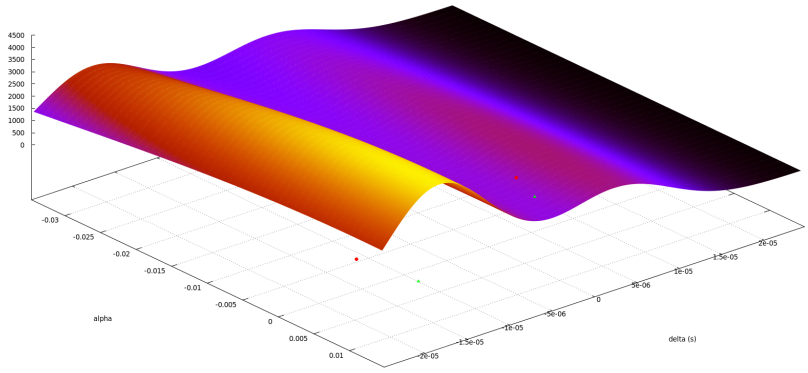
TÉCNICO LISBOA



Real time audio system for SNR improvement with applications to distant speech recognition

Alberto Eduardo Artilheiro Ferreira

Thesis to obtain the Master of Science Degree in

Engineering Physics

Supervisor(s): Professor Diogo Gonçalo Franco Falcão Osório de Alarcão
Professor Pedro Miguel Félix Brogueira

Examination Committee

Chairperson:	Prof. Horácio João Matos Fernandes
Supervisor:	Prof. Diogo Gonçalo Franco Falcão Osório de Alarcão
Member of the Committee:	Prof. João Luís Maia Figueirinhas
Member of the Committee:	Prof. Octávio José Patrício Fernandes Inácio

November 2014

This thesis is dedicated to the memory of my beloved grandfather Alberto Eduardo Ferreira, whose humbleness and wisdom I hope to attain some day.

Acknowledgments

I would like to thank all my family and friends for their wonderful help and motivation not only for this thesis but throughout the years. I would like to thank my parents for their neverending support, help and patience, but whom I worried too much at less blissful times.

I would like to express my special gratitude to my coordinator who lended me equipment, gave me access to the the anechoic chamber and was friendly, supportive, patient, forgiving, attentive and very helpful.

I would also like to thank in particular the teacher Joel Paulo and my friend António Dias for voluntarily borrowing me microphones without which I couldn't have performed the tests. I would like to thank to my friends Beatriz Ferro, Helena Reis, Inês Amaro, Pedro Neves and António Dias for lending me their voices so that I could create the test database.

Resumo

Nesta tese estudou-se o método de separação de fontes DUET de forma a avaliar o seu potencial para vir a funcionar como interface de processamento de sinais de fala em tempo-real de forma a alimentar software de reconhecimento automático de fala e assim obter um sistema de reconhecimento de fala à distância em ambientes com pouco eco e mais que um locutor.

A dispersão dos estimadores de localização no espaço binaural foi estudada e explicada pela primeira vez e a hipótese de estacionariedade local do DUET refutada para sons plosivos e sibilantes. O conhecimento adquirido sobre os factores de erro no espaço binaural permitem uma mais robusta localização dos picos, que é o ponto fulcral do DUET. Este conhecimento é extensível a outros métodos. Novas métricas de confiança e distância foram desenvolvidas para a identificação dos picos em tempo real que dispõe de pouca estatística. Um sistema de *streaming* em tempo real para processamento em paralelo de múltiplas fontes foi desenvolvido.

Palavras-chave: Separação de fontes, DUET, Reconhecimento de Fala, Processamento de áudio, Processamento de Sinal.

Abstract

In this thesis the DUET blind source separation method was studied in order to assess its potential as a real-time front-end for an Automatic Speech Recognition (ASR) engine with which a distant speech recognition engine would work in low reverberation environments with more than one active speaker.

The dispersion of the source position parameters in the binaural space was studied and explained for the first time and the local stationarity hypothesis of DUET refuted for plosive and sibilant speech sounds. The gained knowledge of the error factors in the binaural space allows a more robust peak tracking that is long known to be the crux of DUET. This knowledge is extensible to other methods. New confidence and distance metrics were developed for the peak identification in real-time which has low statistics. A real-time source streaming system for parallel multi-source processing was developed.

Keywords: Blind Source Separation (BSS), DUET, Automatic Speech Recognition (ASR), Audio Processing, Signal Processing.

Contents

Acknowledgments v
Resumo vii
Abstract ix
List of Tables xv
List of Figures xix
Nomenclature xxi
Glossary xxiii

1 Introduction **1**
1.1 Introduction 1
1.2 Motivation 3

2 Basic Concepts **4**
2.1 Linear Acoustics 4
2.1.1 Acoustic wave equation 4
2.1.2 Harmonic plane waves 5
2.1.3 Spherical waves 5
2.1.4 Acoustic intensity 6
2.1.5 Sound levels and Decibel scales 7
2.2 Speech production 8
2.3 Sound perception 8
2.3.1 Human ear 8
2.3.2 Auditory cues for source localization 9
2.3.3 Auditory Masking 9
2.3.4 Ear integration time and hearing frequencies 10
2.4 Signal processing 10
2.4.1 Fourier Transform 10
2.4.2 Discrete Fourier Transform 11
2.4.3 Short-Time Fourier Transform 11
2.4.4 Convolution 12
2.4.4.1 High-speed convolution 12
2.4.5 Cross-correlation 12
2.4.6 Autocorrelation 13
2.5 Window functions 13
2.6 Source Separation 14
2.6.1 Mixing models 14

2.6.2 W-disjoint orthogonality . . . 15
2.6.3 Source separation metrics . . . 15
2.6.3.1 SNR . . . 16
2.6.3.2 Total relative distortion (TRD) . . . 16
2.6.3.3 SDR, SIR and SAR . . . 17

3 State-of-the-art **18**
3.1 Beamforming . . . 19
3.2 Independent Component Analysis (ICA) . . . 19
3.3 TRINICON . . . 21
3.4 Nonnegative Matrix Factorization (NMF) . . . 21
3.5 Computational Auditory Scene Analysis (CASA) . . . 22
3.5.1 Inter-Channel Time, Level and Phase-based methods . . . 23
3.5.1.1 Azimuth Discrimination and Resynthesis (ADRess) . . . 23
3.5.1.2 DUET . . . 23
3.5.2 Other methods . . . 25
3.6 Auxiliary techniques . . . 26
3.6.1 Visual cues . . . 26
3.6.2 Source localization/counting . . . 26
3.6.3 Speech cleaning . . . 26
3.6.3.1 Spectral subtraction . . . 27
3.6.3.2 Wiener filtering . . . 27
3.6.3.3 Dereverberation . . . 28
3.7 Summary . . . 30

4 Development and System Implementation **31**
4.1 DUET fundamentals . . . 31
4.1.1 W-disjoint orthogonality . . . 31
4.1.2 Local stationarity . . . 32
4.1.3 Separation principle . . . 32
4.1.4 Maximum microphone distance . . . 33
4.1.5 Building the histogram . . . 33
4.1.6 Building the masks . . . 33
4.1.7 Applying the masks and rebuilding the signals . . . 34
4.2 DUET Λ space . . . 35
4.2.1 Coordinate system . . . 35
4.2.2 Λ space range . . . 35
4.2.2.1 a-range . . . 35
4.2.2.2 Typical α-range . . . 37
4.2.2.3 δ-range . . . 37

4.2.3 Phase aliasing analysis . . . 38
4.3 DUET implementation/modifications . . . 39
4.3.1 Two-dimensional Λ histogram vs. Λ histogram marginals . . . 39
4.3.2 Heuristic Clustering . . . 39
4.3.3 Λ-Distance metrics . . . 41
4.3.4 Confidence metrics . . . 41
4.3.4.1 Local confidence metrics . . . 41
4.3.4.2 Global/Coordinate confidence metrics . . . 44
4.3.5 Source streaming algorithm . . . 44
4.4 Generic implementation details . . . 45
4.4.1 Adding the inverse transforms back together . . . 45
4.4.2 Source separation degeneracy . . . 46
4.4.3 Source separation metrics changes . . . 46

5 Tests and Results **47**
5.1 Simulation . . . 47
5.2 Experimental setup . . . 48
5.2.1 Sound database creation . . . 48
5.2.2 Recording procedure . . . 49
5.2.2.1 Signal conditioning . . . 51
5.2.2.2 Inter-channel-frequency-gain . . . 51
5.2.2.3 HPF cutoff frequency . . . 52
5.3 Preliminary tests and results . . . 53
5.3.1 Two-dimensional Λ histogram . . . 53
5.3.2 Histogram Λ-dependent variability . . . 54
5.3.3 Λ tails and spurious peaks . . . 55
5.3.3.1 $\hat{\delta}$ tails and bimodality . . . 55
5.3.3.2 $\hat{\alpha}$ spurious high-frequency peaks . . . 56
5.3.4 Reflections by a wall . . . 57
5.3.5 Kurtosis tests . . . 58
5.3.6 Windows . . . 59
5.3.6.1 Window types . . . 59
5.3.6.2 Window size . . . 62
5.3.7 Real-time streaming system . . . 63
5.4 Experimental results . . . 64
5.4.1 Anechoic chamber tests . . . 64
5.4.1.1 (p,q) choice . . . 64
5.4.1.2 Separation tests for central target . . . 65
5.4.1.3 Separation tests for central target and three interfering sources . . . 66

5.4.2 Living room tests . . . 67
5.4.2.1 Two human speakers . . . 67
5.4.2.2 Pink noise interferer . . . 68

6 Conclusions **69**
6.0.2.3 Experimental implementation concerns . . . 70
6.1 Main Achievements . . . 70
6.2 Future Work . . . 71

References **80**

A Figures **81**

B Movement in the $\alpha - \delta$ plane **87**
B.0.0.4 δ-variations . . . 87
B.0.1 Mappings $\Lambda \rightarrow (d_{n1}, d_{n2}) \rightarrow (x, R)$ and $\Lambda \rightarrow (x, R)$ . . . 91
B.0.2 $|\dot{\delta_M}|$ . . . 92

List of Tables

4.1 Behaviour of a and α. 37

5.1 Over-degeneracy count for the anechoic window tests at 1m where $l = low$, $m = mid$, $w. = weak$, $wkr. = weaker$ refer to the variants of the Hamming window. 61

5.2 Separation performance for the dynamic test. 64

5.3 Separation results for two people speaking naturally at around -30° and 30°. SNR_0 shows the SNR value prior to applying the separation and SNR_{IBM} the SNR value obtained using the ideal binary masks. The first test had the speakers at $1\,m$ from the microphones and the second at $1.9\,m$. 68

5.4 Separation results for a person at 30° and a pink noise source at -30°, both at $d \approx 1\,m$. 68

List of Figures

4.1 Coordinate system where the microphones are represented in red and the source in blue. 35

5.1 Straightforward experimental setup showing the omni-directional microphones and simple position measurement through a thread attached to the microphone stand that is compared against one of the lines marked radially in the CD-ROM. 50

5.2 Positions of the recorded sources in the anechoic room with the microphone axis as origin (labelled) and corrected for the coordinate system where the origin is at the center of the microphones. 50

5.3 Measured attenuation for several frequencies with the loudspeaker placed at $1\,m$, 0° . . . 51

5.4 Separation performance for a central target with varying HPF cutoff frequency. 52

5.5 Source placed at $(0.6, 0.2, 0.0)$. In left-to-right reading order, the cumulative histograms of $w = Hann$, $Hamming$, $Hamming_0$ and $Rectangular$. 54

5.6 Source placed at $\mathbf{Z_R}$ with Dtotal=0.0104297 and SNR=19.851dB. 55

5.7 Source placed at $\mathbf{M_R}$ (left) with Dtotal=0.0091 and SNR=20.44dB and source placed at $\mathbf{F_R}$ (right) and Dtotal=0.00874 SNR=20.61dB 55

5.8 Λ histogram for the file M1c1 recorded in the anechoic chamber at position $d = 1\,m$ and 30° (top and bottom right) and a tone at the same position using the simulation (bottom left). The top left image uses a smoothing of $(0.03, 3\,\mu s)$ and the others use $(0.02, 2\mu s)$. The bottom right image has a high-pass filter applied at 400Hz. (p,q) was set to (1,0) on all tests. 56

5.9 Λ histogram for the file M1c1 recorded in the anechoic chamber at position $d = 1\,m$ and 30° (top and bottom right) using the information from a single analysis window of $64\,ms$ where an "s" sound occurred. The smoothing $(0.03, 3\,\mu s)$ was used and (p,q) was set to (1,0). 57

5.10 Λ space resulting in a second peak for closer sources. The image on the left shows the source placed farther apart from the wall ($\theta = 0^o$) and the one on the right, the source closer to the wall ($\theta = 30^o$) with an evident reflection image. 57

5.11 Simulation of the signal F2c2 at $(-0.8, 1.0, 0.0)\,m$ and an image of equal intensity at $(0.8, 1.0, 0.0)\,m$ on the left. Same simulation but with the image moved to $(0.4, 1.0, 0.0)\,m$. 58

5.12 Tests for a rectangular window (left) and Hamming window (right) with several sizes. . . 62

5.13 Number of streams produced in function of accumulated windows information used for each localization decision where the whole data is represented by 303 windows. An additional curve is plotted showing that the streaming system follows a $1/N_{accum}$ tendency. 64

5.14 Separation stats for the central target source with DUET and IBM separation stats. Test performed with 3 sources in the anechoic room at $d = 1m$. 66

5.15 Separation stats for the central target source with DUET and IBM separation stats. Test performed with 3 sources in the anechoic room at $d = 2.5m$. 67

5.16 Separation performed in a bedroom with the same experimental setup as the anechoic case. The mixture features two independent speakers at approximately symmetric $\pm 30^o$ angles. $d \approx 1\,m$ (left) and $d \approx 1.9\,m$ (right). 67

5.17 Source at $d = 1\,m$, 30° and interfering pink noise source at $1\,m$ and -45°. On the left the Λ histogram is shown without spectral subtraction applied. On the right, the iterative spectral subtraction from [Miyazaki et al., 2014] was applied. 68

A.1 The left channel of the two source signals are first shown, followed by the left channel of their mixture (left channel input), shown in grey. Next the blue waveforms are the stream outputs for $N_{accum} = 15$. Finally, the individual sources and mixture are shown again in grey and finally the streams output by the system with $N_{accum} = 33$, the limit at which only two streams (as many as input sources) are estimated. 81

A.2 Separation metrics for the anechoic simulation using standard window types. The target source is placed at 0° and $d_{target} = d_{interferer} = 1\,m$. 82

A.3 Separation metrics for the anechoic simulation using standard window types. The target source is placed at 0° and $d_{target} = d_{interferer} = 2.5\,m$ 82

A.4 Separation metrics for the Hann, Hamming and custom weak and weaker Hamming window types using simulated data with sources at 1m. 83

A.5 Separation metrics for the Hann, Hamming and custom weak and weaker Hamming window types using simulated data with sources at 2.5m. 83

A.6 Separation metrics for the anechoic chamber using different window types. The target source is placed at 0°, 1m away from the microphones, as the interfering source. 84

A.7 Separation metrics for the Hann, Hamming and custom weak and weaker Hamming window types using anechoic chamber data with sources at 1m. 84

A.8 Frame where spurious peaks appear and all are rejected but the true one which is the second tallest peak. On the left the highlighted purple numbers show the only peak that passed the kurtosis test was the second most tall which is the right one. On the right the 2D histogram in Λ space of that frame is shown. 85

A.9 Separation stats for the central target source with the DUET (left) and IBM separation stats for the same condition. Test performed with 2 sources in the anechoic room at $d = 1m$. 85

A.10 Separation stats for the central target source with the DUET (left) and IBM separation stats for the same condition. Test performed with 2 sources in the anechoic room at $d = 2.5m$. 86

A.11 Over-degeneracy counts over individual tests with two sources in the anechoic room for 1m (above) and 2.5m (below). Green bars show the number of tests for an angle where no over-degeneracies occurred, yellow, the number of tests where one over-degeneracy occurred and red, the number of tests where the number of over-degeneracies was equal to two. 86

B.1 Plot of $|\max \dot{\delta_n}|(x, R)$ on the left and maximum relative δ motion on the right during a $64\,ms$ interval, with $v_M = 1m/s$, $c = 340m/s$ and $\Delta = 2\,cm$. 94

Nomenclature

s Sound source signal in the time domain.

T Wave Period.

x Mixture signal in the time domain.

x and X Signal x represented in the time or frequency domain respectively.

Greek symbols

ρ Density.

Roman symbols

p Pressure.

$\mathbf{u}$ Velocity vector.

u, v, w Velocity Cartesian components.

Subscripts

i, j, k, n Indexes.

x, y, z Cartesian components.

Superscripts

$*$ Adjoint.

T Transpose.

Glossary

(M)ADRess	(Modified) Azimuth Discrimination and Resynthesis
ASA	Auditory Scene Analysis
ASR	Automatic Speech Recognition. System that recognizes spoken human words.
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
DSR	Distant Speech Recognition. System capable of performing ASR when sources are far away.
DUET	Degenerate Unmixing Estimation Technique
EQ	Equalizer
FA	Factor Analysis
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
IFA	Independent Factor Analysis
ISA	Independent Subspace Analysis
ISS	Informed Source Separation
MIMO	Multiple-Input Multiple-Output
MOG	Mixture of Gaussians
NMF	Nonnegative Matrix Factorization
SBSS	Semi-Blind Source Separation
STFT	Short-Time Fourier Transform
T-F	Time-Frequency
TQOS	True Quality Oriented System
VAD	Voice Activity Detector
WDO	W-disjoint orthogonal(ity)
i.i.d.	Independent and identically distributed
pdf	Probability density function

Chapter 1

Introduction

1.1 Introduction

The objective of this work is to develop and implement a real-time system capable of improving the Signal to Noise Ratio (SNR) of speech signals by selecting the desired sound source and eliminating as much as possible the others in order to work as a front-end to an Automatic Speech Recognition (ASR) engine.

This domain is called Audio Source Separation (ASS/SS). There are several approaches in this domain in increasing order of the information assumed (and thus scenario specificity): Blind (Audio) Source Separation (BASS/BSS), Semi-Blind Audio Source Separation (SBASS/SBSS) and Informed Source Separation (ISS).

There is no truly "blind" method as something has always to be assumed, thus the names BSS and SBSS are fairly interchangeable. ISS though is more distinct because some additional information exterior to the system is given. Additional information restricts the degrees of freedom of the system which in turn facilitates improvements to its separation capability.

Techniques akin to Spectral Subtraction [Choi, 2004; Romanin et al., 2009] can also be useful as a possible front-end to source separation techniques. One possible line of work can be the real-time integration of Spectral Subtraction and ASS so as to maximize the separation quality given the appropriate environments.

In addition, more techniques do exist that can be employed to improve the SNR of recordings in general if we're given the chance of designing the recording environment ourselves, instead of being given only the recorded mixtures. One such technique is beamforming, in which one uses a set of microphones (usually Linear Microphone Arrays) and proper delays between them to spatially select sounds. The selection is never perfect of course, but the targeted signals are boosted and the others attenuated, thus improving the SNR of the desired signals. There are several beamforming techniques.

Intent of source separation

Source separation can be Audio Quality Oriented or Significance Oriented.

In the former and more demanding case the goal is high quality sound unmixing, usually for later human listening, whilst the latter is done to facilitate accurate information extraction of specific features such as words without paying attention to tone changes, diction, *et cetera*.

Speech Recognition System

The noisy speech is first presented to the ASR *front-end*. There it is broken into very small overlapping segments tipically 10-100ms in length that are quasi-stationary and from which the phonemes will be computed [J.Rennie, 2008].

The denoising component of the front-end usually has a speech detector module, a noise estimation module and a noise removal module.

The speech detector is usually energy-based while the noise estimation module uses frames that were classified as speech-free and from those noise is estimated. Then, the noise removal module uses the noise estimate to clean the speech by using spectral subtraction or Wiener filtering.

The cleaned speech frames are fed to the feature labelling system which evaluates the acoustic model (usually a Gaussian Mixture Model (GMM)) of each speech state given some informative prior based on training data or past states. These labels will then be passed to the system decoder and combined with phonetic, syntactic and grammatical constraints embedded in the language model to compute the speech state posterior sequences, thus converting speech to text.

Usually speech models in the ASR systems are Hidden Markov Models (HMM) and the state sequence search is performed through the Viterbi algorithm.

The HMM models for noisy speech could be trained directly but training data would be exhausted before enough statistics for all combinations of noise with speech were collected. Instead, a clean speech model is trained from the output of a front-end that performs speech cleaning (noise filtering and source selection) (see for example [J.Rennie, 2008][3,4,6,7][1]).

For distances of around 1m and above one enters the most difficult domain of speech recognition, i.e., Distant Speech Recognition (DSR) where using a simple ASR engine is not enough. As distance to the target speaker increases, other effects must be eliminated by the front-end, as SNR can be significantly lowered by interfering sources, reverberation, head orientation and room modes [Wölfel and McDonough, 2009].

In summary, a DSR engine shall contain an ASR back-end and in the front-end, a source separation stage that shall separate the target from interfering sources and perform correction of acoustic distortions.

Sometimes one experiences ridiculous outputs from speech recognition engines. This stems from the fact that the ASR engine or ASR back-end in a DSR engine can assume that competing noise is not dominant versus the desired signal. When the assumption doesn't hold the target speech signal and interferences can yield wrong phonemes which after parsing in the grammatical engine result in different words than expected.

In this thesis the use of a blind source separation technique called DUET for realtime blind source separation and streaming to an ASR system shall be studied and the intended usage is in a living room to give remote voice commands to devices such as the TV or stereo.

[1] This citing format is used when the reader might find the sources useful although they are not of utmost relevance to the discussion at hand. In those cases the first citation is the paper where such claim was made and following it, the references from that very paper the reader should consult.

1.2 Motivation

The motivation for this work derives from the development of a front-end of an ASR system that would possibilitate issuing voice commands for controlling a media set, such as a TV, in a living room with more than one person present and devices which emit sound such as the media set itself, or a radio or other noise sources.

The system should be cheap, small, non-intrusive and reliable. This means the system should work in real-time with low latency and should possess no more than two simple microphones close to each other.

Another goal is to make the system capable at very low SNR values as well. This is useful for instance if the media set is in a loud volume setting and one still wishes to issue voice commands without straining the voice.

As early as 2005, there were already many of such projects, though most used a handheld command with integrated microphone and a press-to-talk button: Accenda, inVoca, PoGo VRC-400 or AgileTV Promptu. At the time the only system that provided hand-free voice command inputs at a distance was the VoiceMe [consortium, 2005].

Distant-talking Interfaces for Control of Interactive TV (DICIT) was an european funded project that lasted from 2006 to 2009 where 5 000 000€ were invested. The final product was a working prototype that required a linear microphone array of 13 microphones spread over 1.92m and two additional microphones placed also 1.92m apart and 32cm above the other array [consortium, 2009]. Although the results seem to be good it is not cheap or compact. Having a low-cost and more discrete alternative would be desirable.

The economic potential for such techonolgy is large. By [consortium, 2005], the economic revenue of the Digital Video Recorder segment belonging to Set-Top Boxes, where voice search and commands would be extremely useful, was projected to reach $14,000,000,000$ by 2009 and keep growing.

In Portugal alone for instance, as of June of 2014, the TV service provider with more clients is NOS, with a total of 1.4 million clients (46.4% of the whole market segment), followed by MEO with 1.3 million clients (41.6% of the whole market segment) [Ribeiro, 2014]. This numbers mean the total number of clients of this kind of service was close to 3.4 million clients and with both companies providing a minimum monthly subscription of 29.99€ and 35€ for NOS and MEO respectively the baseline combined yearly gross earnings are of around 1050 million €.

The domain of application of such technology naturally extends not only to today's TVs which are becoming more and more like media centers that support voice and gesture commands but also to smart home appliances or mobile devices if the system is compact enough as DUET is.

There are several companies studying DSR right now, as Samsung for their SmartTVs, although they started with demos of voice commands performed at a distance and in the final commercial deployment regressed to the handheld remote with inbuilt microphone and press-to-talk button which is proof that this problem is far from solved yet.

Chapter 2

Basic Concepts

2.1 Linear Acoustics

Acoustic waves in inviscid fluids are longitudinal mechanical waves. Fluid particles move back and forth in the direction of propagation of the sound wave creating compression and rarefaction regions. The only restoring force responsible for propagating a wave in inviscid fluids is the pressure change caused by the fluid compressions and expansions.

The two main quantities needed to describe sound waves are the acoustic pressure, p in Pascal, and particle velocity $\vec{u}$, in m/s.

The acoustic pressure p is defined as

$$p = \mathscr{P} - \mathscr{P}_0 \tag{2.1}$$

where $\mathscr{P}$ is the instantaneous pressure and $\mathscr{P}_0$ is the constant equilibrium pressure in the fluid.

The particle velocity $\vec{u}$ is defined as the time derivative of the position vector of the particle's center of mass.

A fluid particle means a volume element large enough to contain millions of molecules so that the fluid may be thought of as a continuous medium, yet small enough that all acoustic variables may be considered nearly constant throughout the volume element.

Assuming that the fluid is homogeneous, isotropic and perfectly elastic, that no external forces act on it, and neglecting viscous energy losses then, considering waves of relatively small amplitude so that changes in density of the medium are small when compared to the equilibrium value, acoustic disturbances can be described resorting only to linear equations.

2.1.1 Acoustic wave equation

The linearized lossless wave equation for the propagation of sound in air is

$$\nabla^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{2.2}$$

where $c^2 = \left(\frac{\partial p}{\partial \rho}\right)_0$ is the speed of sound or the phase speed for acoustic waves and ρ the density of the air, which equals $1.21 kg/m^3$ at a temperature of $20^o C$ and a reference atmospheric pressure of $1\, atm$ corresponding to $1.013 \times 10^5\, Pa$.

2.1.2 Harmonic plane waves

If all the acoustic variables depend on only one spatial coordinate, the phase of any variable is constant on the plane perpendicular to this coordinate and such a wave is called a plane wave. If the coordinate system is chosen so that the wave propagates along x, 2.2 reduces to

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{2.3}$$

The complex form of the harmonic solution for the acoustic pressure of a plane wave with angular frequency $\omega = 2\pi f$, f being the wave's frequency in Hz, is given by:

$$\mathbf{p} = \overbrace{\mathbf{A}e^{j(\omega t - kx)}}^{\mathbf{p}_+} + \overbrace{\mathbf{B}e^{j(\omega t + kx)}}^{\mathbf{p}_-} \tag{2.4}$$

The associated particle velocity can be obtained from the linearized Euler equation:

$$-\rho\frac{\partial \vec{u}}{\partial t} = \nabla p \tag{2.5}$$

so that:

$$\vec{\mathbf{u}} = \left[\frac{\mathbf{A}}{\rho_0 c}e^{j(\omega t - kx)} - \frac{\mathbf{B}}{\rho_0 c}e^{j(\omega t + kx)}\right]\vec{\mathbf{e}}_{\mathbf{x}} \tag{2.6}$$

which means $\vec{\mathbf{u}_+} = \mathbf{p}_+/\rho_0 c$ and $\vec{\mathbf{u}_-} = -\mathbf{p}_-/\rho_0 c$.

For a plane wave travelling in an arbitrary direction,

$$\mathbf{p} = \mathbf{A}e^{j(\omega t - \vec{k}\cdot\vec{r})} \tag{2.7}$$

is the planar wave solution to 2.2 with the wave number $|\vec{k}|$ being $k = \omega/c$.

Since the surfaces of constant phase are given by $\vec{k}\cdot\vec{r} = constant$, $\vec{k} = \nabla(\vec{k}\cdot\vec{r})$ is perpendicular to the surfaces of constant phase and points along the direction of propagation, $\vec{k}$ is called the propagation vector.

2.1.3 Spherical waves

For spherically symmetric pressure fields, the Laplacian operator in spherical coordinates can be written as

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}$$

Applying it to the linearized wave equation 2.2,

$$\frac{\partial^2 p}{\partial r^2} + \frac{2}{r}\frac{\partial p}{\partial r} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{2.8}$$

Rewriting the previous equation with rp as dependent variable

$$\frac{\partial^2 (pr)}{\partial r^2} = \frac{1}{c^2}\frac{\partial^2 (pr)}{\partial t^2} \tag{2.9}$$

This equation has the same set of solutions as the plane wave in pr,

$$rp = f_1(ct - r) + f_2(ct + r) \Leftrightarrow p = \frac{1}{r} f_1(ct - r) + \frac{1}{r} f_2(ct + r)$$

but the first term represents a spherical wave diverging from the origin at speed c and the other a convergent wave. The former can be written in complex form as

$$\mathbf{p} = \frac{\mathbf{A}}{r} e^{j(\omega t - kr)} \tag{2.10}$$

By using Euler's equation, the particle's velocity is given by:

$$\vec{\mathbf{u}} = \left(1 - \frac{j}{kr}\right) \frac{\mathbf{p}}{\rho_0 \, c} \vec{e_r} \tag{2.11}$$

2.1.4 Acoustic intensity

The acoustic intensity I of a sound wave is defined as the average rate of flow of energy through a unit area normal to the direction of propagation. Its fundamental units are watts per square meter (W/m^2). Since the instantaneous rate at which work is done per unit area by one element of fluid on an adjacent element is pu,

$$I = \langle pu \rangle_T = \frac{1}{T} \int_0^T pu \, dt \tag{2.12}$$

where the integration is taken over a time corresponding to the period T of a cycle.

Again, the relationship between p and u must be known. For planar waves there are two solutions according to the positive or negative travelling direction in x, $p = \pm \rho_0 c u$ and

$$I = \pm \frac{1}{2} P_\pm U_\pm = \pm \frac{1}{2} P_\pm^2 / (\rho_0 c)$$

where $P_\pm$ is the pressure amplitude for the positive or negative travelling wave solution and $U_\pm$ the wave velocity amplitude in the $+x$ or $-x$ direction.

It is also possible to express the previous solution in terms of effective quantities. Defining F_e as the effective (root-mean-square) amplitude of a periodic quantity $f(t)$ over the period T,

$$F_e = \sqrt{\frac{1}{T} \int_0^T f^2(t) \, dt}$$

For harmonic quantities $F_e = F/\sqrt{2}$, thus

$$I_\pm = \pm P_e U_e = \pm P_e^2 / (\rho_0 c) \tag{2.13}$$

if the plane wave is moving in the $+x$ or $-x$ direction. This is true for plane waves and approximatedly true for diverging waves at great distances from their sources.

For a harmonic spherical wave

$$I = \frac{1}{T}\int_0^T Pcos(\omega t - kr)Ucos(\omega t - kr - \theta)\,dt = \frac{1}{2}PU\cos\theta = \frac{P^2}{2\rho_0 c} \tag{2.14}$$

which is exactly the same result found for planar waves.

The average rate at which energy flows through a closed spherical surface of radius r surrounding a source of symmetrical spherical waves is

$$\Pi = 4\pi r^2 I = 4\pi r^2 \frac{P^2}{2\rho_0 c} = 4\pi r^2 \frac{(A/r)^2}{2\rho_0 c} = \frac{2\pi A^2}{\rho_0 c} \tag{2.15}$$

a result independent of the radius, which is consistent with the conservation of energy in the so considered lossless medium.

2.1.5 Sound levels and Decibel scales

It is customary to describe sound pressures and intensities with logarithmic scales known as sound levels. This is because audible intensities vary from 10^{-12} to 10 W/m^2 and using a logarithmic scale allows the compression of such range. Also, humans judge the relative loudness of two sounds by the ratio of their intensities, a logarithmic behaviour [Kinsler et al., 1982] known as the Fechner law.

The most common logarithmic scale used for describing sound levels is the decibel scale. The intensity level IL of a sound of intensity I is defined as

$$IL = 10\log\left(\frac{I}{I_{ref}}\right) \tag{2.16}$$

where I_{ref} is a reference intensity and IL is expressed in decibels referenced to I_{ref} ($dB\ re\ I_{ref}$).

As shown previously, both planar and spherical waves have $I = P_e^2/(\rho_0 c)$ thus intensities can be replaced by pressures leading to the sound pressure level expressed in ($dB\ re\ P_{ref}$):

$$SPL = 20\log\left(\frac{P_e}{P_{ref}}\right) \tag{2.17}$$

The standard reference for airborne sounds is $10^{-12}\,W/m^2$ which is approximately the intensity of a $1\,kHz$ pure tone barely audible to a person with unimpaired hearing. In terms of pressure, using the result for planar or spherical waves,

$$P_{e_{ref}} = \sqrt{\rho_0 c I} = 20.4\,\mu Pa \tag{2.18}$$

This value rounded to 20 μPa is often the used reference for sound pressure levels in air. If the measured waves are progressive planar or spherical waves both IL and SPL give the same results, however if standing waves are measured the results can diverge if I is no longer proportional to P_e^2. Since the majority of microphones provide signals proportional to acoustic pressure, SPL is the most used sound level metric.

2.2 Speech production

The main frequency range for speech production ranges from around 300 Hz to 3400 Hz and easily understandable if sampled at 8kHz [Bradbury, 2000], only the sounds like "s" are greatly attenuated.

Another important characteristic of speech is that the silence periods in conversations take over 50% of their duration.

Independently of language, the human anatomy and laws of physics limit the range of speech signals that humans can produce.

In simplified terms (source-filter model) the speech production in humans starts with air being pushed from the lungs that goes through the vocal tract[1] and exits at the mouth. In this model the lungs serve as the excitation source and the vocal tract as filter. The higher the volume of air expelled, the louder the sounds will be.

Phonemes, the basic units of a language phonology, are split into voiced and unvoiced sounds. The vowels are voiced sounds that often have high average energy levels and distinct resonant/formant frequencies. Consonants are usually signals with lower energy and higher frequency than voiced sounds. The vocal cords either produce a voiced (tonal) or unvoiced excitation (noise-like) that is then articulated (filtered) with the mouth and nose.

Voiced sounds are produced by forcing air from the lungs to pass through the vocal cords. This produces vibrations in them with a quasi-periodic behaviour that releases pulses of air called glottal pulses. The pitch of the sound produced is determined by the rate of vibration of the vocal cords. These pulses then pass through the rest of the vocal tract where some frequencies resonate.

Unvoiced sounds are produced by a turbulent air flow forced through the vocal tract, during which the vocal cords do not vibrate and remain open until the sound is produced. Pitch is an unimportant factor in unvoiced sounds since there is no vibration of the vocal cords and consequently, no glottal pulses [Bradbury, 2000]. Plosive sounds for instance are non-stationary [Schluter and Ney, 2001].

Different vocal tracts have different resonant frequencies, making each person have a characteristic voice. The vocal tract constantly changes shape at a very slow rate, modifying the sound as the person speaks.

2.3 Sound perception

Humans can hear sounds from 20 Hz to 20 kHz and detect displacements of the eardrum of one-tenth of the diameter of a hydrogen molecule at 1kHz.

2.3.1 Human ear

The human ear has three main parts: the outer, middle and inner ear. The first deals with oscillations in air, the third with oscillations in the fluids that surround the sensory cells and the middle ear reduces energy losses by matching the impedance of the inner and outer ear through a mechanical lever system.

[1] The vocal tract encompasses throat, tongue, nose and mouth.

The pinna of the outer ear, almost useless in humans, collects sound into the auditory canal which is an approximately straight $2.5\,cm$ long tube closed at the tympanic membrane (eardrum) with about $0.7\,cm$ in diameter. The auditory canal can significantly boost the $2-6\,kHz$ range with a SPL peak of $10\,dB$ at $3\,kHz$. Together with the diffraction of sound waves by the head, the SPL amplification at the eardrum can reach 15 to $20\,dB$ at some frequencies [Kinsler et al., 1982].

The ossicles in the middle ear work as a lever that transmits the vibrations from the eardrum to the oval window in the internal ear and along with the area ratio of 30:1 provide the impedance match from the air in the outer ear to the fluid in the inner ear where the cochlea is located.

These vibrations produce an oscillatory motion of the perilymph (fluid), basilar and tectorial membranes inside the cochlea, with a frequency-dependent peak amplitude localization (known as the tonotopic mapping). The hair cells in the Organ of Corti are flexed by the motions of such membranes, thereby causing the nerve endings to which they are attached to fire. This signals are then sent through the auditory nerve and the information from both ears combined. In summary, the tonotopic arrangement in the inner ear enables us to hear pitch.

