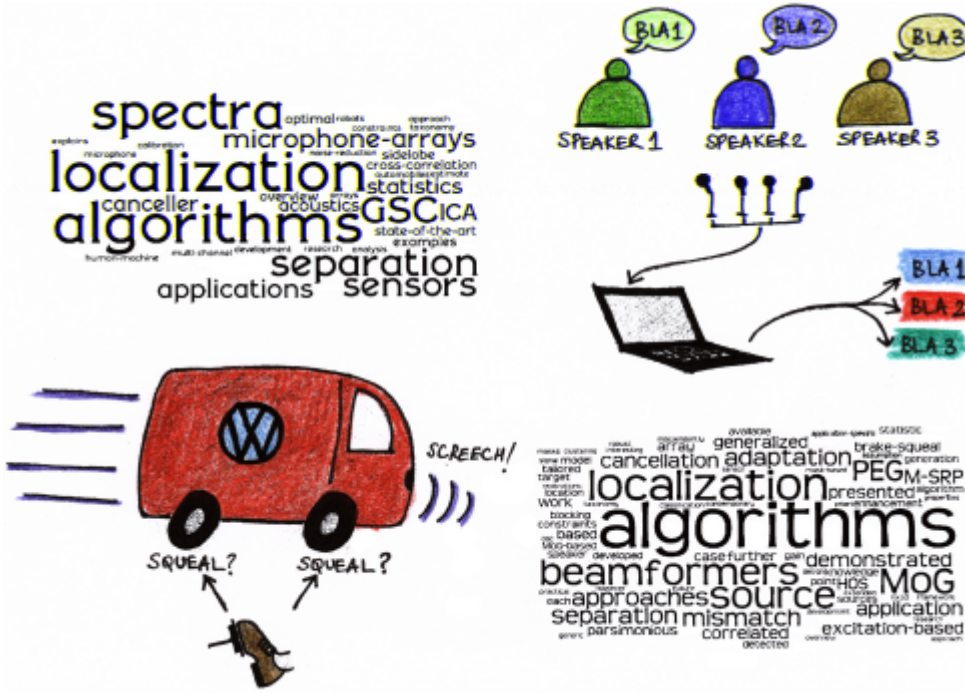
Nilesh Madhu

Acoustic Source Localization:

Algorithms, Applications and Extensions to Source Separation



DER ANDERE VERLAG

Nilesh Madhu

Acoustic Source Localization: Algorithms, Applications and Extensions to Source Separation

DER ANDERE VERLAG 2010

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Gedruckt auf säurefreiem, holzfreiem, chlorfrei (TCF) hergestelltem, unbegrenzt alterungsbeständigem Papier nach ANSI-Z 3948 und DIN/ISO 9706 entsprechend der Forderung des Deutschen Bibliotheksinstituts.

Cover illustrations: Aparna Srinivasan
Word clouds: http://www.wordle.net
Top left: From the introduction
Bottom right: From the conclusions

Lektorat: Der Andere Verlag, Bergweg 1, 25832 Tönning
Tel. (04861) 610 514, Fax (04861) 610 859
E-Mail: talkto@der-andere-verlag.de
Internet: http://www.der-andere-verlag.de
ISBN: 978-3-89959-969-5

Acoustic Source Localization: Algorithms, Applications and Extensions to Source Separation

DISSERTATION
zur Erlangung des Grades
eines Doktor-Ingenieurs
der
Fakultät für Elektrotechnik und
Informationstechnik
an der Ruhr-Universität Bochum

von
Nilesh Madhu

Bochum, 2009

Tag der Promotion	15 Mai 2009
Gutachter der Dissertation	Prof. Dr.-Ing. R. Martin
	Prof. Dr.-Ing. B. Yang
Weitere Prüfer	Prof. Dr.-Ing. habil. H. D. Fischer (Vorsitzender)
	Prof. Dr.-Ing. H. Hudde
	Prof. Dr.-Ing. V. Staudt

ॐ गणेशाय नमः

...तद्विज्ञाय । पुन-रेव वरुणं
पितर-मुपससार ।
अधीहि भगवो ब्रह्मेति ।
तगं होवाच ।
तपसा ब्रह्म विजिज्ञासस्व । तपो ब्रह्मेति ।
स तपोऽतप्यत । स तप-स्तप्त्वा ॥ ...

...Having known that, verily did he (Bhrigu)
approach Varuna, his father, saying
"Venerable sir, teach me Brahman."
To him, Varuna said:
"By Tapas seek you to know Brahman. Tapas
is Brahman".
He then resorted to Tapas, and having
practised Tapas...

– Extract from the Bhriguvalli,
Taittriya Upanishad.

Naan indha aayvuraiyai vetrigaramaaga mudippadharkku
yengal kudumba sumaigalukku edaiyae,
muzhu moochudan yenakku uruthunaiyaaga erundhu
udhvegamum alitha
yenadhu thunaivi Aparna–virkku
naan yen vaazhnaal muzhuvadhum nandri kadan pattirukkiren

Acknowledgements

The following work is the culmination of an enjoyable stay at the Institute of Communication Acoustics, Ruhr-Universität Bochum. My heartfelt thanks to my *Doktorvater* Prof. Dr.-Ing. Rainer Martin for his support, patience and understanding throughout. Your profound depth of knowledge, attention to detail, masterly management style and wonderful sense of humour have always been a source of inspiration to me.

My thanks also go to Prof. Dr.-Ing. Bin Yang for readily agreeing to be the second examiner of my thesis. The quality of my thesis has been significantly improved by your amazingly painstaking corrections. Amazing, because you did it in a very short span of time.

Dr. Rehn: thank you for your constant support and encouragement, and for allowing us the time and the freedom to exercise our research capabilities to the limit. Working on the VW project gave me the much needed field-experience and confidence to handle research on my own.

Ivan: you've been a wonderful mentor and a friend. The 3 months at Microsoft Research were made all the more richer by our daily interaction and for that I am thankful.

To the other members of the IKA too, I owe a debt of gratitude. *Colin:* thanks for introducing me to the intricacies of speech signal processing and for always being available to bounce ideas. *Dirk:* I've presumed a lot on your patient ear and your capabilities as a handyman. I have hugely enjoyed sharing the office with you. Your sunny disposition and your *Fencheltee* made starting work in the morning more tolerable. That, and the fact that I was invited by *Who's Who* and you weren't (until much later). *Christian* and *Andreas:* I've enjoyed our cooking adventures on weekends and miss them terribly. *Christian:* you are right, Linux is a fun OS. *Edith, Renate:* IKA wouldn't be what it is without you around. Thanks for your support – administrative and moral! *Herbert:* thanks for being my Table-Tennis partner. I've enjoyed our afternoon sessions. *To the workshop:* my thanks for the splendid hardware support. *Peter: vielen Dank* for accompanying me on my trips to Wolfsburg. Your presence has always been very reassuring and has soothed many a jangled nerve. Thanks also for sharing your wide experience in the field of audio measurements. There is still so much I have to learn...

To my ex-students and (still) friends *Timo, André, David, Sebastian (M):* thanks for allowing me to learn and grow with you. It has been a rewarding experience. To *Sebastian (G):* we have only worked together for a short while, but your support has been immense.

To the *DAAD:* thanks for bringing me to this beautiful country – Germany! I came, I saw, I stayed!

To *Advaith:* your antics were delightfully refreshing and just what I needed after a long day at work, especially during the last phase of my thesis.

Lastly, to my parents, brother, in-laws and *thatha-pattis*: many thanks for your prayers and support. I cannot even begin to repay you. What I am today, I owe to you.

List of abbreviations

AED	Adaptive eigenvalue decomposition
AIC	Akaike's information criterion
BIC	Bayesian information criterion
DFT	Discrete Fourier transform
DML	Deterministic maximum likelihood
DOA	Direction of arrival
DSB	Delay-and-sum beamformer
EM	Expectation maximization
FIR	Finite impulse response
FTDS	Fourier transform of discrete signals
GCC	Generalized cross-correlation
GSC	Generalized sidelobe canceller
HOS	Higher order statistics
ICA	Independent component analysis
i.i.d	Independently and identically distributed
LCMP	Linearly constrained minimum power
LCMV	Linearly constrained minimum variance
LMS	Least mean square
MLE	Maximum likelihood estimation
(M)MSE	(Minimum) Mean square error
MoD	Mixture of distributions
MoG	Mixture of Gaussians
MUSIC	Multiple signal classification
MVDR	Minimum variance distortionless response
MWF	Multi-channel Wiener filter
pdf	Probability density function
PEG	Parsimonious excitation-based GSC
PHAT	Phase transform
pmf	Probability mass function
RMSE	Root mean square error
SCORE	Spectral correlation
SINR	Signal to interference + noise ratio
SIR	Signal to interference ratio
SML	Stochastic maximum likelihood
SNR	Signal to noise ratio
SOS	Second order statistics
(M-)SRP	(Multi-source) Steered response power
STFT	Short time Fourier transform
TDE	Time delay estimation
TDOA	Time delay of arrival
TTL	Time to live

