can be found below the aliasing frequency ($k = 29$) for each approach. It can be seen that the least squares measure with interchannel phase difference data produces the best performance, as it has most of the time-frequency bins estimated correctly in c). The correlation with interchannel phase difference data in d) performs worse, but has also correct estimates above the aliasing frequency. When using the Weighted Least Squares data in a)-b), most of the estimates in the higher frequencies are incorrect. In this figure it can also be seen that there are frequency bands where the estimation fails almost every time instance with every method. These are most probably the frequencies where the aliasing peaks align over all microphone pairs and thus the correct direction of arrival can not be resolved from the aliased ones. This was discussed in more detail in Sec. 6.2.2.

To obtain more reliable results, the correlation can be performed over a broader

range of frequencies using the frequency buffer as discussed in Sec. 6.2.2. The frequency buffer size Ω means how many frequency bins are included in the correlation above or below the studied bin. For example, buffer size 10 means that the total number of frequency bins is $2\Omega + 1 = 21$. The total buffer size is kept constant over all frequencies and if possible, equal number of frequency bins are used above and beyond the inspected frequency bin. For the highest and lowest bins it is not possible and for example for the highest frequency bin the buffer covers 20 bins below the highest bin.

To show the results with frequency buffers of different sizes, the correlation calculation was performed once more using (54) for interchannel phase difference correlation and the corresponding modifications of (52) for the Weighted Least Squares data correlation and (55) for the least squares measure with Weighted Least Squares/interchannel phase difference data. The buffer size was increased from 1 to 28. The source was located to $\varphi = 50°$ and SNR = 6dB. The obtained root-mean-square error values are plotted in Fig. 25 as a function of the buffer size. It can be seen that with buffer size of 1 the least squares method with interchannel phase difference data still has the lowest root-mean-square error value. When the buffer size is increased, the correlation method using the interchannel phase difference data outperforms all the other methods. It can also be noticed that the correlation method using the Weighted Least Squares data is now performing better than the least squares method with the same data. Based on these results the correlation method with interchannel phase difference data is chosen to be used in Sec. 7.2 where more realistic source signals are used.

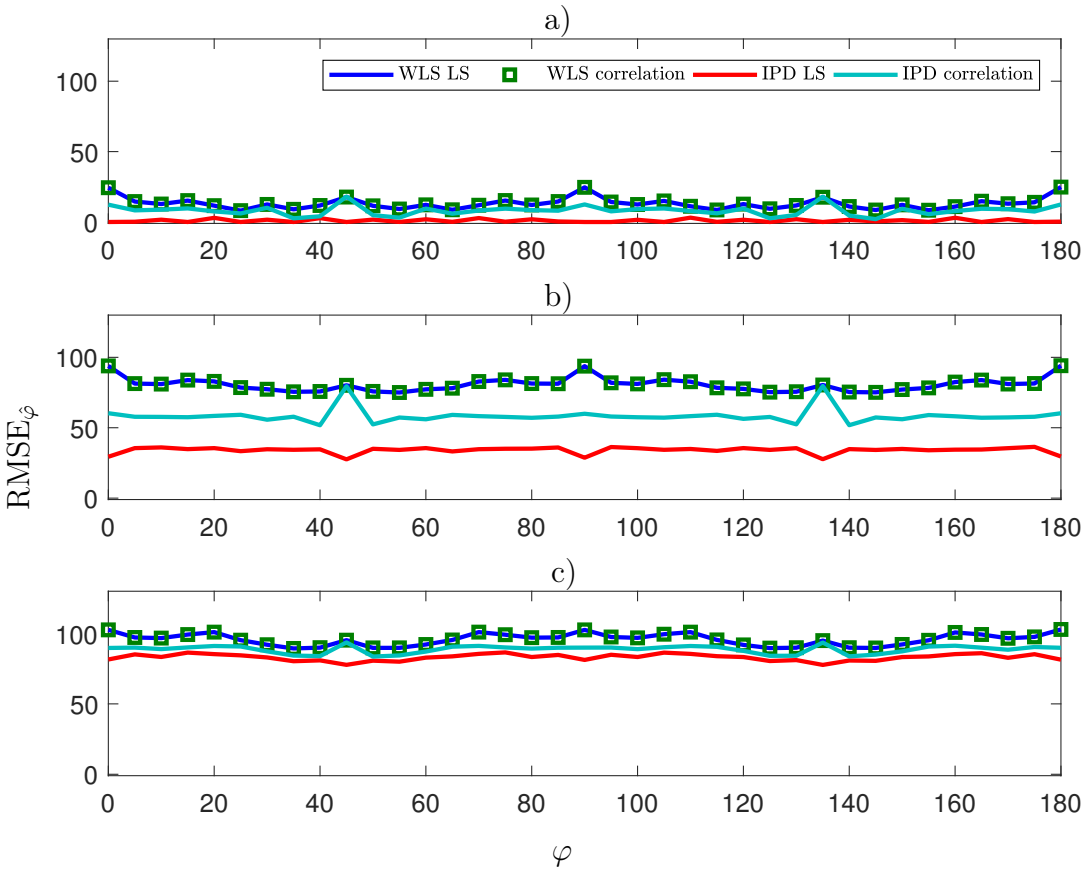

Figure 23: Comparison of the RMSE$_{\hat{\varphi}}$ values of narrowband estimation with correlation/least squares measures and Weighted Least Squares (WLS)/interchannel phase difference (IPD) data. A single white noise source was located from $\varphi = 0°$ to $\varphi = 180°$ with 5° steps. Noise was added to obtain a) 60dB b) 20dB c) 5dB SNRs. The least squares measure with interchannel phase difference data produced the smallest errors.