2.3.2 Auditory cues for source localization

There are several types of auditory cues used by humans to locate a sound in space. Three of them relate to differing sound characteristics of the sounds that reach each ear: Interaural Level Difference (ILD), the level difference between signals detected at both ears, Interaural Phase Difference (IPD), the phase difference of the signals at both ears and Interaural Time Difference (ITD), the time difference it takes for sound to go from the source to each ear. ITD can be used for transitory sound signals and IPD for stationary signals. This use of level and time characteristics is known as Duplex Theory, introduced by Lord Rayleigh [Rayleigh, 1907].

For higher frequencies (above $3000Hz$), ILD is used due to the accentuated diffraction caused by the head which increases the intensity difference between ears. For lower frequencies (under $1500\,Hz$), where phase aliasing is reduced, IPD is used. In the intermediate range between both, sound source localization is more diffuse [Litovsky, 2008].

Another interesting auditory feature is the localization of sound sources in the median plane between ears. Independently from the sound source localization, for narrow-band signals with a centre frequency of $300Hz$ or $3kHz$, the sound source is perceived to be in front of the subject. The sound is otherwise perceived to come from above for narrow-band signals centred at $8kHz$ and from behind for narrow-band sounds centred at 1 or $10kHz$ [Fastl and Zwicker, 2007].

2.3.3 Auditory Masking

Auditory masking refers to the fact that a loud sound hides another within a critical band.

There are two types of masking when two speech signals mix together [Mowlaee, 2010; Moore, 2003; Wang et al., 2008]: energetic and informational masking. The first happens when both speakers produce signals in the same critical frequency band (strong signals mask weak signals). Isolating sources from this

information is particularly difficult. As energetic masking takes place, the separating ability is reduced.

However, there exists as well informational masking, when a listener cannot distinguish the speech of two speakers when both are talking. This can be ameliorated with source separation algorithms that use additional information to separate such signals.

2.3.4 Ear integration time and hearing frequencies

Only above $50\,ms$ are we able to perceive sound frequencies and this *ear integration time* could be compared to the size of the spectral analysis window of the ear and any auditory phenomena of period larger than this is not captured as periodic, thus characterized by a frequency, but as individual sound events. If two acoustic pressure peaks are separated by more than 50ms we hear them separated, but under this interval they start to become more and more smeared until we hear only one thing. This 50ms are thus not coincidently responsible for the lowest limit of frequency hearing of 20Hz. Below this frequency we hear distinct sound events instead of pitches.

The upper limit on the audible frequency range is caused by the physical limitations of the ear.

2.4 Signal processing

2.4.1 Fourier Transform

Sometimes it is useful to analyze signals in the frequency domain instead of the time domain. The Fourier transform $\mathscr{F}$ transforms the signal from the time to the frequency domain and the inverse Fourier transform $\mathscr{F}^{-1}$ does the opposite.

It is possible to describe continuous periodic signals as a series of sines and cosines of harmonic frequencies (Fourier series) up to an energy divergence of zero even for signals such as a square wave, that don't have a continuous derivative, even though overshoot and ringing of the representation near the non-continuous derivative points[2] will occur.

Let $f(t) : \mathbb{R} \to \mathbb{C}$ and throughout this thesis capital letters will be used to represent the same function $F(\nu)$ in the frequency domain[3],

$$\begin{aligned} F(\nu) \equiv \mathscr{F}[f(t)](\nu) &:= \int_{-\infty}^{\infty} f(t) e^{-i2\pi\nu t}\, dt \\ \mathscr{F}^{-1}[F(\nu)](t) &:= \int_{-\infty}^{\infty} F(\nu) e^{i2\pi\nu t}\, d\nu \end{aligned} \tag{2.19}$$

This operator is linear,

$$\mathscr{F}[af(t) + bg(t)](\nu) = a\mathscr{F}[f(t)] + b\mathscr{F}[g(t)] = aF(\nu) + bG(\nu) \tag{2.20}$$

and symmetric,

$$F(\nu) = \mathscr{F}_t[f(t)](\nu) \Rightarrow F(-\nu) = \mathscr{F}_t[f(-t)](\nu) \tag{2.21}$$

[2] http://mathworld.wolfram.com/FourierSeries.html
[3] http://mathworld.wolfram.com/FourierTransform.html

2.4.2 Discrete Fourier Transform

When the support of the function is discrete, because it was sampled at a finite number of intervals, the discrete Fourier transform (DFT) is applied. The sequence of N complex samples of the signal $f[n]$, is transformed into an N-periodic sequence of complex numbers

$$F[k] := \sum_{n=0}^{N-1} f[n] e^{-i2\pi kn/N} = \sum_{n=0}^{N-1} f[n] e^{-i\omega n}, \ k \in \mathbb{Z} \tag{2.22}$$

where $F[k]$ is a complex number that represents the k-th sinusoidal component with frequency k/N through both an amplitude and phase.

This transform is periodic, with period N,

$$F[k+N] = \sum_{n=0}^{N-1} f[n] e^{-\frac{2\pi i}{N}(k+N)n} = \sum_{n=0}^{N-1} f[n] e^{-i\frac{2\pi i}{N}kn} \cancel{e^{-2\pi i n}} = F[k] \tag{2.23}$$

and also symmetric,

$$F[k] = F^*[-k] \tag{2.24}$$

which means the real part of F is an even function and the imaginary part an odd function.

2.4.3 Short-Time Fourier Transform

The Short-Time Fourier Transform (STFT) is a Fourier Transform applied to a small interval of the interest signal to obtain a local representation of its frequency content. This in turn allows the analysis of the signal in the frequency domain throughout time, i.e. a two-dimensional T-F representation.

If the frequencies chosen in a source's mask are changing drastically the decisions between consecutive masks will be incompatible and thus noise is created. This is especially noticeable if very short FFT_N values are chosen or more sources are added.

The problem of finding the optimum STFT reconstruction can be formulated as a minimum squared error problem (thus no phase is considered) where the windowed target STFT $Y^w(lI, k)$ and the estimated windowed STFT $\hat{X}^w(lI, k)$ should minimize the functional

$$\sum_{l=-\infty}^{\infty} \sum_{k=0}^{N-1} \left| \hat{X}^w(lI, k) - Y^w(lI, k) \right|^2 \tag{2.25}$$

where w is the analysis window applied to the signal in the time domain before the DFT to the l-th analysis window, I the hop size and k the frequency index of the DFT.

The solution to 2.25 [Röbel, 2006] satisfies the *perfect reconstruction*[4] [Selesnick, 2009]:

$$\hat{x}[n] = \frac{\sum_{l=-\infty}^{\infty} w(n - lI)\hat{y}^w(lI, n)}{\sum_{l=-\infty}^{\infty} w(n - lI)^2} \tag{2.26}$$

where $\hat{x}$ is the minimum squared error solution to the reconstruction of the modified windowed signals

[4]The envelope of the overlapping windows must be constant.

in the frequency domain Y^w.

Several state-of-the-art iterative methods take into consideration the phase of the content in overlapping windows,[Alsteris and Paliwal, 2005; Bouvrie and Ezzat, 2006; Le Roux et al., 2010; Sturmel and Daudet, 2011; Krawczyk and Gerkmann, 2012], but are much more computationally expensive.

2.4.4 Convolution

The convolution operation is defined as

$$(f * h)(t) \equiv \int_{-\infty}^{+\infty} f(t-\tau)h(\tau)\, d\tau \tag{2.27}$$

There is a very interesting set of properties, often called the convolution theorem:

$$\begin{aligned} \mathscr{F}[f * g] &= \mathscr{F}[f]\mathscr{F}[g] \\ \mathscr{F}[fg] &= \mathscr{F}[f] * \mathscr{F}[g] \\ \mathscr{F}^{-1}[\mathscr{F}(f)\mathscr{F}(g)] &= f * g \\ \mathscr{F}^{-1}[\mathscr{F}(f) * \mathscr{F}(g)] &= fg \end{aligned} \tag{2.28}$$

showing that the convolution of two functions is just their multiplication in the complementary domain.

2.4.4.1 High-speed convolution

The convolution operation $h*f$ is especially slow with large filter sizes N_h due to the amount of calculations required. Even when not extending the signal, calculating one instant of the output in the worst and common case, requires N_h multiplications and $N_h - 1$ sums. The initial terms have less calculations but despising that and the extension of the signal due to the convolution filter the computational complexity of the convolution is $O\left(N_f N_h\right)$, where N_f is the number of samples of the signal f. For sufficiently large filter sizes the calculation of the convolution using the definiton is not feasible for a real-time application.

The advent of the Fast Fourier Transform (FFT)[5] made possible such computations. The high-speed, fast or FFT convolution and relies on the convolution theorem 2.28. The problem is transformed to the complementary domain where the convolution becomes a multiplication and the problem is transformed back to the original domain. The filter size N_h should be bigger than 60 to have performance improvements over the standard convolution (overhead of the FFT transformations). This has precision advantages since the number of successive operations and consequently the round-off errors are reduced [Smith et al., 1997, chapter 18]. The FFT fftw3 library was chosen for its performance.

2.4.5 Cross-correlation

The discrete cross-correlation operation does not commute and is defined as:

$$(f \star g)[t] = \sum_{i=-\infty}^{\infty} f^*[i]\, g[t+i] \tag{2.29}$$

[5] DFT with $O(N \log_2 N)$ complexity.

For sampled signals in the time domain when neither past nor future are known, the signals are assumed to be zero outside the sampling support.

2.4.6 Autocorrelation

The autocorrelation $f_\star$ of f is a particular case of the cross-correlation

$$f_\star[t] \equiv (f \star f)[t] \equiv \sum_{i=-\infty}^{\infty} f^*[i]\, f[\,t+i\,] \tag{2.30}$$

and can be efficiently computed through the Wiener-Khinchin theorem[6].

If there are periodic features in the signal f they will cause the same periodicity in $f_\star$. If the signal is non-periodic and stationary such as white noise, the autocorrelation will be weak and almost flat.

Autocorrelation is also symmetric for negative time values $R_f(-\tau) = R_f^*(\tau)$.

Notice that for the autocorrelation, the first coefficient is always the energy of the chosen signal f:

$$f_\star[0] = \|f\|^2 \tag{2.31}$$

2.5 Window functions

When transforming a signal from the time domain to the frequency domain through the Fourier Transform the signal is extended periodically, being the period the duration of the sampled signal sequence.

Harmonic components whose periods are not sub-multiples of the analysis window of the Fourier Transform will create spectral leakage, i.e., as they do not belong to the basis set of the transform they will be approximated by a combination of elements of the basis. In the time domain, this is seen as the mismatch between the signal value at the boundaries of the window. When the signal in this window is extended periodically it produces leakage because there is a jump at the edge between adjacent periodic extensions of the signal which can be described by high frequency content. Spectral leakage is reduced by macthing as many orders of derivatives of the data weighted by the window as possible, at the window boundary. The easiest way to accomplish this is to set such derivatives to zero or as close to it as possible.

The generalized Hamming window functions are given by:

$$w_N(n) = \alpha - \beta \cos\left(\frac{2\pi\, n}{N-1}\right) \tag{2.32}$$

One of the most used windows in that family is the Hann window, with $\alpha = \beta = 0.5$. Another well known function in that family is the Hamming function with $\alpha = 25/46$ and $\beta = 21/46$. The zero-phase Hamming function $Hamming_0$ is the same as Hamming but has the bigger windows weights at its extremities. Only β changes from the Hamming function to its symmetric value.

The two-decimal places variant of the Hamming window which is supposed to better suppress the side-lobes, $\alpha = 0.54$ and $\beta = 0.46$, and the equiripple version, $\alpha = 0.53836$ and $\beta = 0.46164$. As tiny changes in the Hamming window changed the separation metrics, variations on it were also considered, namely

[6] `http://mathworld.wolfram.com/Wiener-KhinchinTheorem.html`

the weak and weaker Hamming functions for the standard, mid and lower versions. The weak windows subtract 0.1 from the Hamming window β parameter and the weaker Hamming functions subtract twice that value. The mid versions subtract to α half of the value removed from Hamming's β parameter and the low version subtracts that same value so that the minimum of the function matches the Hamming window minimum, the only difference is that the function height is compressed.

The Blackman window function is higher-order than the windows in the Hamming family and the Nutall family even higher-order. This family of functions is given by the expression:

$$w_N(n) = a_0 - a_1 * \cos\left(\frac{2\pi n}{N-1}\right) + a_2 \cos\left(\frac{4\pi n}{N-1}\right) - a_3 * \cos\left(\frac{6\pi n}{N-1.0}\right) \tag{2.33}$$

where the Blackman function uses $a_0 = 7938.0/18608.0$, $a_1 = 9240.0/18608.0$, $a_2 = 1430.0/18608.0$, $a_3 = 0.0$; the BlackmanNutall, $a_0 = 0.3635819$, $a_1 = 0.4891775$, $a_2 = 0.1365995$ and $a_3 = 0.0106411$. and Nutall, $a_0 = 0.355768$, $a_1 = 0.487396$, $a_2 = 0.144232$, $a_3 = 0.012604$.

2.6 Source Separation

2.6.1 Mixing models

The models for the mixtures x_m of signal sources s_n are hereby presented as they will be used later on throughout this thesis.

Each mixture x_m represents a measurement (obtained in the real world through a microphone m) of the several sources n in a particular acoustic environment. Each microphone can also be subjected to additional spurious or periodic noises n_m.

The basic model for source mixing is *instantaneous / linear* mixing (which while appropriate for EEG data for instance, is not appropriate for audio because the sound doesn't propagate nearly instantaneously), where a_{nm} represents the physical process the signal n undergoes to reach the microphone m:

$$x_m(t) = \sum_{n=1}^{N} a_{nm} s_n(t) + n_m(t) \Leftrightarrow \mathbf{x}(t) = A\mathbf{s}(t) + \mathbf{n}(t) \tag{2.34}$$

In this case a_{nm} can represent the attenuation of the signal but in the general (convolutive) model represents the joint impulse response from the devices and the acoustic environment in the path from the source n to the microphone m.

A more accurate model is the *delayed* or *anechoic* mixing that is realistic in rooms without reverberation because it correctly takes into account the time τ_{nm}[7] sound takes to go from the source n to microphone m:

$$x_m(t) = \sum_{n=1}^{N} a_{nm} s_n(t - \tau_{nm}) + n_m(t) \Leftrightarrow \mathbf{x}(t) = A * \mathbf{s}(t) + \mathbf{n}(t) \ , \quad A_{nm} = a_{nm}\delta(t - \tau_{nm}) \tag{2.35}$$

This model is however not appropriate for reverberant environments. The complete model for acoustic

[7] $\tau_{nm} = \frac{d_{nm}}{\langle c_{nm}(T)\rangle} \approx \frac{d_{nm}}{c(T_{amb})}$ where $\langle c_{nm}(T)\rangle$ is the weighted average of the speed of sound on that path, which depends on the temperature. For most scenarios it will be the speed of sound for the ambient temperature.

scenarios is the *convolutive* / *echoic* or *reverberant* model:

$$x_m(t) = \sum_{n=1}^{N} a_{nm} * s_n(t - \tau_{nm}) + n_m(t) \Leftrightarrow \mathbf{x}(t) = A * \mathbf{s}(t) + \mathbf{n}(t) \ , \quad A_{nm} = h_{nm} \tag{2.36}$$

where h_{nm} is the joint impulse response of the medium and devices involved in the path from the source n to the microphone m, which can be approximated in finite discrete form with a Mth-order FIR (Finite-Impulse-Response) filter:

$$x_m[t] = \sum_{n=1}^{N} \sum_{k=0}^{M-1} a_{nmk} s_n[t - \tau_{nm} - k] + n_m[t] \tag{2.37}$$

Determinacy of the mixtures The source separation literature adopted linear algebra nomenclature to describe different cases in terms of number of unknowns and equations. For M microphones and N sources, the setups are called determined if $M = N$, under-determined (or degenerate) when $M < N$ and over-determined for $M > N$.

Time-varying/invariant Notice that the previous mixing models implicitly assumed that the experiment conditions are time-invariant. For time-varying regimes the impulse responses a_{nm} will be functions of time as well.

2.6.2 W-disjoint orthogonality

If any set of sources does never participate in the same Time-Frequency (T-F) bin they are said to respect W-disjoint orthogonality (WDO). Two signals S_1 and S_2 are WDO when

$$\forall_\omega \ S_1(\tau, \omega) S_2(\tau, \omega) = 0 \tag{2.38}$$

Approximately W-disjoint speech signals can be separated using specific methods such as DUET [Rickard and Özgür Yilmaz, 2002]. WDO decreases with increased reverberation and number of active sources. In practice, the approximate WDO for a source S_i is defined in [Yilmaz and Rickard, 2004] by:

$$WDO = \frac{\left\| \Phi_i^0(t,\omega) S_i(t,\omega) \right\|^2 - \left\| \Phi_i^0(t,\omega) Y_i(t,\omega) \right\|^2}{\left\| S_i(t,\omega) \right\|^2} \quad , \quad Y_i(t,\omega) = \sum_{\substack{n=1 \\ n \neq i}}^{N} S_n(t,\omega) \tag{2.39}$$

$$\Phi_i^x = \begin{cases} 1, & 20 \log(|S_i(t,\omega)|/|Y_i(t,\omega)|) > x \\ 0, & \text{otherwise} \end{cases} \tag{2.40}$$

where Φ_i^0 is the chosen Ideal Binary Mask (IBM).

2.6.3 Source separation metrics

There are three types of metrics to assess speech separation: objective metrics, ASR performance and human listening tests. The objective metrics provide reproducible tests without requiring human inter-

vention. Eventually, the performance of the system shall be measured by an ASR backend word error rate but for now only the objective metrics will be discussed.

2.6.3.1 SNR

Signal to Noise Ratio (SNR) is often used to assess source separation [Hummersone et al., 2011]:

$$SNR = 10\log_{10}\left(\frac{\sum_t s^2[t]}{\sum_t \left(\hat{s}[t] - s[t]\right)^2}\right) \tag{2.41}$$

where s is the original signal and $\hat{s}$ the estimated signal, at sample t.

However, because a signal subject to reverberation can be regarded in a simplified form as

$$s_r[t] = s[t] + \sum_{\tau=1}^{T} a[\tau]s[t-\tau] \tag{2.42}$$

where $|a| < 1$ are reflection coefficients and $T << \infty$ is the reverberation time, if reverberation is only considered for the reconstructed signal while s is not subject to the same reverberation, then this will change the denominator of 2.41. In addition, the results will depend on the acoustic setting even if the sources are in the same positions and every other condition is fixed. By comparing the separated signals with the emitted signals recorded individually in the room and not the pre-recorded signals that feed the speaker, the existence of echo by itself no longer changes the value of SNR, although missing part of the echoic energy of the target source or having interfering echoic energy from others will.

Finally, since only the reconstructed sound quality matters, i.e., $\hat{s}$ matches s as closely as possible in shape and the denominator of 2.41 is not invariant to scaling of $\hat{s}$, both signals must be normalized. If distortions happen to the signal, the energy will remain largely unaltered although local characteristics such as the maximum of the signal are more likely to be corrupted. Because of this the energy $E(s) = \|s\|^2$ was chosen as normalization constant. Since $\sum_t \left(\hat{s}[t] - s[t]\right)^2$ is very insensitive to small differences, small variations in energy of the signal will yield small errors in the normalization which lead to small errors in the denominator, thus it is a stable quantity to normalize and evaluate the SNR. Given that $E \propto A^2$ where A is the amplitude of the signal and the normalization occurs in the amplitude of the signal, each point is divided by $\sqrt{E}$, so $E(s) = 1$ in the end. The definition of the normalized SNR is thus:

$$SNR \equiv 10\log_{10}\left(\frac{\|s\|^{-2}\sum_t s^2[t]}{\sum_t \left(\frac{\hat{s}[t]}{\|\hat{s}\|} - \frac{s[t]}{\|s\|}\right)^2}\right) \tag{2.43}$$

where s is the measured signal in the current acoustic setting, when no other sources are active, and $\hat{s}$ the signal separated from the mixture in the same acoustic setting.

2.6.3.2 Total relative distortion (TRD)

Total Relative Distortion (TRD) [Gribonval et al., 2003] can also compare two signals,

$$D_{total}(s_n, \hat{s}_n) = \frac{\|\hat{s}_n\|^2 - |\langle \hat{s}_n, s_n \rangle|^2}{|\langle \hat{s}_n, s_n \rangle|^2} \tag{2.44}$$

where $\langle f, g \rangle = \sum_t f(t)\overline{g}(t)$ is the standard inner product and $\|f\|^2 = \langle f, f \rangle$ the squared norm of f, or its energy and since by convention each source is normalized, $\|s_n\| = 1$.

Since the original signal might not be normalized and normalizing it is more expensive than calculating a normalization-unbiased metric, by using the linearity property of the inner product $\langle f, kg \rangle = \overline{k}\langle f, \tilde{g} \rangle$, being $\tilde{g}$ a normalized signal with $\|\tilde{g}\|^2 = 1$, a scale-invariant version of D_{total} is defined:

$$D_{total}(s_n, \hat{s}_n) \equiv \frac{\|\hat{s}_n\|^2 - |\langle \hat{s}_n, s_n \rangle|^2 / \|s_n\|^2}{|\langle \hat{s}_n, s_n \rangle|^2 / \|s_n\|^2} = \frac{\|\hat{s}_n\|^2 \|s_n\|^2 - |\langle \hat{s}_n, s_n \rangle|^2}{|\langle \hat{s}_n, s_n \rangle|^2} \tag{2.45}$$

This metric proved to be the very best at giving an idea of separation quality and identifying the original and estimated source pair. Values of $D_{total} < 0.2$ indicate a very good separation and at $D_{total} > 0.5$ the separation already starts to have many artifacts to be considered a good separation. This values also seem to have a high correlation with human perception of the test at least in terms of being able to easily tell the quality of the separation before listening to the signal.

2.6.3.3 SDR, SIR and SAR

Numerous studies have been dedicated to the assessment of source separation quality alone. There are two standards today, packaged into open-source toolkits: BSS Eval Toolbox [Gribonval et al., 2003; Vincent et al., 2006, 2007] and PEASS Toolbox [Vincent, 2012]. The model of the former seems to have stabilized, thus using its metrics will guarantee comparability with other studies. The latter is a modified version of the BSS Eval metrics aimed at mimicking human subjective assessment which is not the crux of an ASR interface design.

The three metrics returned by the BSS Eval package are the Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR) and Signal-to-Distortion Ratio (SDR). SDR is an overall measure of the separation performance, SIR is a measure of the suppression of the interference signals and SAR a measure of the naturalness of the separated signals.

The implemented method combines the information from both channels into one output per source with the same time-alignment as the signal in the left channel thus [Vincent et al., 2007]:

$$\begin{aligned}
\text{SIR}_n &= 10 \log \frac{\|s_n^{img} + e_n^{spat}\|^2}{\|e_n^{interf}\|^2} \\
\text{SAR}_n &= 10 \log \frac{\|s_n^{img} + e_n^{spat} + e_n^{interf}\|^2}{\|e_n^{artif}\|^2} \\
\text{SDR}_n &= 10 \log \frac{\|s_n^{img}\|^2}{\|e_n^{spat} + e_n^{interf} + e_n^{artif}\|^2}
\end{aligned} \tag{2.46}$$

where $s_n^{img} + e_n^{spat}$ is the reverberated signal that is measured at the left microphone, e_n^{interf} the error component due to interfering sources and e_n^{artif} the error caused by "blurbing" noise. For a more detailed discussion consult [Vincent et al., 2007].

Chapter 3

State-of-the-art

There isn't an ultimate solution for sound separation just yet as the sheer diversity of existing scenarios and chosen approaches suggest, but a multitude of methods applicable to different settings.

Most BSS methods do not solve directly the delayed or convolutive models in the time domain. Instead, by taking a convolutive mixing problem and solving it in the frequency domain the sound source mixing system becomes instantaneous due to the convolution theorem that states that a convolution in a domain is equivalent to a product in its complementary domain.

Another important and often explored property by BSS methods is signal sparsity. Problems where signal sparsity is bigger are easier to solve than problems where signals are spread over the representation domain and some methods even require signal sparsity as is the case of DUET (which will be explained later).

A signal is *sparse* in a domain if most of its coefficients in that domain are zero or close to it [Burred, 2008]. Sparser signals will have more peaked probability distributions in that domain.

By choosing an appropriate representation domain, signals can be made sparser. There are several measures of sparsity such as the L1-norm, kurtosis or negentropy[1] and geometrically, the sparser the signals are the more the coefficients will be concentrated around the mixing directions (in respect to the set of sensors). Analytically, the *Maximum a posteriori / posterior* (MAP) approach has a smaller penalty for lower sparsity values than a Maximum Likelihood (ML) formulation [Burred, 2008].

For instance, through the Short-Time Fourier Transform (STFT) one can work in the Time-Frequency (T-F) domain that is much sparser than the time domain. Otherwise, logarithmic resolution front-ends such as Constant-Q Transform and Discrete Wavelet Transform, or auditory-based resolution front-ends such as Bark, ERB (Equal Rectangular Bandwidth) and the Mel scale can be used. Other alternatives are the adaptive signal decompositions such as Basis Pursuit and Matching Pursuit [Burred, 2008].

[Aichner, 2007] contains a comprehensive study and mathematical formulation of the sound fields and the properties of sound waves that can be explored in BSS. It is one of the most comprehensive sources on BSS identifying the several problems that are encountered in practice and explains what can be done or not. It also discusses several BSS algorithms.

There are diverse methods specialized in music source separation, which is a demanding task and that is why most are supervised.

Supervised sinusoidal methods [Burred, 2008; Burred and Sikora, 2007] model the sound source spectral envelope and its temporal evolution. The detailed temporal evolution allows to ignore harmonicity constraints and thus separation of the chords and inharmonic sounds is possible.

[1] Relative entropy to the maximal entropy distribution (Gaussian). It is never negative.

Learnt priors for Wiener-based separation [Ozerov. et al., 2005] work for single-channel with one Hidden Markov Model (HMM) for the singing voice and another for the accompaniment.

Since in our case one wants to extract speech sources from the mixture, methods that can only separate music sources like the previous ones are not relevant. There are although, methods that allow separation of generic signals, whether they are musical or not, *e.g.*, Nonnegative Matrix Factorization (NMF). ISS methods should also be discarded because no direct user input or training data is available.

3.1 Beamforming

Beamforming systems use multiple microphones to separate spatially the desired speaker from other sources by choosing the appropriate time delays between the microphone array signals so that constructive interference occurs for the target source [J.Rennie, 2008][10,11].

In [Thiemann and Vincent, 2013], a comparison is done between unconstrained or harmonically constrained state-of-the-art BSS algorithm [Bimbot, 2010; Nakatani et al., 2011] and three state-of-the-art beamforming methods, two of which with post-filtering and one without. The results show a distinctive advantage in SIR improvement for BSS techniques when there's more than one sound source. The harmonic BSS was the method that most improved SIR in the presence of two speakers with a SIR improvement of $20\,dB$, almost twice of the obtained SIR improvement for the unconstrained BSS method and the best beamformer type for an input SNR of $-6\,dB$. Another tendency is that BSS systems increase the separation level with more input channels, whilst beamforming systems tended to level off at 4 channels [Thiemann and Vincent, 2013]. The method with smaller Mel Cepstral Distance for two sources was the unprocessed beamformer, followed by the unconstrained BSS method.

According to Vincent [2005], beamforming doesn't generate pipenoise artifacts, but is limited to determined instantaneous or slightly convolutive mixtures, whilst T-F masking works potentially better for under-determined or strongly convolutive mixtures, but generates artifacts when sources are not W-disjoint.

3.2 Independent Component Analysis (ICA)

ICA works by maximizing an objective measure of statistical independence between the unmixed sources. Some implementations maximize the nongaussianity of the sources, like FastICA, through kurtosis or negentropy; Other implementations minimize the mutual information between sources, while others use the Maximum Likelihood (ML) estimation like the Bell-Sejnowski algorithm [Bell and Sejnowski, 1995], the Natural gradient algorithm [MacKay, 2003] or the ML-based FastICA.

Other class of ICA methods is the Tensorial family that use higher order statistics, such as FOBI (Fourth-Order Blind Identification) and JADE (Joint Approximate Diagonalization of Eigenmatrices). JADE has the advantage of not requiring parameter tuning like traditional ICA, which requires good priors on the source signals [Li et al., 2000].

ICA requires different combinations of signal energies as inputs, otherwise the unmixing matrix would

become singular, *i.e.*, the microphones have to be far apart for better results. ICA also requires sources to be non-gaussian [Knuth, 2005], which is not a limitation for speech signals [Little, 2011].

ICA was originally developed for instantaneous mixtures [Comon, 1994] and doesn't perform well in noisy or very under-determined setups. Mitianoudis and Davies have developed another frequency domain method for determined BSS based on ICA [Mitianoudis and Davies, 2003], where the permutation problem is addressed by means of frequency coupling in the source model.

All blind ICA methods are limited to the recovery of sources with arbitrary scaling and permutation which is a big limitation for Robust ASR in generic real enviroments. A second limitation is that ICA works only for the determined mixing problem as the inversion of the mixing matrix is required. Another concern is that the convergence of most ICA methods is not guaranteed [Yang et al., 2011].

Clustering methods can be used to solve the degenerate problem that ICA can't. These methods allow the separation of mixtures by direct inspection of the scatter plot of the mixtures where sparsity is of utmost importance. Examples of these techniques are the kernel-based angular clustering family [Bofill and Zibulevsky, 2001]. Other methods are based on k-Means or fuzzy C-means, which respectively constitute hard and soft clustering methods [Burred, 2008]. Image processing techniques such as the Hough transform [et al., 1997] can also be applied to the scatter plots to determine the mixing matrix.

Another variant of ICA is Independent Subspace Analysis (ISA) [Casey and Wstner, 2000; Burred, 2008], which removes ICA's restriction of equal number of sources and sensors and has the interesting property of being able to separate a single channel into multiple acoustic features, whether they belong to one or more sources (it is able to separate for example several cellos in a quartet if they are performing different melodic lines), which can be a problem for certain sound sources applications.

Independent Factor Analysis (IFA) [Attias, 1999] is a more general method than ICA as compared by [Rowe, 1999] against a fully Bayesian model where it is also shown that both are particular cases of the general one. One interesting aspect of IFA is the modeling of the noise, which improves the source separation capabilities.

The standard model in IFA is an instantaneous mixing model where sources are modeled as a Mixture of Gaussians (MOG) and noise is considered to be Gaussian. The same author who had introduced IFA developed an extension: IFA with Temporally Structured Sources [Attias, 2000], which no longer considers each source to be i.i.d.[2] in time.

A state-of-the-art method, that happens to be of the Factor Analysis family, called Permutation-Free Infinite-Space Factor Analysis (PF-ISFA) was recently introduced [Nagira et al., 2013]. This method works in the frequency domain thus it is able to separate convolutive mixtures and it has solved the permutation problem in an integrated manner. The authors also compared the performance of their method with Permutation Free ICA and generally obtained better results. They finally state that their future work is on turning the algorithm fast enough for real-time processing.

For a setup with two sources, two microphones and a $RT_{20} = 20\,ms$ this method achieved (SDR, SIR, SAR) = (10.07, 19.86, 11.70) dB and for $RT_{20} = 400\,ms$, a reverberation time much longer than

[2]Independent and identically distributed

the STFT window length of 64 ms, which inevitably degrades the performance, it achieved (SDR, SIR, SAR)=(5.22, 9.81, 9.49) dB. The method thus outperformed Frequency Domain ISFA (FD-ISFA) for all criteria, except SAR, under all conditions, namely, the SIR by 2.82 dB under $RT_{20} = 20\,ms$ and 0.41 dB under $RT_{20} = 400\,ms$. With the mentioned setup it achieved better average SIR except for $RT_{20} = 150\,ms$, which means the interfering signals were more suppressed with PF-ISFA than the ICA counterpart. As to the distortion, both showed similar SDR results but PF-ICA outperformed PF-ISFA in SDR when 4 microphones were used in a determined scenario since the source signals of PF-ISFA are estimated by sampling and SDR is affected by SAR [Nagira et al., 2013].

Although the PF-ISFA method was generally better than ICA it still is limited in the sense that works only for determined cases, besides not being real-time, although that was set as goal for future work.

3.3 TRINICON

The general framework of TRIple-N ICA for CONvolutive mixtures (TRINICON) for adaptive Multiple-Input Multiple-Output (MIMO) in BSS applications, dereverberation and parameter estimation was presented in [Buchner and Aichner, 2004]. This approach uses the 3 fundamental statistical source properties: non-gaussianity, non-whiteness and non-stationarity.

In [Mandel et al., 2010] TRINICON achieved an average SDR of 22.47 dB for the anechoic tests with two speakers, doubling the performance of the second best method to which it was compared: Model-based Expectation-Maximization Source Separation and Localization (MESSL). For the reverberant tests in a room with reverberation time of around 565 ms [Mandel et al., 2010] with two speakers it achieved 3.72 dB, where the best method was the MESSL-G variant with 5.87 dB.

TRINICON was used in DICIT. The problem with TRINICON is that more than 2 microphones are required and works only for determined mixing problems. In DICIT a 1.92m array of microphones was required, which is outside the scope of this project.

3.4 Nonnegative Matrix Factorization (NMF)

Nonnegative Matrix Factorization (NMF) does not aim at statistical independence as it has been proven that under specific conditions, non-negativity is sufficient for separation. In NMF a non-negative matrix representation of the signals matrix which represents the magnitude of the STFT as a product of two low rank non-negative matrices $Y = HW$ [Mowlaee, 2010]. The columns of H are the basis vectors that define the spectral structure of the time-frequency representation of the sources. The rows of W define the temporal structure and the ratios of importance of the sources in the mixture Y. NMF fails when there is too much overlap between signals (energy masking).

One of the state-of-the-art NMF methods is its extension called Positive Semi-Definite Tensor Factorization (PSDTF) [Yoshii et al., 2013] and was presented in the Music Information Retrieval Conference ISMIR 2013. NMF has become popular for music source separation although it deals only with the spectrogram power and discards phase information. The component spectrograms are usually estimated

through Wiener filtering that reuses the phase of the mixture diagram. This causes the phases to be incorrect and high-quality unmixing is impossible because the estimated spectrograms are inconsistend and do not correspond to real-time domain signals. The authors use PSDTF to estimate the source signals directly in the time domain. The separation quality is claimed to be better than any NMF variant. The problem is that although the separation is of a very high quality, the computational cost of the developed Log-Dominant PSDTF (LD-PSDTF) method is even greater than the conventional NMF. The analysis of each mixture signal was reported to take several hours when using a Xeon X5492 (3.4 GHz) processor which deems it inaccessible for a real-time application in the foreseeable future. For NMF variants with added sparsity constraints consult [Virtanen, 2003, 2004].

3.5 Computational Auditory Scene Analysis (CASA)

Listening to a specific speaker in a group where several people are speaking at the same time is often known as the cocktail party problem and humans are experts at solving it. Auditory Scene Analysis (ASA) is a model from psychophysics for human hearing perception. Computational ASA (CASA) algorithms try to mimick human hearing perception given its success at separating sources. See for example [Wang and Brown, 2006; Barker et al., 2006].

ASA which derives many of its grouping principles from Gestalt psychology is mainly concerned with the ability of creating and following an auditory stream from all the incoming sound stimuli, *exempli gratia*, the ability to listen to the cello section in an orchestra.

Since humans use ILD, IPD and ITD for source localization and their ability to segregate sources is greatly impaired when using only one ear, this auditory cues are often used in CASA methods for source localization and prior segregation through T-F masks which can be soft or binary.

Binary Mask A binary mask is a two-dimensional indicator function that selects which T-F bins from a signal should be discarded (0) or mantained (1) for the source reconstruction. The masked signal is then transformed back to the time domain. This method doesn't work well in separation of signals at low SNR values and can't recover the interferent speech at high SNR values either, given the energy masking.

Ideal Binary Mask (IBM) At each frame, for the anechoic case the ideal mask is the one that keeps only the T-F cells where the target signal energy is higher than the interfering energy by an arbitrary threshold. This can be further extended in live environments to incorporate the early and/or late reverberations as part of the desired source [Roman and Woodruff, 2011].

This is considered the upper bound on source separation performance of any and all binary masks [Li and Wang, 2009] and is used in CASA methods as the benchmark for speech separation.

The problem with masking is that spectral characteristics are not completely preserved by discarding T-F bins that contain some of the target source energy and so "musical noise" or "blurbing" can arise.

There are missing data techniques that handle unreliable acoustic information better than a straightforward IBM [Cooke et al., 2001] for ASR. In [Boldt et al., 2010] the authors show how to correct errors in a estimated ideal mask by using an HMM along with the Viterbi algorithm.