Glossary

b	Frame index of the STFT
c	Speed of sound in the medium considered (air)
k, k'	Discrete frequency bin index of the DFT
m, m'	Microphone index
n, l	Discrete time index
$p(a)$	Probability density function of random variable a
q, q'	Source index
$\mathbf{r}_s$	Source location (Euclidean plane)
s, S	Source signal (time, frequency domain)
v, V	Background noise (time, frequency domain)
x, X	Microphone signal (time, frequency domain)
$\mathbf{x}$	Representation of a vector quantity (usually discrete time domain)
$\mathbf{A}$	Propagation matrix (frequency domain)
B	Total number of frames in an observed time segment
$\mathbf{B}$	Spatial blocking matrix (GSC)
$\mathcal{D}_{a,b}\{\cdot\}$	Selection operator, selecting the first a rows and b columns of the matrix operand
$\mathfrak{E}$	Residual error (MMSE calibration approach)
$\mathbf{D}$	Scaling matrix (ICA)
$\mathcal{H}$	Hypothesis
$\mathbf{H}^\dagger$	Moore–Penrose pseudoinverse of matrix $\mathbf{H}$
$\mathcal{I}$	Order of the MoG model
$\mathcal{J}$	Cost function
$\mathbf{J}$	Permutation matrix (ICA)
$\mathcal{L}$	Likelihood function
L	Number of taps in discrete-time
M	Total number of microphones in an array
$\mathcal{N}(\mu, \sigma^2)$	Univariate Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{P}$	Number of free parameters for estimation
P_a	Probability of occurrence of event a
$\mathbf{P}$	Projector onto a space
$\mathbf{P}^\perp$	Projector onto the orthogonal complement ($\mathbf{I} - \mathbf{P}$)
Q	Total number of sources present
$\mathcal{R}$	Region of occupancy of a source
T_{60}	Reverberation time
$\mathcal{W}$	Temporal or spatial window function
$\overset{\vee}{\mathbf{W}}$	Orthonormal matrix (ICA)
$\overset{\circ}{\mathbf{W}}$	Whitening matrix (ICA)
$\mathbf{W}_V$	Noise canceller (GSC)
$\mathbf{X}$	Representation of a matrix quantity in time and frequency domain (additionally, vector quantity in frequency domain)
$\mathcal{Z}$	Hit percentage or instances of an event
$\overset{\circ}{Z}$	Whitened variable in frequency domain (ICA)
η	Smoothing parameter for first-order recursive averaging

Continued...

ς	Step-size for adaptive algorithms
τ	TDOA
ω	Frequency variable (continuous-time Fourier transform)
Υ	Threshold for decision making
θ	Direction of arrival (azimuth)
$\boldsymbol{\Theta}$	Parameter vector
Φ_{xy}	Correlation between variables x and y
$\boldsymbol{\Phi}_{\mathbf{xy}}$	Correlation matrix for vector valued variables $\mathbf{x}$ and $\mathbf{y}$
Ψ_{YY}	Power spectral density of variable Y
$\boldsymbol{\Psi}_{\mathbf{YY}}$	Power spectral density matrix of vector valued variable $\mathbf{Y}$
Ω	Frequency variable (Fourier transform of discrete signals)
Ω_k	k th discrete frequency (DFT)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Microphone or sensor array technology has established itself as a valued tool in many acoustics-based applications utilizing spatial diversity. Examples of such applications are already visible in our day-to-day life, such as the use of microphone arrays for improving human-machine interfaces, video-conferencing, noise-reduction in hands-free systems in automobiles, etc. Additionally, one finds the use of microphone arrays in several non-speech applications such as remote surveillance, fault analysis of machinery, automotive acoustics, and autonomous robots.

In most cases, localization of the acoustic sources is the *primary* step, and serves as a fundamental building block for further signal processing or decision making. It comes as no surprise therefore that this topic has been the subject of significant research activity for a long time and still enjoys considerable interest in the signal processing community. This work is devoted to source localization, both as an end in itself and as the means to an end.

We start, in Chapter 3, by introducing a localization taxonomy and provide an exhaustive overview of the contemporary state-of-the-art algorithms for source localization, drawn from different application areas. We then proceed to derive the relations between these approaches and demonstrate their ultimate dependence on only the multi-channel cross-correlation (second order statistics) – under which umbrella all algorithms are *unified*. We maintain that the differences between the various algorithms lie in the signal *prefiltering* before the computation of the cross-correlation, or on the assumptions made regarding the signals or environment, which explains why some algorithms are more suited to a specific application than others.

In Chapter 4 we illustrate the latter point by developing localization algorithms for two different applications with varying amounts of *a priori* knowledge on the source and environment model. The applications considered are a) the localization of brake squeal in a travelling automobile and b) multiple speaker localization. We shall see that while the statistic utilized in both cases is still the multi-channel cross-correlation, one algorithm cannot be used *in lieu* of the other. We conclude here that when designing a localization algorithm for a specific application, one would do well to consider all the application specific characteristics and constraints available. Simply implementing the generic algorithm for that *class* of problems will not be optimal.

In the second part of the work (Chapters 5–6), we shall consider the use of localization for source separation. Here the specific application considered is that of speaker separation and enhancement in a noisy environment with competing speakers – as exemplified by the cocktail-party problem. We begin with a taxonomy for source separation and briefly discuss the state-of-the-art algorithms. Then we show the influence of source localization in the design of separation algorithms. For this we consider first a simple 2–source 2–microphone

case, where we use independent component analysis (ICA) for separation, and show how using a localization estimate leads to an algorithm with reduced computational complexity. We then proceed to the more general case of Q sources and M microphones and demonstrate how the speaker localization algorithm previously developed contains sufficient information for source separation. This approach is versatile in that it allows for the implementation of a variety of source separation algorithms and can further function in the underdetermined or noisy case – where many state-of-the-art algorithms, including ICA, would fail. In theory, the generalized separation approach we develop does not require the constraint of constant multi-talker activity (required for ICA algorithms) or detectors of target-speaker activity (required for many conventional approaches based e.g., on the generalized sidelobe canceller (GSC)).

Note that a good localization estimate or separation performance presupposes a fully functional sensor array. Correspondingly, the third part of this work is dedicated to algorithms for determining the health of the array, especially when it is used in hostile surroundings, with the goal of performing *in situ* self-tests and removing, if required, degraded sensors from further processing. (The terms channels, microphones and sensors will be used synonymously throughout the text.) The two closely related algorithms presented here make use of the fact that sensors of a compact array should record similar long-term power spectra of the incident signals. Any deviation therefrom is used as an indicator of sensor degradation. In addition to detecting sensor degradation, the algorithms may also be used to perform online gain calibration of the healthy sensors of an array. Such a calibration is useful, among other things, for reducing the effects of sensor-mismatches in practical applications.

The work concludes with some directions for further research.

Essentially, all models are wrong, but some are useful.
–George Box

Chapter 2

Signal Model

Consider an array consisting of M microphones at positions $\mathbf{r}_m$ capturing the signal emitted from a source at position $\mathbf{r}_s$. Figure 2.1 illustrates the general situation in two spatial dimensions (spanned by unit vectors $\mathbf{e}_x$ and $\mathbf{e}_y$). These signals recorded by the microphones may then be expressed in the continuous or the discrete time domain, considering the contribution of all the paths from the source to the individual microphones of the array.

As a point of note, the signals considered in this work are modelled as realizations of stochastic processes with an underlying probability density function, *a priori* information about which may or may not be available. Thus, when we refer colloquially to signals having a specified distribution, or to signals being uncorrelated or statistically independent, we mean the underlying stochastic processes in the strict sense.

2.1 Continuous time model

The signal $x_m(t)$ received at the mth microphone of the array, located at $\mathbf{r}_m = (x_m, y_m, z_m)^T$, due to a source located at $\mathbf{r}_s = (x_s, y_s, z_s)^T$ may be written, in continuous time t, as

$$x_m(t) = a_m(t) * s_0(t) + v_m(t)\,, \tag{2.1}$$

where $s_0(t)$ represents the source waveform, $a_m(t)$ represents the room impulse response from the source position to the microphone m, the $*$ represents the convolution operator, and $v_m(t)$ represents the noise at the microphone. This may be extended to the general case of Q sources as

$$x_m(t) = \sum_{q=1}^{Q} a_{mq}(t) * s_{0q}(t) + v_m(t)\,, \tag{2.2}$$

where $a_{mq}(t)$ now represents the room impulse response from the qth source to the mth microphone.