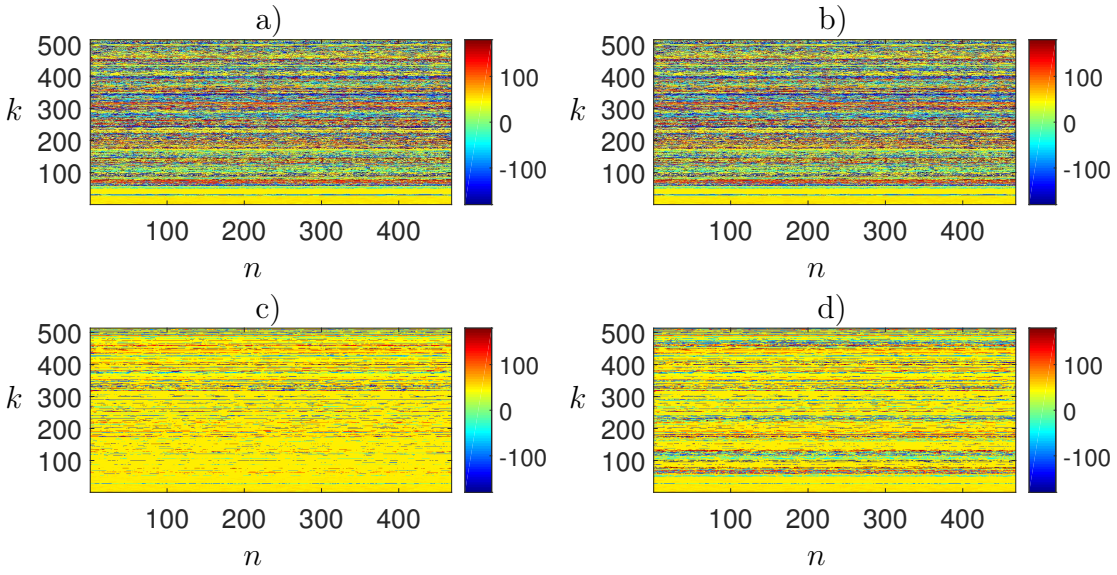

Figure 24: Narrowband direction of arrival estimates as a function of n and k for a noise source at $\varphi = 50°$. Estimates computed with a) least squares/Weighted Least Squares (WLS) b) correlation/WLS c) least squares/interchannel phase difference (IPD) d) correlation/IPD combinations of similarity measure/data source. Best results are seen in c). On specific frequencies the estimation fails often.

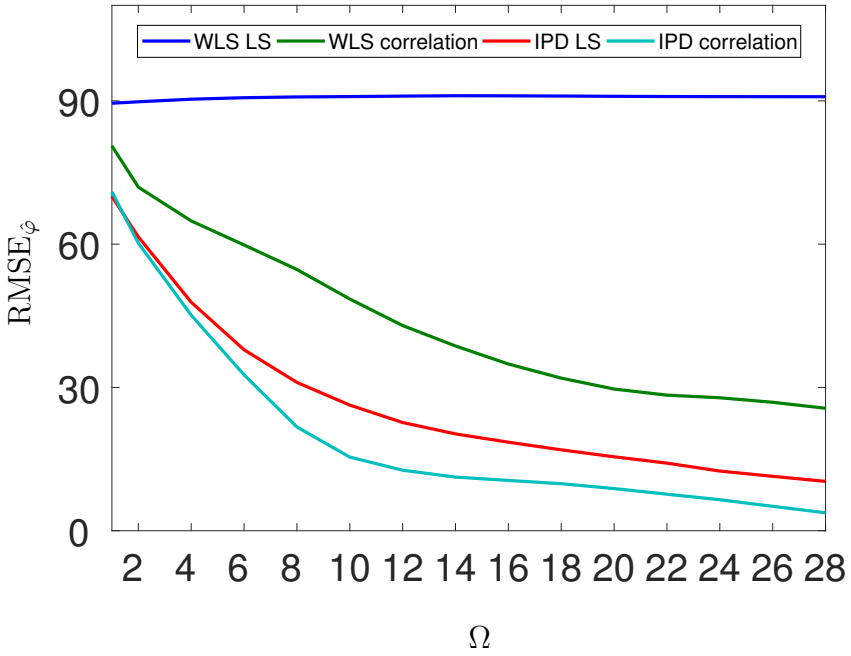

Figure 25: RMSE$_{\hat{\varphi}}$ values as a function of frequency buffer size Ω. Noise source was located at $\varphi = 50°$ and SNR = 6dB. The lowest error can be achieved by using the correlation similarity measure and interchannel phase difference (IPD) data.

7.2 Results with other signals

In Sec. 7.1 the different versions of the correlation method, explained in Sec. 6.2, were tested with a single noise source. Based on those results it was found that the use of wider frequency range is needed in the correlation-based method to produce accurate results. The use of this frequency buffer is a trade-off between frequency resolution and correct estimates. When using the frequency buffer, the correlation method with the interchannel phase difference data was found to produce the smallest root-mean-square error values and in this section it is used with frequency buffer size of $\Omega = 28$. This section presents the results when analyzing more realistic signals like speech. In addition to the correlation-based method, also results for the phase unwrapping ([9], explained in Sec. 5.2), envelope detection ([10], explained in Sec. 5.3) and parameter extrapolation (Sec. 6.1) methods are shown. As a upper reference, the estimates are also calculated using power weighted average of the true directions of arrival. For comparison, the estimates produced by the aliased Weighted Least Squares estimator are shown and they were discussed in more detail in Sec. 4.1. The estimates are first shown for plane wave simulations and after that also for room impulse response simulations. Only the frequencies above the aliasing frequency are discussed in the results as they are of main interest in this thesis.