3.5.1 Inter-Channel Time, Level and Phase-based methods

As stated in [J.Rennie, 2008]: "*Source-to-observation transfer functions in reverberative environments is highly dynamic and difficult to learn; spatially selective algorithms for speech separation offer a promising alternative to dealing with reverberation*".

Just as humans use Interaural Level Difference (ILD) and Interaural Time Difference (ITD) for sound localization, ICTD, ICLD and even ICPD properties are explored. Another name commonly used for ICTD is the Time-Difference of arrival (TDOA). As reported in [J.Rennie, 2008], with the delayed model,

$$x_m(t) = \sum_{n=1}^{N} a_{nm} s_n(t - \tau_{nm}) \tag{3.1}$$

and according to [Brandstein, 2005], if all sources are more than 1 m away from the microphone array, a_{nm} is approximately independent of n[3]. For very close microphones, a_{nm} is independent of m. This means that for far-field sources and tight microphone arrays $\forall_{n,m}\ a_{nm} \approx a$. This is important, for in most scenarios where the gear is away from the person, time becomes the single discriminant factor between sources as discrimination through instensity loses efficacy. [Viste and Evangelista, 2003; Favrot et al., 2006] also present studies related to binaural hearing parameters.

3.5.1.1 Azimuth Discrimination and Resynthesis (ADRess)

In [Barry et al., 2004], a very lightweight phase cancellation method is presented where artificial stereo panning retains the phase and only the amplitude between the channels is changed. The phase cancellation is achieved in the interchannel difference spectrogram but it requires different sources to have different pannings in the stereo mixture for separation. Unfortunately it requires the same signal in both the left and right channels and it shall differ only by panning. This implies ADRess uses an instantaneous mixing model where the time sound takes to travel from the source to each microphone is not considered and so it is ruled out for live environments and applicable only to stereo music mixtures.

An extension for microphone signals called Modified ADRess algorithm was presented in [Cahill et al., 2006] using a third microphone to separate speech sources with the same azimuth but different elevation from noise in a simulated room with two interfering noise sources. This leads to Signal to Interference Ratio (SIR) gains of 10.3 dB to 12.8 dB depending on the azimuths of the sources and the sources used. Signal to Distortion Ratio (SDR) ranged from 7.0 to 13.1 dB. When using 4 noise sources placed vertically and laterally with respect to the target source, using the third microphone doubled the SIR gain.

3.5.1.2 DUET

Degenerate Unmixing Estimation Technique (DUET)'s introduction [Jourjine et al., 2000] brought about a method that with only 2 mixtures and an anechoic model is capable of separating an arbitrary number of sources where the noise for each mixture is modelled as white Gaussian noise [Yilmaz and Rickard, 2004]. The number of clusters in the amplitude-delay histogram yields the number of sources if the signals

[3] This is because $a_{nm} \propto r^{-1}$ in spherical waves and the derivative gets smaller as distances increase.

are W-disjoint orthogonal (WDO) [Rickard and Özgür Yilmaz, 2002].

With [Yilmaz and Rickard, 2004] separation of convolutional mixtures were tested. The conclusion was that even though DUET was developed for anechoic recordings with W-disjoint sources, it works as well with small levels of reverberation and approximate signal W-disjoint orthogonality. The conclusions were that DUET provides perfect estimations of the parameters in the anechoic case but in the echoic scenario the IBM far outperforms DUET. Nevertheless, the non-ideal DUET masks still allow a satisfactory source separation in moderately echoic environments whereas ideal masks provide much better results so there is room for improvement [Yilmaz and Rickard, 2004].

One of the problems with DUET-like methods is the wrapping of phase due to multiple periods bringing about the same observable phase, which creates artifacts. In [Rickard, 2007], the restriction that the maximum interest frequency due to phase ambiguities is half the speed of sound over the maximum microphone spacing is removed. A more efficient alternative to the phase-aliasing was later proposed in [Yang et al., 2011].

Scott Rickard states that according to his eight-year experience working with DUET, multiple speakers talking simultaneously can be succesfully demixed with high-quality using two microphones and DUET [Rickard, 2007]. The system also works with mild reverberation and with low computational cost. This technique positions itself as a strong candidate for any real-time speech separation front-end, if yet another front-end that removes extraneous sources *a priori* is used beforehand.

DUET-DESPRIT [Melia and Rickard, 2007] introduces three extensions to the original DUET method and the study covers the use of DUET with more than two mixtures and in echoic environments.

In 2013 Johannes Traa proposed a new separation algorithm[4] based on DUET [Traa and Smaragdis, 2013; Traa, 2013], which explores ICPD in a model which allows phase-wrapping and is claimed by the authors to be robust in reverberant scenarios. In 2014, a study using the previous algorithm, which is very similar to DUET, determined that the optimal distance between two microphones is not superior to $5\,cm$ and that in that scenario the ICPD is much more informative than the ICLD, which becomes almost irrelevant [Johannes et al., 2014]. This study shows that separation through ICLD tendentially improves when the distance between the microphones increases and the opposite is true for the ICPD, though the peak performance for ICPD-based separation occurs around $3\,cm$. Combining both features and placing the microphones at that distance yields the best separation performance though the increase from the ICPD-based separation has moderate gains. Whilst the ICPD yields for a $3\,cm$ inter-microphone distance around $13\,dB$ of SDR, the combined ICPD-ICLD separation yields $14\,dB$ of SDR and in SIR the increase is from $20\,dB$ to $22\,dB$[5] in a simulated room with $T_{60} = 20\,ms$ and the speakers placed in the middle of the room, two meters apart, which attains a similar performance to the PF-ISFA.

DUET seems thus a method of choice for a restrictive scenario where only 2 microphones and real-time requirements are in place. Regarding the real-time performance of DUET, in [Rickard et al., 2001] a realtime version of DUET is devised that runs 5 times faster than real time on a 750MHz CPU in a laptop through an online ML gradient search method to track the mixing parameters. They assume

[4] Which will be henceforth designated by Traa.

[5] This values are approximated since they're taken from plots.

however that they know the number of sources which can't be done in a real setting thus the histograms will still be used.

3.5.2 Other methods

In [Mandel et al., 2010] several methods were compared, namely Model-Based Expectation-Maximization Source Separation and Localization (MESSL) [Mandel et al., 2010], the method developed by Sawada for convolutive mixtures and here referred to by the same name [Sawada et al., 2007], Mouba [Mouba et al., 2006], DUET and TRINICON. The tests were performed by placing two microphones at the ear positions of a dummy head. Notice that for DUET this should not even be tested since the diffraction by the head ILD changes with frequency and the microphones are placed farther apart than they should considering that the phase-aliasing extension was not used [Yang et al., 2011].

For the anechoic tests with two speakers the TRINICON far outperformed all the other methods with an SDR of 22.47 dB, approximately doubling the performance of any other method, but for the echoic test with two sources, MESSL-G was the best method, with SDR=5.87 dB while TRINICON achieved only 3.72 dB and DUET 2.35 dB.

TRINICON doesn't work however for the degenerate cases and with the anechoic test and three sources, MESSL-G achieved an SDR=8.41 dB and Mouba an SDR=6.34 dB, followed by Sawada and DUET, with SDR=4.98 and SDR=4.38 dB respectively.

MESSL is state-of-the-art, but very expensive computationally, as is one SSL, Variational EM SSL (VESSL) [Deleforge et al., 2013] that achieves even higher performance through ICLD alone by learning the high-dimensional manifold mapping from interaural variables to 3D-space variables.

In [Mandel et al., 2010], MESSL outperformed Mouba, Sawada and DUET, by this order. They were comparing DUET in a non-free-field condition with large microphone spacing and without the phase aliasing extension [Yang et al., 2011]. It is a surprise that DUET works at all. A new comparison should be done in a more fair setting with at least such DUET extension implemented to have fair results.

Additionally, in [Mouba et al., 2006] they claim that they're doing a real-time implementation but provide no timings. In MESSL, 2.5s of audio with 2 sources require 80s on a 1.86GHz Xeon and Mouba uses a similar methodology based on EM which means that the order of magnitude of the execution time of both should be close, which is not feasible for real-time thus this methods weren't chosen. Sawada also uses EM and frequency permutation alignment through autocorrelation so it is also computationally expensive in the same order of magnitude as the MESSL and Mouba's methods.

This methods were not tried because even if the method was able to run at half-real-time speed, it won't be enough to run in real-time, so a method that is guaranteed to be really light is preferred. The method of choice in an environment with two close microphones and small/moderate room reverberation time is DUET given the real-time requirements on comodity hardware.

3.6 Auxiliary techniques

3.6.1 Visual cues

In [Knuth, 2005], it is explained how to restrict the instantaneous mixing matrix A elements by noting that they are dependent only on the attenuation in the sound path $A_{mn} = \frac{1}{4\pi r_{mn}^2}$, which means that with better priors on the sources positions, a better A_{mn} posterior $p(A_{mn}|Positions_{Sources}, Positions_{Microphones})$ is reached, which can improved even by knowing only the maximum distance of the sources to the microphones. Better yet, would be to have at least priors on the angular distribution of the sources.

Use of additional cues for tracking sources is not unheard of as well. These cues could be used for counting the speakers, as well as tracking visually the positions of the sources, such as the human speakers mouths, because the source positions are essential in any source separation problem. Using this information along with visual identification of who's speaking directly at the device, as commonly humans do when they address someone, can help selecting the desired source in a multi-user system and even allow selection of a speaker-dependent trained model.

Using additional cues such as lip reading can might improve word recognition rates for speech signals. [Dong et al., 2005] show that adding visual cues added 12dB to the SNR of a source separation system.

3.6.2 Source localization/counting

Since the crux of DUET is identifying and localizing the unknown number of sources in the histogram, if a more robust method for that task is available and is computationally inexpensive it could be used.

Arberet et al. [2011] presented a method for anechoic environments that is capable of counting up to 5 sources using a chi-squared test and a confidence metric that penalizes outliers. The method determines the number of sources N without specifying *a priori* information.

For the instantaneous case the method counts up to 7 sources with precision. In the anechoic case a strange behaviour occurs: with 2 sources the percentage of correct N determinations is only 80% whilst it is again 100% for 3 and 4 sources and 90% for 5 sources. It is completely unusable with more sources (10% for 6 sources and 0% for 7). The spatial location of sources with DEMIX is much better than with DUET for any SNR level as reported in [Arberet et al., 2011] although the lower performance for 2 sources seems strange.

3.6.3 Speech cleaning

One way to enhance speech in a single-channel scenario is to train dictionaries of noise and speech but the model of the noise can be innacurate and online training can be computationally prohibitive. A common alternative is to assume that noise varies more slowly compared to speech and that speech is temporally sparse. In this case accurate Voice Activity Detectors (VAD) are required to estimate noise when there is no speech and keeping track of the minimum level of spectrum at the several frequencies is required. Based on the estimates of the noise the speech can be extracted using spectral subtraction, Wiener filtering or variants of either. The speech extraction is done by controlling the gain of the noisy

spectrum [Cho, 2013].

Other statistical filters that model the acoustic background and speech and can be applied to isolate the target speech [Frey et al., 2001; Kristjansson et al., 2004; Deng et al., 2001] were considered the state-of-the-art in [J.Rennie, 2008].

3.6.3.1 Spectral subtraction

Spectral subtraction relies on having a good estimation of the power spectra of the noise. Considering one source s, noise n and the measured signal x at time t, x(t) = s(t) + n(t), spectral subtraction tries to clean the signal by removing the estimative[6] of the noise power spectra in the frequency domain:

$$|\hat{S}(\omega)|^{\alpha} = |X(\omega)|^{\alpha} - \beta|\hat{N}(\omega)|^{\alpha} \tag{3.2}$$

where $\omega = 2\pi f$, f is the frequency, $\hat{S}$ is the estimated source in the frequency domain[7], X the measured signal in the frequency domain and $\hat{N}$ the estimated noise in the frequency domain.

If $\beta > 1$ oversubtraction is employed and if $\alpha = 2$ it is called power spectral subtraction [Cho, 2013].

Usually noise is estimated when the speaker is silent or has no energy at a certain frequency. The estimate comes from the average of noise magnitude for that frequency $|\hat{N}(\omega)| = E[|N(\omega)|]$, assuming the noise spectrum varies more slowly than the speaker's voice.

One of the problems with this approach is that phase is not considered [Evans et al., 2005]. Another problem is a wrong estimative of the noise and in that case Spectral Subtraction creates artifacts like musical noise. One common technique to minimize such artifacts is to oversubtract the estimate of the noise power spectrum and to prevent the spectrum from going below a minimum level called the spectral floor [Anuradha R. Fukane, 2011].

A recent development in spectral subtraction was the musical noise-free theory iterative spectral subtraction which has much reduced artifacts [Khan and Hasan, 2008; Inoue et al., 2010; Miyazaki et al., 2012; Kondo, 2014; Miyazaki et al., 2014] and was used in the real tests.

3.6.3.2 Wiener filtering

can be used to denoise signals and to perform source separation [Roux and Vincent, 2013] and uses as a metric the Minimum Mean Square Error (MMSE) to restore the underlying clean signals [Ephraim, 1992]. As an example, the estimated source spectrum using a Wiener filter[Cho, 2013] is:

$$\hat{S}(\omega) = \frac{E\left[|S(\omega)|^2\right]}{E\left[|S(\omega)|^2\right] + E\left[|N(\omega)|^2\right]} X(\omega) \tag{3.3}$$

The Wiener filter can be interpreted in speech enhancement as a Time-Frequency (T-F) mask/filter where the value of each T-F cell c represents $\frac{E_{Target}}{E_{Total}}(c)$ [Mowlaee, 2010].

[6] Terms with a hat indicate an estimation.
[7] Capital letters represent a quantity in the frequency domain.

The problem of mismatch of the signals extracted from different T-F bins when using Wiener filtering can be alleviated by smoothing the parameters of the filter thus reducing artifacts at the expense of separation capability. According to [Cov, 2012], temporal covariance smoothing [Yu et al., 2008] was the smoothing method that achieved the best trade-off between artifacts and interference. Consistent Wiener Filtering [Roux and Vincent, 2013] could also reduce artifacts and interference by using the redundancy of the Short-Time Fourier Transform (STFT) coefficients to enforce consistency of the reconstructed signals in the time domain. Cho [2013] shows that square-root Wiener-filtering is the power spectral subtraction method, with over-subtraction if $\alpha > 1$.

3.6.3.3 Dereverberation

Reverberation can be modeled as the convolution of a source by a linear filter called the Room Impulse Response (RIR), or Acoustic Impulse Response (AIR), $h(t)$. This function depends on the room characteristics and the positions of the source and the receiver in the room.

The Reverberation Time of the room $(R)T_{60}$[8], which measures the time from direct sound cessation to a SPL reduction of 60dB caused by reflections in the environment, is often used to describe the RIR.

The most accurate methods to measure T_{60} are non-blind and provide an excitation signal from which through correlation operations RT_{60} is estimated [Farina and Righini, 1997; Farina, 2007]. The most accurate, exponential sine sweep, is recommended for silent spaces and Maximum Length Sequence (MLS) for noisy spaces [Stan et al., 2002]. Estimating RT_{60} can also be done blindly (passively), when emitting excitation signals is out of question, by analyzing ambient signals onsets and decays [Vesa, 2004; Vesa and Harma, 2005], or comparing the late reverberation tail to an exponentially damped uncorrelated Gaussian noise model [Ratnam et al., 2003].

In room acoustics the RIR is usually decomposed into the direct path component, early reflections and late reverberation. The early reflections can in humans improve perception due to increased audible signal energy but change the signal $s(t)$ statistics and are characterized by sparse peaks in $h(t)$. In normal circumstances they are caused by the lowest-order reflections in the room and have thus strong correlations to the source and sensor positions in the room. The late reverberation is caused by higher-order reflections and gives rise to a diffuse field that has very low correlation to the sound source location. This same diffuse field presents stochastic characteristics and noise-like behaviour and is accurately modelled by plane waves with random phases and amplitudes in the framework of Statistical Room Acoustics [Naylor and Gaubitch, 2010].

One interesting property of the late reverberation is that since it has no geometrical correlation to the source location anymore, by knowing the profile of the late reverberation for a source position, the late reverberation suppression can use the same information to suppress late reverberations from a source at any location in the room. In [Tsilfidis, 2011; Tsilfidis and Mourjopoulos, 2011] this effect was explored by using as calibration signal a simple handclap.

[8] Usually calculated for the octave band centered at 500 Hz [Vesa, 2004].

Dereverberation techniques Finding an inverse impulse response G by simply inverting the filter in the frequency domain

$$G(k) = \frac{1}{H(k)} \tag{3.4}$$

doesn't work in practice since $h(t)$ is usually nonminimum phase [Bouchard et al., 2006; Perrin, 2012] and thus the inverse filter is not stable. Several methods exist that use a regularization mechanism to constrain the inverse filter search problem [Bouchard et al., 2006].

The common single-channel solution to the dereverberation problem is to find an approximation to the inverse filter by Least-Squares (LS) minimization. In a single channel system with one speaker the approximate dereverberation solution to the system is given by the Moore-Penrose pseudo-inverse but the error of the estimated filter does not converge to zero due to the nonminimum phase impulse response [Pedersen et al., 2007]. The Multiple Input/Output Inverse Theorem (MINT) allows the exact inverse of the impulse response to be found but requires more input channels [Pedersen et al., 2007; Miyoshi and Kaneda, 1988].

There are many approaches to speech dereverberation: channel shortening [Zhang et al., 2010], maximization of the skewness of the Linear Predictive (LP) residual [Mosayyebpour et al., 2010], or even combined late reverberation supression through a blind deconvolution combined with spectral subtraction [Furuya and Kataoka, 2007].

No methods were found that were capable of suppressing the early reverberation of multiple simultaneously active sources given the extremely different impulse response characteristics for their different positions. Most are not capable of suppressing the late reverberation for multiple active sources either. For instance, with LP Coding (LPC), estimating one speaker's vocal tract is possible but combinations of multiple speakers in a degenerate setting is not.

Taking the stochastic properties of the late reverberation (uncorrelated to the direct sound)[Naylor and Gaubitch, 2010, Chapter of Statistical Room Acoustics] and knowing that it is largely invariant to speaker position [Tsilfidis and Mourjopoulos, 2011; Tsilfidis, 2011], through a method that tracks the ambient sound statistics without explicit dependencies on the source(s), should be able to suppress the late reverberation of multiple active sources. This was experimented by implementing [Sadjadi and Hansen, 2012] and although it seemed to work, the results will be included in future work as the parameters need to be optimized.

Using today's methods, the suppression of reverberation before the BSS stage is only possible for the late reverberation part. Unfortunately, according to [López et al., 2013], applying a dereverberation stage prior to a BSS method does not improve separation considerably. This statement should be taken in moderation, but for BSS methods that perform separation through localization it should be largely true as the late reverberant part has no geometrical correlation to the source position so it should not introduce much bias in the source localization and separation algorithm.

3.7 Summary

There are methods providing separation with great quality (mainly used in music) like the NMF family, but usually at great computational cost [Yoshii et al., 2013]. Increasing the detail of the models significantly increases computational costs and this raises the question of which statistical or physical phenomena one wishes to model, or approximate, as it was seen in the section 3.5.1, for instance, the instantaneous mixing matrix coefficients can be considered in specific situations to be equal.

The biggest problem with the ICA methods is not signal gaussianity but the lack of guaranteed convergence for most variants [Yang et al., 2011] and the restriction to non-degenerate solutions.

High quality source separation can be achieved by using statistical source models defined in the log spectral domain and represented by spectral HMMs even with monaural mixes [Roweis, 2003; Radfer et al., 2010]. [Sawada et al., 2007] was used as source location model in [Nakatani et al., 2011] that was considered in [Thiemann and Vincent, 2013] to be state-of-the-art in BSS and demonstrated extremely consistent separation results.

One is still far from fully-general real-time audio-quality-oriented systems. Methods based on time-frequency masking can separate more realistic mixtures but have more artifacts and interferences due to not considering the spectral characteristics when segregating the bins according to the sources. Methods based on sinusoidal modelling have a more artificial timbre but less interference.

Methods that use T-F masking are one of the most successful approaches to degenerate unmixing problems. A promising real-time approach of that type is DUET [Jourjine et al., 2000], where it is assumed that signals are sparse and thus only one source dominates each T-F bin. Other options that are less restrictive in that aspect but more computationally expensive are available [Abrard and Deville, 2003; Puigt and Deville, 2006], but according to [Yilmaz and Rickard, 2004], DUET can still work in slightly non-ideal conditions.

Since the Head-Related Transfer Function (HRTF) shows a better sensivity when the array of sensors is pointed straight at the source, in [Valin et al., 2003] the robot head was actively rotated, improving the SNR for each of the two speakers in $10\,dB$. This just goes to show how the efforts in source selection systems can not only be concentrated on the algorithms themselves. The right physical recording setup can have dramatic effects on the performance of the system and should not only be taken into account as well as it should be studied in detail.

In this thesis a real-time deterministic system based on DUET was developed and implemented. DUET was chosen as the BSS method given its good properties, particularly given the constraints of using only 2 microphones (extensible to more if required) for separating degenerate signals in real-time, and thus able to perform sound source separation in a real live environment such as a living room with several people and audio background generated by a radio or a TV.

The method can be improved if combined with other stages that preprocess its input, namely late dereverberation and spectral subtraction, or provide additional cues such as number or location of sources.

Alternatives similar to DUET in that they also use W-disjointness of the sources to separate T-F bins through ICLD, ICPD or ICTD cues could have been considered as well.

Chapter 4

Development and System Implementation

4.1 DUET fundamentals

DUET adopts a noisy anechoic mixing model for two microphones and assimilates the differences in amplitude and delay between the channels into the second microphone model, since only relative differences matter [Jourjine et al., 2000]:

$$\begin{aligned} x_1(t) &= \sum_{n=1}^{N} s_n(t) + n_1(t) \\ x_2(t) &= \sum_{n=1}^{N} a_n s_n(t - \delta_n) + n_2(t) \end{aligned} \tag{4.1}$$

where the noise models are represented by independent Gaussian signals.

DUET requires W-disjoint orthogonal (WDO) source signals and different spatial signatures. Because there are two microphones, the latter means that any sources at the same azimuth and distance from the microphone pair won't be separable.

4.1.1 W-disjoint orthogonality

Two functions are *W-disjoint orthogonal* if given a window function $w(t)$ the supports of the windowed Fourier transforms of the functions are disjoint. The windowed Fourier transform[1] of $s(t)$ is defined as

$$S^w(\tau, \omega) := \mathscr{F}^w[s](\tau, \omega) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} w(t - \tau) s(t) e^{-i\omega t} dt \tag{4.2}$$

In DUET, to separate the sound sources, they must be W-disjoint orthogonal:

$$\forall_{(\tau,\omega)} \forall_{j \neq k} \quad S_j^w(\tau, \omega) S_k^w(\tau, \omega) = 0 \tag{4.3}$$

This allows the creation of binary masks M which are plain indicator functions with which a source can be separated from the mixture:

$$M_j(\tau, \omega) := \begin{cases} 1 & S(\tau, \omega) \neq 0 \\ 0 & otherwise \end{cases} \tag{4.4}$$

[1] If $w(t) = 1$, the windowed Fourier transform $\mathscr{F}^w$ is represented by $\mathscr{F}$.

$$\forall_{(\tau,\omega)}\ S_j^w(\tau,\omega) = M_j(\tau,\omega)X_1^w(\tau,\omega) \tag{4.5}$$

where X_1^w represents the transform of the signal from microphone 1.

4.1.2 Local stationarity

By the time shift property of the Fourier transform, true for $w(t) = 1$ and approximately true for other windows [Özgür and Scott, 2004, eq. 17], [Balan et al., 2000], state that

$$\begin{aligned}\mathscr{F}^w[s_j(t-\delta)](\tau,\omega) &= \exp(-i\omega\delta)\mathscr{F}^w[s_j(t)](\tau-\delta,\omega)\\ &\approx \exp(-i\omega\delta)\mathscr{F}^w[s_j(t)](\tau,\omega)\end{aligned} \tag{4.6}$$

where the delay of the transformed function is assumed negligible. This means that the solutions in the histograms aren't exact and will vary according to the delay δ.

For phase shifts small in relation to the window size the previous relation will hold even if the window has finite support. As stated in [Rickard, 2007, p. 221] this is often called the narrowband assumption in array processing but here *local stationarity* is more fitting.

DUET requires 4.6 to hold, even if w has finite support up to a maximum delay Δ/c between sensor signals, *id est*, $\forall_n\ |\delta_n|c \leq \Delta$ for the method to work.

4.1.3 Separation principle

By the W-disjoint property of the signals, for a single active source j at a certain ω :

$$\begin{bmatrix}X_1^w(\omega)\\ X_2^w(\omega)\end{bmatrix} \approx \begin{bmatrix}1 & \cdots & 1\\ a_1e^{-i\omega\delta_1} & \cdots & a_Ne^{-i\omega\delta_N}\end{bmatrix}\begin{bmatrix}S_1^w(\omega)\\ \vdots\\ S_N^w(\omega)\end{bmatrix} \overset{W\,disj.}{=} \begin{bmatrix}1\\ a_je^{-i\omega\delta_j}\end{bmatrix} S_j^w(\omega) \tag{4.7}$$

By this observation the most important quantity in DUET is $F(\omega,\tau)$:

$$F(\omega,\tau) = \frac{X_2^w(\omega)}{X_1^w(\omega)} = ae^{-i\omega\delta_j} \tag{4.8}$$

This is a complex quantity thus two quantities can be extracted, in this case the delay and attenuation between the microphones for the active source j can be obtained, forming the *amplitude-phase function* Λ [Yang et al., 2011]:

$$\begin{aligned}\Lambda = (a_j,\delta_j)_\omega &= \left(\|F(\omega,\tau)\|,\ Im\left\{\log F^{-1}(\omega,\tau)\right\}/\omega\right)\\ &= \left(\|F(\omega,\tau)\|,\ -\arctan\left(\frac{Im\{F(\omega,\tau)\}}{Re\{F(\omega,\tau)\}}\right)/\omega\right)\end{aligned} \tag{4.9}$$

Then each bin is assigned to a binary mask of one of the sources on the T-F domain.

This means the first step is to work in the T-F domain, find the clusters in the $a-\delta$ plane and per

bin for that time slice, construct the masks that select each source in the T-F histogram based on this clusters and possibly past information (temporal statistics) and convert back to the time domain.

4.1.4 Maximum microphone distance

In the standard DUET method one of the limitations is that the spacing of the microphones has a maximum value that depends on the maximum frequency analyzed in the signal. This is because there is phase-aliasing when determining a phase not smaller than $|\pi|$.

Take a phasor[2] $z(\phi) = e^{i\phi}$. Recovering a negative, null or positive phase from a complex number can only be done on the unit circle, which means that $|\phi| < \pi$ for the phase estimator to match the true phase.

Since DUET recovers the phase from $F = ae^{-i\omega\delta}$, phase aliasing does not occur for $|\omega\delta| < \pi$. Solving for $\Delta \geq c|\delta|$,

$$\omega_{max}\Delta/c < \pi \Leftrightarrow \Delta < \frac{c}{2f_{max}} \tag{4.10}$$

This limits the maximum distance between microphones to Δ where f_{max} is the maximum frequency of interest, according to which the Nyquist frequency is chosen and Δ defined.

For $f_{max} = 8kHz$ and 20^oC, the maximum microphone separation is $\Delta \approx \frac{343}{16000} \approx 2.1\,cm$

4.1.5 Building the histogram

To build the Λ histogram each T-F can have its own weight, given that more energetic frequency-bins are probably more reliable (more likely to be source energy instead of noise) and higher frequencies have smaller error, there are more optimal ways to weight the histogram, therefore, weighing parameters p and q are introduced for determining a weighted histogram H where each T-F bin has the weight [Rickard, 2007]:

$$|X_1^w(\tau,\omega)X_2^w(\tau,\omega)|^p\omega^q \tag{4.11}$$

Various options for the values of p and q can be adopted. Examples are [Yang et al., 2011]: $(p,q) = (0,0)$ for the original DUET setting, $(1,0)$ motivated by the symmetric attenuation estimator, $(1,2)$, motivated by the ML delay estimator, $(2,0)$ to reduce the delay estimator bias and $(2,2)$ for mixtures with low SNR.

After the histogram is built it needs to be smoothed. In this particular implementaion the smoothing was performed by the convolution of the histogram with a two-dimensional Gaussian kernel of diagonal covariance.

4.1.6 Building the masks

After having identified the clusters in the two-dimensional histogram, each cluster should correspond to a sound source since all frequencies from that source will, in theory, have the same $\alpha - \delta$ coordinates, which are coordinates that depend only on the geometric placement of the microphones and the source.

[2] In page 229 of [Rickard, 2007] there's an alternative explanation based on the Nyquist theorem.

The ML solution for the mask construction [Rickard, 2007, p. 227]:

$$J(\tau,\omega) := \underset{k}{\operatorname{argmin}} \frac{\left|\hat{a}_k e^{-i\hat{\delta}_k \omega} X_1^w(\tau,\omega) - X_2^w(\tau,\omega)\right|^2}{1+\hat{a}_k^2} \tag{4.12}$$

where the estimated mask $\hat{M}_n$ for the source $\hat{S}_n$ is

$$\hat{M}_n(\tau,\omega) = \begin{cases} 1 & J(\tau,\omega) = n \\ 0 & otherwise \end{cases} \tag{4.13}$$

Conversely, since each T-F bin has its own Λ coordinate, one of the formulas in section 4.3.3 could be used to assign it to the closest cluster, or even reject outliers, silencing noise T-F bins that are far from any cluster[3].

4.1.7 Applying the masks and rebuilding the signals

There are two ways to apply the masks. The simplest method is the Single-Channel (SC) reconstruction. This is a perfectly acceptable solution for closely spaced microphones, that have approximately the same energy, when the noise is low or there isn't much energy masking:

$$\hat{S}_n^w(\tau,\omega) = \hat{M}_n(\tau,\omega) X_1^w(\tau,\omega) \tag{4.14}$$

The Maximum Likelihood (*ML*) based reconstruction uses both microphone signals X_1^w and X_2^w in the STFT domain, where $\hat{M}_n$ is the mask to rebuild the source n:

$$\hat{S}_n^w(\tau,\omega) = \hat{M}_n(\tau,\omega) \frac{X_1^w(\tau,\omega) + \hat{a}_n e^{in\hat{\delta}_n \omega} X_2^w(\tau,\omega)}{1+\hat{a}_n^2} \tag{4.15}$$

The ML method can improve separation by combining the information from both channels when the estimated cluster positions are reliable.

After applying the binary masks either with the SC or ML reconstruction the inverse FFT is applied and another section of the signal is rebuilt in the time domain.

[3] This idea is not new [Aziz-Sbaï et al., 2012].

4.2 DUET Λ space

4.2.1 Coordinate system

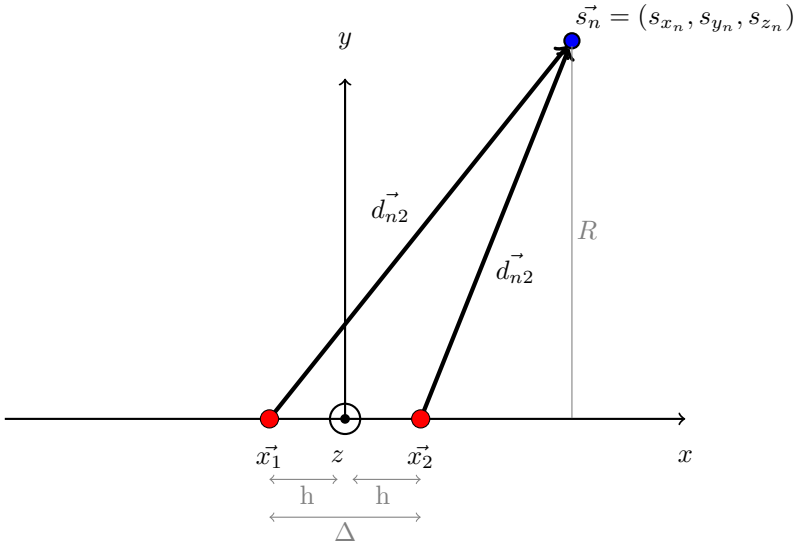


Figure 4.1: Coordinate system where the microphones are represented in red and the source in blue.

A cartesian system (x, y, z) with origin placed at the center of the microphone system (fig. 4.1) was chosen. The mixtures x_1 and x_2 are recorded at the microphone positions $\vec{x_1} = (-h, 0, 0)$ and $\vec{x_2} = (h, 0, 0)$ and the source s_n is placed at position $\vec{s_n} = (s_{x_n}, s_{y_n}, s_{z_n})$ where the index n can be dropped when the meaning is clear.

Given the rotational symmetry of (α, δ) about the x axis, if the position in the Λ space is known, its corresponding position in the conventional 3D-space can only be determined to lie on a circle (x, R), where $R = \sqrt{y^2 + z^2}$. It is thus often convenient to work in that manifold where the bijective $(x, R) \leftrightarrow (\alpha, \delta)$ map exists.

Finally, for tests where the z coordinate is negligible, the source position can be designated in polar coordinates on the $x-y$ plane through distance $d = \sqrt{s_x^2 + s_y^2}$ to the origin and angle $\theta = \tan^{-1}(s_x/s_y)$.

4.2.2 Λ space range

4.2.2.1 a-range

By modelling the acoustic sources as omnidirectional point sources, and considering no reflections, waves propagate spherically, $p = \frac{A}{r}$, where A is the reference amplitude at the origin, r the distance travelled and p_m the pressure at the microphone m. Considering a source placed at (s_x, s_y, s_z) in the room, $R = \sqrt{s_y^2 + s_z^2}$ and $h \equiv \Delta/2$:

$$a = \frac{p_2}{p_1} = \frac{A/d_{n2}}{A/d_{n1}} = \frac{d_{n1}}{d_{n2}} = \frac{\sqrt{(s_x + \Delta/2)^2 + s_y^2 + s_z^2}}{\sqrt{(s_x - \Delta/2)^2 + s_y^2 + s_z^2}} = \frac{\sqrt{(s_x + \Delta/2)^2 + R^2}}{\sqrt{(s_x - \Delta/2)^2 + R^2}} \tag{4.16}$$

Physically no two objects can simultaneously occupy the same space so $d_{n1} > 0$ and $d_{n2} > 0$, thus $a \in]0, \infty[$, where the limits 0 and ∞ would occur only if $d_{n1} = 0$ or $d_{n2} = 0$ respectively.

The minimum would occur when the numerator is 0 which is possible if $s_x = -\Delta/2$ and $R = 0$. The extreme values of a will be found when $R = 0$ because it is a summed component in both numerator and denominator, however, in all practical scenarios if the gear is placed at a distance from the sources (such as over a TV), $R > 0$ all the time.

Lets find the extreme points of a:

$$\frac{\partial a}{\partial R} = 0 \Leftrightarrow R = 0 \tag{4.17}$$

thus there is only one extreme point in $a(R)$, it occurs for $R = 0$. Increasing R brings a closer to 1.