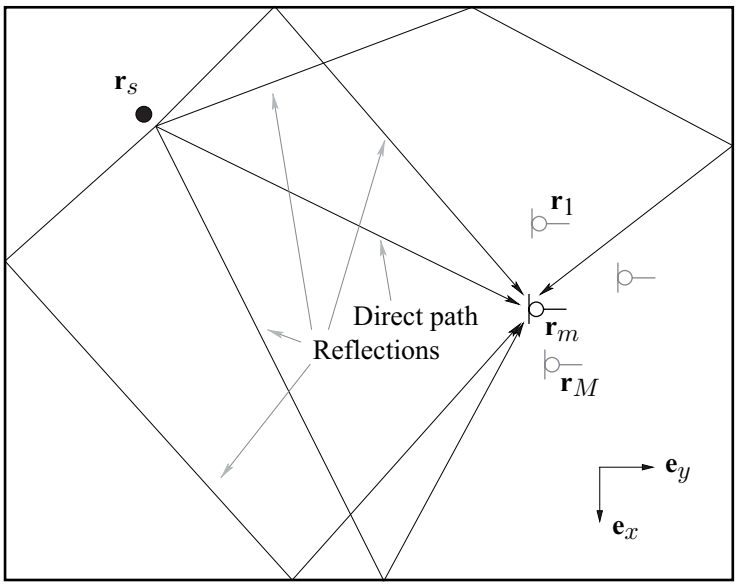


Figure 2.1: Acoustic signal paths for a particular microphone in the x–y plane. $\mathbf{r}_s$ and $\mathbf{r}_m$ denote the locations of the source and the mth microphone respectively. Each path from the source to the microphone m may be represented by a (frequency-dependent) gain and delay of the source signal. The direct path possesses the least delay. The sum of all the paths constitutes the impulse response of the room for the particular source and microphone position.

2.2 Discrete time representation

Since we shall mostly deal with the digital representations of the microphone and source signals, the concept of continuous time serves only to clarify some basic ideas. Consequently, relations (2.1) and (2.2) will now be extended to the discrete time case. To simplify the discussion, we approximate the room impulse responses by finite impulse response (FIR) filters of order $L-1$. For the single source case, we now have an impulse response vector

$$\mathbf{a}_m = \big(a_m(0)\,, a_m(1)\,,\, \ldots\,,\, a_m(L-1)\,\big)^T \tag{2.3}$$

and we may write the signal at microphone m as

$$x_m(n) = \mathbf{a}_m^T \mathbf{s}_0(n) + v_m(n)\,, \tag{2.4}$$

where n is the discrete time index, $\mathbf{s}_0(n) = \big(s_0(n)\,, s_0(n-1)\,,\, \ldots\,,\, s_0(n-L+1)\,\big)^T$ is the sampled source signal, and $v_m(n)$ is the sampled noise signal. For the multi-source scenario we define the impulse response vectors from source q to microphone m as

$$\mathbf{a}_{mq} = \big(a_{mq}(0)\,, a_{mq}(1)\,, \ldots\,, a_{mq}(L-1)\,\big)^T \tag{2.5}$$

and obtain

$$\begin{pmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{pmatrix} = \begin{pmatrix} \mathbf{a}_{11}^T & \cdots & \mathbf{a}_{1Q}^T \\ \vdots & \ddots & \vdots \\ \mathbf{a}_{M1}^T & \cdots & \mathbf{a}_{MQ}^T \end{pmatrix} \begin{pmatrix} \mathbf{s}_{01}(n) \\ \vdots \\ \mathbf{s}_{0Q}(n) \end{pmatrix} + \mathbf{v}(n)\,, \tag{2.6}$$

where $\mathbf{v}(n) = \big(v_1(n)\,, v_2(n)\,, \ldots\,, v_M(n)\,\big)^T$ is the vector of noise signals.

2.3 Frequency domain formulation

Equations (2.4) and (2.6) may also be formulated in the frequency domain using the Fourier transform of discrete signals (FTDS) [91], [122, Chap. 3]. Provided that the Fourier transforms of all signals under consideration exist, we obtain the frequency domain equivalent of (2.4) as

$$X_m(\Omega) = A_m(\Omega)\, S_0(\Omega) + V_m(\Omega)\,, \tag{2.7}$$

where $\Omega = 2\pi f/f_\mathrm{s}$ denotes the normalized frequency variable and f_s is the sampling rate. In the multiple source case we obtain, with $\mathbf{X}(\Omega) = (X_1(\Omega)\,, X_2(\Omega)\,, \ldots, X_M(\Omega))^T$ and $\mathbf{V}(\Omega) = (V_1(\Omega)\,, V_2(\Omega)\,, \ldots, V_M(\Omega))^T$,

$$\begin{aligned}\mathbf{X}(\Omega) &= \begin{pmatrix} A_{11}(\Omega) & \cdots & A_{1Q}(\Omega) \\ \vdots & \ddots & \vdots \\ A_{M1}(\Omega) & \cdots & A_{MQ}(\Omega) \end{pmatrix} \begin{pmatrix} S_{01}(\Omega) \\ \vdots \\ S_{0Q}(\Omega) \end{pmatrix} + \begin{pmatrix} V_1(\Omega) \\ \vdots \\ V_M(\Omega) \end{pmatrix} \\ &= \mathbf{A}(\Omega)\, \mathbf{S}_0(\Omega) + \mathbf{V}(\Omega)\ .\end{aligned} \tag{2.8}$$

2.4 Simplified model for localization

The localization algorithms considered in the present work assume dominance of the direct path. Consequently, each $A_{mq}(\Omega)$ may be written as

$$A_{mq}(\Omega) = |A'_{mq}(\Omega)\,| e^{-\jmath\,\Omega f_\mathrm{s} \tau_{mq}(\mathbf{r}_q)} + A''_{mq}(\Omega)\ , \tag{2.9}$$

where $|A'_{mq}| (\gg |A''_{mq}|)$, represents the gain along the direct path and $A''_{mq} \in \mathbb{C}$ indicates the net gain and phase smearing caused by the reflections along the indirect paths. $\tau_{mq}(\mathbf{r}_q)$ represents the *absolute* time delay of the signal from source q to the microphone m along the direct path. Then, (2.8) takes the form

$$\begin{aligned}\mathbf{X}(\Omega) &= \begin{pmatrix} |A'_{11}(\Omega)\,| e^{-\jmath\,\Omega f_\mathrm{s} \tau_{11}(\mathbf{r}_1)} & \cdots & |A'_{1Q}(\Omega)\,| e^{-\jmath\,\Omega f_\mathrm{s} \tau_{1Q}(\mathbf{r}_Q)} \\ \vdots & \ddots & \vdots \\ |A'_{M1}(\Omega)\,| e^{-\jmath\,\Omega f_\mathrm{s} \tau_{M1}(\mathbf{r}_1)} & \cdots & |A'_{MQ}(\Omega)\,| e^{-\jmath\,\Omega f_\mathrm{s} \tau_{MQ}(\mathbf{r}_Q)} \end{pmatrix} \mathbf{S}_0(\Omega) \\ &\quad + \begin{pmatrix} A''_{11}(\Omega) & \cdots & A''_{1Q}(\Omega) \\ \vdots & \ddots & \vdots \\ A''_{M1}(\Omega) & \cdots & A''_{MQ}(\Omega) \end{pmatrix} \mathbf{S}_0(\Omega) + \mathbf{V}(\Omega) \\ &= \mathbf{A}'(\Omega)\, \mathbf{S}_0(\Omega) + \mathbf{A}''(\Omega)\, \mathbf{S}_0(\Omega) + \mathbf{V}(\Omega)\ .\end{aligned} \tag{2.10}$$

For the subsequent discussion we shall suppress the explicit dependence of the time delays on the source position for the purpose of convenience (i.e., $\tau_{mq} = \tau_{mq}(\mathbf{r}_q)$), and reintroduce the notation where disambiguation may be required. The *propagation matrix* $\mathbf{A}'(\Omega)$ is directly related to the geometric arrangement of the sources and the sensors and thus is key to solving the localization and separation problem. The vectors $\mathbf{A}''(\Omega)\, \mathbf{S}_0(\Omega)$ and $\mathbf{V}(\Omega)$ constitute disturbances. While the former term is obviously correlated with the source signals, the latter is typically modelled as being statistically independent of the source signals. To simplify the discussion we will frequently neglect the contribution $\mathbf{A}''(\Omega)\, \mathbf{S}_0(\Omega)$ of the indirect paths or subsume it into the definition of noise. In this case, we have

$$\mathbf{A}(\Omega) \approx \mathbf{A}'(\Omega)\ .$$

For convenience, we introduce a reference point $\mathbf{r}_0$ that may coincide, for example, with one of the microphone locations. Then, we define the source signal spectra $S_q(\Omega)$ that are received at this reference point when only the direct path is considered as

$$S_q(\Omega) = |A'_{0q}(\Omega)|e^{-\jmath\,\Omega f_\mathrm{s}\tau_{0q}}S_{0q}(\Omega) \tag{2.11}$$

and rewrite (2.10) in terms of the source signal vector

$$\mathbf{S}(\Omega) = \left(S_1(\Omega), \ldots, S_Q(\Omega)\right)^T \tag{2.12}$$