The double talk scene simulation that has been used throughout the thesis, like in Sec. 3.2.1, was used for the comparison in the simulations. The used microphone array is the 4-microphone rectangular array shown in Fig. 3, with $r = 9$cm. The estimated directions of arrival with different methods are shown in Fig. 26 as color coded values as a function of time and frequency. The plot a) shows the power weighted true directions of arrival as a reference and b) the Weighted Least Squares estimates. The phase unwrapping results in c) show that the method works well when the single source assumption is satisfied, e.g., right before 2 seconds only the source at $\varphi = -92°$ is active, indicated by the blue color at all frequencies. However, when there are both sources active the unwrapping becomes more complicated and this method fails often. For example right before 4 seconds there are some completely wrong estimates. The extrapolation method in plot e) shows similar results as only one estimate is produced for the higher frequencies. The extrapolation fails also when both of the sources are active but the failure results in an estimate between the true source locations, i.e., light green color here.

The results of the envelope method show good reliability as most of the computed estimates are either of the correct directions of arrival. The very highest frequencies were not included in the estimation because they were left out of the highest 1/3-octave band filter. It seems that the areas where the blue coded source is strong in the reference, are also blue in the envelope estimates and the other areas are yellow. The same can be noticed for the interchannel phase difference correlation estimates in f). Based on these results the envelope and correlation-based methods yield the best results in this double talk scenario when using a plane wave simulation. The difference between these two is that the envelope method has fixed frequency bands

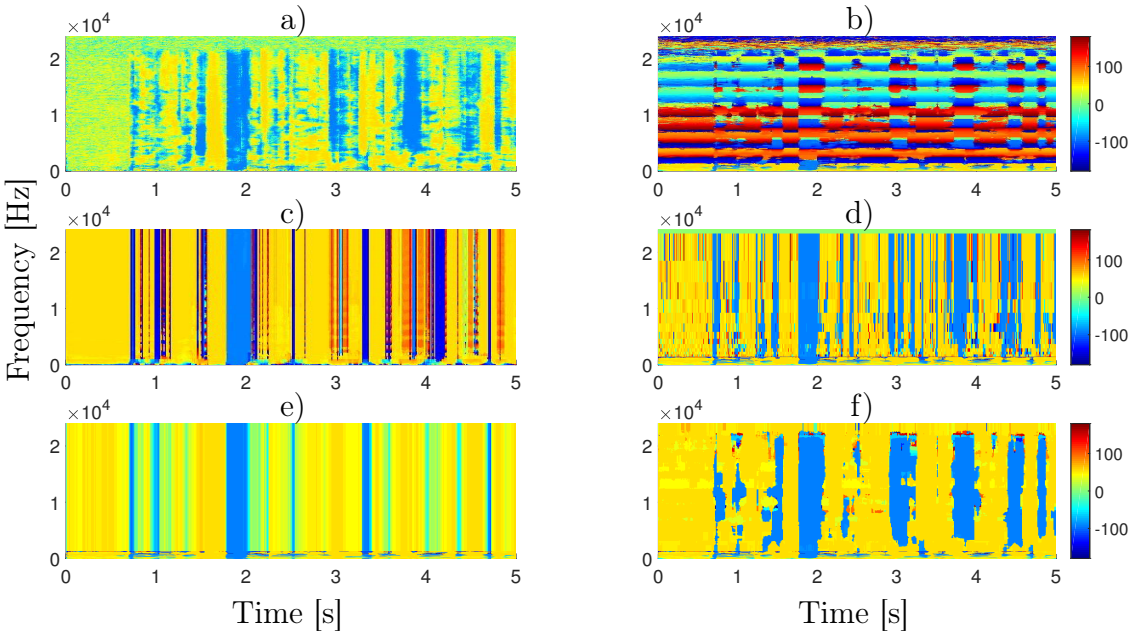

Figure 26: direction of arrival estimation with different methods using a plane wave simulation and SNR = 60dB. a) Power weighted true directions of arrival. b) Aliased Weighted Least Squares estimation. c) Phase unwrapping. d) Envelope detection. e) Extrapolation. f) interchannel phase difference correlation with $\Omega = 28$ bins.

inside which one estimate is produced. In the interchannel phase difference correlation the frequency buffer size is also fixed but this does not produce fixed frequency resolution in the estimates.

To obtain more realistic microphone signals for the direction of arrival estimation, a room impulse response simulation was made similarly to Sec. 4.2.1. Microphone array is the same as for the plane wave simulation above. The most important difference to the plane wave simulation is that there are reflections and late reverberation according to the reverberation time $T_{60} = 0.25$s. The plane wave assumption is still valid in this simulation for most of the frequencies. The results of this simulation are shown in Fig. 27 as color coded direction of arrival estimates as a function of time and frequency. The reference in plot a) has not changed when compared to the plane wave simulation. In b) it can be seen that the estimates of the Weighted Least Squares method have become noisier. This affects also the phase unwrapping and extrapolation results in c) and e), respectively. For both of these there are now more incorrect estimates. The methods still work when a single source is clear enough, i.e., right before 2 seconds mark, but the overall accuracy degradation is noticeable especially for the phase unwrapping. The accuracy degradation can be seen as there are more color codes other than the yellow and blue. The estimates with the envelope method have become significantly more noisy. The interchannel phase difference correlation results in f) seem to be closest to the reference, but also for that method there are now more wrong estimates than was the case with