Now varying only s_x we obtain:

$$\frac{\partial a}{\partial s_x} = \frac{(s_x + h)\frac{d_{n1}}{d_{n2}} - \frac{d_{n1}}{d_{n2}}(s_x - h)}{d_{n2}^2} \tag{4.18}$$

$$\begin{aligned}
\frac{\partial a}{\partial s_x} = 0 &\Leftrightarrow \frac{d_{n2}}{d_{n1}}(s_x + h) = \frac{d_{n1}}{d_{n2}}(s_x - h) \wedge d_{n2} \neq 0 \Leftrightarrow d_{n2}^2(s_x + h) = d_{n1}^2(s_x - h) \wedge d_{n2} \neq 0 \Leftrightarrow \\
&\Leftrightarrow (s_x - h)^2(s_x + h) + R^2(\cancel{s_x} + h) = (s_x + h)^2(s_x - h) + R^2(\cancel{s_x} - h) \wedge d_{n2} \neq 0 \Leftrightarrow \\
&\Leftrightarrow (s_x^2 - 2hs_x + h^2)(s_x + h) + 2hR^2 = (s_x^2 + 2hs_x + h^2)(s_x - h) \wedge d_{n2} \neq 0 \Leftrightarrow \\
&\Leftrightarrow 2[-2hs_x^2 + (s_x^2 + h^2)h] + 2hR^2 = 0 \wedge d_{n2} \neq 0 \Leftrightarrow s_x^2 = h^2 + R^2 \wedge d_{n2} \neq 0 \\
&\Leftrightarrow s_x = \pm\sqrt{h^2 + R^2} \wedge d_{n2} \neq 0
\end{aligned} \tag{4.19}$$

As seen above, the extrema points of a occur for $x = \pm D$ where

$$D = \sqrt{h^2 + R^2} \tag{4.20}$$

When $R = 0 \Leftrightarrow D = h$, the extrema solutions reduce to placing the source on top of one of the microphones as seen earlier, however, the general solution shows an interesting symmetry:

$$a(x = D, R) = \frac{\sqrt{D^2 + 2hD + h^2 + R^2}}{\sqrt{D^2 - 2hD + h^2 + R^2}} = \frac{\sqrt{2D^2 + 2hD}}{\sqrt{2D^2 - 2hD}} = \sqrt{\frac{D^2 + hD}{D^2 - hD}} \overset{D \neq 0}{=} \sqrt{\frac{D + h}{D - h}} = \max a(R) \tag{4.21}$$

$$a(x = -D, R) = \frac{\sqrt{D^2 - 2hD + h^2 + R^2}}{\sqrt{D^2 + 2hD + h^2 + R^2}} \overset{D \neq 0}{=} \sqrt{\frac{D - h}{D + h}} = \min a(R) \tag{4.22}$$

Since $h > 0$, by definition 4.20, $D \geq h > 0$, thus the previous solutions are general and the relationship between 4.21 and 4.22 is

$$\max a = \frac{1}{\min a} \tag{4.23}$$

This suggests a symmetric parameter α defined as:

$$\alpha = a - \frac{1}{a} \tag{4.24}$$

Because $a \in [0, \infty[$, $\max \alpha$ can be found when the difference grows large between a and $\frac{1}{a}$. This happens when a grows. The opposite term is relevant for the minimum since the minimum possible value

of a is 0:

$$\begin{aligned}\max\alpha &= \max\left(a-\frac{1}{a}\right) = \max a - \frac{1}{\max a}\\ \min\alpha &= \min\left(a-\frac{1}{a}\right) = \min a - \frac{1}{\min a}\end{aligned} \tag{4.25}$$

The following limit is more stringent than necessary since $\frac{1}{a}$ changes very rapidly in relation to a. By using the property 4.23,

$$\begin{aligned}\lim_{\max a\to\infty} \max\,\alpha &= \max a\\ \lim_{\min a\to 0} \min\,\alpha &= -\frac{1}{\min a} = -\max a\end{aligned} \tag{4.26}$$

the symmetry of α's range was proved.

The table 4.1 summarizes the behaviour of a and α.

Table 4.1: Behaviour of a and α.

$d_{n1} < d_{n2}$	$d_{n1} = d_{n2}$	$d_{n1} > d_{n2}$
$0 < a < 1$	$a = 1$	$1 < a < \infty$
$-\max a < \alpha < 0$	$\alpha = 0$	$0 < \alpha < \max a$

The broadest range for α is $\alpha \in]-\max a, \max a[$.

Adopting the parameter α there is a symmetry in the behaviour resulting from the distances d_{n1} and d_{n2}, which can be swapped in $\alpha = \frac{d_{n1}^2 - d_{n2}^2}{d_{n1} d_{n2}}$ resulting only in a signal change. This is equivalent to swapping the signals of the two microphones, as mentioned in [Rickard, 2007].

4.2.2.2 Typical α-range

The second-order Taylor expansion for $\max a$ in the limit $h/R \to 0$ is

$$\max a \approx 1 + \frac{h}{R} + \frac{h^2}{2R^2} \tag{4.27}$$

Since the variations in a are bounded by $\min R$ and usually people are at least $1\,m$ away from the microphones and thus from the microphone array axis, $a \approx 1$, thus to get the α-range needed for the histogram a rough estimate for $\max a$ in a typical real scenario is determined:

According to 4.21, if $\min R = 1\,m$ and $\Delta = 5\,cm$ which is already a big value for Δ since better separation is attainable for smaller values of Δ, as the report [Johannes et al., 2014] suggests,

$$\max a \approx 1 + 0.05 + \frac{0.05^2}{2} = 1.05125 \to \max\alpha \approx 0.1 \tag{4.28}$$

4.2.2.3 δ-range

$$\begin{aligned}c\,\delta = d_{n2} - d_{n1} &= \sqrt{(s_x - \Delta/2)^2 + s_y^2 + s_z^2} - \sqrt{(s_x + \Delta/2)^2 + s_y^2 + s_z^2}\\ &\equiv \sqrt{(s_x - \Delta/2)^2 + R^2} - \sqrt{(s_x + \Delta/2)^2 + R^2}\end{aligned} \tag{4.29}$$

Again the same strategy can be used as for calculating the range of a.

The maximum occurs when $R = 0$,

$$\max \delta = \frac{1}{c} \max \left(|s_x - \Delta/2| - |s_x + \Delta/2| \right) = \frac{\Delta}{c} \tag{4.30}$$

and it is a symmetric quantity.

4.2.3 Phase aliasing analysis

In order to avoid phase-aliasing problems (see 4.1.4), DUET is implemented in a way that corresponds to the region where Traa's phase plot doesn't wrap around. If phase aliasing occurs with DUET and without the extension [Yang et al., 2011], the delay estimates become noisy whilst with Traa phase cross-overs occur.

Since a signal can be decomposed into a series of *sin* and *cos* functions, consider $s(t) = sin(\omega t) + sin(k\omega t)$, which implies

$$s(t - \delta) = sin(\omega t - \underbrace{\omega\delta}_{\phi_\omega}) + sin(k\omega t - \underbrace{k\omega\delta}_{\phi_{k\omega}}) \tag{4.31}$$

DUET uses the delay δ to which the phase ϕ is proportional :

$$\phi_\omega \equiv \omega \, \delta \tag{4.32}$$

The phase can be modeled as a sum of two parts, the principal solution $\tilde{\phi_\omega}$ where $|\tilde{\phi_\omega}| < \pi$ and n_ω turns on the unit circle:

$$\phi_\omega = 2\pi n_\omega + \tilde{\phi_\omega} \tag{4.33}$$

The estimated phase $\hat{\phi_\omega}$ can be obtained from the polar angle of $F(\omega, \tau)$ (eq. 4.8). However, estimating the argument of a complex number recovers only the principal solution,

$$\hat{\phi_\omega} \equiv \hat{\tilde{\phi}}_\omega \tag{4.34}$$

If a frequency is delayed by more than one period, spatial aliasing in the binaural histogram will occur, and even if the estimative of $\tilde{\phi_\omega}$ was exact, spatial aliasing can corrupt the whole histogram. To see this notice that the error of the delay depends then on the number of periods:

$$\hat{\delta} = \frac{\hat{\phi}_\omega}{\omega} = \frac{\hat{\tilde{\phi}}_\omega}{\omega} \tag{4.35}$$

The error of $\hat{\delta}$ is then:

$$\epsilon_\delta = \hat{\delta} - \delta = \frac{\hat{\tilde{\phi}}_\omega - (2\pi n_\omega + \tilde{\phi_\omega})}{\omega} = \frac{\epsilon_{\tilde{\phi_\omega}} - 2\pi n_\omega}{\omega} \propto \frac{2\pi n_\omega}{\omega} \tag{4.36}$$

The error formula for the $\hat{\delta}$ estimation is very insightful into the behaviour of DUET.

If there is no phase estimation error $\epsilon_{\tilde{\phi_\omega}}$ and the phase is contained in the principal branch, $\hat{\delta} = \delta$. If the no phase aliasing condition is met but there is an estimation error on the phase ($\epsilon_{\tilde{\phi_\omega}}$) the error in

the delay (4.36) is smaller for higher frequencies. If there is phase aliasing, even if there is no estimation error in the principal phase ($\epsilon_{\tilde{\phi}_\omega}$), there is an error in $\hat{\delta}$ proportional to the number of periods n_ω and the error is frequency-dependent (4.36). The relative error of $\hat{\delta}$ in relation to the estimated value itself is

$$\epsilon_{\delta_r} \equiv \frac{\epsilon_{\hat{\delta}}}{\hat{\delta}} = \frac{\hat{\delta} - \delta}{\hat{\delta}} = \frac{\epsilon_{\tilde{\phi_\omega}} - 2\pi n_\omega}{\hat{\phi}_\omega} \approx -2\pi n_\omega / \hat{\phi}_\omega \tag{4.37}$$

Considering no phase estimation error the relative error caused by spatial aliasing even with $n_\omega = 1$ is at least 2 in magnitude since the estimated phase is unitary which means the spreads in $\hat{\delta}$ can be quite relevant, as reported in [Yang et al., 2011].

By comparison, phase aliasing occurs for each source at the frequencies above the first wrap-around in the phase-frequency plot of the Traa method. The errors in the estimates of the Traa method are not smaller at higher frequencies since the phase estimates are not divided by the frequency to obtain the delay as in DUET.

4.3 DUET implementation/modifications

4.3.1 Two-dimensional Λ histogram vs. Λ histogram marginals

The peak search and smoothing by convolution operations in the two-dimensional Λ space are considerably more computationally expensive when performed in a two-dimensional histogram (scale with $O(B_\alpha B_\delta)$[4]) instead of on its marginals $\hat{\alpha}$ and $\hat{\delta}$, where they scale by $B_\alpha + B_\delta$.

It was studied first if performing the clustering could be accurately performed using only the marginals of the Λ space. It turns out that theoretically it is possible to do so since even if two clusters share either α or δ by evaluating the likelihood of a cluster at each combination of $\hat{\alpha}$ and $\hat{\delta}$ clusters.

However, while such information can be disambiguated, the two-dimensional histogram is in fact one of DUET's advantages since it is a sparser representation domain than its marginals. For reliable peak tracking in any but the most simple conditions, this sparsity becomes important. That is why all the operations are performed in the two-dimensional histogram, except the measure of confidence which is analytically represented by calculations over the Λ marginals.

4.3.2 Heuristic Clustering

As stated in [Rickard, 2007, p. 226], tracking the peaks is not easy, it is in fact the crux in DUET. Even after trying weighted k-means, model-based peak removel and peak tracking but no single technique appeared to be appropriate in all settings and often manual identification of the peaks is required according to [Rickard, 2007].

The heuristic clustering was devised as a simple means to find the clusters on the histogram efficiently from the histogram data and uses the histogram bins and not the underlying data which limits the precision of the localization to the bins size unlike methods such as *k-means*.

[4] B_α and B_δ represent respectively the number of α bins and δ bins.

The method has two stages. The first is a pre-filter which scans for peaks in the histogram. If the histogram is chosen to be boundless, which means all the counts beyond the histogram space are accrued to the nearest bin in the border, the borders won't be scanned as possible peaks as it is assumed that this is not the intended search space.

To be considered a peak, the current bin must have a minimum height which sets the noise threshold. The height must be larger than any of the four adjacent bins on the 2D cartesian grid by a margin which is also adjustable:

$$\begin{cases} bin(\alpha,\delta) > bin(\alpha \pm d\alpha \quad , \ \delta \qquad) + margin \\ bin(\alpha,\delta) > bin(\alpha \qquad\quad , \ \delta \pm d\delta) + margin \end{cases} \tag{4.38}$$

which sets both the minimum absolute value for the discrete derivative and the maximum for the negative second derivative around the peak.

The second stage is the aggregation stage. This step is only required if the histogram resolution is larger than the irregularities in the data distribution around the peak.

Even in the simulation the peak shape can be irregular meaning that the same sound source, which should have only one peak, can exhibit multiple smaller peaks around the peak's maximum. This only happens when data is being overfitted, which means, in this case, that the resolution of the histogram is larger than it should, so variations in the Λ distribution lead to such artifacts. This stage thus takes two parameters: $\Delta\alpha_{min}$ and $\Delta\delta_{\min}$ and if any peaks from the previous stage are closer both in α and δ than this two values the peak with lower height is merged into the bigger one in the peak list for that particular histogram.

This can be sometimes an efficient alternative to the smoothing of the histogram, which requires many more operations. Other times it can be useful to use both, as the aggregation stage is capable of eliminating saddle points.

Spatial signature of the clusters Similar to the Rayleigh criterion in spectroscopy, since we're using mode identification instead of a GMM model for peak separation, $\forall_{j\neq k}\ (a_j \neq a_k) \vee (\delta_j \neq \delta_k)$, the peaks are identifiable as long as they differ by an amount between 2σ and 3σ [Schilling et al., 2002] due to the dispersion of the Λ parameters around the true values, if the dispersion in the histogram follows a normal (as it approximately does when we use the smoothing Gaussian Kernel even if the underlying distribution is not gaussian). If the peaks were two equal normals they only need to be apart by 2σ.

Alternatively, unimodality is a strong way to aggregate datapoints to the same source in a histogram under the gaussian smoothing (assuming the shape of the histogram becomes gaussian which it does for small data variances and wide gaussian smoothing) as to have a 36% dip between two equal normal densities they must be separated by 3σ [Schilling et al., 2002].

4.3.3 Λ-Distance metrics

The chosen distance from a bin at (α, δ) to a cluster k at (α_k, δ_k) with standard deviations $(\sigma_{\alpha_k}, \sigma_{\delta_k})$[5] is a modified Mahalanobis distance for a diagonal covariance matrix with scaling factors (s_α, s_δ):

$$D(\alpha,\delta)_k = \sqrt{\frac{(\alpha-\alpha_k)^2}{\sigma^2_{\alpha_k}} s_\alpha + \frac{(\delta-\delta_k)^2}{\sigma^2_{\delta_k}} s_\delta} \equiv \sqrt{D^2_{k_\alpha}(\alpha) + D^2_{k_\delta}(\delta)} \tag{4.39}$$

For a bivariate Gaussian distribution with mean $\boldsymbol{\mu} = (\alpha_k, \delta_k)^T$ and diagonal covariance matrix,

$$E[D^2_k] = E[D^2_{k_\alpha}] + E[D^2_{k_\delta}] = s_\alpha + s_\delta \tag{4.40}$$

Choosing the standard Mahalanobis distance $s = (1/D, 1/D)$ would mean the expected value would be 1 for each of the one-dimensional distances D_{k_α},D_{k_δ} and 2 for the two-dimensional D_{k_δ} so the unbiased choice is $1/N$ for each of the stretching factors where N is the number of dimensions.

With this choice, if the bin and the cluster are aligned in all but one direction d, D_k measures the distance to the cluster in that direction in σ_d units.

The additional use for the stretching factor which is to incorporate extra information such as the confidence information (section 4.3.4). Either local $(\Gamma_{\alpha_k}, \Gamma_{\delta_k})$ or global confidence metrics $(\Gamma_\alpha, \Gamma_\delta)$ can be used interchangeably on the following stretching factors.

A different alternative is $s = (\frac{\Gamma_\alpha}{\Gamma_\alpha+\Gamma_\delta}, \frac{\Gamma_\delta}{\Gamma_\alpha+\Gamma_\delta})$ which is the dimensional-compression metric. The expected value of its square is well defined

$$E[D^2_k] = \frac{\Gamma_\alpha + \Gamma_\delta}{\Gamma_\alpha + \Gamma_\delta} = 1 \tag{4.41}$$

making it useful as a statistical tool as the unbiased distance $s = (1/2, 1/2)$ but with the additional advantage of compressing the data in a dimension according to its confidence, thus a dimension with lower confidence has less weight in the distance effectively reducing the distance penalties on noisy dimensions.

4.3.4 Confidence metrics

Confidence metrics were developed to aid in the histogram peaks identification and separation. They determine if a peak (local confidence) or a dimension (global/coordinate confidence) is reliable or not.

4.3.4.1 Local confidence metrics

The local confidence metric is used to classify a peak as reliable or not. Assuming that the appropriate amount of smoothing is used (i.e., the smoothing kernel hides meaningless fluctuations in the data distribution under a smooth shape with one mode) and given that the smoothing kernel is a 2D Gaussian function of diagonal covariance, then, if the data variance for that peak is well behaved (much smaller

[5] Which are set to the standard deviations of the smoothing kernel.

covariance), the shape of the peak should remain Gaussian. The larger the data covariance is, located in the region of the peak, the larger the data spread and deviation from the Gaussian shape will occur for that peak, resulting in a spreading of the shape along the strongest covariant direction and the flatter the distribution will be, when compared to the smoothing (Gaussian) kernel.

Assuming the reliable peaks are the ones that deviate the least from the Gaussian shape of the smoothing kernel a procedure to evaluate that shape is required.

Several methods could be used such as fitting the whole two-dimensional histogram with a Gaussian Mixture Model (GMM) and the resulting model's characteristics would be evaluated at the peak but this is too computationally expensive. Fitting a single Gaussian to the peak under its domain would be a less computationally expensive option but wouldn't give better results than performing a statistical test restricted to the data respective to the peak domain.

For simplicity, the first measure that was tried out was the chi-squared error from the normalized peak to the smoothing kernel, with completely unusable results. Therefore, the next step was to use the excess kurtosis, which measures the shape of a distribution.

Two-dimensional versions of kurtosis exist but since DUET's $\hat{\alpha}$ and $\hat{\delta}$ estimated parameters show independent behaviours[6], the smoothing kernel shall have a diagonal covariance so that each axis in the two-dimensional histogram can be evaluated independently. The only advantage of a two-dimensional kurtosis test would be the detection of deviations along a diagonal, but with any reasonable smoothing size, deviations along a diagonal can be detected as well, in at least one of the axis (α or δ), through the two independent one-dimensional kurtosis tests.

As is common practice in the machine learning subject, models should be constrained as much as possible, to reduce the number of degrees of freedom of the system. For a real-time application, the same is true but the objective is to reduce the computational complexity of the problem.

Since the peak mode is already known, the kurtosis was constrained to use it as the parameter $\boldsymbol{\mu}$. Also, given that the smoothing kernel is calculated up to $\pm 3\sigma$ along each direction there is no point in calculating the kurtosis farther from this range, besides being a moment highly sensitive to variations in the tails and DUET is inherently noisy, especially for real-world data, so this variations should be neglected.

A further step was to constrain as well the kurtosis with the smoothing σ value but this provided really bad results, with very few true positives. This is explained by the underlying unknown data distribution adding to the smoothed distribution's covariance. The conclusion is that using any comparison other than the mode to the original profile is not a good choice.

Another problem stems from the fact that if another Gaussian from another sound source is too closely located and if it mixes with the contemplated source's one, in the 3σ range, it will result in a flatter distribution and thus the otherwise considered positive peak can now be classified as a false negative. However, this only lowers the acceptance rate. Since it doesn't provide false negatives, this method is very robust as will be seen in the results provided in the next chapter. In fact, under simulations, as long as the amount of smoothing is adequate and the size of the histogram is enough so that the

[6]Physically, unlike $\hat{\alpha}$, $\hat{\delta}$ is not sensitive to imbalances of inter-channel gain for instance.

smoothing of the peaks doesn't show up as discretized, this local confidence metric correctly identified all peaks/fake peaks even when the only true peak was smaller than the main peak of the histogram. This particular case occurred for a two-dimensional histogram with 9 unreliable peak and one reliable peak. The true peak rejection rate was really low and unreliable peaks were all rejected using the kurtosis (β_2) thresholds of decision as $|\beta_2(\alpha)| < 0.25$ and $|\beta_2(\delta)| < 0.2$ (as α tends to deviate more than δ from the proper distribution even for reliable peaks).

This approach works because, generally, peaks coming from unreliable data, unlike the ones generated from reliable data, show large spreads and no particular average.

Otherwise since one assumes the error to be non-stationary we can accumulate several frames and this erratic peaks should not add up as much as the sources.

The peaks in the histogram will always have negative excess kurtosis since they result from spreading, which was smoothed with a Gaussian kernel, thus the final distribution is flatter than the Gaussian kernel shape.

By definition, given a probability density function (pdf) H, the kurtosis will be[7]

$$Kurt = \frac{k\sum(x-\mu)^4\tilde{H}}{(k\sum(x-\mu)^2\tilde{H}(x))^2} \tag{4.42}$$

where $k > 0$ is the peak scaling constant, $\tilde{H}$ the histogram normalized to have the same peak height as the Gaussian kernel function, μ the population mean and x the dimension upon which kurtosis is being measured. Since this is for a pdf and not for a sample as one doesn't know how many points are in the peak region, the histogram shape should be smooth to be a good approximation of the pdf.

Assuming that $H = k\tilde{H}$ wasn't normalized thus removing the normalization bias means that the excess kurtosis is calculated through:

$$Kurt \equiv k\frac{\sum(x-\mu)^4\tilde{H}(x)}{\left(\sum(x-\mu)^2\tilde{H}(x)\right)^2} - 3 \tag{4.43}$$

Note that kurtosis is calculated by fixing the mean as the mode of the cluster. It would also be possible to constrain the second moment but since the distribution might be far from the expected this wouldn't measure the kurtosis anymore.

Calculation is done up to 3σ since the size of the smoothing kernel in bins is used as reference to calculate the kurtosis. This is reasonable since the smoothing kernel at $\pm 3\sigma$ covers 0.9946 of the pdf for a diagonal covariance whilst it covers 0.9973 in the 1D case.

Calculating the kurtosis beyond this range would make the outcome very sensitive to the peak tails and since the condition for separation of peaks resides in good kurtosis values (no GMM in the analysis range) then a larger range would result in many positives being rejected.

Evaluation of the 1D kurtosis through the α and δ marginals of the Λ histogram was also tested but since the marginal quantities offer a less sparse representation of the peak locations in Λ space (less peaked), the probability that different peaks interact is greater and thus kurtosis becomes inneffective.

[7] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

This means that for a local confidence metric the best kurtosis-based method is the individual evaluation of the kurtosis of $\hat{\alpha}$ and $\hat{\delta}$ in the two-dimensional Λ space histogram.

Skewedness could be used as well but kurtosis alone already gave very good results in almost 100% of the cases. Some atypical skewed distributions might exist that have a similar kurtosis to the Gaussian function but given the tails that show up in the DUET histogram, even for peaks corresponding to true sound sources, using skewedness would eventually reject them thus increasing the false positives rate.

4.3.4.2 Global/Coordinate confidence metrics

An additional way to measure confidence is by using the fact that the distributions in the histogram are always as flat or flatter than the used smoothing, and summing the volume of the two-dimensional histogram in the $\pm 4\sigma$ vicinity of the detected peaks, after rejecting low peaks derived from noise, provides a measure of the spread of the histogram in that variable. If there is too much spread, then it is not a reliable variable to use.

The global confidence metric for a dimension γ is defined as

$$\Gamma_\gamma = \sum_k \Gamma_\gamma(\gamma_k) \tag{4.44}$$

where k is the cluster and

$$\Gamma_{x_k} = \frac{\int_{x_k-3\sigma_{x_k}}^{x_k+3\sigma_{x_k}} H_x(x)\,dx}{\int_{\min x}^{\max x} H_x(x)\,dx} \tag{4.45}$$

If using this as a local confidence metric small peaks have a higher chance of being merged to bigger peaks which can be an advantage if that peak wasn't a source by itself. The kurtosis-confidence-based metric would have the same effect but with other behaviour obviously.

4.3.5 Source streaming algorithm

To run online a streaming-system was developed. The final product should use this streaming-system to feed the ASR with commands and process them according to ID of the speaker and relevance.

The streams are produced under the proposed Life-Death-or-Merge-Birth (LDMB) algorithm which first analyzes if a stream should remain alive through the correlation between the overlapping portion of the windowed signals that exhibit continued activity, and lastly, position. If there isn't activity for more than a $T_{persistence}$ period the stream either dies and then should be processed/finish processing by the ASR or is merged to another stream (content spread in the histogram if the stream content was too small to be considered autonomous). After all this if there are still unprocessed audio clusters this are considered new streams and the process continues. Also since the instantaneous separation provided bad results, statistics were accumulated and used in the streaming system.

4.4 Generic implementation details

4.4.1 Adding the inverse transforms back together

Since overlapping windows can be used, a way to add them back into a single contiguous output is required. This is implemented via a swap-and-add algorithm.

To write the data the complete buffer must be passed from the processing stage to the output stage and a blocking function takes care of writing the overlapped output windows to a contiguous output stream. If an asynchronous function was used there would have to be synchronization with the caller routine and the complexity of the writer function would be much larger but could be implemented through a state machine that advances the output and releases the buffers as soon as they are processed. This is essential for reutilization of the buffers. Since the program must run in real-time the same buffers must be used over and over.

A synchronous implementation was chosen and it is limited to 50% inter-buffer overlap.

This synchronous function works by writing all the data it can until a new overlap occurs and no more data is available. At that point it returns and stores its state. Only after the next call with a new buffer, the previous buffer can be reused if there is any overlapping. The number of samples of interval between windows is FFT_{slide} and the size of the FFT transform FFT_N. If $FFT_{slide} = FFT_N$ after each call the buffer can be reutilized already. The code works for $FFT_N/2 < FFT_{slide} < FFT_N$.

Listing 4.1: Pseudo-code to perform real-time output

```
if (overlapping_buffers == 2)
  {
    b = new_buffer;
    while (i < FFT_N)
      {
        //o[pos] = a[i] + b[i-FFT_slide];
        for (uint source = 0; source < o.sources(); ++source)
          o(source, pos) = (*a)(source,i) + (*b)(source,i-FFT_slide);
        ++pos;
        ++i;
      }
    RELEASE(a);
    i = FFT_N-FFT_slide;
    a = b;
    overlapping_buffers = 1;
  }

// overlapping_buffers == 1
while (i < FFT_slide)
  {
    // o[pos] = a[i]
    for (uint source = 0; source < o.sources(); ++source)
      o(source, pos) = (*a)(source, i);
    ++pos;
    ++i;
  }
overlapping_buffers = 2;
// Now wait for new call with new_buffer
```

4.4.2 Source separation degeneracy

Given an unknown number of true sources N and rebuilt sources N_e, DUET can have degeneracies in its outputs and D_{total} was the chosen metric to evaluate degeneracy as it is always able to identify the pair of sources that are closest unlike SIR that starts to break down as output sources get worse[8], used as foundation for pairing criteria (total SIR of the outputs) in the BSS Eval Toolbox.

Given the set of all estimated source signals $\hat{s}_i$, $i \in \{n \in \mathbb{Z} : 1 < n < N_e\}$ and original source signals s_n, $n \in \{j \in \mathbb{Z} : 1 < j < N\}$ a pairing $\hat{s}_i \leftrightarrow s_n$ is the solution for

$$\underset{i,n}{\operatorname{argmin}} \, D_{total}(\hat{s}_i, s_n) \tag{4.46}$$

If $\hat{s}_i \leftrightarrow s_n \wedge \hat{s}_{i'} \leftrightarrow s_n$ where $i \neq i'$, the output has estimated source degeneracy. Conversely, if $\hat{s}_i \leftrightarrow s_n \wedge \hat{s}_i \leftrightarrow o_{n'}$ where $n \neq n'$ the output has original source degeneracy, where both can be counted by the number of occurrences of each degeneracy type.

The outputs can also be classified into over-degenerate ($N_e > N$ where $N_e - N$ estimated source degeneracies forcefully occur), under-degenerate ($N_e < N$ where $N - N_e$ original sources degeneracies forcefully occur), or intrinsically degenerated (when degeneracies occur and they shouldn't have, i.e., any estimated or original source degeneracies for $N_e = N$, original sources degeneracies for $N_e > N$ or estimated source degeneracies for $N_e < N$).

While over and under-degeneracies are acceptable, having intrinsical degeneracies is not as it means the system is not performing the proper source separation. This can happen for $N_e = N$ if two or more Λ clusters are very close together and one of the sources has much less energy than the other(s).

4.4.3 Source separation metrics changes

The *BSS Eval Toolbox* did not work right out of the box, requiring additional parts from the Internet. It is only meant for scenarios where the number of rebuilt and true sources are the same. Neither degenerate solutions are taken into account or support for at least 24-bit *.wav* files is available, which is essential to mitigate quantization errors for weak sound signals, therefore the toolkit was extended[9].

The optimum criterium chosen for source pairing was first changed from the total output SIR to the maximum SIR per each source to allow over and under-degenerate solutions, which should be reliable as DUET only works with sources that are dissimilar (W-DO assumption). However, using SIR as source pairing criteria provided unstable results. If sources full of garbage existed, sometimes the BSS Eval would select them over sources that were very similar to the original sources. This means that SIR is not a good metric to quantify signal similarity. BSS Eval was changed to use D_{total} instead which guarantees the closest source match is actually chosen. Given that intrinsical degeneracies might occur even for the $N_e = N$ case, the choice of D_{total} over the total SIR of the outputs should be used instead.

[8] It was found that even selected completely garbled sources when better sources (close to the original signal) were available.

[9] The modified code will be made available online in a GPL license.

Chapter 5

Tests and Results

5.1 Simulation

To test the DUET implementation under consistent and easy to reproduce conditions an anechoic sound simulator was designed[1] that allows the arbitrary positioning of a set of sound sources whose signals are input from *.wav* files.

The model chosen for the simulation tests was the simplest possible, the delayed mixing model and spherical waves with omnidirectional microphones, which means that only the direct paths are considered from each source to each microphone and the amplitude of the captured signal (pressure) follows $p \propto r^{-1}$, as seen in the chapter 2. The maximum simulated sound pressure due to the n-th source on the m-th microphone is defined as p_{nm}, and its normalized counterpart $\tilde{p}_{nm}$:

$$p_{nm} \equiv \frac{A_n}{d_{nm}} \ , \quad A_n \equiv \max_t |s_n[t]| \tag{5.1}$$

where d_{nm} is again the distance between source n and microphone m.

Since the global normalization level of the inputs[2] is a free variable, to minimize information loss, normalization was chosen so that the strongest signal propagated to any of the microphones corresponds to a maximum amplitude of 1, the maximum *.wav* output range with floats through the *libsndfile* library.

The normalization procedure can thus be the following:

$$\begin{aligned} p_{max} &\equiv \max_{n,m} p_{nm} = \max_{n,m} \frac{A_n}{d_{nm}} \\ \tilde{A_n} &= \frac{A_n}{p_{max}} \\ \tilde{p}_{nm} &\equiv \frac{\tilde{A_n}}{d_{nm}} \end{aligned} \tag{5.2}$$

The positions of the two microphones were always fixed at $(x, y, z) = (-\Delta/2, 0, 0)$ and $(\Delta/2, 0, 0)$ where Δ is set according to the maximum sampling frequency f_s which is determined from the chosen source *.wav* files.

Each sound sample from each source reaches each microphone with a delay t_{nm}:

$$t_{nm} = \frac{d_{nm}}{c}$$

[1] A first implementation was done in Python using Numpy but still it took too long to simulate (*circa* 20s for 8s of audio). Thus, a C++ version was built, which ran the same test under 100ms.

[2] Each *.wav* file can have its own normalization.

where d_{nm} is the distance from the source n to the microphone m and c is the speed of sound in the room.

The values of the delays won't normally be multiples of the sampling period. Thus the solution is to interpolate the signals or accept a spatial resolution grid for the sources meaning they'll be effectively placed on top of the points on this grid and not on the chosen position. The spatial resolution of this grid can be calculated though. Since the sampling interval is $\Delta t_s = f_s^{-1}$, then $\delta d = c f_s^{-1}$. Typical values for the sampling frequency for speech recognition are 8 and 16 kHz which results in approximately 2 or $4cm$ resolution for the sources with $c = 340\ m/s$. Since the sampling frequency is relatively high for the speech content the linear interpolation was chosen which is simply a first-order Taylor expansion.

An example of a configuration file for the C++ simulator is:

```
# Choose the audio file folders
sources_folder = sounds_database
output_folder  = sounds
# speed of sound (m/s)
c = 340
# Sources position: (syntax: "source_file.wav = x y z")
mike.wav = 0.5 5 0
mike2.wav = -1.5 5 0
```

First-hand experience showed that this format, where the position of each source must be declared along with its source file is much less error-prone than declaring sources and placements separately.

5.2 Experimental setup

Because in a first approximation sound waves behave linearly, the superposition principle is valid allowing the tests signals to be recorded separatedly. The source separation scenarios are then created by digitally summing the different recorded source signals. The advantage of this method is that it allows the calculation of the metrics which would not be possible to calculate without knowing the original signals. The disadvantage with this method is that at every recording there is noise added by the microphones and pre-amp system that increases as more sources are mixed. The alternative is to measure all the signals at once, which is less flexible.

5.2.1 Sound database creation

A lot of research was done to find a speech database appropriate for a rigorous study, however most databases are not free (TIMIT, the most widely used requires a special license) and the free ones had dubious recording quality, such as VoxForge, where most entries suffer from either reverberation, clipping or distortion from strong plosive sounds or background noise from cars, other speakers, animals or fans. For the samples that seemed to be of good quality, the recording process varied with the user who donated the speech data and was not rigorously documented, i.e. if any filtering was applied, thus a small database was built specially for this purpose.

Clean and anechoic 48kHz recordings of four subjects, two male (M1 and M2) and two female (F1 and F2) were performed in a room with small reverberation time. The subjects were recorded at 30cm from

a Zoom H2 to avoid reverberation effects but not directly towards it to avoid using a pop filter. Each subject read five sentences designated: a1,a2,b,c1,c2,c3 which were phrases from a portuguese scientific magazine, chosen to have approximately the same duration. The longest phrase, a2, lasts for 10.5 seconds with the speaker F2 and only 8 seconds with the speaker M1, due to variations in speaking style. The opposite behaviour occurred for the shortest sentence, c1, which lasts at least 4 seconds for the speaker M1 and 5.3 seconds for the speaker F2. Additionally, each speaker was allowed to choose and speak up to three voice commands. In general this sequence of commands lasted for 8 seconds with a total activity period of around 2 seconds. Such unique user sentences per user was designated v, for voice command. An additional sentence was recorded for subject F2, again read from a book, and lasts 8.5 seconds with a total inactive period under 0.4 seconds.

The file name format chosen for the database entries was *[genre][genre_subject_number][phrase].wav*, e.g., M1a2.wav, M2b.wav and F1c1.wav.

Unlike previous DUET studies, the different phrase sizes of this database, typical to natural speech, add a difficulty level not present when DUET analyzes continuous sequences of words and source signals with the same duration. With the previous DUET studies, the clusters in the two-dimensional histogram have roughly the same height. With this more realistic setup, some clusters have one third of the height of others in the cumulative histogram, for the same distance from the microphones.

5.2.2 Recording procedure

For the real data recordings all the database files were extended with $1.0\,s$ of silence before the signal and $0.5\,s$ of silence after the signal. If needed, the initial "silent" second of each recording can be used to establish the noise profile and perform spectral subtraction for basic noise removal if needed, without requiring a VAD. This initial "silent" segment is discarded from signal analysis.

Each signal was then emitted at the several positions in the anechoic room by a using one channel of the AV40 M-audio monitor and measured with a omni-directional ECM 8000 pair of measurement microphones from Behringer. The soundcard used was an ECHO GINA 3G PCI.

For the live tests only the ECM 8000 microphone was available and the speaker used to emit sounds was a 10€ mitsal speaker that was dismounted to have only one channel working (pontual source). The soundcard was a Roland QUAD-CAPTURE UA-55 USB-2.0[3].

For all the tests in the anechoic chamber the speaker was placed at $s_z = -5\,cm$ [4]. For the tests in the live environment and simulations it was possible to place it at $s_z = 0\,cm$.

Since higher values of $|s_z|$ only reduces the contrast between different source s_x, s_y positions in Λ-space, each source placement is referred to by the distance projected in the $x-y$ plane, d, and ϕ, the angle to the y axis projected in the $x-y$ plane where positive angles represent positions with positive x values.

The microphones were placed with $\Delta = 2\,cm$ in a vertical beam pointing forward but due to their size, their heads are offset in $+\vec{e_y}$ by $13.0\,cm$. A string marked with the relevant distances is used to measure

[3]This card produced too much noise in Phantom Power (+48V) amplification.

[4]This value doesn't have to be zero and in this particular case, for distances of $1\,m$ and above it is completely negligible.