as

$$\begin{aligned}\mathbf{X}(\Omega) = &\begin{pmatrix} \left|\frac{A'_{11}(\Omega)}{A'_{01}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\tau_{01}-\jmath\,\Omega f_\mathrm{s}\tau_{11}} & \cdots & \left|\frac{A'_{1Q}(\Omega)}{A'_{0Q}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\tau_{0Q}-\jmath\,\Omega f_\mathrm{s}\tau_{1Q}} \\ \vdots & \ddots & \vdots \\ \left|\frac{A'_{M1}(\Omega)}{A'_{01}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\tau_{01}-\jmath\,\Omega f_\mathrm{s}\tau_{M1}} & \cdots & \left|\frac{A'_{MQ}(\Omega)}{A'_{0Q}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\tau_{0Q}-\jmath\,\Omega f_\mathrm{s}\tau_{MQ}} \end{pmatrix}\mathbf{S}(\Omega) \\ &+ \begin{pmatrix} \breve{A}_{11}(\Omega) & \cdots & \breve{A}_{1Q}(\Omega) \\ \vdots & \ddots & \vdots \\ \breve{A}_{M1}(\Omega) & \cdots & \breve{A}_{MQ}(\Omega) \end{pmatrix}\mathbf{S}(\Omega) + \mathbf{V}(\Omega)\end{aligned} \tag{2.13}$$

$$\begin{aligned}= &\begin{pmatrix} \left|\frac{A'_{11}(\Omega)}{A'_{01}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{11}} & \cdots & \left|\frac{A'_{1Q}(\Omega)}{A'_{0Q}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{1Q}} \\ \vdots & \ddots & \vdots \\ \left|\frac{A'_{M1}(\Omega)}{A'_{01}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{M1}} & \cdots & \left|\frac{A'_{MQ}(\Omega)}{A'_{0Q}(\Omega)}\right| e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{MQ}} \end{pmatrix}\mathbf{S}(\Omega) \\ &+ \begin{pmatrix} \breve{A}_{11}(\Omega) & \cdots & \breve{A}_{1Q}(\Omega) \\ \vdots & \ddots & \vdots \\ \breve{A}_{M1}(\Omega) & \cdots & \breve{A}_{MQ}(\Omega) \end{pmatrix}\mathbf{S}(\Omega) + \mathbf{V}(\Omega)\,,\end{aligned} \tag{2.14}$$

where

$$\breve{A}_{mq} = \frac{A''_{mq}}{|A'_{0q}|}e^{\jmath\,\Omega f_\mathrm{s}\tau_{0q}}$$

represents the normalized indirect components. $\Delta\tau_{mq} = \tau_{0q} - \tau_{mq}$ is the *relative* time delay or *time delay of arrival* (TDOA) with respect to the reference point. If the reference point is close to the array and the array is well calibrated (i.e., the microphones do not exhibit significantly deviant amplitude and phase characteristics), and both the reference point and array are in the farfield of the sources, we may further simplify (2.13) and obtain [122, Chap. 12]

$$\begin{aligned}\mathbf{X}(\Omega) = &\begin{pmatrix} e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{11}} & \cdots & e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{1Q}} \\ \vdots & \ddots & \vdots \\ e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{M1}} & \cdots & e^{\jmath\,\Omega f_\mathrm{s}\Delta\tau_{MQ}} \end{pmatrix}\mathbf{S}(\Omega) \\ &+ \begin{pmatrix} \breve{A}_{11}(\Omega) & \cdots & \breve{A}_{1Q}(\Omega) \\ \vdots & \ddots & \vdots \\ \breve{A}_{M1}(\Omega) & \cdots & \breve{A}_{MQ}(\Omega) \end{pmatrix}\mathbf{S}(\Omega) + \mathbf{V}(\Omega)\,.\end{aligned} \tag{2.15}$$

The direct path contributions in (2.15) encode the spatial positions of the Q sources in terms of TDOAs. The indirect paths and noise contributions constitute disturbances. When we consider only the qth source, and assume farfield and anechoic conditions, the matrix $\mathbf{A}(\Omega)$ reduces to a vector – the *propagation vector* [122, Chap. 12], which might be written, explicitly parametrized by the source location $\mathbf{r}_{s_q}$ of the qth source and the reference point $\mathbf{r}_0$, as

$$\mathbf{A}(\mathbf{r}_{s_q}, \mathbf{r}_0, \Omega) = \left(e^{\jmath\,\Omega f_s \Delta\tau_{1q}(\mathbf{r}_{s_q}, \mathbf{r}_0)}, \ldots, e^{\jmath\,\Omega f_s \Delta\tau_{Mq}(\mathbf{r}_{s_q}, \mathbf{r}_0)}\right)^T . \tag{2.16}$$

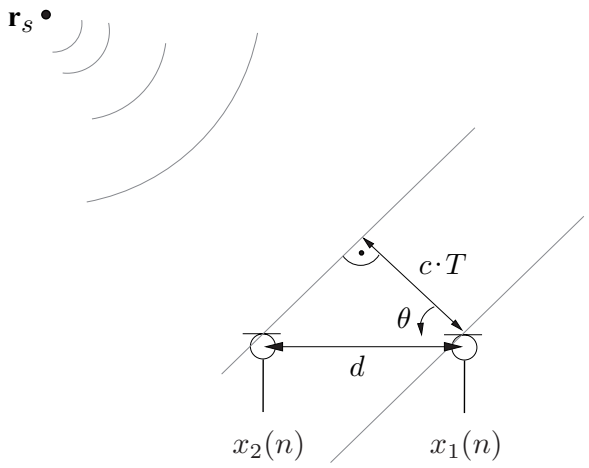


Figure 2.2: TDOA for a microphone pair in the farfield of a source. It can be assumed that the source signal arrives as plane waves at the microphone array.

In the case of a single source we may omit the source index q and obtain

$$\mathbf{A}\left(\mathbf{r}_s, \mathbf{r}_0, \Omega\right) = \left(e^{\jmath\,\Omega f_s \Delta\tau_1(\mathbf{r}_s, \mathbf{r}_0)}, \ldots, e^{\jmath\,\Omega f_s \Delta\tau_M(\mathbf{r}_s, \mathbf{r}_0)}\right)^T . \tag{2.17}$$

The farfield scenario is illustrated in two dimensions in Figure 2.2 for the simple case of a single microphone pair and anechoic, noiseless transmission. In this case the difference of TDOAs allows us to infer the direction of arrival (DOA) (or, equivalently, the angle of incidence) θ from the following relation between θ and the delay difference T:

$$T = \Delta\tau_1 - \Delta\tau_2 = \frac{d\cos(\theta)}{c}\,, \tag{2.18}$$

where d is the microphone distance and c denotes the speed of sound. This relation is evident from the geometric considerations in Figure 2.2.

2.5 Practical realizations

A real-time realization of localization and enhancement algorithms requires the processing of short (windowed) segments of the input signals, and we cannot compute the FTDS, on the basis of which the models in the previous sections have been developed. Therefore, we consider the frame-wise spectral representation of the signal as obtained from the K–point discrete Fourier transform (DFT) on overlapped, windowed segments of the discrete time

domain signal. We shall denote this operation as the short-time Fourier transform (STFT), and the corresponding domain as the STFT domain. The STFT representation $X(k,b)$ of a discrete time signal $x(n)$ is obtained as

$$X(k,b) = \sum_{n=1}^{K} \mathcal{W}(n)\, x(bO+n)\, e^{-\jmath\, 2\pi n \frac{k}{K}}, \quad k = 0,1,\ldots,K \tag{2.19}$$

where k is the discrete frequency bin index, O indicates the frame shift (in samples) between the frames, b is the frame index, and $\mathcal{W}(n)$ is the window function.

Using this instead of the FTDS, we may write (2.10) as:

$$\begin{aligned}\mathbf{X}(k,b) &\approx \begin{pmatrix} |A'_{11}(k)\,|e^{-\jmath\,\Omega_k \tau_{11}} & \cdots & |A'_{1Q}(k)\,|e^{-\jmath\,\Omega_k \tau_{1Q}} \\ \vdots & \ddots & \vdots \\ |A'_{M1}(k)\,|e^{-\jmath\,\Omega_k \tau_{M1}} & \cdots & |A'_{MQ}(k)\,|e^{-\jmath\,\Omega_k \tau_{MQ}} \end{pmatrix} \mathbf{S}_0(k,b) \\ &\quad + \begin{pmatrix} A''_{11}(k) & \cdots & A''_{1Q}(k) \\ \vdots & \ddots & \vdots \\ A''_{M1}(k) & \cdots & A''_{MQ}(k) \end{pmatrix} \mathbf{S}_0(k,b) + \mathbf{V}(k,b) \\ &= \mathbf{A}'(k)\,\mathbf{S}_0(k,b) + \mathbf{A}''(k)\,\mathbf{S}_0(k,b) + \mathbf{V}(k,b)\,.\end{aligned} \tag{2.20}$$

where $\Omega_k = 2\pi k f_\mathrm{s}/K$ is the kth discrete frequency. The approximation is a result of truncating the support of the signals to finite length. Despite this approximation, we shall use this model subsequently to good effect. Under assumptions of direct path dominance, we subsume the reverberation components in (2.20) into the definition of the noise to yield:

$$\begin{aligned}\mathbf{X}(k,b) &\approx \begin{pmatrix} |A'_{11}(k)\,|e^{-\jmath\,\Omega_k \tau_{11}} & \cdots & |A'_{1Q}(k)\,|e^{-\jmath\,\Omega_k \tau_{1Q}} \\ \vdots & \ddots & \vdots \\ |A'_{M1}(k)\,|e^{-\jmath\,\Omega_k \tau_{M1}} & \cdots & |A'_{MQ}(k)\,|e^{-\jmath\,\Omega_k \tau_{MQ}} \end{pmatrix} \mathbf{S}_0(k,b) + \mathbf{V}(k,b) \\ &\triangleq \mathbf{A}(k)\,\mathbf{S}_0(k,b) + \mathbf{V}(k,b)\,.\end{aligned} \tag{2.21}$$

This is the model we shall often use in this work.