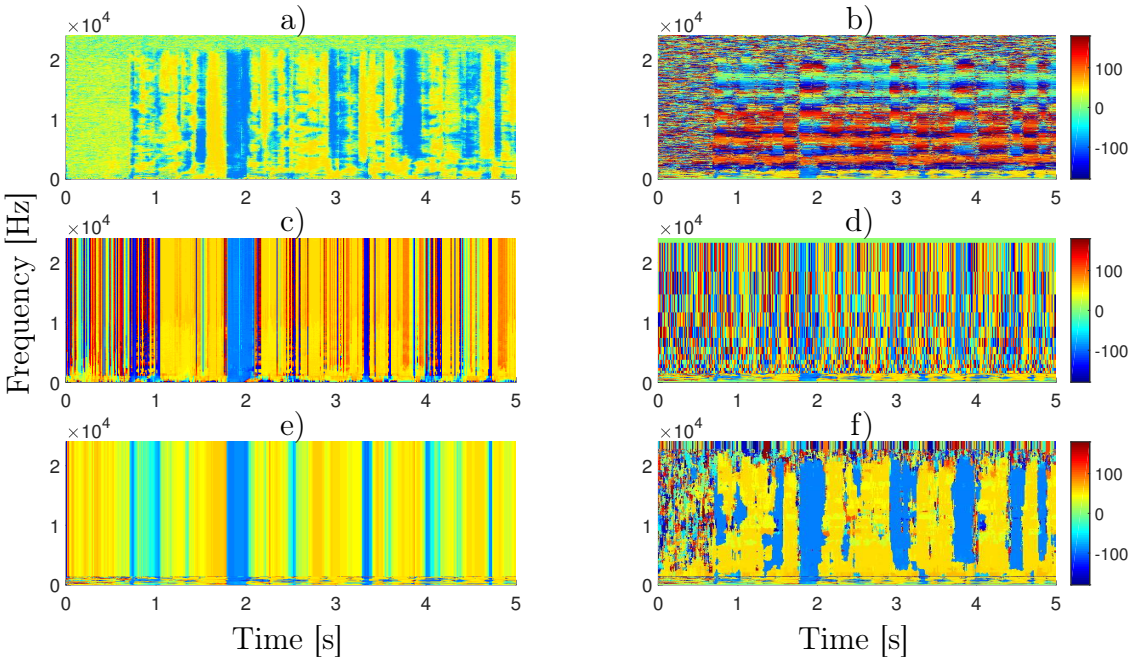

Figure 27: direction of arrival estimation with different methods using a room impulse response simulation and SNR = 60dB. a) Power weighted true directions of arrival. b) Aliased Weighted Least Squares estimation. c) Phase unwrapping. d) Envelope detection. e) Extrapolation. d) interchannel phase difference Correlation with frequency buffer size 28 bins.

the plane wave simulation. The results with the room impulse response simulation show that including the reflections and reverberation makes the direction of arrival estimation more prone to errors when using the methods here.

7.3 Listening test

The direction of arrival estimation accuracy of the phase unwrapping, envelope detection, extrapolation and interchannel phase difference correlation were compared in Sec. 7.2. In those results it was noticed that the extrapolation fails often when the single source assumption is not satisfied. The interchannel phase difference correlation showed promising results even with two active sources. Because these two methods were proposed in this thesis and there are no subjective results on their estimation performance, a listening test was organized.

The listening test was organized as a MUSHRA test [65], where the participants evaluated the accuracy and stability of the spatial image. The reference was the power weighted true directions of arrival. The lower anchor was created by having all the sources in a mono signal that was played by all the loudspeakers so that there is no localization of the sources. The compared methods were the Weighted

Least Squares, extrapolation and the interchannel phase difference correlation, all explained in Secs. 3.2, 6.1 and 6.2.2 respectively. The estimation was done by using microphone signals from a room impulse response simulation similarly to Sec. 4.2.2.

The first test item contained a single pink noise source that was high-pass filtered to contain only frequencies above 8kHz, the direction of arrival for this was $\varphi = 35°$. The second item ("BB" in results) had a bass guitar ($\varphi = -145°$) in the lower frequencies and a bird sound ($\varphi = 63°$) that only contained frequencies above the aliasing frequency. The third item ("DTB" in results) was a double talk scene ($\varphi = 35°$,$\varphi = -150°$) added with the same bird sound ($\varphi = -38°$) as in the second item. The direction of arrival estimates for the second item are shown in Fig. 28 as an example. Plots a)-d) show the color coded estimates for the reference, Weighted Least Squares, extrapolation and interchannel phase difference correlation methods respectively. It should be noticed that in plot c) the extrapolation uses practically only the direction of arrival estimates for the bass guitar (blue color). The diffuseness was estimated using (22) similarly to Sec. 4.2.2.