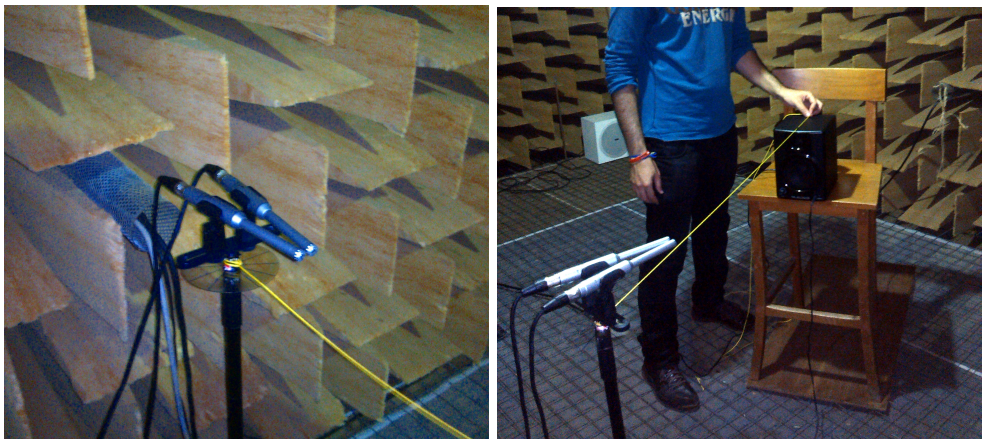


Figure 5.1: Straightforward experimental setup showing the omni-directional microphones and simple position measurement through a thread attached to the microphone stand that is compared against one of the lines marked radially in the CD-ROM.

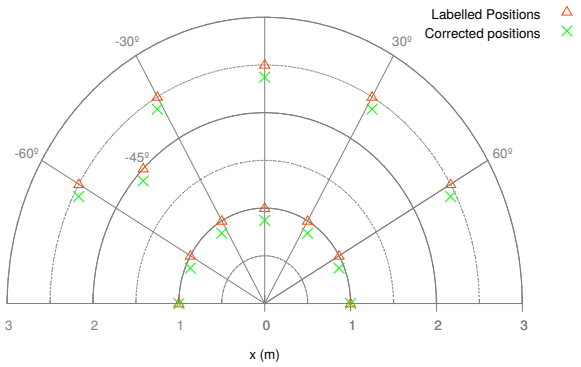


Figure 5.2: Positions of the recorded sources in the anechoic room with the microphone axis as origin (labelled) and corrected for the coordinate system where the origin is at the center of the microphones.

both distance and angle to the y axis by the alignment to the radial stripes marked at increments of 30° in a CD-ROM.

This means that the recordings labelled as having a distance d to the microphone array centre and ϕ degrees are actually slightly deviated as shown in figure 5.2, but this does not change the results in any significant way, so the measurements are still referred by their uncorrected values, although the theoretical cluster positions obtained geometrically shown in the two-dimensional histograms as red dots use this correction.

Given the high volume of recordings, they were organized in hierarchic fashion by: *room/experiment-name/source-distance/source-angle/[gender][gender-subject][sentence].wav* allowing a script to generate the source combination tests that can acount for source selection through customizable regexp patterns and additionally exclude combinations like the same gender, sentence or subject.

Several scripts then execute each test and parse its results into $(\sum_n X_n) \pm \left(\sqrt{\frac{\sum_n ((X_n - E[X])^2)}{N-1}} \right)$ where X is the random variable and X_n its observation for the trial n in N trials.

Given that a test might fail and the errors are already large due to different lingustic variations, degenerate tests are excluded from the averages without any noticeable change in moment, as in many cases, even though there might appear more estimated sources N_e than real sources N, separation still

occurs and sometimes with a not very large decrease in signal reconstruction.

5.2.2.1 Signal conditioning

All the real data tests were recorded at 44.1kHz and only then downsampled to 16kHz totalling 928 manually recorded files which requires 2Gb of disk space, without counting discarded tests and tests performed outside the anechoic chamber.

A High Pass Filter was applied to the 16kHz recordings using the equalizer (EQ) in audacity with 4000 points, -20dB gain at 200Hz and 0dB gain from 250Hz and above to better match the DUET output, that is capable of applying an ideal bandpass filter (where frequencies outside the chosen band are discarded), and in this case, a highpass filter of 250Hz was used. Though there is ringing in the output from the ideal filter and the filters are not the same, the input and output have now more similar energy profiles which provide more accurate, though not necessarily higher SNR, SDR, SIR and SAR values.

To calibrate the gain of the pre-amp for each microphone either a pink noise emitter or a tone emitter placed at 1m and 0° in the anechoic room would provide a sensible setup for easy "factory" calibration since the angular error is guaranteed to be small and no reflections alter the balance of the system, however, given that the microphones are not balanced, it was determined that the most accurate way to find the balance factor is through the average balance of all available speaker signals emmited through a studio monitor at 1m and 0°.

5.2.2.2 Inter-channel-frequency-gain

Without applying any balance factor to all the files, through the 100, 333, 440, 1000, 2000, 4000, 6000, 8000, 10000 and 15000 Hz tones emission at 1m and 30° in the anechoic chamber the attenuation a was calculated by determining the squared root ratio of the energy of the left and right channels as the signal amplitude A is proportional to $\sqrt{E}$ where E is the signal energy. Ideally, all the frequencies should have $a = 1$:

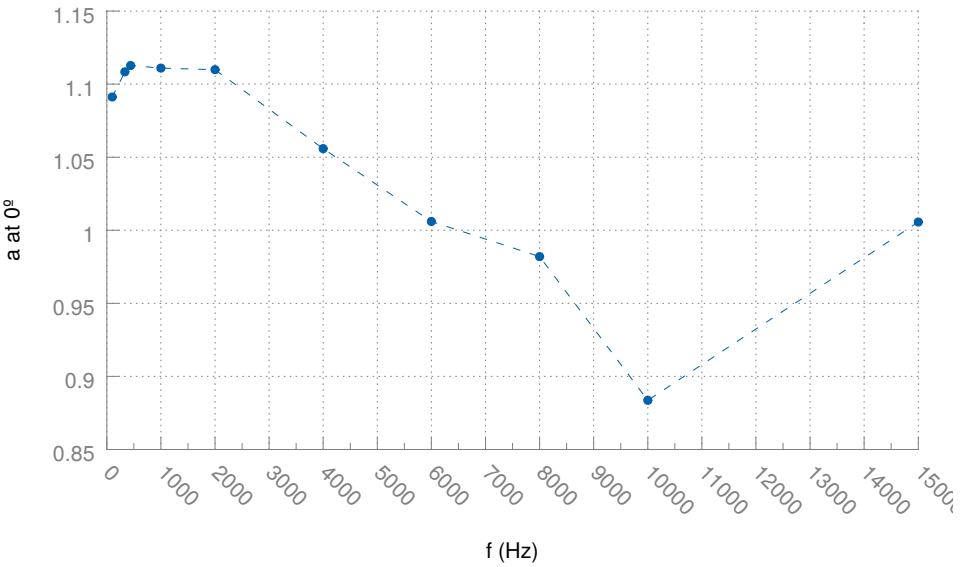


Figure 5.3: Measured attenuation for several frequencies with the loudspeaker placed at $1\,m$, 0°

Figure 5.3 shows an imbalance of 11% at low frequencies and −12% at high frequencies which gives a variation of 23% balance change in the speech spectrum range. Depending on the spectrum of the speech, for the same position, different Λ distributions are observed.

Given that this was later reproduced in a live environment with a different soundcard though with a curve that was the inverse of this (microphones were swapped), this is due to the microphones or cables. This frequency-dependent attenuation is going to cause spread in $\hat{\alpha}$. A function to interpolate this curve was used to correct $\hat{\alpha}$ in real-time[5] but although the $\hat{\alpha}$ spread was reduced it didn't improve significantly. This could have been due to the sparse frequency tests performed of a at 0° together with the noise that was higher in one of the channels and thus the correction factor might have to be adapted from stationary tone signals to instantaneous $\hat{\alpha}$ estimations by training the a_0 (attenuation at 0°) directly from speech data and the loudspeaker at the same position. This was not studied further but there is clearly potential for improvement.

5.2.2.3 HPF cutoff frequency

Low frequencies don't matter much for speech perception and can be detrimental to the system since DUET is more accurate for higher frequencies thus an ideal digital high-pass filter which discards spectral content below the high-pass filter (HPF) cutoff frequency was used. Also, as more frequencies are discarded by DUET the faster it runs.

The metrics were calculated for the target source at 1m, 0° with an interfering source at 1m and 30° and the HPF cutoff frequency was increased in intervals of 20Hz. The subjects M1 and F1 were tested for the sentences a1 and c1 at the target position and the subjects M2 and F2 spoke the sentences a2 and c2 at the 30° position.

The original signal was not filtered so that only the separation output is changed and still some positive variations occur. The plot of the ratios SNR, SDR, SIR and SAR is shown next for the varying HPF cutoff frequency:

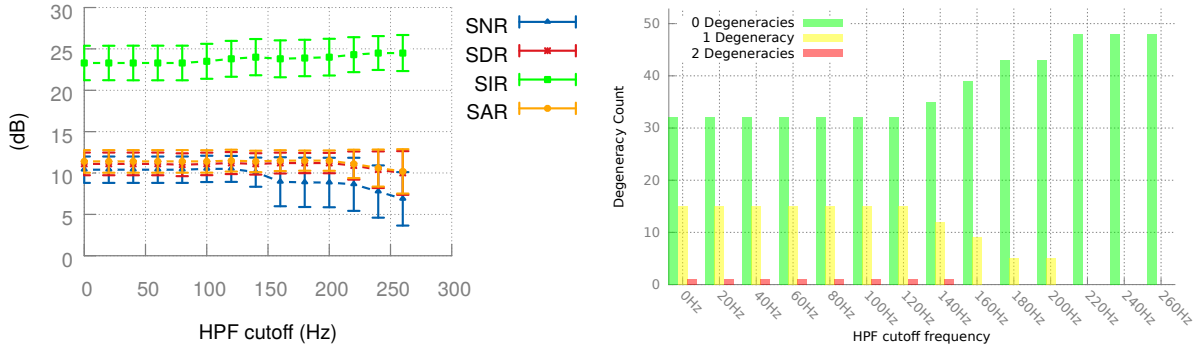


Figure 5.4: Separation performance for a central target with varying HPF cutoff frequency.

To guarantee a reduced degeneracy count the ideal DUET HPF cutoff frequency was set at 250Hz as

[5]Both channels can be corrected accurately by attenuating frequencies in the channel where they are stronger but this takes twice the resources of correcting just one channel (which should have equivalent separation performance, except for weak signals that might see their quantization errors amplified), or simply a, which won't improve separation results since there is now a bias between the $\hat{\alpha}$ of the samples and the corrected $\hat{\alpha}$.

stated in the beginning of chapter 5.

An alternative to an HPF filter that for sure eliminates the over-degeneracies is over-sampling the spectrum signal which will reduce the spectral leakage, by zero-padding the windowed time signals.

5.3 Preliminary tests and results

5.3.1 Two-dimensional Λ histogram

To evaluate the histogram's validity one source was simulated in a fixed position and the cumulative histogram was drawn. This test is then independent of the separation masks since there is only one source present, and independent of W-disjoint orthogonality, thus the peak in the histogram should appear on top of the simulated (α_1, δ_1) coordinates. At the same time, this tests validated the calculation of the parameters α and δ, which form the basis for the DUET analysis.

This simulation test was repeated with three different sound source positions, from very close to the microphone axis to very far away, respectively closer and farther from a typical usage, as well as in a common situation.

The audio test file in this section was *mike.wav*, a clean anechoic recording of a human speaker done at a sampling frequency of 22050 Hz with 191258 frames. The chosen window was a 64ms Hann window with hop size of 50% FFT_N. The weighing DUET exponents were set to conservative values of $p = 1$ and $q = 1$, which tend to "kill" secondary peaks. For these tests the histogram with all the sample data and was accumulated (cumulative histogram) and no smoothing was applied. Δ was set to the maximum distance that doesn not create phase-aliasing.

The red dots in every two-dimensional histogram show the theoretical (α, δ) for each source and the green triangles indicate what (α, δ) clusters DUET found. In this case one single histogram, the cumulative histogram of the 8s of audio, was processed.

The histogram used bin sizes of $(d\alpha, d\delta) = (0.001, 0.2\mu s)$ and the source was placed for the **c**lose setup at symmetrical azimuths: $\mathbf{C}_L = (-0.6, 0.2, 0.0)(m) \rightarrow \Lambda = (-0.046263, 22\mu s)$, $\mathbf{C}_R = (0.6, 0.2, 0.0)(m) \rightarrow \Lambda = (0.046263, -22\mu s)$ in two separate symmetrical experiments.

A recommended range for the $\mathbf{C_R}$ (**C**lose source placement at the microphones **r**ight) experiments is $\alpha \in [0, 0.1]; \delta \in [-30\mu s, -10\mu s]$.

As can be seen in figure 5.5, the histogram shapes change with the window function. The rectangular window shows a sharper peak than the Hann window, but the same kind of tail-shaped α leakage with increasing values of $|\alpha|$. This means that if the source is closer to microphone 1 and thus $\alpha < 0$, then the tail still moves away from 0 towards even more negative α values whilst the opposite sources to the right have tails in the opposite direction.

The Hann and Hamming windows show the same kind of peak shape. For this case of a single sound source, the quality of the reconstruction was apparently the same for ML_{Hann} and $ML_{Hamming}$. Also, except for the slightly more prolonged tails and lower peak height of the cumulative histogram, the histograms generated with both windows were very similar.

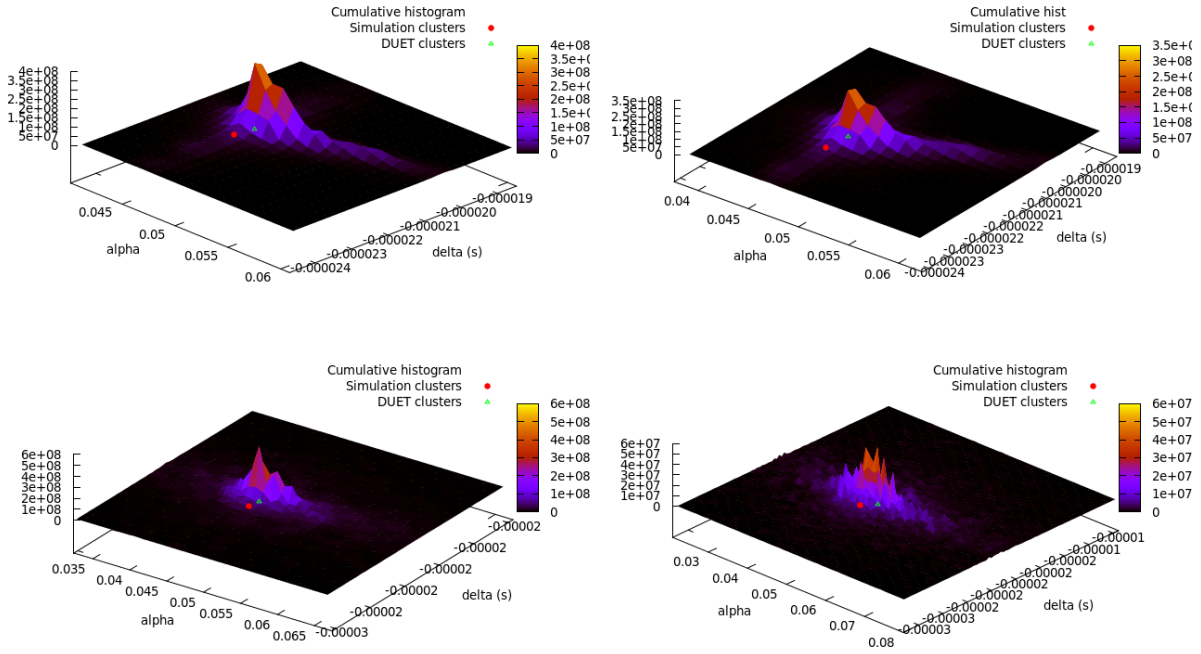


Figure 5.5: Source placed at $(0.6, 0.2, 0.0)$. In left-to-right reading order, the cumulative histograms of $w = Hann$, $Hamming$, $Hamming_0$ and $Rectangular$.

The zero-phase Hamming window, Hamming_0, provides a more symmetric histogram cluster with balanced tails. This window is however not appropriate for the reconstruction of the signals since the window's weights are bigger in the extremities and smaller in the centre, thus reconstructing the signal from it with masks will generate a lot of artifacts. However, it is interesting to note that the resulting histogram shows higher symmetry. When histograms are built with low statistics, this added symmetry can be useful in identifying clusters since an assymmetry in a cluster means a mixture of clusters. However, the width of the resulting cluster will also be correspondingly larger. The end benefits are therefore not conclusive without more simulation tests.

5.3.2 Histogram Λ-dependent variability

Having studied the effect of using different types of windows on the histogram shape, the simulation tests showed ill-defined cluster behaviour for $R \to 0$. This effect is caused by using linear interpolation to delay the signals in the simulation stage.

A 64ms long Hamming window was used [Özgür and Scott, 2004] and the position of the single source was $\mathbf{Z_R} = (0.6, 0.0, 0.0)$, $\mathbf{M_R} = (0.6, 3.0, 0.0)$ and $\mathbf{F_R} = (0.6, 5.0, 0.0)$, for in-axis (zero) placement, medium and far placements to the right of the microphones.

The conclusion is that as the sources get closer to the microphone axis, $\hat{\alpha}$ becomes more ill-defined though δ behaviour remains largely unchanged.

The tails seem to change direction according to the position of the source in the graph and seem to be sensitive as well to the interpolation applied to the signals in the simulation since the histogram changed slightly when the relative delay and interpolation of the right channel was used instead of the

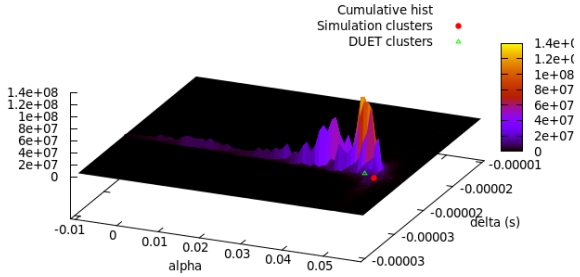


Figure 5.6: Source placed at $\mathbf{Z_R}$ with Dtotal=0.0104297 and SNR=19.851dB.

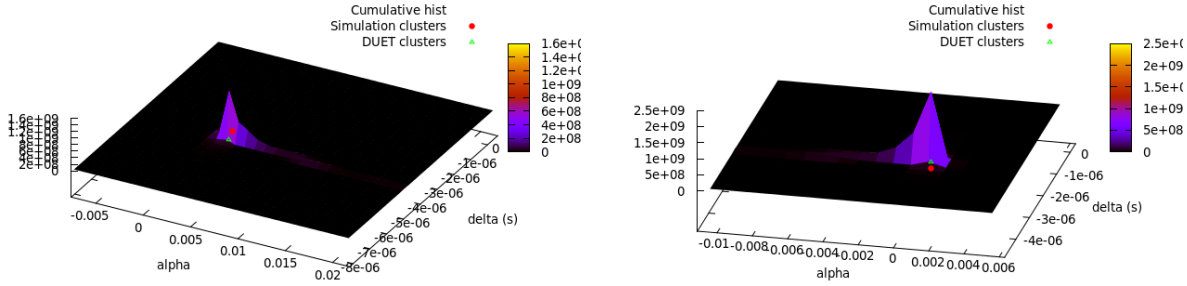


Figure 5.7: Source placed at $\mathbf{M_R}$ (left) with Dtotal=0.0091 and SNR=20.44dB and source placed at $\mathbf{F_R}$ (right) and Dtotal=0.00874 SNR=20.61dB

interpolation of both the channels through the absolute delay values.

If smoothing is used, its amount should become bigger as the sources get closer to the axis to hide meaningless variations in the histogram, such as the distorted tails that appear in the $\mathbf{Z_R}$ case.

5.3.3 Λ tails and spurious peaks

DUET's $\hat{\delta}$ and $\hat{\alpha}$ estimators in Λ inherently have biased or erratic behaviours in specific conditions that produce tails in the Λ histogram or produce spurious peaks. This section analyzes such behaviour, since it is critical to fully understand the implications of the DUET model.

5.3.3.1 $\hat{\delta}$ tails and bimodality

To demonstrate the effects of the asymmetric tails in the $\hat{\delta}$ parameter in the Λ histogram, generated by a recording of a male speaker in the anechoic room (M1c1) at $1m$ and 30°, (fig. 5.8, top left) shows an asymmetric tail behaviour in $\hat{\delta}$ that could possibly be attributed to the local stationarity approximation, less than perfect sound absorption in the walls of the anechoic room, or any other experimental imbalance caused by the equipment, since this effect is much less pronounced if the same test is replicated in the anechoic simulation. However, by reducing the smoothing size (5.8, top right) a multimodal $\hat{\delta}$ behaviour becomes evident.

Using the anechoic simulation and a source emitting a low-frequency tone of 100Hz at the same position (fig. 5.8, bottom left) reveals the same behaviour, this time bimodal, of $\hat{\delta}$.

A closer look revealed that the problem is simply caused by spectral leakage. For 1024 window points and 16kHz sample rate, 100Hz does not belong to the basis set of matched filters in the FFT (the closest frequencies are 93.750Hz and 109.375Hz). The occurrence of leakage added to the fact that the delay is calculated by dividing the phase by ω creates a multimodal behaviour in $\hat{\delta}$ predominant at low frequencies since the signal frequency is of the same order of magnitude of the frequency resolution Δf.

To verify this, a high-pass $400Hz$ filter was applied to the anechoic signal removing most of the peaks in fig. 5.8 but the one that is closest to the theoretical peak marked with the red dot (fig. 5.8, bottom right).

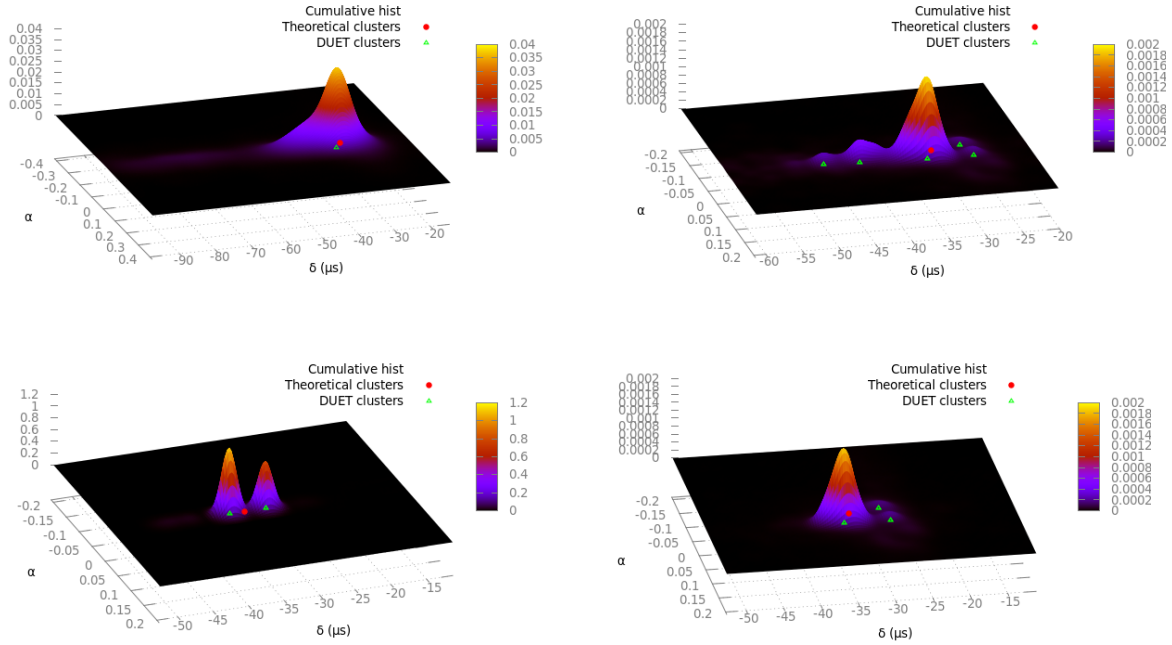


Figure 5.8: Λ histogram for the file M1c1 recorded in the anechoic chamber at position $d = 1\,m$ and 30° (top and bottom right) and a tone at the same position using the simulation (bottom left). The top left image uses a smoothing of $(0.03, 3\,\mu s)$ and the others use $(0.02, 2\mu s)$. The bottom right image has a high-pass filter applied at 400Hz. (p,q) was set to (1,0) on all tests.

5.3.3.2 $\hat{\alpha}$ spurious high-frequency peaks

A new behaviour in the Λ space that is easy to miss was found, since it happens for very specific individual frames and given the erratic peak locations produced, when statistics are accumulated, such effect disappears. For low statistics though, which is the domain of the real-time implementation, this peaks can be highly disruptive to a source tracking system that doesn't know the true number of sources. This behaviour was found to occur only for unvocalized sounds such as "s" and plosive sounds, which are non-stationary.

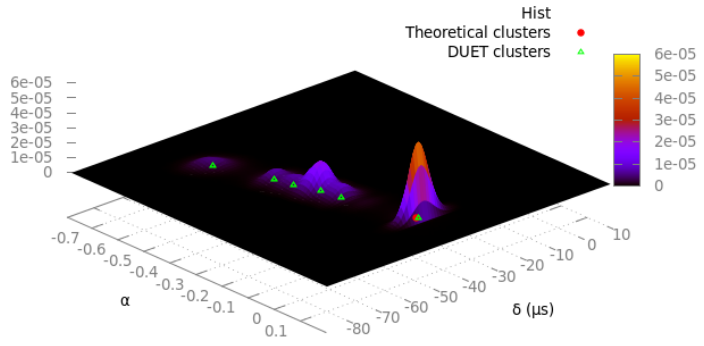


Figure 5.9: Λ histogram for the file M1c1 recorded in the anechoic chamber at position $d = 1\,m$ and 30° (top and bottom right) using the information from a single analysis window of $64\,ms$ where an "s" sound occurred. The smoothing $(0.03, 3\,\mu s)$ was used and (p,q) was set to (1,0).

This effect also disappears if a low-pass filter of $4\,kHz$ is applied to the histogram input data. However, such processing inhibits the correct assignment of sibilant data to sources with the same δ parameter. A better approach is to classify each peak as spurious or not ("true") and this can be done for instance through the developed kurtosis-based peak confidence metric.

5.3.4 Reflections by a wall

The anechoic room was used to study wall reflection effects in the Λ space. For this effect a small wooden rectangular surface was placed in the room at the plane $x = 223.5\,cm$. Given the relatively small surface size, the rectangular wooden board was placed at a position that maximized the specular reflection from the source towards the microphones. In the microphone coordinate system, three of its vertices are found at $(223.5, 22.0, -61.5)\,cm$, $(223.5, 22.0, 5.5)\,cm$ and $(223.5, 86.6, 5.5)\,cm$. The figure 5.10 shows the results from the wall reflection with a source at $d = 2.5\,m$ and $\theta = \{0^o, 30^o\}$.

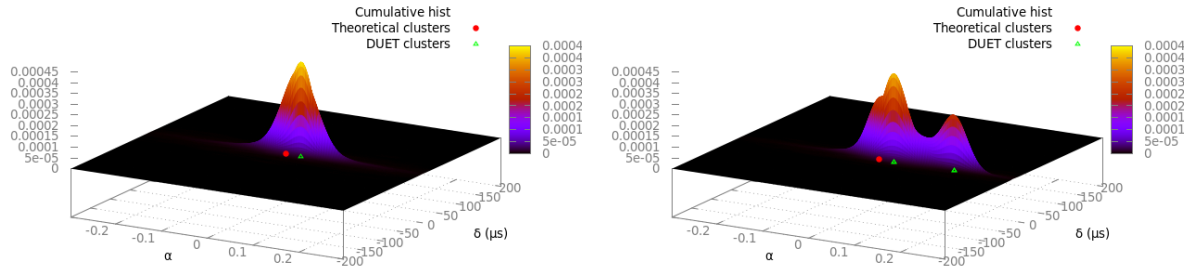


Figure 5.10: Λ space resulting in a second peak for closer sources. The image on the left shows the source placed farther apart from the wall ($\theta = 0^o$) and the one on the right, the source closer to the wall ($\theta = 30^o$) with an evident reflection image.

Figure 5.10 shows as expected that sources that are closer to a wall create a stronger reflected image and the ones that are farther apart a weaker one. This means that if the room has walls with similar

reflection characteristics, the sources to the right of the microphones will be more reflected by the right wall and the ones on the left by the wall on the left if the microphones are centered in the room.

Simulations show the same results in 5.10 but in a more evident manner. When a reflection occurs, given that any specular reflection should not come from a place much farther apart from the microphones than the original source or its effects would become negligible, its signal will be a copy of the original and fit inside the analysis window of 64 ms. This means the F quantity will be a combination of the original source with its reflections and as the simulation shows (fig. 5.11), the estimated Λ won't match either of the true source positions and $\hat{\alpha}$ tends to have a better discrimination than $\hat{\delta}$.

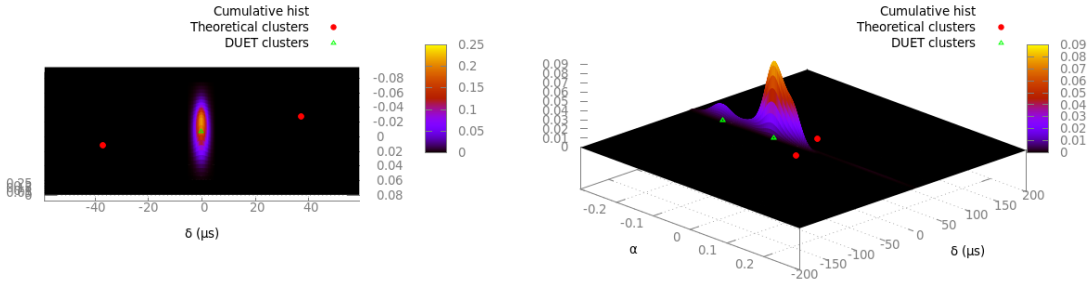


Figure 5.11: Simulation of the signal F2c2 at $(-0.8, 1.0, 0.0)\,m$ and an image of equal intensity at $(0.8, 1.0, 0.0)\,m$ on the left. Same simulation but with the image moved to $(0.4, 1.0, 0.0)\,m$.

5.3.5 Kurtosis tests

A test was performed with $\Delta = 2cm$ and one of the male samples placed at $(0.9, 1.0, 0.0)\,m$ in the simulation with the smoothing standard deviations set to 2×10^{-2} and $5\,\mu s$. The histogram bin sizes were set to 5×10^{-3} and $1, \mu s$.

If only one frame is analyzed at a time, with very high probability, even the true peak is rejected as non-gaussian. If, however, 2 frames are accumulated, the dispersion is reduced and most of the time the peak is detected close to the simulated reported position, as being Gaussian if the threshold of decision is set to $|kurt(\alpha)| < 0.2$ and $|kurt(\delta) < 0.2|$. With these thresholds, the Gaussian peaks will very rarely be discarded, in the first manually checked 50 consecutive frames.

Notice that this only works if the histogram is smooth (histogram bins are set to be much smaller than the smoothing standard deviations σ_α and σ_δ). With the chosen histogram resolution above, further reducing the smoothing, *e.g.* $\sigma_\delta = 3\,\mu s$ would make the histogram irregular and lead to an ineffective kurtosis test (only negatives).

When spurious $\hat{\alpha}$ peaks arise due to non-stationary speech sounds the application of the kurtosis tests rejects them.

In fact, even more interestingly, the only peak that is not rejected is situated very close to the simulated position and is sometimes not the strongest peak (see fig. A.8). This means that for working reliably, noisy spurious peaks can be rejected independently of their level, even if they are stronger than the reliable portion of the signal. This approach is quite interesting compared to a rejection based approach on the assumption that noisy components would have values below a certain floor.

This method still works if one keeps increasing the number of accumulated frames. Nevertheless, if the source is moved away from the 1m distance to 3m distance the kurtosis classification criteria worsens a lot. This happens when peaks are no longer locally separated in the histogram but they do heavily mix in the histogram (locally the histogram peaks can no longer be modelled by a single Gaussian and a GMM model would be required which rules out kurtosis).

A local confidence metric (such as kurtosis or for instance the spectral content of each peak) is very important because when the frequencies are higher (this problem disappears with a low pass filter with cutoff frequency of 4kHz), the error in the local stationarity assumption grows, causing spurious alpha peaks but if it seems that the whole 2D histogram information should be rejected or at least the α information should, actually, a portion of the signal (low frequencies) won't suffer from this approximation problem thus such proper parts can be still estimated reliably with a local peak confidence metric such as kurtosis and no useful information is thrown out.

5.3.6 Windows

5.3.6.1 Window types

Several different types of windows were tested. First the Hann and Hamming, Blackman, Nutall, BlackmanNutall and Rectangular windows were used. The different window types were tested in regards to target speaker performance at 0° versus an interfering source at azimuthal angles of 30°, 60° and 90° when both speaker and target sources were at $d = 1\,m$ or $d = 2.5\,m$.

Since windows need to be robust to any gender combination, all speech content and speaker characteristics, the different window types were assessed with the permutations of the phrases a2 and c2 for all the four speakers, two male and two female, excluding only the possibility of having the same subject or sentence as both target and interference which results in 24 tests per angle. All the windows had 1024 points, or 64ms. All the windows used a hop size of 50% of the window size and $(p, q) = (1, 0)$. For the simulations the high pass filter was set to a low value of $50\,Hz$ and for the anechoic tests at $250\,Hz$ due to the microphone unbalance.

First the tests were performed for the anechoic simulation which is not affected by experimental problems such as unbalanced microphones or angular frequency response. Then, to prove the consistency of the results, the tests were repeated for $d = 1m$ in the anechoic case with real data which already is sensitive to such problems.

Given the stability of the errors across source placement and window types, which suggests the results are consistent and the variations in the data caused only by the variations in the different speech combinations which provide a realistic range of separation quality, the error bars were removed from the plots to improve readability.

For the simulation tests at both $1\,m$ and $2.5\,m$ the sample standard deviation $\mathcal{S}$ in the range $[1.3, 1.7]\,dB$ for SDR, SNR and SAR of all window types except the Rectangular window with $\mathcal{S} \in [1.2, 1.5]\,dB$. For the anechoic tests the variability was slightly higher, with $\mathcal{S} \in [1.5, 1.9]\,dB$. For SIR, in all settings $\mathcal{S} \in [1.7, 2.2]\,dB$.

Simulation with standard windows Figure A.2 shows the separation results for $d = 1m$ and the standard window types Nutall, Blackman, the hybrid BlackmanNutall, Rectangular, Hann and Hamming. The Blackman family was chosen to test if windows other than Hamming it provided good or better solutions to the BSS problem, to extend the research in [Yilmaz and Rickard, 2004]. Figure A.3 shows the same tests, for $d = 2.5\,m$.

The Hamming window with two-decimal places and equiripple versions were also tested though the first had worse results than the standard Hamming and the equiripple version even worse results so they were promptly abandoned.

Data in figures A.2 and A.3 shows that for $d = 1m$ and $d = 2.5m$ the Hamming window provides the best results, while Hann comes as a close second, having only the same SIR as the Hamming window for $1\,m$ and 30°.

Overall, the SNR, SDR and SAR show the same absolute difference tendencies between all window types for all angles. Compared to the Hamming window, the Hann window has an average performance drop of $0.2\,dB$, followed by the Blackman window, with a more significant performance drop of around $1\,dB$. The following best windows were the Rectangular, BlackmanNutall and Nutall with performance drops greater than $1.2\,dB$, $2.0\,dB$ and $2.2\,dB$, respectively.

In terms of SIR, either two or three groups with different performance arise. For $d = 1m$ only the Rectangular window has a significantly reduced performance (less 1.5dB than the Hamming) when compared to all others that have only up to $-0.5\,dB$ than the Hamming window. For $d = 2.5m$ The Hamming, Hann and Blackman windows provide similar performance, while BlackmanNutall and Nutall have close to a drop of $0.5\,dB$ and finally, with a performance drop of $1.8\,dB$, comes the Rectangular window.

As sound sources get closer (smaller interfering source angle) beyond a point where their T-F bins start to occupy the same region in Λ space, DUET's separation performance will drop, since the classification of bins in the common region through their Λ coordinate becomes impossible. The crossover of this region is showcased in figure $A.2$ For $d = 1\,m$ where most windows have similar performance for all the tested angles, except for SIR that shows a more accentuated SIR growth between 30° and 60° (less T-F mixing occurs in Λ space) and then performance mostly stabilizes. There is one additional factor, that for $\pm 90^o$, the tails in α in the simulation grow due to numeric errors in the time-delay signal interpolation. This can reduce the performance for sources placed on the microphone array axis.

For all the tests the Rectangular window had a hard time rejecting interfering sources (low SIR) and didn't seem to improve the separation results when the interfering source angle increased.