I

Source Localization

Discovery consists of seeing what everyone has seen and thinking what nobody has thought.
– Albert Szent-Gyorgyi

Chapter 3

Localization Overview

3.1 Localization approach taxonomy

The multi-channel approaches to acoustic source localization may broadly be divided into two main classes: indirect and direct. Indirect approaches to source localization are usually two-step methods: first, the relative time delays $\Delta\tau_{mq}$ for the various microphone pairs are determined and then the source location is found as the intersection of the corresponding half-hyperboloids centered around the respective microphone pairs. Direct approaches, on the other hand, generally scan the so-called *candidate* source positions and pick the Q most likely candidates – thus performing the localization in a single step.

In the following discussion, we shall start by considering the simple case of two microphones and a single source. Further, without loss of generality, we shall assume the source to be in the farfield, with the source wavefront propagating as plane waves and impinging upon the microphone pair with the corresponding time delay of arrival T (Fig. 2.2). The extension to microphone arrays with more than two microphones is the subject of the later sections. We shall first present the indirect approaches, followed by the direct ones.

3.2 Indirect localization approaches

Indirect approaches explicitly estimate the time delays of arrival (TDOA) before performing the actual localization task. Early approaches [64, 41, 50] to time delay of arrival estimation consider an anechoic, farfield signal model:

$$\begin{aligned} x_1(t) &= s(t+\Delta\tau_1) + v_1(t) \\ x_2(t) &= s(t+\Delta\tau_2) + v_2(t)\ . \end{aligned} \tag{3.1}$$

By means of an LMS-type algorithm the approaches of [41, 112] then explicitly adapt a time delay T to minimize the mean square error (MSE), also termed *cost function*, between the microphone signals, i.e.,

$$\widehat{T} = \underset{T}{\operatorname{argmin}}\ \mathrm{E}\left\{(x_1(t) - x_2(t+T))^2\right\} \tag{3.2}$$

$$= \underset{T}{\operatorname{argmin}}\ \mathrm{E}\left\{x_1^2(t)\right\} + \mathrm{E}\left\{x_2^2(t+T)\right\} - 2\mathrm{E}\left\{x_1(t)\,x_2(t+T)\right\}. \tag{3.3}$$

Note that the first two terms of (3.3) represent the signal power at the two channels and, for stationary input signals, are independent of T. Therefore, they do not contribute to the cost function, simplifying the expression to

$$\widehat{T} = \underset{T}{\operatorname{argmax}}\ \mathrm{E}\left\{x_1(t)\, x_2(t+T)\right\}. \tag{3.4}$$

Thus, *minimizing* the mean square error may be seen to be equivalent to *maximizing* the cross-correlation between the microphone signals.

However, approaches that explicitly search the optimal time delay in the time domain suffer from two drawbacks: first, in discrete time systems, the time delay could contain fractional sample shifts that require some form of interpolation and lead to more complicated estimation procedures [25]; secondly, the approaches are based on an overly simplistic signal model.

Further improvements along this direction lead to LMS-type algorithms [97] where, instead of a delay parameter T, the microphone signal $x_2(t)$ is *filtered* by a filter $h(t)$ such that an approximate solution to

$$h_{\mathrm{opt}}(t) = \underset{h(t)}{\operatorname{argmin}}\ \mathrm{E}\left\{\left(x_1(t) - h(t) * x_2(t)\right)^2\right\} \tag{3.5}$$

is found. When the direct path is dominant, an estimate $\widehat{T}$ of the time delay can be obtained as the abscissa of the largest peak of $h_{\mathrm{opt}}(t)$. Besides MSE, other optimization criteria may be used to compute the optimal filter impulse response $h(t)$. These lead to a family of TDOA estimation algorithms that fit into the general framework presented in Section 3.2.1 below.

In the discrete time implementation of this algorithm, the optimal filter is estimated by the normalized LMS approach (see, e.g., [122, Chap. 13]), the update equations for which may be written as

$$\begin{aligned} e(n) &= x_1(n - L_D) - \mathbf{h}^T(n)\, \mathbf{x}_2(n) \\ \mathbf{h}(n+1) &= \mathbf{h}(n) + \varsigma e(n)\, \frac{\mathbf{x}_2(n)}{\|\mathbf{x}_2(n)\|^2}, \end{aligned} \tag{3.6}$$

where ς is the step size and $L_D > \lceil f_s d/c \rceil$ is an integral sample delay[1] required in order to preserve causality in the case of negative TDOA values. After convergence, the TDOA estimate (in samples) is obtained as the *difference* between the abscissa of the maximum value of $\mathbf{h}_{\mathrm{opt}}$ and L_D. Usually, as the TDOA may consist of fractional sample shifts, some interpolation of $\mathbf{h}_{\mathrm{opt}}$ is done around the peak value (see, e.g., [69]).

3.2.1 Generalized cross-correlation (GCC)

Generalized cross-correlation (GCC) [64] is the term given to the framework that encompasses a wide range of approaches to TDOA estimation. The block diagram of a generalized cross-correlator is shown in Fig. 3.1.

The optimal time delay estimate $\widehat{T}$ is obtained as

$$\begin{aligned} \widehat{T} &= \underset{T}{\operatorname{argmax}}\ \mathrm{E}\left\{\big(x_1(t) * h_1(t)\big)\big(x_2(t+T) * h_2(t)\big)\right\} \\ &= \underset{T}{\operatorname{argmax}}\ \Phi^g_{x_1x_2}(T)\ , \end{aligned} \tag{3.7}$$

[1] The operator $\lceil x \rceil$ rounds x to the nearest upper integer value.

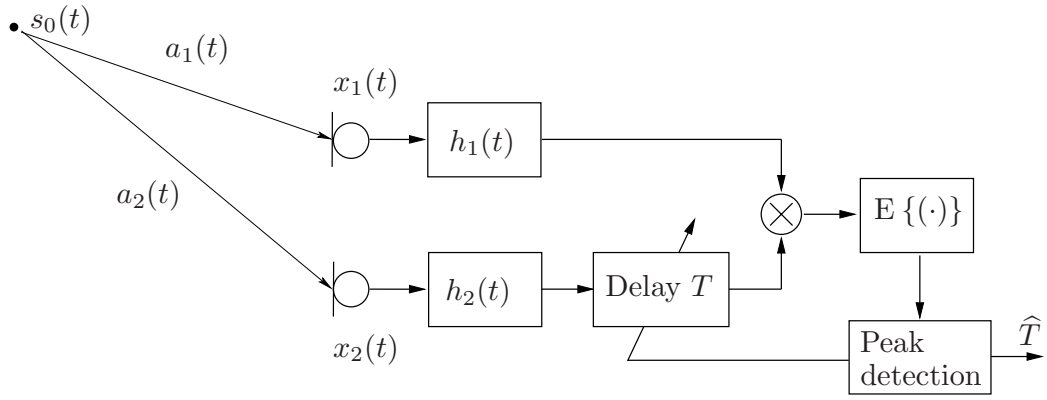


Figure 3.1: Block diagram of the generalized cross-correlation (GCC) method for the estimation of the TDOA

where $\Phi^g_{x_1x_2}(T)$ denotes the *generalized cross-correlation* (GCC) function. The filters $h_m(t)$ are chosen according to the particular optimization criterion considered.