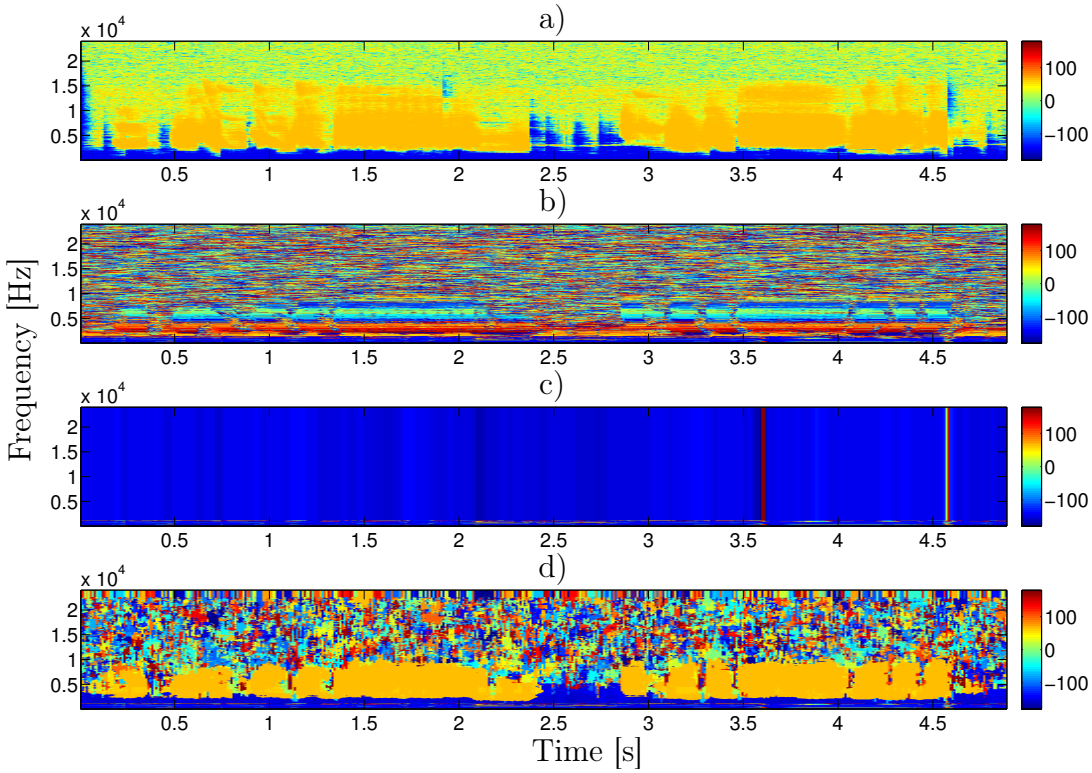

Figure 28: direction of arrival estimates for a room impulse response simulation with two sources; a bird $\varphi = 63°$ and bass guitar $\varphi = -145°$. The estimation methods were a) Power weighted true directions of arrival b) Weighted Least Squares c) extrapolation d) interchannel phase difference correlation. These direction of arrival estimates were used in the synthesis of the second listening test item.

After performing the estimation procedure for each item with all the methods, a DirAC synthesis was made for a 5.0 loudspeaker setup. In the test there were 8 expert listeners who were asked to grade the hidden reference to 100, the anchor to 0 and the rest of the items in between these values, based on their spatial image quality when compared to the reference.

The results including all participants are shown in Fig. 29. The middle point of each vertical bar shows the mean of the given grades and the height describes the 95% confidence interval. It can be seen that the interchannel phase difference correlation method (pink cross) achieved grades closest to the hidden reference (black line). In fact, it was even confused with the reference in some items by some listeners. The Weighted Least Squares estimator (red box) achieved the second best grades and the extrapolation (blue box) was slightly worse. The first test item ("Noise") with a single noise source was found confusing by the listeners. This is because, e.g., the localization of the Weighted Least Squares estimated version was also clear and its aliased direction of arrival was not too far from the reference direction of arrival.

The clearest difference between the compared methods is seen in the second item ("BB"), probably because it is easier for the listeners to evaluate the sound field and notice the failures that the Weighted Least Squares and extrapolation methods produce. In addition, only having two sources makes it easier to obtain correct estimates with the correlation method, so its grade is relatively high. In the third item ("DTB") the sound field becomes more complicated and this made the evaluation a bit more difficult but the same grading order of the methods can be seen. However, the total grades for all items show that the interchannel phase difference correlation led to a sound field perceptually closest to the reference. The grades for the Weighted Least Squares method are clearly lower than for the correlation method in all items. This means that the frequency resolution reduction is less of an issue than using the aliased direction of arrival estimates, as was assumed Sec. 6.2.2. Because in all items there were either multiple sources or then the source did not have energy below the f_a, the extrapolation received even worse grades than the Weighted Least Squares method.

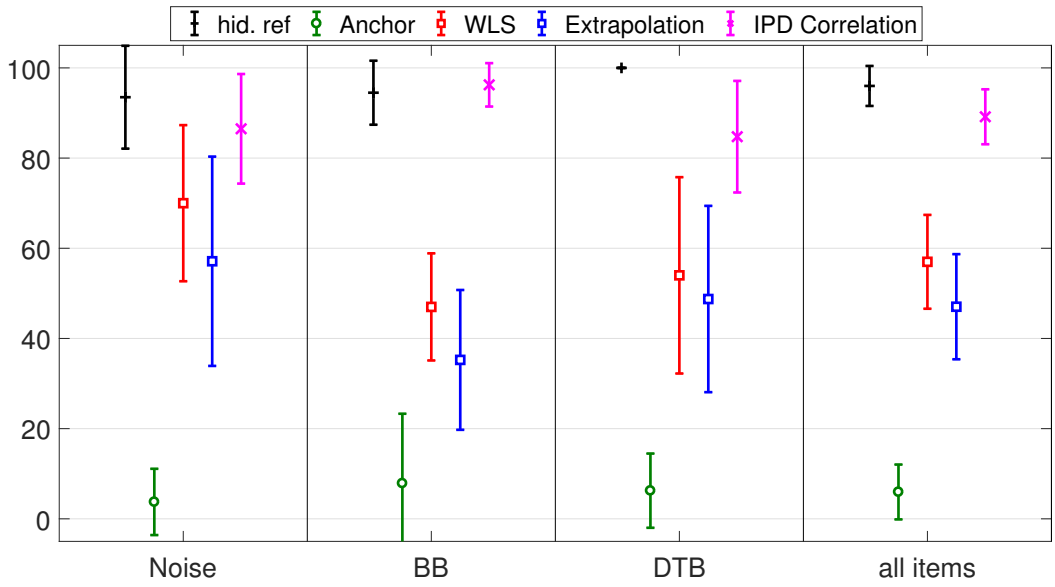

Figure 29: Listening test results expressed as vertical lines for each test item. Middle point of the line indicates the mean of the answers and the height of the line denotes the 95% confidence interval. The interchannel phase difference (IPD) correlation can achieve almost equivalent spatial image quality with the reference. Extrapolation did not produce good results because its broadband single source assumption was not satisfied.