Simulation with the generalized Hamming family variants Figure A.4 shows the simulation results for $d = 1\,m$ and the developed variants of the Hamming window. This variants were developed because it was observed that $\hat{\alpha}$ suffers high deviations for high frequencies (production of spurious peaks with $\hat{\alpha}$ error several orders of magnitude bigger than the value itself) which can only be caused by a failure in the non-stationarity approximation for a. This hypothesis is validated by the fact that plosive sounds are non-stationary [Schluter and Ney, 2001]. Since the Hamming window weighs more the center of the

window, this problem is even more accentuated. To make the approximation more valid the Hamming window was modified to have a wider shape in the time domain to make it closer to $W = 1$ which validates the local stationarity assumption 4.6 and increases the relevant time-span of the window, augmenting the correlation between channels with higher inter-channel delays. This rationale lead to the development of the weak and weaker windows.

For the 24 for tests per angle, at $d = 1\,m$, 2 over-degeneracies occurred at 60° for the Blackman, BlackmanNutall and Nutall windows. For 90°, 0, 1 and 2 over-degeneracies were measured for those same windows respectively. For $d = 2.5\,m$, 1 over-degeneracy occurred for the Nutall, Hann, Hamming, weak, weaker, midweak and lowweaker Hamming windows for 90°. The Blackman window had 2 over-degenerate tests for 90°.

For $d = 1\,m$ all the modified windows provided better separation than the Hamming window in terms of SNR, SDR and SAR.

For both the simulated distances distances, SNR, SDR and SAR the midweakerHamming window has almost the same performance boost over the Hamming window that the Hamming window has over the Hann ($0.2\,dB$). The lowweakerHamming window provides a smaller gain ($0.07 - 0.08\,dB$) in SNR, SDR and SAR but, unlike the midweakerHamming window, secures a SIR as good as the Hamming window (for the interfering source placed at $1\,m$ and 30°) or better by up to $0.07\,dB$, the same improvement that would be obtained by using a Hamming window instead of a Hann window (figure A.5).

Anechoic window tests For the anechoic tests the occurred over-degeneracies are shown in table 5.1.

Table 5.1: Over-degeneracy count for the anechoic window tests at 1m where $l = low$, $m = mid$, $w. = weak$, $wkr. = weaker$ refer to the variants of the Hamming window.

Angle	Rectangular	Blackman	Nutall	Hann	Hamming	lw.	lwkr.	mw.	mwkr.	w.	wkr.
60°	4	0	0	0	1	1	2	2	2	3	1
90°	2	2	2	2	2	1	1	1	1	1	1

With the anechoic tests for the standard windows where the electrical noise was non-neglegible and stronger than on one of the pre-amps and the inter-channel-frequency-gain was also biased the figures A.6 and A.7 still show comparable results to the simulations which suggests the consistency of the results for a real application.

The biggest difference that was observed was that with increasing distance between target and interfering sources the separation performance increase is in general more consistent than in the simulation results with the exception of the Rectangular window SIR (fig. A.6), that differs from all the others in the sense that is more prone to leakage problems.

Unlike the simulation tests, for the anechoic tests which suffered from the aforementioned equipment issues, the lowweakerHamming window seems to be preferrable to the midweakerHamming window as it provides similar SNR,SDR and SAR values and for the SIR has better performance than the Hamming window for 60° and 90° while providing a SIR for 30° that is as good as the Hann window.

It is to be expected that for very small source distances, the Hamming window which has been designed to have optimum rejection characteristics achieves better separation than the lowweaker variant, which

is simply a non-optimized transformation of the Hamming window aimed at reducing the error in the attenuation estimation of plosive sounds. However, as distance grows, this factor loses relevance and the error in the attenuation estimation gives an edge to some of the modified windows. These previous preliminary studies show the potential for the optimization of windows where a wider window in the time domain can be used to separate sources that are far apart and when they are not, the Hamming window or another untestested variant can be used (or a parametric window transformation can be used to optimally adapt the window to the setting in real-time since the location of the clusters is known and the speed at which they move is bounded (eq. B.36)), realizing that for speech, there are several key window aspects that need to be optimized.

5.3.6.2 Window size

For this section the sizes of the tested windows were: 100, 200, 400, 800, 1024, 2000, 4000, 6000, 8000, 10000, 12000 and 15000 samples. The files a2 for both males were used as target source and the files c1 for both females were used as interfering source. Both sources were placed at $d = 1\,m$, with $\theta_{target} = 0^o$ and $\theta_{interferer} = 30^o$. Exceptionally, since the evaluation of the window influence over the whole spectrum was desired, no input HPF filter was applied and DUET's ideal HPF cutoff frequency was set to a low value of $50\,Hz$ so it closely matches that condition although it is not so small that it exacerbates the multimodal $\hat{\delta}$ behaviour since no over-sampling was applied.

The Hamming window used the hop size of $50\%FFT_N$ and the rectangular window no overlap whatsoever.

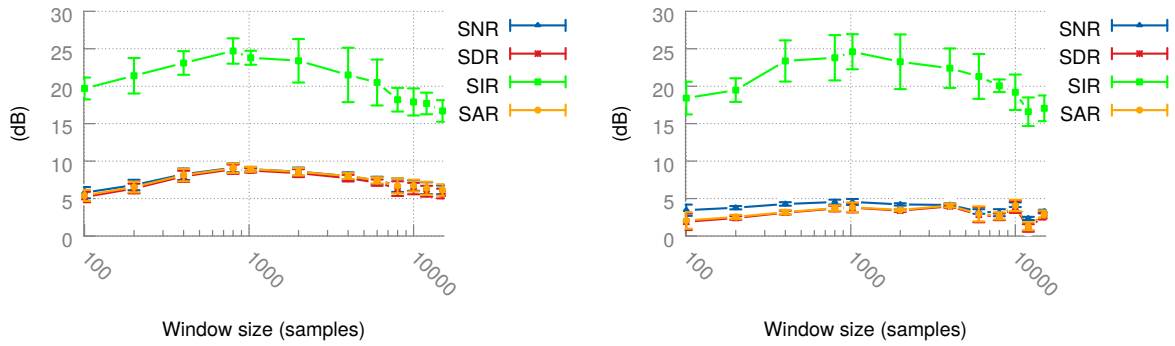


Figure 5.12: Tests for a rectangular window (left) and Hamming window (right) with several sizes.

No degeneracies occurred for any test and besides the optimal behaviour, which occurs for the Hamming window at 1024 points (64ms), which confirms the results in [Yilmaz and Rickard, 2004], the optimal rectangular window size was 900 samples (around 56ms), which was missed by [Yilmaz and Rickard, 2004].

Another interesting result is that by dramatically increasing the window sizes one can find a resonance point where separation quality increases. This point occurs for 10000 samples or 0.625 seconds and is reproducible with the IBM meaning the behaviour is not primarily caused by the localization of sources but their W-disjointness. At this regime one would expect the W-disjoint to be very bad, however there must be another smaller effect providing this local gain in separation quality. Exactly what this effect is, is not known, but using bigger window sizes means that the frequency resolution is higher and 10000

points or 0.625 seconds might be the optimal size at which the probability of this narrower frequency bins intersect less, other than the 1024 points windows.

These plots also don't show that for the rectangular window one can hear the target speaker clearly over a weaker background noise that sounds like an echo that changes character at every STFT analysis frame. If a way to efficiently eliminate this "virtual echo" is found then the signal separation metrics will increase dramatically and processing can be performed with a much higher computational efficiency since the number of analysis windows is reduced tenfold while the processing time of a 10000 point FFT is almost the same as that of a 1000 point FFT. Additionally this effect was occurring only for 10000 points. With slightly different window sizes the background noise breaks down to intrusive noise and the humanly assessed separation quality degrades quickly.

5.3.7 Real-time streaming system

The source streaming system was tested with the real-time streaming mode that uses low statistics. The number of accumulated windows N_{accum} for the calculation of a Λ histogram is shown below, along with the number of created streams (only two streams are desired, one for each source). For 303 Hamming windows of 64ms and 50% window overlap, the number of produced streams was summarized in the figure 5.13.

The sources used were M1a2 at $(1.0, 1.0, 0.0)\,m$ and F2c2 at $(-0.8, 1.0, 0.0)\,m$ with the simulation with the input filtered with the band-pass filter with band $200 - 8000\,Hz$. The simulation was used here to avoid the microphones inbalance which could lead to biased results. The Gaussian smoothing kernel size was set to $(0.06, 6\,\mu s)$. (p, q) were set as usual to $(1, 0)$. The cluster aggregation post-filter was set to a resolution of $(0.015, 1.5\,\mu s)$ and the maximum numbers of sources and clusters were set to 20, the maximum number of simultaneously active streams to 100, the overlap correlation metric was disabled, stream merging based on minimum number of active blocks was disabled and the maximum inactivity time for the streams (persistence time) before stream "death" was set to $0.5\,s$. The modified *Lambda*-distance of maximum 0.5 between consecutive cluster modes was used as decision threshold for position continuity.

The lowest number of accumulated windows that produced two streams was $N_{accum} = 33$. This slightly varies from test to test according to the speech characteristics. For speaker positions this number increases as both speakers get closer to each other, however a tendency of $1/N_{accum}$ exists, which can be useful to estimate how many accumulated frames are required to make reliable decisions. This behaviour was not studied but is believed to be caused by a statistical limit which is derived from the sample standard deviation for a gaussian, which is $\propto 1/\sqrt{N_{accum}}$. However, since the distance is calculated between a pair of gaussians they must both be inside the chosen radius of $D_k = 0.5$ thus it is the product of the error of each mode that gives the N_{accum}. By increasing the Λ distance threshold from 0.5, the total number of streams produced should decrease but it should still follow the same $1/N_{accum}$ law, what changes is the scaling factor, that was 170 for the case shown in figure 5.13. The performance of the system is not ideal at that minimum N_{accum} limit where the number of streams is the number of sources though, it is only the limit for which only two streams are produced but the table 5.2 shows that

performance still increases with additional information and around 3 s of speaker activity are necessary in anechoic conditions to accuratly track the speaker with optimal performance.

Table 5.2: Separation performance for the dynamic test.

N_{accum}	33	110	303
SNR_{x_1} (dB)	12.80	14.00	14.00
SNR_{x_2} (dB)	8.23	8.80	8.80

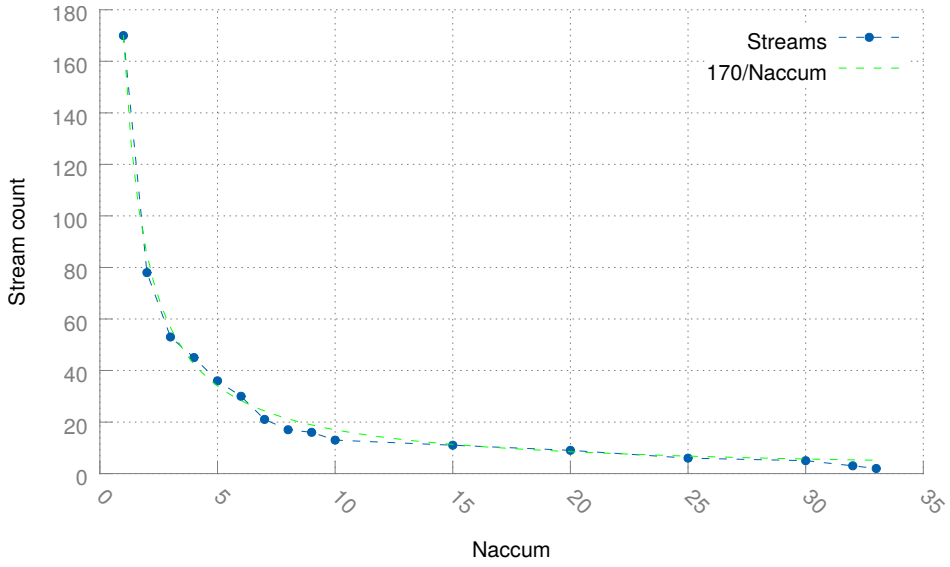


Figure 5.13: Number of streams produced in function of accumulated windows information used for each localization decision where the whole data is represented by 303 windows. An additional curve is plotted showing that the streaming system follows a $1/N_{accum}$ tendency.

$N_{accum} = 33$ to achieve only two streams represents 1.056 s of continued activity required for a source to be located in this particular test.

Additionaly, some streams were selected to be shown in the time domain in figure A.1.

The program was designed and optimized for real-time processing and this is attested in its runtime of 0.52s (excluding the initial memory allocation times). The test with $N_{accum} = 33$ and 9.7s of audio ran in under 0.55s with a single thread in an Intel Core i5-3570K demonstrating that the system has real-time capabilities.

5.4 Experimental results

5.4.1 Anechoic chamber tests

5.4.1.1 (p,q) choice

Several tests were done to find the best combination of (p,q) and it was determined both for unbalanced[6] and corrected measurements, although correcting occasionaly changes the histogram shape (non-linear

[6]Balancing the pre-amps on the soundcard was impossible so all the files were balanced digitally by finding the average balance factor using the speaker measurements at 1m and 0° as using tones was more innacurate. This yielded the corrected measurements.

behaviour shows up for real-world data which means not being able to record the data balanced could have impaired slightly the results), that:

(0,0) provides an extremely wide dispersion on α as does (0,1) and (0,2) to diminishing degree. This is because with real-world data noise is very influent if no penalty is used (p=0). (1,2) also suffered from this problem though less than both (0,1) and (0,2) since the term w^q is still of higher order than the energy $|X_1(\tau,\omega)X_2(\tau,\omega)|^p$ of the bin thus this term does not aniquilate the noisy bins that are very deviated as the frequency is the relevant factor and not the energy.

Besides this two settings that work are (1,0) and (1,1) though the first has less spreading in α again for the same reason as explained before.

The worst setting of all was (2,0) that generates many fake peaks and generally very bad histograms because there are two competing behaviours. Using q=0 has the same extremely wide α behaviour as observed in (0,0) and p=2 is very selective in space (resulting in non-smooth histograms) but due to the fact that the input is noisy real-world data this variations times the very wide underlying α behaviour generates many peaks even though they are usually close to the true source positions in the histogram.

The setting (2,1) already suffered less from this problem and was very close to an acceptable solution as the underlying spread in α for q=1 is lower than q=0 as seen in (0,0) and (0,1) and generally estimated one more source (with a very weak peak) than the number of sources when using the same settings as the better (p,q) settings such as (1,0) and (1,1).

Another good solution is (2,2), which ends up selecting with q=2 a narrower significative bandwith which means the spread due to unbalancing in the microphone pairs inter-channel-frequency-gain is less relevant. However, due to inbalances, some asymetries/shoulders in the peaks arise sometimes and if the sources have differing energies it makes them harder to identify as the weaker sources will be much more easily masked in the histogram than with (1,0).

The conclusion is that when using very good equipment (constant inter-channel-frequency-gain) and a balanced gain setup at the moment of recording and in noisy conditions where identifying sources is a tough challenge, identifying one reliably at all is better than erroneous source tracking and (2,2) should be used (and probably as the gear is more balanced (2,1) also becomes a more acceptable solution).

For unbalanced volumes/speaking activities and less noisy regimes, (1,0) is preferred as it doesn't penalize weaker sources so much and is smoother (p=1 is less agressive in spatial resolution which results in smoother histograms because any inter-channel-frequency-dependent imbalance is not boosted as much as with p=2).

5.4.1.2 Separation tests for central target

The tests were performed with all the combinations of the sentences a1,a2,c1,c2 and c3 for the central target source at 1m and 0° and one interfering source at varying angles with excluding tests where both target and interfering source are spoken by the same subject or speak the same sentence and the results are shown in figures A.9 and A.10.

There was a large assymetry for positive and negative angles as the over-degeneracy plots show and

the varying SIR performance. This asymmetry is rooted in the experimental setup since feeding DUET with swapped left and right channel inputs gives the opposite results. The experimental setup and source combinations were symmetric and the tests were performed inside a anechoic room thus the only possible cause for the asymmetry is the equipment used and in fact as was previously seen, the inter-channel-frequency-gain is not constant and one of the microphones generates more noise than the other. The over-degeneracies are also shown for the different tests in fig. A.11.

The IBM mask is dependent only on sources energies and uses only one channel for reconstrction. Its evenness further suggests that the cause of the DUET performance asymmetry was caused by the uneven inter-channel-frequency-gain.

The IBM results show worse results than the DUET approach but the IBM masks were built using only the information from the left channel thus indicating that the ML masked source reconstruction of DUET did provide a gain over a SC approach.

5.4.1.3 Separation tests for central target and three interfering sources

Again, in the same conditions as before but this time testing only the phrases a1 and c1 for the target speaker and symmetric angles for both interfering sources that speak either a2 or c2 and tests where the same subject or sentence appears at multiple positions are excluded. This leaves 96 tests per angle for which 1 and 14 over-degeneracy tests occurred for the interfering sources at 30° and 60° respectively.

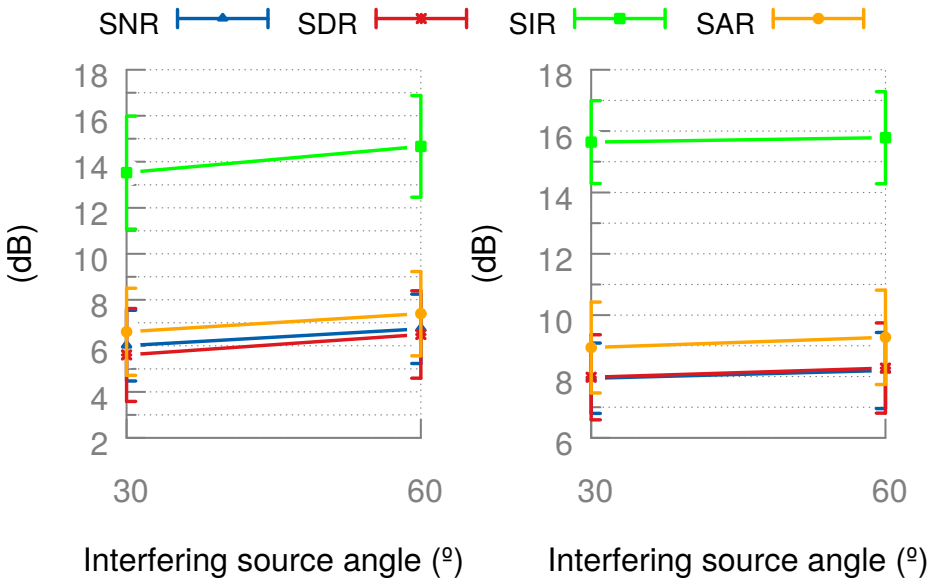


Figure 5.14: Separation stats for the central target source with DUET and IBM separation stats. Test performed with 3 sources in the anechoic room at $d = 1m$.

As the angle of the interferers is increased from 30° to 60° the separation improves with DUET whilst the separation with IBM remains largely unnaffected which means there is more cluster mixing in Λ space for 30°. The separation results achieved by DUET are close to $2.5\,dB$ and $2\,dB$ below the ideal masks.

Although the separation performance greatly degrades with 3 sources, it was noticed that in some cases it was humanly impossible to understand some of the speakers and after applying DUET, although the outputs had many distortion artifacts, it became easy to understand what was said. Further studies

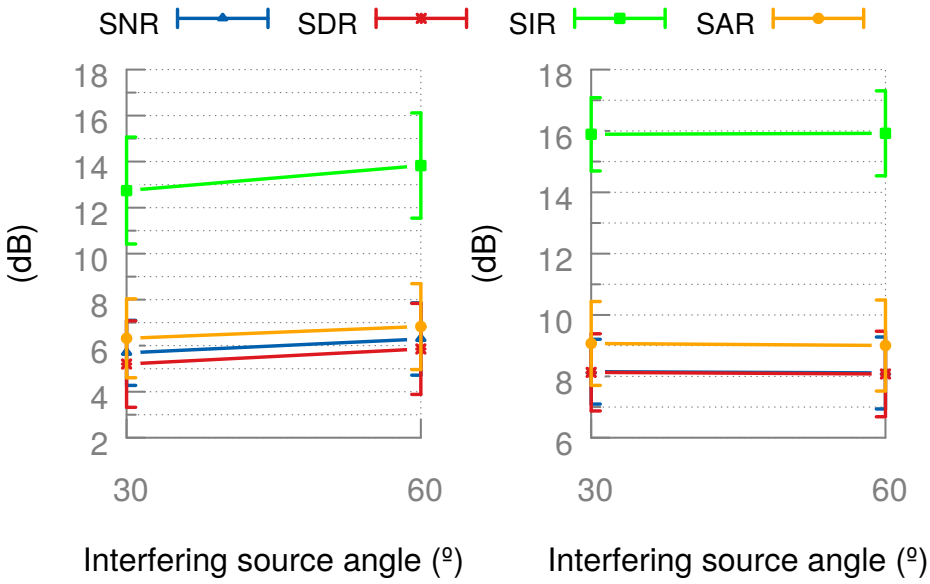


Figure 5.15: Separation stats for the central target source with DUET and IBM separation stats. Test performed with 3 sources in the anechoic room at $d = 2.5m$.

with the processing by an ASR system must be done though to evaluate the potential of the developed method as DSR streaming pre-processing interface to an ASR system.

5.4.2 Living room tests

5.4.2.1 Two human speakers

The two tests were performed by recording two people speaking naturally in a $4.0 \times 3.5 \times 2.8$ m bedroom, at either $d \approx 1\,m$ or $d \approx 1.9\,m$. The test subjects were standing at close to 30° and -30° angles.

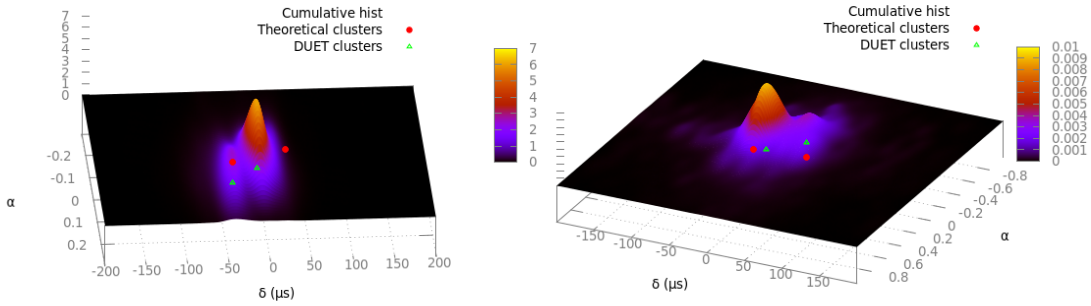


Figure 5.16: Separation performed in a bedroom with the same experimental setup as the anechoic case. The mixture features two independent speakers at approximately symmetric $\pm 30^o$ angles. $d \approx 1\,m$ (left) and $d \approx 1.9\,m$ (right).

The SNR improved for both signals for the test where the sources where at $1\,m$. For $d \approx 1.9\,m$ the sources are very far from the microphones and the reflections start to affect the signal too much for a reconstruction to be useful for ASR. This means there is still future work to do although DUET is separating the sources in a real environment with low reverberation time for $d = 1\,m$. Changing the reconstruction criteria from the ML to a Λ-distance that is capable of discarding outliers could be

Table 5.3: Separation results for two people speaking naturally at around -30° and 30°. SNR_0 shows the SNR value prior to applying the separation and SNR_{IBM} the SNR value obtained using the ideal binary masks. The first test had the speakers at $1\,m$ from the microphones and the second at $1.9\,m$.

$d\,(m)$	SNR_0	SNR (dB)	SNR_{IBM} (dB)	SDR (dB)	SIR (dB)	SAR (dB)
1	2.15	4.10	7.08	5.41	13.73	6.29
1	2.56	6.62	11.60	6.46	11.60	8.34
1.9	2.90	1.88	10.16	0.87	3.32	6.20
1.9	1.79	0.68	11.66	-0.89	1.98	4.40

interesting to test in the $d \approx 1.9\,m$ case. In both tests, the coupling effect seen in the test with one wall inside the anechoic room and in the simulation where none of the sources is estimated at the right position occurs (fig. 5.16). For $d = 1\,m$ this issue was less severe.

5.4.2.2 Pink noise interferer

A pink noise interfering source at $1\,m$ and -45° to test if DUET could cope with it. It turns out that given its random nature (non-stationary), DUET is not able to pick it up thus only a cluster arises even though the energy of the fan-like noise is higher than the speaker. This can be seen in fig. 5.17. This has also other implications. The late reverberant tail which has noise-like characteristics will also slightly improve the SDR and SIR of the direct path signal although it won't make the identification of peaks easier in the Λ histogram, since as seen in the figure 5.17, the non-stationary signal did not cause significant changes in the Λ histogram shape.

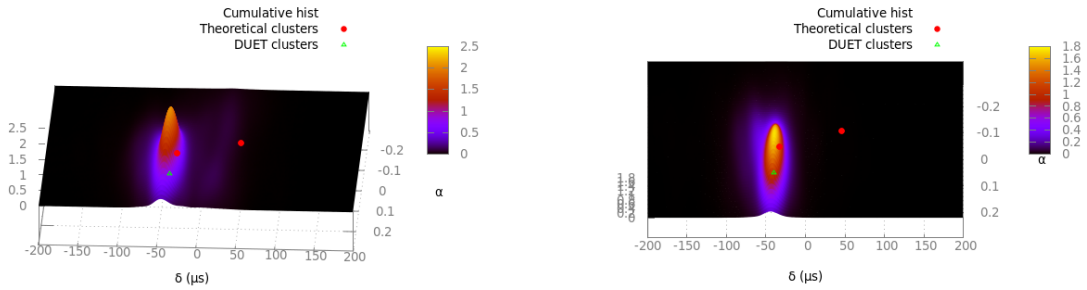


Figure 5.17: Source at $d = 1\,m$, 30° and interfering pink noise source at $1\,m$ and -45°. On the left the Λ histogram is shown without spectral subtraction applied. On the right, the iterative spectral subtraction from [Miyazaki et al., 2014] was applied.

Table 5.4: Separation results for a person at 30° and a pink noise source at -30°, both at $d \approx 1\,m$.

Method	SNR (dB)	SDR (dB)	SIR (dB)	SAR (dB)
DUET	3.06	1.80	2.03	16.75
Spectral sub. [Miyazaki et al., 2014] + DUET	5.04	8.33	25.19	8.43

By using a spectral subtraction module (implemented according to [Miyazaki et al., 2014]), where non-optimal $\eta = 0.7$, $\beta = 3.3$, $\alpha = 0.0$ and 20 iterations were used, the interference was largely eliminated increasing dramatically the signal quality. This merging of spectral subtraction is thus crucial for real scenarios with DUET.

Chapter 6

Conclusions

In this work the DUET method was explored in further detail. The crux of DUET is the peak tracking task in Λ space which was now significantly improved. One of the problems with this task is that the behaviour of the Λ space estimators and consequent tail production had not previously been studied in detail and understood.

The quasi-erratic $\hat{\alpha}$ behaviour that only shows up at specific frames when the Λ histogram is calculated for individual windows was discovered and its cause, non-stationarity of high-frequency speech sounds (plosives), determined. This challenges the fundamental local stationarity hypothesis of DUET [Jourjine et al., 2000; Rickard et al., 2001; Yilmaz and Rickard, 2004; Rickard, 2007; Özgür and Scott, 2004; Melia and Rickard, 2007; Yang et al., 2011] since the local stationarity error is neglegible only for the delay of the signals (given the high frequency), but not for the attenuation.

This behaviour, along with the predominantly bimodal $\hat{\delta}$ behaviour at low frequencies caused by leakage are the causes for tail production in the Λ space when high statistics are available (cumulative Λ histogram of $8\,s$) and is transformed into a much more serious problem with low statistics (individual or few histograms) with the production of spurious peaks and peak movement due to variations in the speech spectrum which critically reduce the peak tracking abilities, especially for the real-time peak tracking that has less statistics.

Both of these problems can be fixed. The $\hat{alpha}$ can be solved very efficiently, even with computational reduction times by discarding all the frequencies above 4kHz in the histogram. This completely eliminates the spurious peaks caused by non-stationary plosive sounds. It doesn't mean however that the output has that cutoff frequency. After having identified the clusters in the Λ space with this reduced bandwidth all the T-F bins up to the Nyquist frequency can still be separated and in this case, the source with that δ will receive the plosive sounds that come from $\hat{\delta}$. If however there are two sources with similar δ and different α, this will cause such non-stationary sounds to be attributed almost indiscriminately. For this application scenario a local confidence metric can be used to assess if a peak is spurious or not. If it is, this content can then be removed from all the estimated sources masks that have similar δ. A studied local confidence metric was the one-dimensional kurtosis evaluated over $\hat{\alpha}$ and $\hat{\delta}$ that had excellent results in the anechoic conditions even identifying spurious peaks that were stronger than the "true" one.

With the new Λ space understandings of what causes spurious peaks and tails, it should be possible to track peaks more accurately and hopefully execute DUET in realtime, in very cheap hardware, without an histogram and without knowing the number of sources [Rickard et al., 2001].

The behaviour of DUET under reflections was also studied. It was shown that the reflection images force in general a wrong localization of the original source and in general only $\hat{\alpha}$ mantains any

discriminating ability between source and reflection image.

Variants of the Hamming window aimed at reducing the local stationarity error for non-stationary signals have been tested and showed performance improvements in specific conditions (far away sources and high source sparsity), over the original window. Neglecting those variants, the optimal window was found to be the $64\,ms$ Hamming window, confirming the results in [Yilmaz and Rickard, 2004] even under a broader search scenario.

Although it has been shown that the method works in a real scenario when the human speakers were at $1\,m$ from the microphones, for increasing distances the energy of the reflections can be so large that it shadows the direct path (critical distance exceeded) and source separation is greatly degraded and thus further studies have to be made.

6.0.2.3 Experimental implementation concerns

The chosen equipment was completely standard off-the-shelf material and not laboratory-grade equipment, given the practical aim of the research. Had such higher quality equipment been used the results would have been even better.

However, after this transition sometimes the recorded signal had big fluctuations around the 0-level in the order of 1s. This variations are probably due to power fluctuations since the board is only powered through USB which makes its recording somewhat noisy (besides this big fluctuations). Such fluctuations were then removed by a high-pass filter with cutoff frequency in the 30Hz (to cut perfectly at 20Hz given the roll-off).

The electronic noise was assumed to be caused mainly/altogether by the soundcard since when trading microphones, the recordings with the better microphones generated the same perceived amount of noise though with much more natural sounding results (flatter frequency response).

The electronic quasi-white noise (in fact is stronger in low and high frequencies) is effectively a cause for concern since it disables DUET altogether so signal processing was required (Spectral Subtraction). This also possibilitates the use of cheaper microphones / preamp-stage. In fact it possibilitates the use of cheaper preamp stage since that is what's generating the noise. The cheaper microphones would be mainly chosen by the quality of the frequency response required which is not that much for ASR.

6.1 Main Achievements

Different windows were tested and a class of here named weakHamming windows was found that provided better separation under at least anechoic simulations but it is expected that the behaviour is mantained in the echoic scenario. From those, the lower weak family was found to provide equal SNR improvement as the weak windows but lower audible distortion.

A new heuristic method to detect peaks was used, running in linear time to the number of histogram bins. The method uses the fact that the most energetic peaks are the most relevant and so it gradually discards weak peaks if they're no longer able to fit on the stack. This delivers maximum peak identification count with minimal energy loss.

A global confidence metric based on the ratio of measured energy allows us to know if all the peaks were detected very quickly without resorting to a GMM model.

A previously unspoken of weakness in the attenuation parameter for high frequencies was discovered and a local confidence metric over the 2D histogram data based on uni-dimensional kurtosis was derived to cope with it and it achieved very good unreliable peak detection with acceptable true positive rejection.

New Mahalanobis-based distances were proposed to measure distances in the Λ plane.

A streaming mechanism was developed LDMB clusters position, and overlapping windowed signal correlation. Spectral Subtraction was used to clean the inputs.

The real-time multi-frame with heuristic source tracking and LDMB source streaming was developed.

The BSS Eval was modified to allow degenerate solutions and the optimality criteria for assignment had to be changed from maximal average SIR to Dtotal as maximal SIR per source was unable to assign correctly the estimates even in good separation scenarios.

6.2 Future Work

The Hamming variants showed that there is a potential improvement to be done by using specialized windows. The study of other windows, including parametric ones is not new, [Gaddipati et al., 2010].

As more cluster mixing occurs in the two-dimensional Λ histogram this metric cannot be used, but another one could be the frequency content of the cluster. If it is all from a very narrow-band it is a spurious peak and should be discarded or assimilated to a "true" peak.

The spectral subtraction coefficients used could be optimized according to the formulas in [Miyazaki et al., 2014].

The next step to increase robustness for sources found at great distances in echoic environments is to find the optimal microphone spacing, akin to the study in [Johannes et al., 2014], using the phase-aliasing correction [Yang et al., 2011], since reverberation in a real environment couples the object and image clusters in the Λ space although $\hat{\alpha}$ seems to have better discrimination under that coupling and $\hat{\alpha}$ becomes more informative as the distance between microphones grows.

References

(2012). Covariance smoothing and consistent wiener filtering for artifact reduction in audio source separation.

Abrard, F. and Deville, Y. (2003). Blind separation of dependent sources using the 'time-frequency ratio of mixtures' approach. *ISSPA 2003. Paris, France: IEEE, July 2003.*

Aichner, R. (2007). Acoustic blind source separation in reverberant and noisy environments.

Alsteris, L. D. and Paliwal, K. K. (2005). Some experiments on iterative reconstruction of speech from stft phase and magnitude spectra. In *INTERSPEECH*, pages 337–340.

Anuradha R. Fukane, S. L. S. (2011). Different approaches of spectral subtraction method for enhancing the speech signal in noisy environments. *International Journal of Scientific & Engineering Research, Vol. 2, Issue 5, May 2011, ISSN 2229-5518.*

Arberet, S., Gribonval, R., and Bimbot, F. (2011). A robust method to count and locate audio sources in a multichannel underdetermined mixture.

Attias, H. (1999). Independent factor analysis.

Attias, H. (2000). Independent factor analysis with temporally structured sources.

Aziz-Sbaï, S. M., Aïssa-El-Bey, A., and Pastor, D. (2012). Contribution of statistical tests to sparseness-based blind source separation. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–15.

Balan, R., Rosca, J., Rickard, S., and O'Ruanaidh, J. (2000). The influence of windowing on time delay estimates. In *Proceedings CISS.*

Barker, J., Coy, A., Ma, N., and Cooke, M. (2006). Recent advances in speech fragment decoding techniques. *Proceedings of the Ninth International Conference on Spoken Language Processing (ICSLP).*

Barry, D., Lawlor, B., and Coyle, E. (2004). Sound source separation: Azimuth discrimination and resynthesis. *Proc. of DAFx.*

Bell, A. J. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation, 7:1129-1159.*

Bimbot, A. O. V. (2010). A general flexible framework for the handling of prior information in audio source separation.

Bofill, P. and Zibulevsky, M. (2001). Underdetermined Blind Source Separation using sparse representations. *Signal Processing, Vol. 81.*

Boldt, J., Pedersen, M., Kjems, U., Christensen, M., and Jensen, S. (2010). Error-correction of binary masks using Hidden Markov Models. *Proc. IEEE int. Conf. Acoust., Speech, Signal Processing, 2010, pp. 4722-4725.*

Bouchard, M., Norcross, S. G., and Soulodre, G. A. (2006). Inverse filtering design using a minimal-phase target function from regularization. In *Audio Engineering Society Convention 121.* Audio Engineering Society.

Bouvrie, J. V. and Ezzat, T. (2006). An incremental algorithm for signal reconstruction from short-time fourier transform magnitude. In *INTERSPEECH.*

Bradbury, J. (2000). Linear predictive coding. *Mc G. Hill.*

Brandstein, M. S. (2005). *A framework for speech source localization using sensor arrays.* PhD thesis, Brown University.

Buchner, H. and Aichner, R. (2004). Trinicon: A versatile framework for multi-channel blind signal processing. *Proc. of ICASSP.*

Burred, J. and Sikora, T. (2007). Monaural sources separation from musical mixtures based on time-frequency timbre models. *Proc. Int. Conf. on Music Information Retrieval (ISMIR), Vienna, Austria, September 2007.*

Burred, J. J. (2008). Musical source separation: Principles and state of the art.

Cahill, N., Cooney, R., Humphreys, K., and Lawlor, R. (2006). Speech source enhancement using a Modified ADRess algorithm for applications in mobile communications. *Proc. of the AES 121st Conv.*

Casey, M. and Wstner, A. (2000). Separation of mixed audio sources by independent subspace analysis. *Proc, Int. Computer Music Conference (ICMC), Berlin, Germany.*

Cho, E. (2013). Source separation tutorial mini-series i, speech enhancement algorithms.