We may also write the generalized cross-correlation function $\Phi^g_{x_1x_2}(T)$ in the frequency domain as

$$\Phi^g_{x_1x_2}(T) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H_1(\omega)\, H_2^*(\omega)\, \Psi_{X_1X_2}(\omega)\, e^{\jmath\,\omega T}\, d\omega\,, \tag{3.8}$$

where $\Psi_{X_1X_2}(\omega)$ represents the cross-power spectral density of signals $x_1(t)$ and $x_2(t)$ and ω represents the continuous frequency variable. Further, defining $G(\omega) = H_1(\omega)\, H_2^*(\omega)$ and rewriting (3.8), we obtain:

$$\Phi^g_{x_1x_2}(T) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega)\, \Psi_{X_1X_2}(\omega)\, e^{\jmath\,\omega T}\, d\omega\,. \tag{3.9}$$

The GCC-based approaches to TDOA estimation may be summarized in this framework as in Table 3.1. The term $\Gamma(\omega)$ in Table 3.1 represents the *coherence* between the microphone signals at frequency ω, and is defined as:

$$\Gamma(\omega) = \frac{\Psi_{X_1X_2}(\omega)}{\sqrt{\Psi_{X_1X_1}(\omega)\, \Psi_{X_2X_2}(\omega)}}\,. \tag{3.10}$$

The terms $\Psi_{SS}(\omega)$ and $\Psi_{V_mV_m}(\omega)$, $m = 1, 2$, in the definition of the Eckart weighting represent, respectively, the power spectral density of the source signal and that of the noise at the corresponding microphone.

Note that the development so far has assumed perfect knowledge of the cross- and auto-power spectral density of the source signals and the noise. However, in practice, these quantities have to be estimated from a fixed time record of observations.

3.2.2 Adaptive eigenvalue decomposition (AED)

The adaptive eigenvalue decomposition (AED) and its variants [12, 34] are fairly recent approaches to TDOA estimation. The block diagram of the basic AED algorithm is shown in

Table 3.1: Generalized Cross-Correlation (GCC) weighting functions

Weighting function $G(\omega)$	**Approach**
1	Regular Cross-Correlation (CC)
$\frac{1}{\sqrt{\Psi_{X_1X_1}(\omega)\,\Psi_{X_2X_2}(\omega)}}$	Smoothed Coherence Transform (SCOT)
$\frac{1}{\Psi_{X_2X_2}(\omega)}$	Roth (Wiener–Hopf weighting)
$\frac{\lvert\Gamma(\omega)\rvert^2}{1-\lvert\Gamma(\omega)\rvert^2}$	Hannan–Thomson (Maximum Likelihood estimate)
$\frac{\Psi_{SS}(\omega)}{\Psi_{V_1V_1}(\omega)\,\Psi_{V_2V_2}(\omega)}$	Eckart weighting
$\frac{1}{\lvert\Psi_{X_1X_2}(\omega)\rvert}$	Phase Transform (PHAT)

Fig. 3.2. In the noiseless case, the signal model of (2.1) reduces to

$$\begin{aligned} x_1(t) &= a_1(t) * s_0(t) \\ x_2(t) &= a_2(t) * s_0(t)\ . \end{aligned} \tag{3.11}$$

The aim, then, is to find optimal, energy-constrained filters $h_m(t)$, $(m \in \{1,2\})$, that minimize

$$\mathrm{E}\left\{e^2(t)\right\} = \mathrm{E}\left\{\left(h_1(t) * x_1(t) - h_2(t) * x_2(t)\right)^2\right\}\ . \tag{3.12}$$

In the light of Wiener–Hopf filtering [49], this can be seen as an attempt to match the two microphone signals. Under certain conditions [133] and using the commutative property of linear convolution, the signals can be exactly matched when

$$\begin{aligned} h_1(t) &= \beta a_2(t) \quad \text{and} \\ h_2(t) &= \beta a_1(t)\ , \end{aligned} \tag{3.13}$$

where β is a scaling factor. The TDOA may then be computed as the difference between the abscissae of the largest values of the respective optimal filters.

Formulating (3.11) and (3.12) in the discrete time domain as in (2.4), we get

$$\begin{aligned} x_1(n) &= \mathbf{a}_1^T \mathbf{s}_0(n) \\ x_2(n) &= \mathbf{a}_2^T \mathbf{s}_0(n)\ , \end{aligned} \tag{3.14}$$

where the room impulse responses $\mathbf{a}_m$ have been modeled as FIR filters of length L. Rewriting equation (3.12) using (3.14), we obtain

$$\mathrm{E}\left\{e^2(n)\right\} = \mathrm{E}\left\{\left(\mathbf{h}_1^T\mathbf{x}_1(n) - \mathbf{h}_2^T\mathbf{x}_2(n)\right)^2\right\}\ , \tag{3.15}$$

where

$$\mathbf{h}_m = \left[h_m(0)\,,\ h_m(1)\,,\ldots,h_m(L_{\mathrm{AED}}-1)\right]^T \tag{3.16}$$

$$\mathbf{x}_m(n) = \left[x_m(n)\,,\ x_m(n-1)\,,\ldots,x_m(n-L_{\mathrm{AED}}+1)\right]^T\ . \tag{3.17}$$

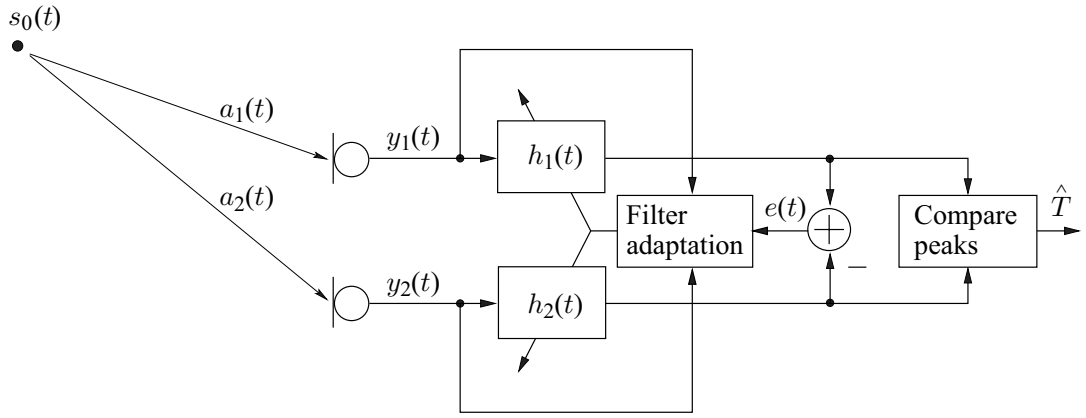


Figure 3.2: Block diagram of the AED algorithm (in continuous time)

We shall rewrite equation (3.15) more compactly as

$$\mathrm{E}\left\{e^2(n)\right\} = \mathbf{h}^T \mathbf{\Phi}_{\mathbf{xx}} \mathbf{h}\,, \tag{3.18}$$

where

$$\mathbf{h} = \left[\mathbf{h}_1^T,\ -\mathbf{h}_2^T\right]^T \tag{3.19}$$

$$\mathbf{x}(n) = \left(\mathbf{x}_1^T(n)\,,\ \mathbf{x}_2^T(n)\right)^T \tag{3.20}$$

$$\mathbf{\Phi}_{\mathbf{xx}} = \begin{pmatrix} \mathrm{E}\left\{\mathbf{x}_1\mathbf{x}_1^T\right\} & \mathrm{E}\left\{\mathbf{x}_1\mathbf{x}_2^T\right\} \\ \mathrm{E}\left\{\mathbf{x}_2\mathbf{x}_1^T\right\} & \mathrm{E}\left\{\mathbf{x}_2\mathbf{x}_2^T\right\} \end{pmatrix}. \tag{3.21}$$

It is now easy to see that, when the filter vector $\mathbf{h}$ is energy-constrained, the optimal solution $\mathbf{h}_{\mathrm{opt}}$ to the minimization problem is the eigenvector corresponding to the zero eigenvalue of $\mathbf{\Phi}_{\mathbf{xx}}$ in the noiseless case, or the eigenvector corresponding to the smallest eigenvalue when the microphone noises are spatially and temporally uncorrelated and independent of the source signals [115]. When the filters $\mathbf{a}_m$ do not have any common zeros, we obtain

$$\begin{aligned} \mathbf{h}_1 &= \beta \mathbf{a}_2 \quad \text{and} \\ \mathbf{h}_2 &= \beta \mathbf{a}_1\,, \end{aligned} \tag{3.22}$$

provided $L_{\mathrm{AED}} = L$. The TDOA may then be obtained as explained previously. For convenience, the energy of the filter vector is constrained to unity. Thus, the scale factor is determined and the following adaptive algorithm for the iterative update of the filter vector $\mathbf{h}(n)$ results:

$$\mathbf{h}(n+1) = \frac{\mathbf{h}(n) - \varsigma\, \mathbf{x}(n)\, \mathbf{x}^T \mathbf{h}(n)}{\|\mathbf{h}(n) - \varsigma\, \mathbf{x}(n)\, \mathbf{x}^T(n)\, \mathbf{h}(n)\,\|}\,, \tag{3.23}$$

where ς denotes the stepsize parameter, and $\|\cdot\|$ denotes the Euclidean norm.

Note that if we fix any one filter in (3.12) by, say, the Kronecker's delta function, the system reduces to the LMS based approach with a single adaptive filter, i.e., the discrete time version of (3.5). In the frequency domain, this corresponds to an implementation of the GCC with the Roth weighting. This relation between the AED and the GCC is elaborated in the next section.