8 Future work and conclusions

This section ends the thesis by providing a brief look on the possible future work and by concluding the work that was done.

8.1 Future work

The presented correlation-based method was found to reduce the negative effects of spatial aliasing problem. However, there are still many aspects that should be examined in more detail to make sure it is an applicable solution. First of all, the proposed version can not achieve the original time-frequency resolution when using the frequency buffer, thus making it similar to the envelope detection method. These two should be compared in a listening test to find out if there is a perceptual difference. The effect of reducing the size of the buffer should be examined in different scenarios to obtain a possible optimal size. In this work only a single array structure was used, the 4-microphone planar array. For this there are multiple different wrapping patterns because the orientations and distances of the microphone pairs are different. The assumption is that by adding more microphones the combined pattern would become even more unique and vice versa. It should be examined if there is an optimal array structure to be used in the correlation approach. In addition, when using a linear array, it should be studied if only the differences in microphone distances can provide the required uniqueness and with what frequency buffer size. The method was only developed to cover the azimuth angles but in the future also the elevation angle should be included in the analysis. This could be possible by creating the aliasing and wrapping patterns that include these angles also, producing one more dimension to the patterns. After these, the method should also be tested with real recordings to see how much the nonidealities, e.g., in the microphone placements affect the accuracy.

Another issue is the computational complexity of the correlation approach. Because of the large search space, finding the best match is computationally heavy. This becomes a problem especially when the elevation angle is included in the estimation and the search space increases. There are two things that could reduce the search space size. First, if the used array is symmetrical, the lowest correlation is found on the opposite direction of the correct one, as was seen in the Fig. 17. This feature can be utilized by only perfoming the correlation on half of the space. If the absolute value of the lowest correlation is higher than the highest correlation, it is known that the highest correlation in the whole space is on the opposite direction to what the lowest correlation point would indicate. The second solution to decrease the computational load would be to decrease the wrapping pattern resolution, e.g., to every 2°. This might still be a perceptually acceptable reduction. By combining the search space splitting and pattern resolution reduction, a significant reduction in computational complexity could be achieved. In addition, the correlation approach could be used similarly to the envelope method, i.e., for fixed frequency bands. Also the correlation could be only used for time-frequency bins that contain only the

direct sound, i.e., bins with low diffuseness.

The proposed extrapolation method was found to be an efficient solution that can produce good results when there is only a single source with energy below the f_a. If a detection method for these conditions is applied, the extrapolation could be used for single time instances instead of the correlation method to reduce the computational load of using only the correlation method.

8.2 Conclusions

This thesis has examined the spatial aliasing problem with direction of arrival estimation when using sparse microphone arrays of omnidirectional microphones. The problem was approached in time-frequency domain using Directional Audio Coding processing framework. When using narrowband state of the art direction of arrival estimators like Estimation of Signal Parameters via Rotational Invariance Techniques or Weighted Least Squares, spatial aliasing causes ambiguity for the direction of arrival estimates. It was noticed that the ambiguity leads to perceptually undesired effects in the synthesized sound field because of the wrong direction of arrival estimates. The current solutions to spatial aliasing, like phase unwrapping and envelope detection can decrease the amount of wrong estimates but with the cost of decreased frequency resolution of the direction of arrival estimates. In short, the contributions of this thesis were the development and testing of the extrapolation and wide-/narrowband correlation-based methods to reduce the negative effects of spatial aliasing. In the following, a short summary of each method is presented.

In the case there is only a single source signal, the decreased resolution is not a major problem because there is no need to separate the directions of arrival for different frequency bands. It was found that if the direction of arrival parameter can be estimated correctly below the spatial aliasing frequency, the aliased estimates can be discarded by extrapolating this correct estimate to the higher frequencies. This proposed extrapolation method is an efficient way to obtain correct estimates for the frequencies above the aliasing frequency without the aliasing effects. For the situation when there is no correct direction of arrival estimate below the aliasing frequency, a correlation-based method was developed. In this method the aliasing pattern of the microphone array is utilized. The aliasing pattern is the frequency, direction of arrival and microphone array dependent matrix of aliased direction of arrival estimates. This can be utilized by calculating the correlation between the aliasing pattern and the obtained aliased direction of arrival vectors for each frequency. In simulation tests the maximum of the summed correlation over all frequencies was noticed to point out a reliable broadband estimate. This estimate corresponds to the most dominant source even if all of the source's frequency content is above the aliasing frequency. In simulations the narrowband estimation with this method was not reliable because in the pattern there are ambiguities for each frequency, i.e., the correlation maximum does not point out only the true direction

of arrival when only considering one frequency bin. To reduce the ambiguities of the narrowband estimation, a 2-step estimation approach was suggested. In this 2-step correlation-based method the broadband correlation estimates are used as a priori information for the narrowband correlation.