Choi, E. H. C. (2004). Noise robust front-end for asr using spectral subtraction, spectral flooring and distribution mapping.

Comon, P. (1994). Independent component analysis, a new concept?

consortium, D. (2005). Summary of speech technology and market opportunities in the tv and set-top box markets: hands-free remote control systems.

consortium, D. (2009). Dicit acoustic woz data.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication, vol. 34, no. 3, pp. 267-285.*

Deleforge, A., Forbes, F., and Horaud, R. (2013). Variational em for binaural sound-source separation and localization. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 76–80.

Deng, L., Acero, A., Jiang, L., Droppo, J., and Huang, X. D. (2001). High-performance robust speech recognition using stereo training data. *Proceedings of the 2001 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Dong, L., Foo, S. W., and Lian, Y. (2005). A two-channel training algorithm for hidden markov model and its application to lip reading. *EURASIP Journal on Applied Signal Processing 2005:9, 1382-1399*.

Ephraim, Y. (1992). Statistical-model-based speech enhancement systems. *Proceedings of the IEEE, vol. 80, no. 10, pp. 152-1555*.

et al., J. K. L. (1997). Feature extraction approach to blind source separation. *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*.

Evans, N. W. D., Mason, J. S., Liu, W. M., and Fauve, B. (2005). On the fundamental limitations of spectral subtraction: An assessment by automatic speech recognition.

Farina, A. (2007). Impulse response measurements. Technical report.

Farina, A. and Righini, F. (1997). Software implementation of an mls analyzer with tools for convolution, auralization and inverse filtering. In *Audio Engineering Society Convention 103*. Audio Engineering Society.

Fastl, H. and Zwicker, E. (2007). *Psychoacoustics, Facts and Models*.

Favrot, A., Erne, M., and Faller, C. (2006). Improved cocktail-party processing. *Proc. of DAFx*.

Frey, B. J., Deng, L., Acero, A., and Kristjansson, T. (2001). Algonquin: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*.

Furuya, K. and Kataoka, A. (2007). Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(5):1579–1591.

Gaddipati, P., Dave, N., Rao, P., and Velmurugan, R. (2010). Improving time-frequency sparsity for audio spatialization by time-adaptive windowing. In *Communications (NCC), 2010 National Conference on*, pages 1–5. IEEE.

Gribonval, R., Vincent, E., Févotte, C., and Benaroya, L. (2003). Proposals for performance measurement in source separation. In *in Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation (ICA2003)*, pages 763–768.

Hummersone, C., Mason, R., and Brookes, T. (2011). Ideal binary mask ratio: a novel metric for assessing binary-mask-based sound source separation algorithms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2039–2045.

Inoue, T., Saruwatari, H., Takahashi, Y., Shikano, K., and Kondo, K. (2010). Theoretical analysis of iterative weak spectral subtraction via higher-order statistics. In *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, pages 220–225. IEEE.

Johannes, T., Minje, K., and Paris, S. (2014). Phase and level difference fusion for robust multichannel source separation.

Jourjine, A., Rickard, S., and Özgür Yilmaz (2000). Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. *Proc. ICASSP2000, Vol. 5, Pages 2985-2988, June 5-9, Istanbul, Turkey.*

J.Rennie, S. (2008). Graphical models for robust speech recognition in adverse environments.

Khan, M. R. and Hasan, T. (2008). Iterative noise power subtraction technique for improved speech quality. In *Electrical and Computer Engineering, 2008. ICECE 2008. International Conference on*, pages 391–394. IEEE.

Kinsler, L. E., Frey, A. R., Coppens, A. B., and Sanders, J. V. (1982). *Fundamentals of ACOUSTICS, third edition.*

Knuth, K. H. (2005). Informed source separation: A baeyesian tutorial.

Kondo, K. (2014). A computationally restrained and single-channel blind dereverberation method utilizing iterative spectral modifications.

Krawczyk, M. and Gerkmann, T. (2012). Stft phase improvement for single channel speech enhancement. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pages 1–4. VDE.

Kristjansson, T., Hershey, J., and Attias, H. (2004). Single microphone source separation using high resolution signal reconstrucion. *Proceedings of the 2004 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Le Roux, J., Kameoka, H., Ono, N., and Sagayama, S. (2010). Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency. In *Proc. of International Conference on Digital Audio Effects DAFx*, volume 10.

Li, Y., Powers, D., and Peach, J. (2000). Comparison of blind source separation algorithms. *2000 WSES; p18-21, Mastorakis(2000) Advances in Neural Networks and Applications, WSES.*

Li, Y. and Wang, D. L. (2009). On the optimality of ideal binary time-frequency masks. *Speech Communication, vol. 51, no. 3, pp. 230-239.*

Litovsky, R. (2008). *Binaural Hearing.* Cochlear.

Little, M. A. (2011). Mathematical foundations of nonlinear, non-gaussian, and time-varying digital speech signal processing. In *Advances in Nonlinear Speech Processing*, pages 9–16. Springer.

López, N., Maazaoui, M., Grenier, Y., Richard, G., and Bourmeyster, I. (2013). Does dereverberation help multichannel source separation? a case study. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5. IEEE.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (2010). Model-based expectation-maximization source separation and localization. *IEEE Trans. on Audio,Speech and Language Proc., vol. 18, pp. 382-394.*

Melia, T. and Rickard, S. (2007). Underdetermined blind source separation in echoic environments using desprit. *EURASIP Journal on Applied Signal Processing*, 2007(1):90–90.

Mitianoudis, N. and Davies, M. (2003). Audio source separation of convolutive mixtures. *IEEE Trans. on Speech and Audio Processing, vol. 11, pp. 489-497.*

Miyazaki, R., Saruwatari, H., Inoue, T., Shikano, K., and Kondo, K. (2012). Musical-noise-free speech enhancement: Theory and evaluation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4565–4568. IEEE.

Miyazaki, R., Saruwatari, H., Nakamura, S., Shikano, K., Kondo, K., Blanchette, J., and Bouchard, M. (2014). Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction. *Signal Processing*, 102:226–239.

Miyoshi, M. and Kaneda, Y. (1988). Inverse filtering of room acoustics. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(2):145–152.

Moore, C. (2003). *An introduction to the psychology of hearing*. San Diego, CA: Academic Press.

Mosayyebpour, S., Sayyadiyan, A., Zareian, M., and Shahbazi, A. (2010). Single channel inverse filtering of room impulse response by maximizing skewness of lp residual. In *Signal Acquisition and Processing, 2010. ICSAP'10. International Conference on*, pages 130–134. IEEE.

Mouba, J., Marchand, S., et al. (2006). A source localization/separation/respatialization system based on unsupervised classification of interaural cues. *A Source Localization/Separation/Respatialization System Based on Unsupervised Classification of Interaural Cues*, pages 233–238.

Mowlaee, P. (2010). New strategies for single-channel speech separation.

Nagira, K., Otsuka, T., and Okuno, H. G. (2013). Nonparametric bayesian sparse factor analysis for frequency domain blind source separation without permutation ambiguity. *EURASIP Journal on Audio, Speech, and Music Processing 2013, 2013:4.*

Nakatani, T., Araki, S., Yoshioka, T., and Fujimoto, M. (2011). Joint unsupervised learning of hidden markov source models and source location models for multichannel source separation.

Naylor, P. A. and Gaubitch, N. D. (2010). *Speech dereverberation*. Springer.

Ozerov., A., Phillippe., O., Gribonval, R., and Bimbot, F. (2005). One microphone singing voice separation using source-adapted models. *Proc. IEEE Workshop on Applicaitons of Signal Processing to Audio and Acoustics (WASPAA), New Paltz. USA. 2005.*

Özgür, Y. and Scott, R. (2004). Blind separation of speech mixtures via time-frequency masking. *Signal Processing, IEEE Transactions on*, 52(7):1830–1847.

Pedersen, M. S., Larsen, J., Kjems, U., and Parra, L. C. (2007). A survey of convolutive blind source separation methods. *Multichannel Speech Processing Handbook*, pages 1065–1084.

Perrin, G. (2012). Multipoint optimization of a loudspeakeer impulse response.

Puigt, M. and Deville, Y. (2006). A time-frequency correlation-based blind source separation method for time-delay mixtures. *Proc. of ICASSP.*

Radfer, M. H., Wong, W., Dansereau, R. M., and Chan, W. Y. (2010). Scaled factorial hidden Markov models: a new technique for compensating gain differences in model-based single channel speech separation. *Proc. ICASSP-2010, pp. 1918-1921.*

Ratnam, R., Jones, D. L., Wheeler, B. C., O'Brien Jr, W. D., Lansing, C. R., and Feng, A. S. (2003). Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892.

Rayleigh, L. (1907). On our perception of sound direction. *Philosophical Magazine, vol. 6, pp. 214-232.*

Ribeiro, S. (2014). Nos perdeu 100 mil clientes num ano. http://www.sol.pt/noticia/107677. Publisher: Jornal SOL, accessed on September 2014.

Rickard, S. (2007). The DUET blind source separation algorithm. *Blind Speech Separation, S. Makino, T. W. Lee and H. Sawada, Eds. Springer.*

Rickard, S., Balan, R., and Rosca, J. (2001). Real-time time-frequency based blind source separation. In *Proc. ICA*, pages 651–656.

Rickard, S. and Özgür Yilmaz (2002). On the approximate w-disjoint orthogonality of speech. 1:I–529.

Röbel, A. (2006). Analysis/resynthesis with the short time fourier transform. Technical report.

Roman, N. and Woodruff, J. (2011). Intelligibility of reverberant noisy speech with ideal binary masking. *The Journal of the Acoustical Society of America*, 130(4):2153–2161.

Romanin, M., Marchetto, E., and Avanzini, F. (2009). A spectral subtraction rule for real-time dsp implementation of noise reduction in speech signals.

Roux, J. L. and Vincent, E. (2013). Consistent wiener filtering for audio source separation. *IEEE Signal Processing letters, vol. 20, no. 3, March 2013.*

Rowe, D. B. (1999). Bayesian blind source separation. *IEEE Transactions Signal Processing.*

Roweis, S. T. (2003). Factorial models and refiltering for speech separation and denoising. *Proc. EUROSPEECH-2003, pp. 1009-1012.*

Sadjadi, S. O. and Hansen, J. H. (2012). Blind reverberation mitigation for robust speaker identification. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4225–4228. IEEE.

Sawada, H., Araki, S., and Makino, S. (2007). A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 139–142. IEEE.

Schilling, M. F., Watkins, A. E., and Watkins, W. (2002). Is human height bimodal? *The American Statistician*, 56(3):223–229.

Schluter, R. and Ney, H. (2001). Using phase spectrum information for improved speech recognition performance. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 133–136. IEEE.

Selesnick, I. W. (2009). Short-time fourier transform and its inverse.

Smith, S. W. et al. (1997). *The Scientist and Engineers Guide to DSP*. California Technical Pub. San Diego.

Stan, G.-B., Embrechts, J.-J., and Archambeau, D. (2002). Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262.

Sturmel, N. and Daudet, L. (2011). Signal reconstruction from stft magnitude: a state of the art. In *International Conference on Digital Audio Effects (DAFx)*, pages 375–386.

Thiemann, J. and Vincent, E. (2013). An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement.

Traa, J. (2013). Multichannel source separation and tracking with phase differences by random sample consensus.

Traa, J. and Smaragdis, P. (2013). Blind multi-channel source separation by circular-linear statistical modeling of phase differences.

Tsilfidis, A. (2011). *Signal processing methods for enhancing speech and music signals in reverberant environments*. PhD thesis, PhD Thesis, University of Patras.

Tsilfidis, A. and Mourjopoulos, J. (2011). Blind single-channel suppression of late reverberation based on perceptual reverberation modeling. *The Journal of the Acoustical Society of America*, 129(3):1439–1451.

Valin, J.-M., Michaud, F., Rouat, J., and Létourneau, D. (2003). Robust sound source localization using a microphone array on a mobile robot. In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 2, pages 1228–1233. IEEE.

Vesa, S. (2004). *Estimation of reverberation time from binaural signals without using controlled excitation.* PhD thesis, HELSINKI UNIVERSITY OF TECHNOLOGY.

Vesa, S. and Harma, A. (2005). Automatic estimation of reverberation time from binaural signals. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 3, pages iii/281–iii/284 Vol. 3.

Vincent, E. (2005). Blind audio source separation, a review of state-of-the-art techniques. *Centre for Digital Music, Electronic Engineering Department, Queen Mary, University of London.*

Vincent, E. (2012). Improved perceptual metrics for the evaluation of audio source separation. In *Latent Variable Analysis and Signal Separation*, pages 430–437. Springer.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469.

Vincent, E., Sawada, H., Bofill, P., Makino, S., and Rosca, J. P. (2007). First stereo audio source separation evaluation campaign: data, algorithms and results. In *Independent Component Analysis and Signal Separation*, pages 552–559. Springer.

Virtanen, T. (2003). Sound source separation using sparse coding with temporal continuity objective. *Proc. Int. Computer Music Conference (ICMC), Singapore.*

Virtanen, T. (2004). Separation of sound sources by convolutive sparse coding. *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA), Jeju, Korea.*

Viste, H. and Evangelista, G. (2003). On the use of spatial cues to improve binaural source separation. *Proc. of DAFx.*

Wang, D. and Brown, G. (2006). Computational auditory scene analysis: Principles, algorithms, and applications. *Wiley-IEEE Press.*

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (2008). Speech perception of noise with binary gains. *The Journal of the Acoustical Society of America, vol. 124, no. 4, pp. 2303-2307.*

Wölfel, M. and McDonough, J. (2009). *Distant Speech Recognition.* Wiley Online Library.

Yang, W., Özgür, Y., and Zhengfang, Z. (2011). Phase aliasing correction for robust blind source separation using duet. *IEEE Transaction on signal processing.*

Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Proc., vol. 52, pp. 1830-1847.*

Yoshii, K., Tomioka, R., Mochihashi, D., and Goto, M. (2013). Beyond nmf - time-domain audio source separation without phase reconstruction. *14th International Society for Music Information Retrieval Conference (ISMIR 2013).*

Yu, G., Mallat, S., and Bacry, E. (2008). Audio denoising by time-frequency block thresholding. *IEEE TSP*.

Zhang, W., Habets, E. A., and Naylor, P. A. (2010). On the use of channel shortening in multichannel acoustic system equalization. In *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*.

Appendix A

Figures

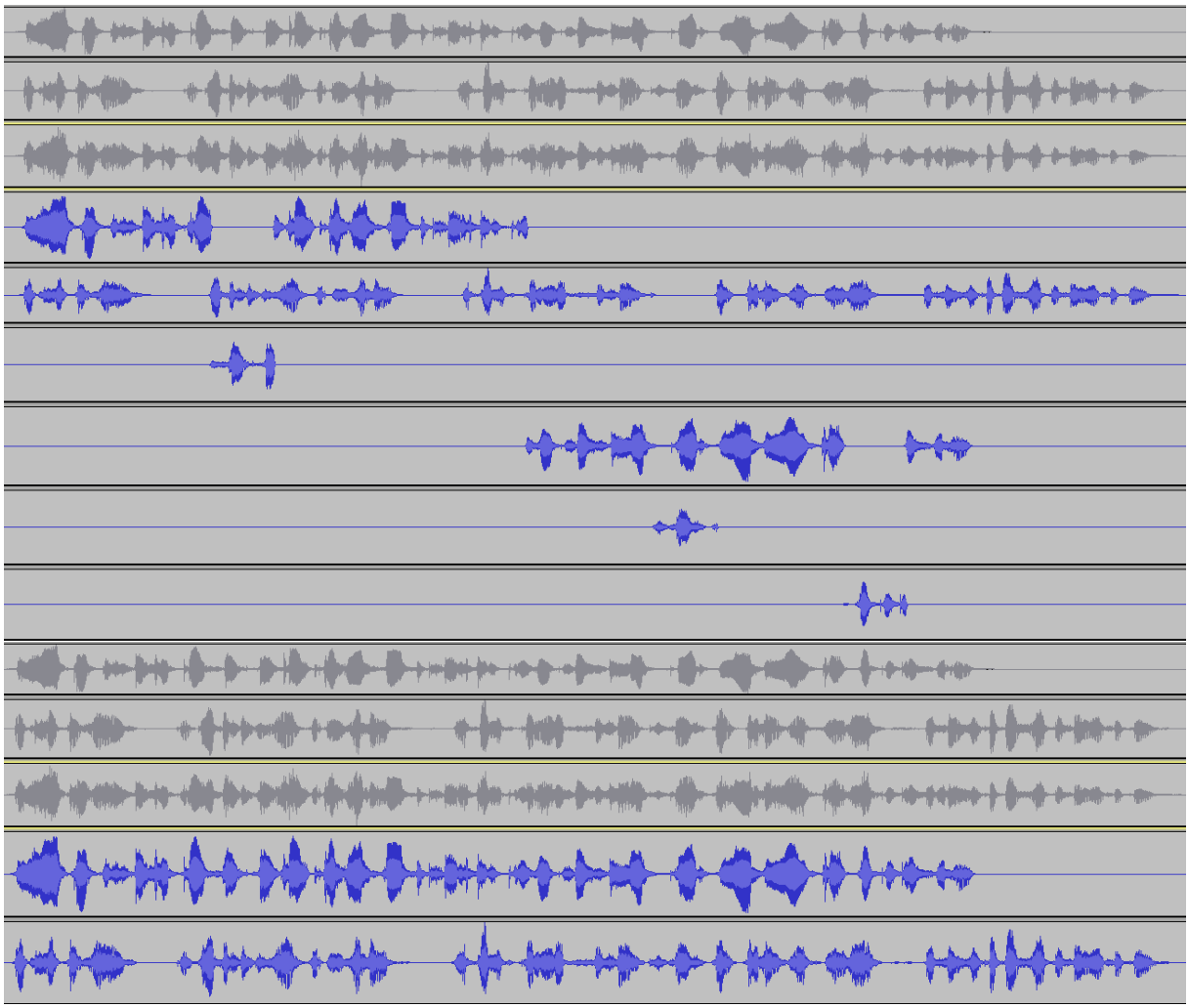


Figure A.1: The left channel of the two source signals are first shown, followed by the left channel of their mixture (left channel input), shown in grey. Next the blue waveforms are the stream outputs for $N_{accum} = 15$. Finally, the individual sources and mixture are shown again in grey and finally the streams output by the system with $N_{accum} = 33$, the limit at which only two streams (as many as input sources) are estimated.

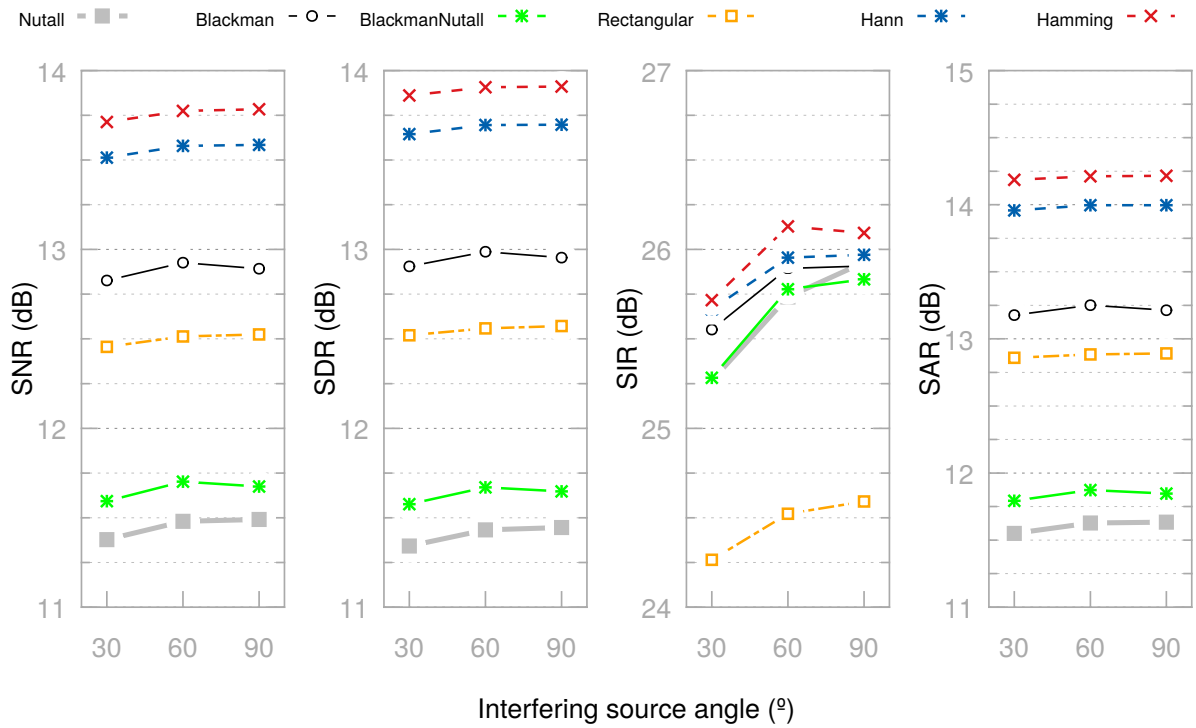


Figure A.2: Separation metrics for the anechoic simulation using standard window types. The target source is placed at 0° and $d_{target} = d_{interferer} = 1\,m$.

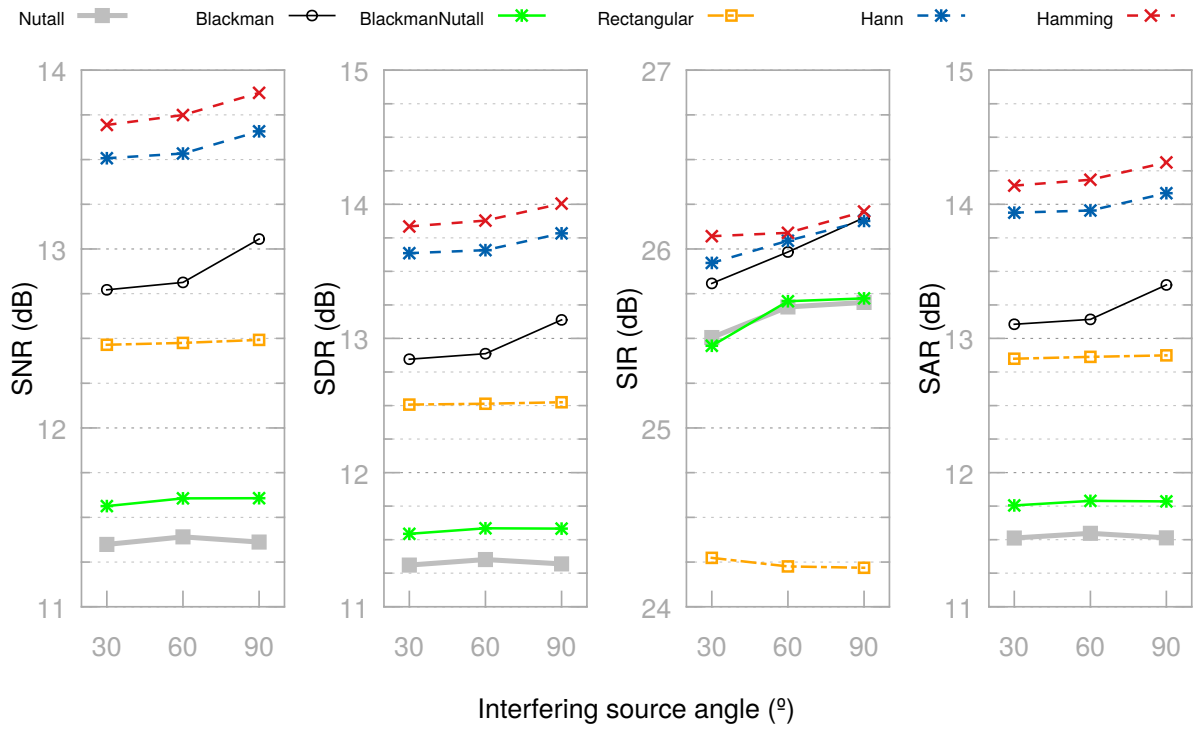


Figure A.3: Separation metrics for the anechoic simulation using standard window types. The target source is placed at 0° and $d_{target} = d_{interferer} = 2.5\,m$

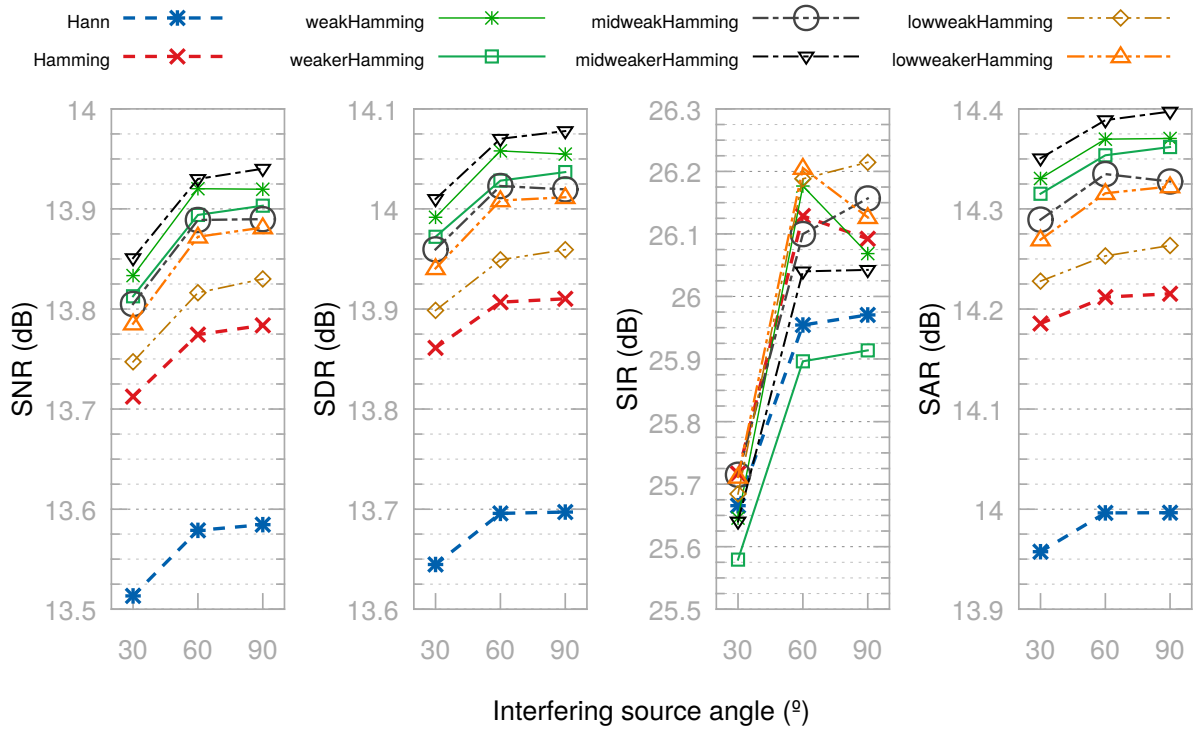


Figure A.4: Separation metrics for the Hann, Hamming and custom weak and weaker Hamming window types using simulated data with sources at 1m.

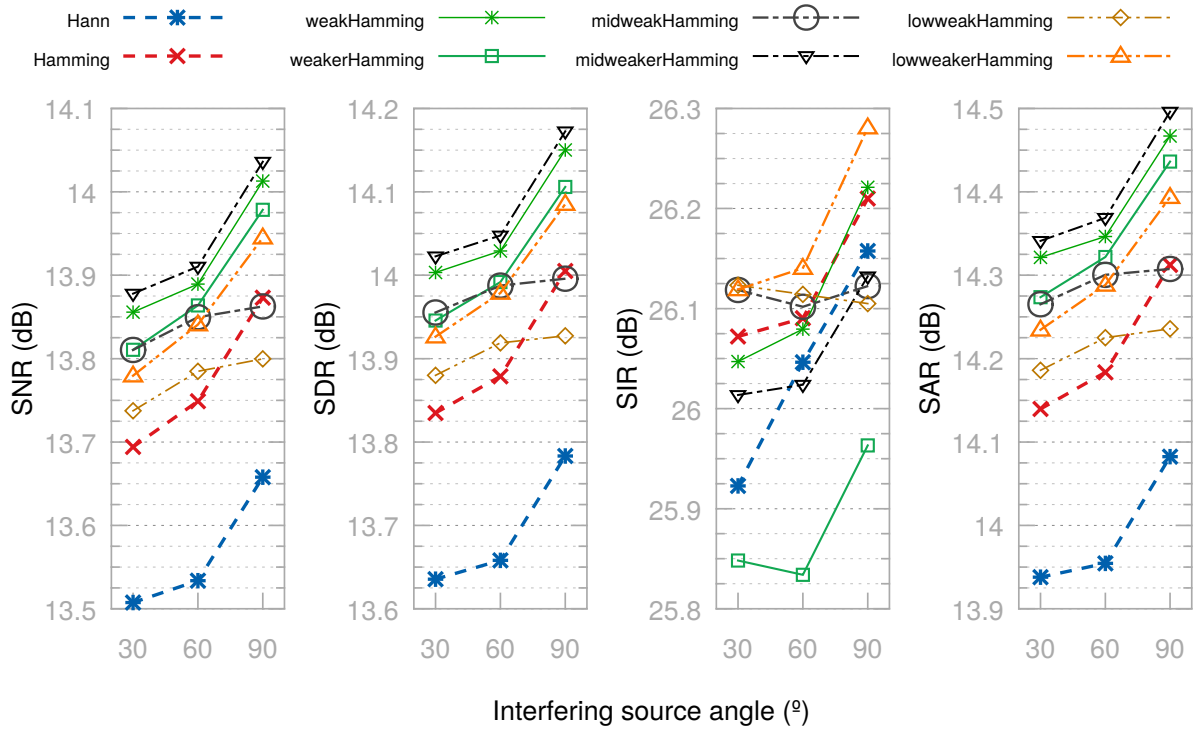


Figure A.5: Separation metrics for the Hann, Hamming and custom weak and weaker Hamming window types using simulated data with sources at 2.5m.

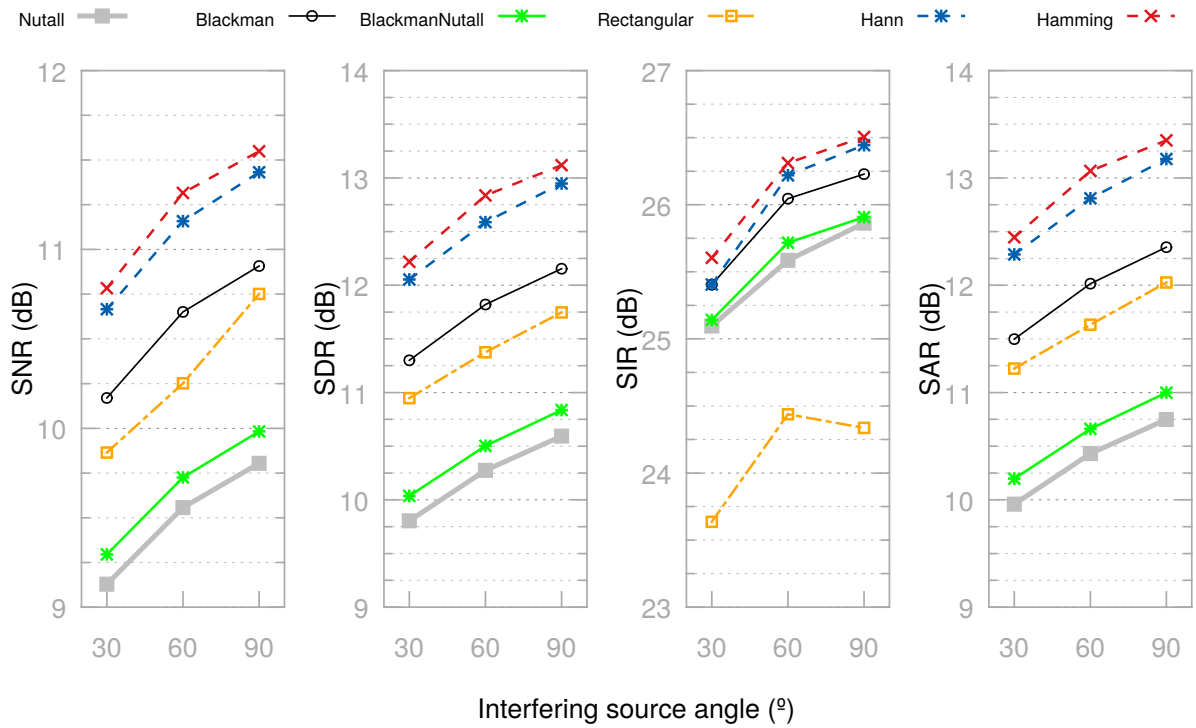


Figure A.6: Separation metrics for the anechoic chamber using different window types. The target source is placed at 0°, 1m away from the microphones, as the interfering source.

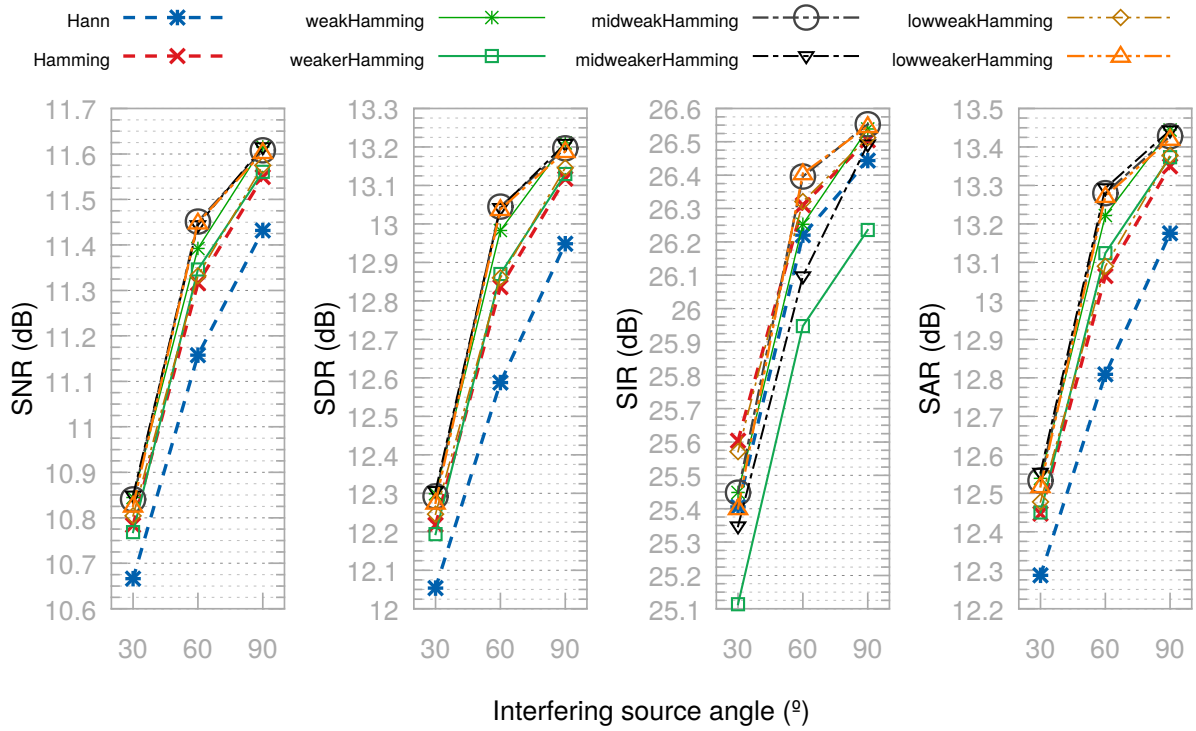


Figure A.7: Separation metrics for the Hann, Hamming and custom weak and weaker Hamming window types using anechoic chamber data with sources at 1m.