3.2.3 Equivalence of AED and GCC

Consider that we want to match two microphone signals $x_1(n)$ and $x_2(n)$, such that their mean square error is minimized. We shall formulate the optimization in a manner similar to (3.15), and estimate two filters $h_1(n)$ and $h_2(n)$ that minimize:

$$\begin{aligned}\mathcal{J}(h_1, h_2) &= \mathrm{E}\left\{ \left(x_1(n) * h_1(n) - x_2(n) * h_2(n) \right)^2 \right\} \\ &= \mathrm{E}\Big\{ \Big(\sum_{l=-\infty}^{\infty} x_1(n-l)\, h_1(l) - \sum_{l=-\infty}^{\infty} x_2(n-l)\, h_2(l) \Big)^2 \Big\}. \end{aligned} \tag{3.24}$$

Now if, say, $h_1(n)$ is fixed to be the Kronecker's delta function, i.e.,

$$h_1(n) = \delta(n)\,,$$

then (3.24) reduces to the estimation of the single filter:

$$\mathcal{J}(h) = \mathrm{E}\Big\{ \big(x_1(n) - \sum_{l=-\infty}^{\infty} x_2(n-l)\, h(l) \big)^2 \Big\} \tag{3.25}$$

We may now optimize this for any value l, by setting the derivative of the cost function to 0 and solving for the corresponding parameter:

$$\begin{aligned}\frac{d\mathcal{J}(h)}{dh(l')} &= -2\,\mathrm{E}\left\{ \left(x_1(n) - \sum_{l=-\infty}^{\infty} x_2(n-l)\, h(l) \right) x_2(n-l') \right\} \\ &= -2 \left(\Phi_{x_1 x_2}(l') - \sum_{l=-\infty}^{\infty} h(l)\, \Phi_{x_2 x_2}(l'-l) \right) \qquad \stackrel{!}{=} 0 \end{aligned} \tag{3.26}$$

Recognizing that the above equation has the form of a convolution, we may take its Fourier transform yielding, in the *frequency* domain, the optimal frequency response:

$$H(\Omega) = \frac{\Psi_{X_1 X_2}(\Omega)}{\Psi_{X_2 X_2}(\Omega)}, \tag{3.27}$$

which, with respect to (3.9) and Table 3.1, may be recognized as the Roth-weighted cross-power spectral density. In this case the AED algorithm is fully *equivalent* to the GCC using the Roth weighting.

In its defense, the AED algorithm in its standard form should be able to better cope with reverberant environments due to the additional degree of freedom afforded by the second adaptive filter. Consequently we expect the AED to yield better results as compared to the GCC-Roth or the equivalent LMS approach as the amount of reverberation increases. However under low SNR conditions this additional degree of freedom may prove to be a drawback as the errors in the filter estimates would be cumulative, leading to an inferior performance of the AED. A comparison of the AED with the LMS-based time delay estimation (TDE) algorithm, as for example in Section 3.4.1, seems to validate this assumption.

3.2.4 Information theoretic approach to TDOA estimation

As shown by [117], the principle of information maximization can be also used to approach the problem of TDOA estimation. The signal model considered for this method is the noiseless variant of that in (3.1). The idea is to find the time delay T that maximizes the mutual information $\mathfrak{I}\left(x_1(t)\,;x_2(t+T)\right)$ between the microphone signals $x_1(t)$ and $x_2(t+T)$:

$$\widehat{T} = \underset{T}{\operatorname{argmax}}\ \mathfrak{I}\left(x_1(t)\,;x_2(t+T)\right)\ . \tag{3.28}$$

The mutual information between two random variables a and b is given in [30] as:

$$\mathfrak{I}(a;b) = \mathfrak{H}\left(a\right) + \mathfrak{H}\left(b\right) - \mathfrak{H}\left(a,b\right)\ , \tag{3.29}$$

where $\mathfrak{H}\left(a\right)$ represents the *entropy* of the random variable a and $\mathfrak{H}\left(a,b\right)$ the joint entropy of a and b. Assuming an underlying stationary, zero-mean stochastic process to be the generator of the source signal, the mutual information between the two microphone signals for a time shift T may be written as

$$\mathfrak{I}\left(x_1(t)\,;x_2(t+T)\right) = \mathfrak{H}\left(x_1(t)\right) + \mathfrak{H}\left(x_2(t+T)\right) - \mathfrak{H}\left(x_1(t)\,,x_2(t+T)\right). \tag{3.30}$$

The entropy for Gaussian random variables is well known (cf. [30]) to be proportional to the logarithm of its variance (neglecting a constant offset). Thus, under the assumptions that the signals have a Gaussian density,

$$\begin{aligned}
\mathfrak{H}\left(x_1(t)\right) &\propto \ln\left(\Phi_{x_1x_1}(0)\right)\\
\mathfrak{H}\left(x_2(t+T)\right) &\propto \ln\left(\Phi_{x_2x_2}(0)\right)\\
\mathfrak{H}\left(x_1(t)\,,x_2(t+T)\right) &\propto \ln\left(\det\left(\boldsymbol{\Phi}_{\mathbf{xx}}\right)\right),
\end{aligned} \tag{3.31}$$

where $\mathbf{x}(t) = \left(x_1(t)\,,\ x_2(t+T)\right)^T$ and $\boldsymbol{\Phi}_{\mathbf{xx}}$ is given by

$$\begin{aligned}
\boldsymbol{\Phi}_{\mathbf{xx}} &= \mathrm{E}\left\{\mathbf{x}(t)\,\mathbf{x}^T(t)\right\}\\
&= \begin{pmatrix} \mathrm{E}\left\{x_1^2(t)\right\} & \mathrm{E}\left\{x_1(t)\,x_2(t+T)\right\}\\ \mathrm{E}\left\{x_1(t)\,x_2(t+T)\right\} & \mathrm{E}\left\{x_2^2(t+T)\right\}\end{pmatrix}\\
&= \begin{pmatrix} \Phi_{x_1x_1}(0) & \Phi_{x_1x_2}(T)\\ \Phi_{x_1x_2}(T) & \Phi_{x_2x_2}(0)\end{pmatrix}.
\end{aligned} \tag{3.32}$$

Consequently,

$$\begin{aligned}
\widehat{T} &= \underset{T}{\operatorname{argmax}}\ \mathfrak{I}\left(x_1(t)\,;x_2(t+T)\right)\\
&= \underset{T}{\operatorname{argmin}}\ \ln\left(\det\left(\boldsymbol{\Phi}_{\mathbf{xx}}\right)\right)\\
&= \underset{T}{\operatorname{argmax}}\ \Phi_{x_1x_2}(T)\ ,
\end{aligned} \tag{3.33}$$

where the simplifications above follow as the first two terms in (3.30) are independent of T according to (3.31). Thus, when the signals are Gaussian distributed, maximizing the mutual information is equivalent to maximizing the *cross-correlation* between the microphone signals. In contrast to GCC, however, the optimization criterion in its general form in (3.28) can exploit the possible non-Gaussian structure of the source signals.

3.2.5 Extension to multiple microphone pairs

The above sections described various approaches to estimate the TDOA using a pair of sensors. Using more than one pair of microphones increases the spatial diversity afforded to the localization system and, consequently, may be exploited to localize the source in more than one spatial dimension and to improve the localization accuracy.

The simplest way to extend the two-channel method to an M channel ($M > 2$) array is to obtain the TDOA estimate T_p for *all* $p \in \{1, 2, \ldots, \binom{M}{2}\}$ microphone pairs, using any of the GCC approaches, e.g., the Roth weighted GCC estimate [37], simple cross-correlation [14], the PHAT estimate [19], or the multi-channel AED approach, where an estimate of the impulse response from the source to each microphone is first obtained, from which the TDOA between all microphone pairs may be computed as in the AED approach. Once this is done, one method of obtaining the source position is by solving the non-linear equation relating the vector of obtained TDOA estimates, the geometry of the array and the source location [24, 36, 54]

$$\widehat{\mathbf{r}}_s = \mathcal{F}(T_1, T_2, \ldots, T_{M(M-1)/2}, \mathbf{r}_1, \ldots, \mathbf{r}_M)\,, \tag{3.34}$$

with $\mathbf{r}_m$ being the spatial co-ordinates of the mth microphone. The methods of [75, 9, 11] additionally consider weighting the contributions of the TDOA estimates from the various pairs according to some optimality criteria when computing the 'averaged' source location, thus improving the location estimate.

Another interesting approach to source localization using multiple microphones is that suggested in [27, 26], where the concept of linear prediction is extended to the spatial case. For the two microphone case, the cost function derived for this approach reduces to that for the information theoretic approach of Section 3.2.4.

Methods to cope with multiple sources are detailed in [105, 106] where the distance structure of peaks in the correlation function and graph-theoretic considerations are exploited. Other approaches project the cross-correlation function $\Phi_{x_m x_{m'}}(\tau)$ of all pairs (m, m') onto a common co-ordinate system, e.g., a one-dimensional direction of arrival value [81], or a two-dimensional grid [19], or the surface of a hemisphere (the accumulative correlation of [14]) and so on, creating, in essence, a histogram of likelihood values over candidate source locations. The source location then corresponds to the most likely candidate position.