When there are multiple sources, the estimation should be made separately for each frequency bin. For this case a correlation-based method utilizing the wrapping patterns of each microphone pair was developed. The wrapping pattern contains frequency, direction of arrival and microphone pair dependent interchannel phase difference values. When using this pattern in the correlation, the ambiguities are likely to be distributed differently for each microphone pair. Because of this, the summation of the correlations for each microphone pair was found to provide more reliable results than the use of the aliasing pattern. In low-noise plane wave simulations this method produced very low root-mean-square error values. However, the narrowband estimation is not possible for each frequency and direction of arrival under noisy conditions. To improve the robustness of the correlation, the use of a frequency buffer was introduced. This way more information around the inspected frequency bin can be included in the correlation. In simulations the use of the correlation-based method with the frequency buffer showed good reliability even with higher noise levels, reflections and reverberation.

The use of the frequency buffer causes reduced frequency resolution for the direction of arrival estimates. The perceptual effect of this reduction was also studied in this thesis. With objective and subjective measures, it was found that the reduction of the frequency resolution reduces also the perceptual quality of the spatial image. However, in an organized listening test, the correlation method with the frequency buffer achieved a major perceptual improvement when compared to using the aliased Weighted Least Squares estimator. Based on this, it can be stated that perceptually more favorable results can be achieved when accepting the frequency resolution reduction to achieve correct direction of arrival estimates.

References

[1] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*, Jun 2006.

[2] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, 2007.

[3] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," in *Audio Engineering Society Convention 50*, Mar 1975.

[4] E. Benjamin and T. Chen, "The native b-format microphone," in *Audio Engineering Society Convention 119*, Oct 2005.

[5] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 984–995, Jul 1989.

[6] O. Thiergart, W. Huang, and E. A. P. Habets, "A low complexity weighted least squares narrowband doa estimator for arbitrary array geometries," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 340–344, 2016.

[7] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Transactions on Signal Processing*, vol. 57, pp. 1383–1395, April 2009.

[8] K. Itoh, "Analysis of the phase unwrapping algorithm," *Appl. Opt.*, vol. 21, pp. 2470–2470, Jul 1982.

[9] K. Chen, J. T. Geiger, and W. Kellermann, "Robust audio localization with phase unwrapping," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 471–475, 2017.

[10] M. Kratschmer, O. Thiergart, and V. Pulkki, "Envelope-based spatial parameter estimation in directional audio coding," in *Audio Engineering Society Convention 133*, Oct 2012.

[11] L. L. Beranek and T. J. Mellow, *Acoustics*. Oxford: Academic Press, 2012.

[12] V. Abagnali, V. Abagnali, and G. Fabbri, *Sound waves: propagation, frequencies and effects*. Hauppauge, N.Y.: Nova Science Publishers, 2012.

[13] ISO 2533:1975. Standard Atmosphere. Switzerland: International Organization for Standardization. 1975. 108 pages.

[14] P. Filippi, *Acoustics : basic physics, theory, and methods*. San Diego: Academic Press, 1999.

[15] F. Jacobsen, "Sound intensity and its measurement". Proceedings of Fifth International Congress on Sound and Vibration. Auburn, AL: The International Institute of Acoustics and Vibration (IIAV). 1997.

[16] A. Politis and V. Pulkki, "Acoustic intensity, energy-density and diffuseness estimation in a directionally-constrained region," *CoRR*, vol. abs/1609.03409, 2016.

[17] A. Spanias, T. Painter, V. Atti, and J. Candy, *Audio Signal Processing and Coding*. Wiley, 2007.

[18] J. G. Proakis and D. G. Manolakis, *Digital signal processing : principles, algorithms and applications*. Upper Saddle River (NJ): Pearson/Prentice Hall, 4th ed., 2007.

[19] Kuttruff, H.: *Room Acoustics*. London: Taylor and Francis, 4th ed., 2000.

[20] O. Thiergart, G. Milano, T. Ascherl, and E. A. P. Habets, "Robust 3d sound capturing with planar microphone arrays using directional audio coding," in *Audio Engineering Society Convention 143*, Oct 2017.

[21] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–529–I–532, May 2002.

[22] F. Jacobsen, "The diffuse sound field," Report No. 27, Ph.D. dissertation, Technical University of Denmark. 1979.

[23] C.-H. Jeong, "Diffuse sound field: challenges and misconceptions," in *Internoise conference*, 2016.

[24] G. Del Galdo, J. Ahonen, M. Taseska, O. Thiergart, and V. Pulkki, "The diffuse sound field in energetic analysis," *Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2141–2151, 2012.

[25] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–65, 1940.

[26] B. Moore, *An Introduction to the Psychology of Hearing*. U.S.: Academic Press, 1989.

[27] G. von Békésy, *Experiments in hearing*. New York: Mcgraw Hill, 1960.

[28] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge (MA): MIT Press, rev. ed., 1997.

[29] Wallach, H., Newman, E. B., and Rosenzweig, M. R. The precedence effect in sound localization. The American Journal of Psychology LXII. 1949.

[30] V. Pulkki and M. Karjalainen, *Communication acoustics : an introduction to speech, audio, and psychoacoustics*. Chichester: Wiley, 2014.

[31] M. Kleiner, *Electroacoustics*. Boca Raton, FL: Taylor & Francis, 2013.

[32] O. Thiergart, M. Kratschmer, M. Kallinger, and G. Del Galdo, "Parameter estimation in directional audio coding using linear microphone arrays," in *Audio Engineering Society Convention 130*, May 2011.

[33] J. Ahonen, G. Del Galdo, M. Kallinger, F. Küch, V. Pulkki, and R. Schultz-Amling, "Planar microphone array processing for the analysis and reproduction of spatial audio using directional audio coding," in *Audio Engineering Society Convention 124*, May 2008.