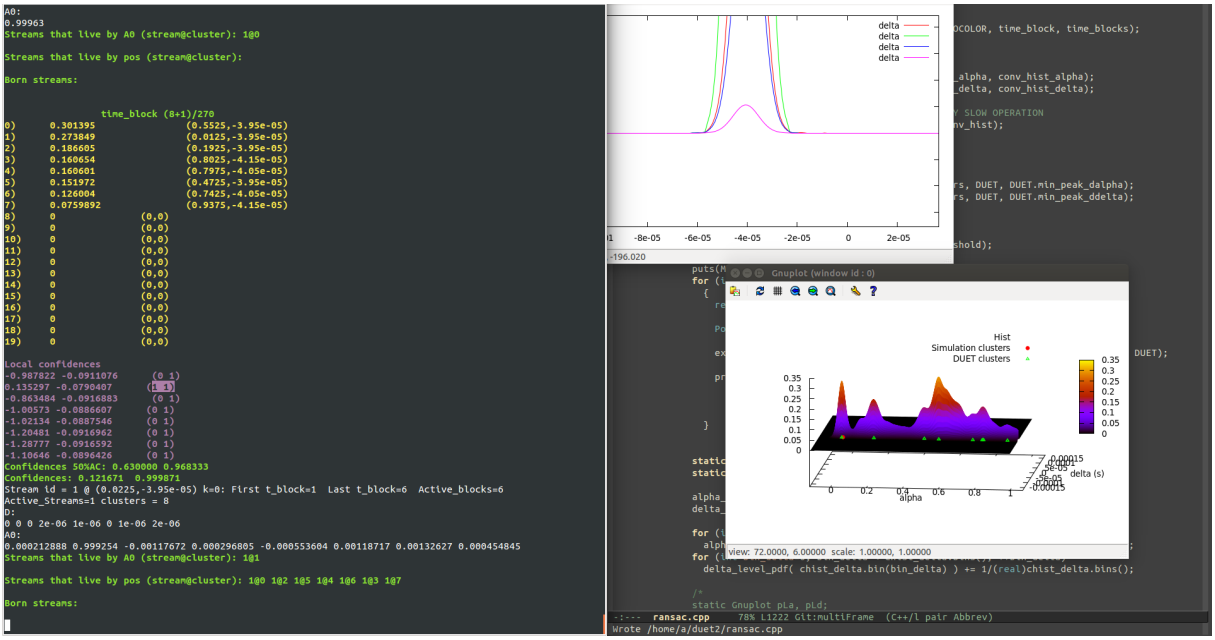


Figure A.8: Frame where spurious peaks appear and all are rejected but the true one which is the second tallest peak. On the left the highlighted purple numbers show the only peak that passed the kurtosis test was the second most tall which is the right one. On the right the 2D histogram in Λ space of that frame is shown.

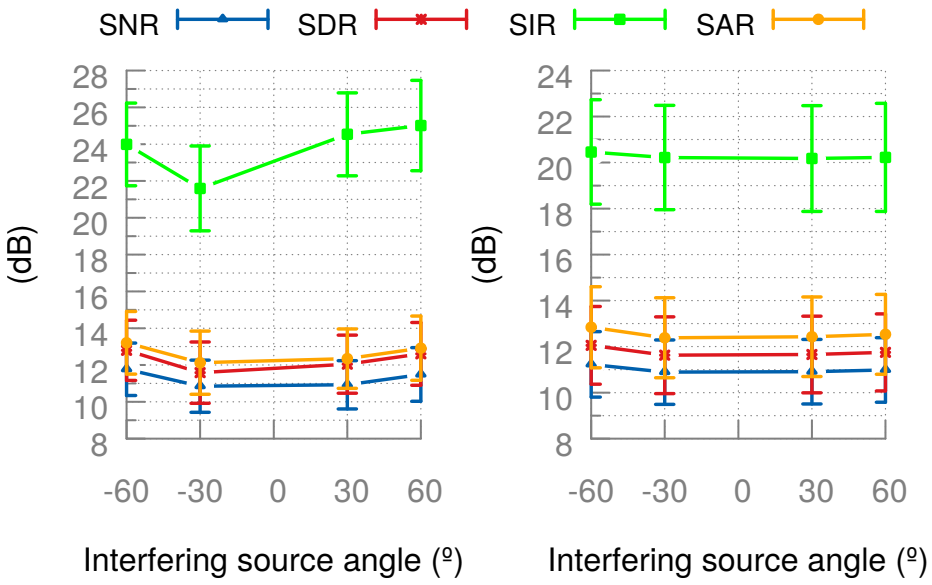


Figure A.9: Separation stats for the central target source with the DUET (left) and IBM separation stats for the same condition. Test performed with 2 sources in the anechoic room at $d = 1m$.

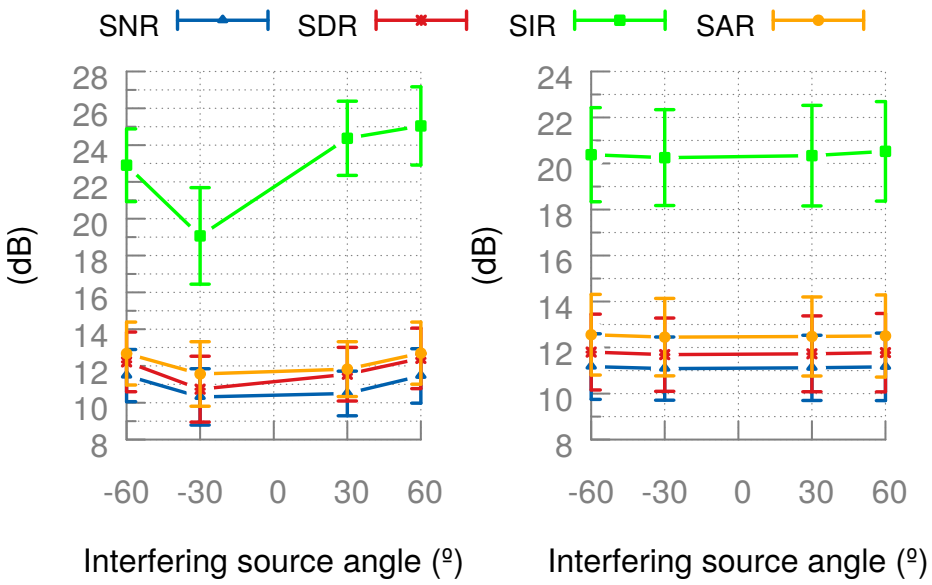


Figure A.10: Separation stats for the central target source with the DUET (left) and IBM separation stats for the same condition. Test performed with 2 sources in the anechoic room at $d = 2.5m$.

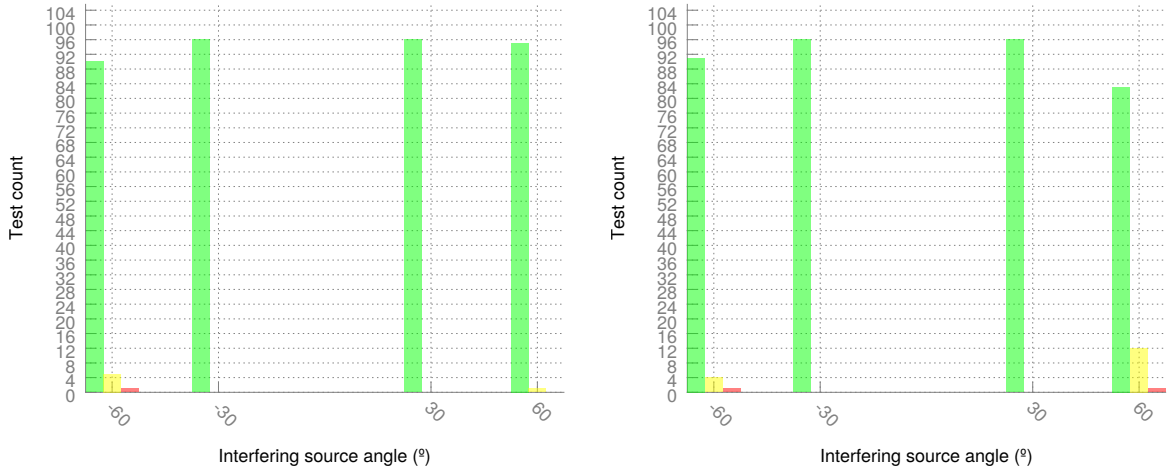


Figure A.11: Over-degeneracy counts over individual tests with two sources in the anechoic room for 1m (above) and 2.5m (below). Green bars show the number of tests for an angle where no over-degeneracies occurred, yellow, the number of tests where one over-degeneracy occurred and red, the number of tests where the number of over-degeneracies was equal to two.

Appendix B

Movement in the $\alpha - \delta$ plane

To track the sources one shall estimate, based on the last known position, where the source will be in the future, given a maximum speed.

Tracking the source up to second order allows determining it's current speed, which, if smaller than the configured maximum speed of any source, will reduce even more the search area.

This requires determining the speed of parameters α and δ. The next section is concerned with finding the maximum value for those parameters in function of the sound source's speed in 3D-space so as to bound them and consequently find the limit of the region where the source might be in the future.

Again, the delayed model shall be used in the next section.

B.0.0.4 δ-variations

By the definition of x_1 and x_2 in the DUET method, δ_n is the delay between the sound reaching microphones 1 and 2, thus, being c as the speed of sound and t_{nm} the time required for sound to travel the distance d_{nm}, from the source n to the microphone m:

$$\delta_n \equiv t_{n2} - t_{n1} = \frac{d_{n2} - d_{n1}}{c} \tag{B.1}$$

$$\begin{aligned} \delta_n &= \frac{d_{n2} - d_{n1}}{c} = \frac{|\vec{s_n} - \vec{x_2}| - |\vec{s_n} - \vec{x_1}|}{c} \\ &= \frac{\sqrt{(s_x - \Delta/2)^2 + s_y^2 + s_z^2} - \sqrt{(s_x + \Delta/2)^2 + s_y^2 + s_z^2}}{c} \end{aligned} \tag{B.2}$$

Having quantified $\delta_n(\vec{s_n})$ the next step is to determine how fast it changes, which is done by differentiating B.2 with respect to time:

$$
\begin{aligned}
\dot{\delta_n} &= \frac{\partial \delta_n}{\partial s_x}\dot{s_x} + \frac{\partial \delta_n}{\partial s_y}\dot{s_y} + \frac{\partial \delta_n}{\partial s_z}\dot{s_z} \\
&= \frac{s_x\dot{s_x} + s_y\dot{s_y} + s_z\dot{s_z}}{c}\left(\frac{1}{\sqrt{(s_x-\Delta/2)^2+s_y^2+s_z^2}} - \frac{1}{\sqrt{(s_x+\Delta/2)^2+s_y^2+s_z^2}}\right) \\
&\qquad - \frac{1}{c}\frac{\Delta}{2}\left(\frac{1}{\sqrt{(s_x-\Delta/2)^2+s_y^2+s_z^2}} + \frac{1}{\sqrt{(s_x+\Delta/2)^2+s_y^2+s_z^2}}\right)\dot{s_x} \\
&= \frac{s_x\dot{s_x} + s_y\dot{s_y} + s_z\dot{s_z}}{c}\underbrace{\left({d_{n2}}^{-1} - {d_{n1}}^{-1}\right)}_{\gamma_-} - \frac{1}{c}\frac{\Delta}{2}\underbrace{\left({d_{n2}}^{-1} + {d_{n1}}^{-1}\right)}_{\gamma_+}\dot{s_x} \\
&\equiv \frac{s_x\dot{s_x} + s_y\dot{s_y} + s_z\dot{s_z}}{c}\gamma_- - \frac{1}{c}\frac{\Delta}{2}\gamma_+\dot{s_x}
\end{aligned}
\tag{B.3}
$$

Next, one attributes an upper and lower bound to $\dot{\delta_n}$, or an upper bound to $|\dot{\delta_n}|$, assuming the maximum source's speed in each direction is symmetrical.

$$
\ddot{\delta_n} = \frac{\partial \ddot{\delta_n}}{\partial s_x}\dot{s_x} + \frac{\partial \ddot{\delta_n}}{\partial s_y}\dot{s_y} + \frac{\partial \ddot{\delta_n}}{\partial s_z}\dot{s_z} + \frac{\partial \ddot{\delta_n}}{\partial \dot{s_x}}\ddot{s_x} + \frac{\partial \ddot{\delta_n}}{\partial \dot{s_y}}\ddot{s_y} + \frac{\partial \ddot{\delta_n}}{\partial \dot{s_z}}\ddot{s_z} \tag{B.4}
$$

To that end, calculating the extrema points is required, i.e., where the gradient is $\vec{0}$:

$$
\nabla\dot{\delta_n} = \left(\frac{\partial \dot{\delta_n}}{\partial \dot{s_x}}, \frac{\partial \dot{\delta_n}}{\partial \dot{s_y}}, \frac{\partial \dot{\delta_n}}{\partial \dot{s_z}}\right) = \left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-},\ s_y,\ s_z\right)\frac{\gamma_-}{c} \tag{B.5}
$$

$$
\nabla\dot{\delta_n} = \left(\frac{\partial \dot{\delta_n}}{\partial \dot{s_x}}, \frac{\partial \dot{\delta_n}}{\partial \dot{s_y}}, \frac{\partial \dot{\delta_n}}{\partial \dot{s_z}}\right) = (0,0,0) \Leftrightarrow \begin{cases} s_x\gamma_- = \frac{\Delta}{2}\gamma_+ \\ s_y\gamma_- = 0 \\ s_z\gamma_- = 0 \end{cases} \tag{B.6}
$$

$$
\Leftrightarrow \begin{cases} s_x\gamma_- = \frac{\Delta}{2}\gamma_+ \\ s_y = 0 \vee d_{n2} = d_{n1} \\ s_z = 0 \vee d_{n2} = d_{n1} \end{cases} \quad \Leftrightarrow \begin{cases} s_x\gamma_- = \frac{\Delta}{2}\gamma_+ \\ s_y = 0 \vee s_x = 0 \\ s_z = 0 \vee s_x = 0 \end{cases} \tag{B.7}
$$

Since the setup will use two microphones facing towards the room which gives maximal azimutal discrimination, $s_y > 0$. Thus in this setup, $d_{n2} = d_{n1}$ which is only true if $s_x = 0$, which means the source must be in the midplane between the two microphones and s_z can take any value. The extrema points are found then in this setup for $s_x = 0 \wedge s_y > 0$, which is the bissector semi-plane in front of the two microphones.

Constrained estimation Since $|\dot{\delta_n}|$ increases with $|\dot{\vec{s_n}}|$ in magnitude, as B.3 shows, because it is one of its dual coordinates, a constraint on its maximum value is necessary. Two constraints can be given to the system: the maximum size of the deployment room and the maximum speed of the sound sources.

Since sound sources will in general be people talking, there's a maximum speed at which they can move and this value is even more constrained when indoors, and even more in typical usage scenarios.

Since the aim is to specify a search perimeter based on maximum speed given it's last known location, the problem is specified as[1]:

$$\max_{\dot{\vec{s_n}}} \dot{\delta_n}(\vec{s_n}, \dot{\vec{s_n}})|_{\vec{s_n}=\vec{const.}} \equiv \max \dot{\delta_n}(\dot{\vec{s_n}})|_{\vec{s_n}}, \quad v_M = \sqrt{\dot{s_x}^2 + \dot{s_y}^2 + \dot{s_z}^2} \tag{B.8}$$

where v_M is the configured maximum speed for a sound source. It would be possible to study the problem under the constraint $v_M \leq \sqrt{\dot{s_x}^2 + \dot{s_y}^2 + \dot{s_z}^2}$ instead using Karush-Kuhn-Tucker (KKT) conditions but this gives no additional information since eq. B.3 shows as stated earlier that $|\dot{\delta_n}|$ increases with $|\dot{\vec{s_n}}|$. Additional spatial constraints can be used though:

$$x_{min} < s_x < x_{Max} \ ; \ y_{min} < s_y < y_{max} \ ; \ z_{min} < s_z < z_{max} \tag{B.9}$$

and a global maximum of $|\dot{\delta_n}|$ could be calculated in this context but such value is of no use since the point is to restrict as much as possible with the available information, i.e., last known position, maximum source speed and possibly last known speed, the location of the source in the future.

Since $\dot{\delta_n}$ is a function in 6 dimensions (3 of space and 3 of speed) the approach is to determine the speed that maximizes the function magnitude at each position in space thus mapping the maximum delay speed to each location in space.

As such the lagrangean system to solve is the following:

$$g = \sqrt{\dot{s_x}^2 + \dot{s_y}^2 + \dot{s_z}^2} - v_M \tag{B.10}$$

$$\begin{aligned} \mathcal{L} &= \dot{\delta_n}(\dot{\vec{s_n}}) + \lambda g \\ &= \frac{s_x \dot{s_x} + s_y \dot{s_y} + s_z \dot{s_z}}{c} \gamma_- - \frac{1}{c}\frac{\Delta}{2}\gamma_+ \dot{s_x} + \lambda \sqrt{\dot{s_x}^2 + \dot{s_y}^2 + \dot{s_z}^2} \end{aligned} \tag{B.11}$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \dot{s_x}} &= 0 \\ \frac{\partial \mathcal{L}}{\partial \dot{s_y}} &= 0 \\ \frac{\partial \mathcal{L}}{\partial \dot{s_z}} &= 0 \\ g &= 0 \end{cases} \tag{B.12}$$

Obviously, as defined before,

$$\sqrt{\dot{s_x}^2 + \dot{s_y}^2 + \dot{s_z}^2} = v_M \tag{B.13}$$

whilst s_y and s_z behave symmetrically

[1] One looks for a maximum only since the quantity is symmetric to reversal of velocity.

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \dot{s_x}} = 0 \\ \frac{\partial \mathcal{L}}{\partial \dot{s_y}} = 0 \\ \frac{\partial \mathcal{L}}{\partial \dot{s_z}} = 0 \end{cases} \Leftrightarrow \begin{cases} \frac{s_x}{c}\gamma_- + \lambda\frac{\dot{s_x}}{v_M} - \frac{1}{c}\frac{\Delta}{2}\gamma_+ = 0 \\ \frac{s_y}{c}\gamma_- + \lambda\frac{\dot{s_y}}{v_M} = 0 \\ \frac{s_z}{c}\gamma_- + \lambda\frac{\dot{s_z}}{v_M} = 0 \end{cases} \tag{B.14}$$

and by simple elimination, λ can be found

$$\begin{aligned} \frac{s_y + s_z}{c}\gamma_- + \lambda\frac{\dot{s_y} + \dot{s_z}}{v_M} &= 0 \\ \Leftrightarrow \lambda &= -\frac{s_y + s_z}{\dot{s_y} + \dot{s_z}}\frac{v_M}{c}\gamma_- \end{aligned} \tag{B.15}$$

which by replacing back gives:

$$\begin{aligned} \dot{s_y} = -s_y\frac{v_M}{c}\frac{1}{\lambda}\gamma_- &= s_y\frac{v_M}{c}\Big(\frac{\dot{s_y} + \dot{s_z}}{s_y + s_z}\frac{c}{v_M}{\gamma_-}^{-1}\Big)\gamma_- = s_y\Big(\frac{\dot{s_y} + \dot{s_z}}{s_y + s_z}\Big) \\ \Leftrightarrow \dot{s_y}(s_y + s_z) &= s_y(\dot{s_y} + \dot{s_z}) \\ \Leftrightarrow \dot{s_y} &= \frac{s_y}{s_z}\dot{s_z} \end{aligned} \tag{B.16}$$

and back-substituting in λ yields

$$\lambda = -\frac{s_z}{s_z}\frac{s_y + s_z}{(s_y/s_z + 1)\dot{s_z}}\frac{v_M}{c}\gamma_- = -\frac{s_z}{\dot{s_z}}\frac{v_M}{c}\gamma_- \tag{B.17}$$

Insertion of λ in the equation of the x coordinate then gives

$$\begin{aligned} &\frac{s_x}{c}\gamma_- + \left(-\frac{s_z}{\dot{s_z}}\frac{v_M}{c}\gamma_-\right)\frac{\dot{s_x}}{v_M} - \frac{1}{c}\frac{\Delta}{2}\gamma_+ = 0 \\ \Leftrightarrow &s_x\gamma_- - \frac{s_z}{\dot{s_z}}\gamma_-\dot{s_x} - \frac{\Delta}{2}\gamma_+ = 0 \\ \Leftrightarrow &\frac{s_z}{\dot{s_z}}\gamma_-\dot{s_x} = s_x\gamma_- - \frac{\Delta}{2}\gamma_+ \\ \Leftrightarrow &\dot{s_x} = \frac{\dot{s_z}}{s_z}\frac{1}{\gamma_-}\left(s_x\gamma_- - \frac{\Delta}{2}\gamma_+\right) \vee \cancel{\frac{s_z}{\dot{s_z}}\gamma_- = 0} \end{aligned} \tag{B.18}$$

But since the system is being solved for $\dot{\vec{s_n}}$ the second solution is impossible.

A vector is finally obtained:

$$\begin{aligned} \dot{\vec{s_n}} &= \left(\left(\frac{s_x}{s_z} - \frac{\Delta/2}{s_z}\frac{\gamma_+}{\gamma_-}\right)\dot{s_z}\ ,\ \ \frac{s_y}{s_z}\dot{s_z}\ ,\ \ \dot{s_z}\right) \\ &= \left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\ ,\ \ s_y\ ,\ \ s_z\right)\frac{\dot{s_z}}{s_z} \end{aligned} \tag{B.19}$$

but $\dot{s_x}$ is still free so $g = 0$ is used, which results in a normalization of the vector and scaling by v_M:

$$\begin{aligned} v_M &= \frac{\dot{s_z}}{s_z}\sqrt{\left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\right)^2 + s_y^2 + s_z^2} \\ \Leftrightarrow \dot{s_z} &= \frac{s_z v_M}{\sqrt{\left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\right)^2 + s_y^2 + s_z^2}} \end{aligned} \tag{B.20}$$

Thus,

$$\underset{\dot{\vec{s_n}}}{\operatorname{argmax}}\ \dot{\delta_n}(\vec{s_n},\dot{\vec{s_n}}) = \left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\ ,\quad s_y\ ,\quad s_z\right)\overbrace{\frac{v_M}{\sqrt{\left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\right)^2 + s_y^2 + s_z^2}}}^{\Gamma} \tag{B.21}$$

Taking the optimal speed vector for the maximal phase difference speed and reintroducing in $\dot{s_n}(\vec{s_n},\dot{\vec{s_n}})$ maps such function from 6 to 3 variables becoming only a function of the real space variables:

$$\begin{aligned}
\max \dot{\delta_n}(\vec{s_n}) &= \frac{s_x\dot{s_x} + s_y\dot{s_y} + s_z\dot{s_z}}{c}\gamma_- - \frac{1}{c}\frac{\Delta}{2}\gamma_+\dot{s_x} \\
&= \left\{\left(s_x\gamma_- - \frac{\Delta}{2}\gamma_+\right)\left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\right) + \left(s_y^2 + s_z^2\right)\gamma_-\right\}\frac{\Gamma}{c} \\
&= \left\{\left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\right)^2 + s_y^2 + s_z^2\right\}\frac{\Gamma}{c}\gamma_- \\
&= \gamma_-\frac{v_M}{c}\sqrt{\left(s_x - \frac{\Delta}{2}\frac{\gamma_+}{\gamma_-}\right)^2 + \underbrace{s_y^2 + s_z^2}_{R^2}}
\end{aligned} \tag{B.22}$$

To calculate quantities in the Λ space the real space must be mapped onto it, thus, the following maps will be used $(a,\delta) \to (d_{n1}, d_{n2}) \to (x, R)$; $(a,\delta) \to (x, R)$ and having such mappings it is possible to calculate all the relevant quantities in the Λ space. Notice that no mapping from $(a,\delta) \to (x, y, z)$ is made since such information is lost under symmetry of the system, thus it's only possible to recover variables under the appropriate system of coordinates that matches the symmetries of the system such as (x, R).

B.0.1 Mappings $\Lambda \to (d_{n1}, d_{n2}) \to (x, R)$ and $\Lambda \to (x, R)$

As seen before, with the spherical wavefront model one has:

$$\begin{aligned}
\delta &= (d_{n2} - d_{n1})/c \\
a &= \frac{d_{n1}}{d_{n2}}
\end{aligned} \tag{B.23}$$

Doing the mapping $(a,\delta) \to (d_{n1}, d_{n2})$ is a matter of solving the system under the already known restrictions of $a > 0$, $d_{n1} > 0$ and $d_{n2} > 0$:

$$\begin{cases} d_{n1} &= a\, d_{n2} \\ d_{n2} &= c\,\delta + d_{n1} \end{cases} \Leftrightarrow \begin{cases} d_{n1} &= a\frac{c\,\delta}{1-a} \\ d_{n2} &= \frac{c\,\delta}{1-a} \end{cases} \tag{B.24}$$

Thus the mapping $(a,\delta) \to (d_{n1}, d_{n2})$ was found.

Now, starting with the very definition of d_{n1} and d_{n2} the $(d_{n1}, d_{n2}) \to (x, R)$ and $\Lambda \to (x, R)$ mappings are determined:

$$\begin{cases} d_{n1} &= \sqrt{(x+h)^2+R^2} \\ d_{n2} &= \sqrt{(x-h)^2+R^2} \end{cases} \tag{B.25}$$

Squaring both equations and subtracting one by the other yields x:

$$d_{n2}^2 - d_{n1}^2 = x^2 - 2xh + h^2 - x^2 - 2xh - h^2 \Leftrightarrow d_{n2}^2 - d_{n1}^2 = -4xh \Leftrightarrow \tag{B.26}$$

$$x = \frac{d_{n1}^2 - d_{n2}^2}{4h} = \frac{d_{n2}^2(a^2-1)}{4h} = \frac{c^2\delta^2}{(1-a)^2}\frac{(a^2-1)}{4h} \tag{B.27}$$

R^2 comes directly from one of the source-mic distances:

$$\begin{aligned} R^2 &= d_{n2}^2 - (x-h)^2 \\ &= d_{n2}^2 - x^2 + 2xh - h^2 \\ &= d_{n2}^2 - \left[\frac{d_{n2}^2(a^2-1)}{4h}\right]^2 + 2\left[\frac{d_{n2}^2(a^2-1)}{4h}\right]h - h^2 \\ &= d_{n2}^2\left(1 + \frac{a^2-1}{2}\right) - \left[\frac{d_{n2}^2(a^2-1)}{4h}\right]^2 - h^2 \\ &= -d_{n2}^4\left[\frac{(a^2-1)}{4h}\right]^2 + d_{n2}^2\frac{a^2+1}{2} - h^2 \\ &= -\left(\frac{c\,\delta}{1-a}\right)^4\left[\frac{(a^2-1)}{4h}\right]^2 + \left(\frac{c\,\delta}{1-a}\right)^2\frac{a^2+1}{2} - h^2 \end{aligned} \tag{B.28}$$

The expression can be simplified considering that it seems close to the form $-(\alpha-\beta)^2$. In such case $\beta = h$ and $2\alpha\beta = d_{n2}^2\frac{a^2+1}{2} \Leftrightarrow \alpha = d_{n2}^2\frac{a^2+1}{4h}$. Considering that the α term doesn't match the square in the formula it must again be completed ($-(a^2+1)^2 + y = -(a^2-1)^2 \Leftrightarrow y = 4a^2$), which multiplied by $d_{n2}^4/(4h)^2$ gives

$$\begin{aligned} R^2 &= \frac{a^2 d_{n2}^4}{4h^2} - \left(d_{n2}^2\frac{a^2+1}{4h} - h\right)^2 \\ &= \frac{a^2}{4h^2}\left(\frac{c\,\delta}{1-a}\right)^4 - \left(\left(\frac{c\,\delta}{1-a}\right)^2\frac{a^2+1}{4h} - h\right)^2 \end{aligned} \tag{B.29}$$

γ_- and γ_+ can also be expressed in Λ-space:

$$\begin{aligned} \gamma_- &= \frac{1}{d_{n2}} - \frac{1}{d_{n1}} = \frac{1}{d_{n2}} - \frac{1}{a\,d_{n2}} = \frac{a-1}{a\,d_{n2}} = \frac{a-1}{a}\,\frac{1-a}{c\,\delta} = -\frac{(a-1)^2}{a\,c\,\delta} \\ \gamma_+ &= \frac{1}{d_{n2}} + \frac{1}{d_{n1}} = \frac{a+1}{a\,d_{n2}} = \frac{a+1}{a}\,\frac{1-a}{c\,\delta} = -\frac{a^2-1}{a\,c\,\delta} = \frac{1-a^2}{a\,c\,\delta} = -\frac{\alpha}{c\,\delta} \end{aligned} \tag{B.30}$$

B.0.2 $|\dot{\delta_M}|$

The maximum value of the magnitude of $\dot{\delta_M}$ is calculated in the Λ space, taking the previous results:

$$\frac{\gamma_+}{\gamma_-} = \frac{{d_{n2}}^{-1} + {d_{n1}}^{-1}}{{d_{n2}}^{-1} - {d_{n1}}^{-1}} = \frac{d_{n1}+d_{n2}}{d_{n1}-d_{n2}} = \frac{a+1}{a-1} \tag{B.31}$$

$$\max\ \dot{\delta_n} = \gamma_- \frac{v_M}{c} \sqrt{\left(s_x - h\frac{\gamma_+}{\gamma_-}\right)^2 + R^2} \tag{B.32}$$

$$R^2 = d_{n1}^2 - (x+h)^2 \tag{B.33}$$

$$\begin{aligned}
\max \dot{\delta_n} &= \gamma_- \frac{v_M}{c} \sqrt{\left(s_x - h\frac{a+1}{a-1}\right)^2 + d_{n1}^2 - (x+h)^2} \\
&= \gamma_- \frac{v_M}{c} \sqrt{\cancel{x^2} - 2hx\frac{a+1}{a-1} + h^2\left(\frac{a+1}{a-1}\right)^2 + a^2 d_{n2}^2 - \cancel{x^2} - 2hx - h^2} \\
&= \gamma_- \frac{v_M}{c} \sqrt{-2h\left[\frac{d_{n2}^2(a^2-1)}{4h}\right]\left(\frac{a+1}{a-1}+1\right) + a^2 d_{n2}^2 + h^2\left(\left(\frac{a+1}{a-1}\right)^2 - 1\right)} \\
&= \gamma_- \frac{v_M}{c} \sqrt{-\cancel{2}\left[\frac{d_{n2}^2(a^2-1)}{\cancel{4}}\right]\left(\frac{\cancel{2}a}{a-1}\right) + a^2 d_{n2}^2 + h^2\left(\frac{(\cancel{a^2} + 2a + \cancel{1}) - (\cancel{a^2} - 2a + \cancel{1})}{(a-1)^2}\right)} \\
&= \gamma_- \frac{v_M}{c} \sqrt{d_{n2}^2\left[(1-a^2)\frac{a}{a-1} + a^2\right] + 4h^2\frac{a}{(a-1)^2}} \\
&= \gamma_- \frac{v_M}{c} \sqrt{d_{n2}^2\Big[(1-a^2)a(a-1) + a^2(a-1)^2\Big]\frac{1}{(a-1)^2} + 4h^2\frac{a}{(a-1)^2}} \\
&= \gamma_- \frac{v_M}{c} \frac{1}{|a-1|} \sqrt{d_{n2}^2\Big[a^2 - a - \cancel{a^4} + a^3 + \cancel{a^4} - 2a^3 + a^2\Big] + 4h^2 a} \\
&= \gamma_- \frac{v_M}{c} \frac{1}{|a-1|} \sqrt{d_{n2}^2\Big[-a^3 + 2a^2 - a\Big]\frac{a}{a} + 4h^2 a} \\
&= \gamma_- \frac{v_M}{c} \frac{\sqrt{a}}{|a-1|} \sqrt{d_{n2}^2\Big[-a^2 + 2a - 1\Big] + 4h^2} \\
&= \gamma_- \frac{v_M}{c} \frac{\sqrt{a}}{|a-1|} \sqrt{4h^2 - (a-1)^2\, d_{n2}^2}
\end{aligned} \tag{B.34}$$

Using the $\Lambda \to (d_{n1}, d_{n2})$ mapping B.24, expanding d_{n2} and using γ_- in Λ-space B.30 further simplifies into

$$\begin{aligned}
\max \dot{\delta_n} &= \gamma_- \frac{v_M}{c} \frac{\sqrt{a}}{|a-1|} \sqrt{4h^2 - (a-1)^2 \frac{c^2\delta^2}{(1-a)^2}} \\
&= \gamma_- \frac{v_M}{c} \frac{\sqrt{a}}{|a-1|} \sqrt{\Delta^2 - c^2\delta^2} \\
&= -\frac{(a-1)^2}{a\, c\, \delta} \frac{v_M}{c} \frac{\sqrt{a}}{|a-1|} \sqrt{\Delta^2 - c^2\delta^2} \\
&= -\frac{v_M}{c} \frac{|a-1|}{c\,\delta} \sqrt{\frac{\Delta^2 - c^2\delta^2}{a}}
\end{aligned} \tag{B.35}$$

Due to the symmetric behaviour regarding speed, the desired quantity is actually $|max\dot{\delta_n}|$:

$$|\max \dot{\delta_n}| = \frac{v_M}{c} \overbrace{\left|\frac{1-a}{c\,\delta}\right|}^{|d_{n2}^{-1}|} \sqrt{\frac{\Delta^2 - c^2\delta^2}{a}} \tag{B.36}$$

The singularity points are the microphones midplane ($x = 0$) and $(x, R) = (\pm h, 0)$ which is physically impossible because the microphones and the sources cannot be in the same place at the same time.

Since $d_{n1} = d_{n2}$ in the microphone midplane, $a = 1$ and $\delta = 0$ thus determining the original source position is impossible and consequently it is impossible to estimate the maximum speed in that region due to the singularity.

Since physically $|\delta| < \Delta/c$, this solution is guaranteed to be real and well behaved outside the midplane.

Over the microphone axis $\delta = \pm\Delta/c$ since the source cannot be placed between the microphones so there is no $\dot{\delta_n}$. Otherwise, as expected, the maximum rate of change of δ_n increases as the source is closer to the microphone pair (the same absolute path, if carried closer to the array has a bigger relative motion in relation to the array).

In figure B.1, the plot of the maximum speed on the Λ space, given a constant speed of $v_M = 1\,m/s = 3.6\,km/h$ and $\Delta = 2\,cm$ and relative δ motion during a $64\,ms$ time-span of a typical window, $|\max \dot{\delta_n}|(x, R)/\delta_n(x, R) * 0.064$, are shown for the position (x, R):

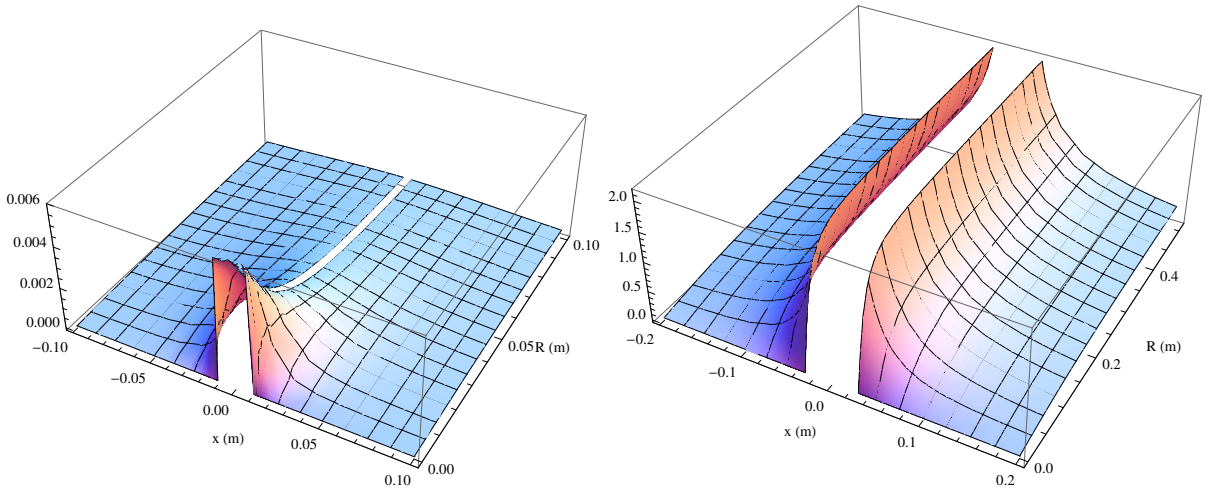


Figure B.1: Plot of $|\max \dot{\delta_n}|(x, R)$ on the left and maximum relative δ motion on the right during a $64\,ms$ interval, with $v_M = 1m/s$, $c = 340m/s$ and $\Delta = 2\,cm$.

This makes tracking feasible only for big azimuths, as the speed of the source in the Λ space quickly grows close to the bissector plane of the microphones. Additionally, the relative δ motion is only inferior to 100% for $|s_x| > \Delta$ during a typical $64\,ms$ analysis window.