[34] J. Eargle, *The microphone book*. Woburn, MA: Focal Press, 2001.

[35] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, 1973.

[36] C. Faller, Parametric coding of spatial audio. Ph.D. thesis, Lausanne, Switzerland: EPFL. 2004.

[37] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric time-frequency domain spatial audio*. Hoboken: Wiley, 2017.

[38] V. Pulkki and J. Merimaa, "Spatial impulse response rendering ii: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc*, vol. 54, no. 1/2, pp. 3–20, 2006.

[39] C. Faller and V. Pulkki, "Directional audio coding: Filterbank and stft-based design," in *Audio Engineering Society Convention 120*, May 2006.

[40] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and b-format microphone array for directional audio coding," in *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*, Mar 2007.

[41] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[42] F. Fahy, *Sound Intensity*. CRC Press, rev. 2 ed., 2002.

[43] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 285–288, 2009.

[44] J. Ahonen, G. Del Galdo, M. Kallinger, F. Küch, V. Pulkki, and R. Schultz-Amling, "Analysis and adjustment of planar microphone arrays for application in directional audio coding," in *Audio Engineering Society Convention 124*, May 2008.

[45] A. Politis, S. Delikaris-Manias, and V. Pulkki, "Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6–10, April 2015.

[46] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, p. 026503, May 2006.

[47] A. A. Khan, *Digital signal processing fundamentals*. Hingham, MA: Da Vinci Engineering Press, 2005.

[48] Y. Bar-Shalom, F. Palimieri, A. Kumar, and H. M. Shertukde, "Analysis of wide-band cross correlation for time-delay estimation," *IEEE Transactions on Signal Processing*, vol. 41, pp. 385–, Jan 1993.

[49] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Transactions on Signal Processing*, vol. 41, pp. 525–533, Feb 1993.

[50] R. Cusani, "Performance of fast time delay estimators," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 757–759, May 1989.

[51] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, Aug 1976.

[52] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, pp. 1497–1498, Oct 1973.

[53] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE Spectrum*, vol. 8, pp. 62–70, April 1971.

[54] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 187–190 vol.1, Apr 1997.

[55] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. Englewood Cliffs (NJ): Society for Industrial and Applied Mathematics, 1974.

[56] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, Mar 1986.

[57] A. J. Weiss and B. Friedlander, "Effects of modeling errors on the resolution threshold of the music algorithm," *IEEE Transactions on Signal Processing*, vol. 42, pp. 1519–1526, Jun 1994.

[58] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, pp. 336–339, Apr 1983.

[59] A. Mhamdi and A. Samet, "Direction of arrival estimation for nonuniform linear antenna," in *2011 International Conference on Communications, Computing and Control Applications (CCCA)*, pp. 1–5, March 2011.

[60] O. A. Oumar, M. F. Siyau, and T. P. Sattar, "Comparison between MUSIC and ESPRIT direction of arrival estimation algorithms for wireless communication systems," in *The First International Conference on Future Generation Communication Technologies*, pp. 99–103, Dec 2012.

[61] K. Chen, J. T. Geiger, W. Jin, M. Taghizadeh, and W. Kellermann, "Robust phase replication method for spatial aliasing problem in multiple sound sources localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.

[62] A. Baldi, F. Bertolino, F. Ginesu. Phase Unwrapping Algorithms: A Comparison. In: Jacquot P., Fournier JM. (eds) Interferometry in Speckle Light. Springer, Berlin, Heidelberg. 2000.

[63] K. Johnson, *Acoustic and Auditory Phonetics*. Wiley, 2003.

[64] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Englewood Cliffs (N.J.): PTR Prentice Hall, 1993.

[65] ITU, "ITU-R Rec. BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[66] M. S. Amin, Ahmed-Ur-Rahman, Saabah-Bin-Mahbub, K. I. Ahmed, and Z. R. Chowdhury, "Estimation of direction of arrival (doa) using real-time array signal processing," in *2008 International Conference on Electrical and Computer Engineering*, pp. 422–427, Dec 2008.

[67] V. V. Reddy, A. W. H. Khong, and B. P. Ng, "Unambiguous speech doa estimation under spatial aliasing conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 2133–2145, Dec 2014.

[68] G. Del Galdo, O. Thiergart, F. Kuech, M. Taseska, and D. Sishtla, "Optimized parameter estimation in directional audio coding using nested microphone arrays," in *Audio Engineering Society Convention 127*, Oct 2009.

[69] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of different microphone array configurations for hands-free speech recognition in noisy and reverberant environment," Feb 2000.

[70] R. Krämer and O. Loffeld, "Presentation of an Improved Phase Unwrapping Algorithm Based on Kalman Filters Combined with Local Slope Estimation," in *ERS SAR Interferometry* (T. D. Guyenne and D. Danesy, eds.), vol. 406 of *ESA Special Publication*, p. 253, Mar 1997.

[71] V. V. Reddy and A. W. H. Khong, "Direction-of-arrival estimation of speech sources under aliasing conditions," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, Apr 2015.

[72] L. Marple, "Computing the discrete-time ldquo;analytic rdquo; signal via fft," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2600–2603, Sep 1999.

[73] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, Upper Saddle River, NJ, USA: Prentice Hall Press, 3rd edition. 2009.

[74] R. Fisher, *Statistical Methods for Research Workers*. Biological monographs and manuals, Hafner, 1958.

[75] Correlation coefficients, Matlab documentation. The MathWorks Inc. 2018. Available on:https://se.mathworks.com/help/matlab/ref/corrcoef.html. Cited: 27.2.2018.