3.3 Direct localization approaches

For the indirect approaches, source localization is the result of a two-step approach. First, an estimate of the TDOA is obtained (using an array of two or more microphones) and then, based on the knowledge of the geometry of the array and the time delay estimates, the source position is estimated. Direct approaches, on the other hand, perform TDOA estimation and source localization in one step. Most direct algorithms scan a set of candidate source locations (the so-called *search space*) and then pick the most likely position as an estimate of the source location. This approach makes it easier to incorporate multiple microphones in the optimization criterion. As in the case of the indirect approaches, the algorithms belonging to the direct class may be formulated in the time or the frequency domain.

We discuss localization algorithms – initially designed for narrowband sources – in the frequency domain. The extension to the wideband case could be of the straightforward *incoherent* kind, where the narrowband location estimates at the center frequency of each subband are averaged over all the subbands [67, 128], or they could be of the *coherent* sort, where the data in all the subbands are collectively used when scanning the candidate locations [67, 124, 55, 68, 136]. We will also develop their link to the GCC framework (when posed in each subband in the frequency domain, as in Table 3.1). Unless mentioned otherwise, we shall consider a single source located at $\mathbf{r}_s = (x_s, y_s, z_s)^T$, in a three-dimensional Cartesian co-ordinate system.

3.3.1 Steered response power beamforming

The steered response power (SRP) beamforming approach [33, 90] searches for the candidate source position that maximizes the output power of a filter-and-sum beamformer steered in that direction. While the optimization criterion is of the broadband type, it is again instructive to expand it into a frequency domain formulation. The output power spectral density of the filter-and-sum beamformer may be written as in [122, Chap. 12]:

$$\begin{aligned}\Psi_{\widehat{S}\widehat{S}}\left(\mathbf{r},\Omega\right) &= \mathrm{E}\left\{\left|\mathbf{H}^H\left(\mathbf{r},\Omega\right)\mathbf{X}\left(\Omega\right)\right|^2\right\} = \mathbf{H}^H(\mathbf{r},\Omega)\boldsymbol{\Psi}_{\mathbf{XX}}(\Omega)\mathbf{H}(\mathbf{r},\Omega) \\ &= \mathbf{H}^H(\mathbf{r},\Omega)\begin{pmatrix}\Psi_{X_1X_1}(\Omega) & \cdots & \Psi_{X_1X_M}(\Omega)\\ \vdots & \ddots & \vdots \\ \Psi_{X_MX_1}(\Omega) & \cdots & \Psi_{X_MX_M}(\Omega)\end{pmatrix}\mathbf{H}(\mathbf{r},\Omega)\,,\end{aligned} \tag{3.35}$$

where the beam is directed towards $\mathbf{r}$ and $\mathbf{H}(\mathbf{r},\Omega) = \big(H_1(\mathbf{r},\Omega),\ldots,H_M(\mathbf{r},\Omega)\big)^T$ is the corresponding vector of the beamforming filter frequency responses. $\Psi_{X_mX_{m'}} \stackrel{\triangle}{=} \mathrm{E}\left\{X_m(\Omega)X^*_{m'}(\Omega)\right\}$ denotes the cross-power spectral density of channels m and m'. Then, the source location $\widehat{\mathbf{r}}_s$ is found as

$$\widehat{\mathbf{r}}_s = \underset{\mathbf{r}}{\operatorname{argmax}}\ \frac{1}{2\pi}\int_{-\pi}^{\pi}\Psi_{\widehat{S}\widehat{S}}(\mathbf{r},\Omega)\,\mathrm{d}\Omega\,. \tag{3.36}$$

We may expand (3.35) as

$$\begin{aligned}\Psi_{\widehat{S}\widehat{S}}(\mathbf{r},\Omega) &= \sum_{m,m'}H^*_m(\mathbf{r},\Omega)H_{m'}(\mathbf{r},\Omega)\Psi_{X_mX_{m'}}(\Omega) \\ &= \sum_m |H_m(\mathbf{r},\Omega)|^2\Psi_{X_mX_m}(\Omega) \\ &\quad + \sum_{\substack{m,m'\\ m\neq m'}}H^*_m(\mathbf{r},\Omega)H_{m'}(\mathbf{r},\Omega)\Psi_{X_mX_{m'}}(\Omega) \\ &= \sum_m \Psi_{X_mX_m}(\Omega) \\ &\quad + \sum_{\substack{m,m'\\ m\neq m'}}e^{\jmath\,\Omega f_{\mathrm{s}}T_m(\mathbf{r})-\jmath\,\Omega f_{\mathrm{s}}T_{m'}(\mathbf{r})}\Psi_{X_mX_{m'}}(\Omega)\,,\end{aligned} \tag{3.37}$$

where for the last step the delay-and-sum beamformer with

$$\mathbf{H}\left(\mathbf{r},\Omega\right) = \left(e^{-\jmath\,\Omega f_{\mathrm{s}}T_1(\mathbf{r})},\ e^{-\jmath\,\Omega f_{\mathrm{s}}T_2(\mathbf{r})},\ \ldots,\ e^{-\jmath\,\Omega f_{\mathrm{s}}T_M(\mathbf{r})}\right)^T \tag{3.38}$$

was assumed, where $T_m(\mathbf{r})$ corresponds to $-\Delta\tau_m$ in (2.17), for any chosen reference point $\mathbf{r}_0$ and hypothesized source location $\mathbf{r}$. For such an expansion of (3.35) it may be seen that the first term is independent of the source location and the second term sums over the cross-power spectral density of all $\binom{M}{2}$ microphone pairs.

Similar to the GCC in the case of two channels, the cross-power spectral densities may be weighted according to the criteria outlined in Table 3.1. Thus we obtain, for instance, the SRP-PHAT approach

$$\widehat{\mathbf{r}}_s = \underset{\mathbf{r}}{\operatorname{argmax}} \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{\substack{m,m' \\ m \neq m'}} e^{\jmath\,\Omega f_\mathrm{s} T_m(\mathbf{r}) - \jmath\,\Omega f_\mathrm{s} T_{m'}(\mathbf{r})} \frac{\Psi_{X_m X_{m'}}(\Omega)}{|\Psi_{X_m X_{m'}}(\Omega)|}\, d\Omega\,, \tag{3.39}$$

which may be seen as an extension of the GCC-PHAT to the M microphone case.

Compared with GCC, the SRP method provides additional degrees of freedom that allow us to smooth over microphone pairs instead of over frequency. In fact, the cost function (3.35) may be evaluated for each frequency separately. Thereby, the method can be easily extended to localize multiple sources with disjoint frequency spectra.

3.3.2 Minimum mean square error (MMSE) approach

This approach was developed in [72] and extended in [77] and is based on the model of (2.15). The idea behind this approach is to search for appropriate phase compensation factors $e^{\jmath\,\Omega f_\mathrm{s} T_m(\mathbf{r})}$ for each channel m such that the mean-squared error between the phase compensated signals of all pairs is minimized. Note that from the localization point of view, the phase compensation factors are parametrized by the candidate source positions. This may be expressed as

$$\begin{aligned} \widehat{\mathbf{r}}_s &= \underset{\mathbf{r}}{\operatorname{argmin}} \sum_{\substack{m,m' \\ m \neq m'}} \mathrm{E}\left\{ \left| X_m e^{\jmath\,\Omega f_\mathrm{s} T_m(\mathbf{r})} - X_{m'} e^{\jmath\,\Omega f_\mathrm{s} T_{m'}(\mathbf{r})} \right|^2 \right\} \\ &= \underset{\mathbf{r}}{\operatorname{argmin}}\ (M-1) \sum_{m} \Psi_{X_m X_m} - \sum_{\substack{m,m' \\ m \neq m'}} e^{\jmath\,\Omega f_\mathrm{s}(T_m(\mathbf{r}) - T_{m'}(\mathbf{r}))} \Psi_{X_m X_{m'}}\,. \end{aligned} \tag{3.40}$$

It may be seen that the first term is again independent of any phase compensation factors and may therefore be neglected, leading to the following simplified cost function

$$\mathcal{J}(\mathbf{r}, \Omega) = - \sum_{\substack{m,m' \\ m \neq m'}} e^{\jmath\,\Omega f_\mathrm{s}(T_m(\mathbf{r}) - T_{m'}(\mathbf{r}))} \Psi_{X_m X_{m'}}\,. \tag{3.41}$$

Thus, the MMSE approach is fully *equivalent* to the SRP approach and also falls under the umbrella of the GCC. Again, as in (3.39), it is possible to weight the cross-power spectral density using various other criteria. One (rather heuristic) weighting which gives good results is suggested in [77]:

$$\mathcal{J}(\mathbf{r}, \Omega) = - \sum_{\substack{m,m' \\ m \neq m'}} e^{\jmath\,\Omega f_\mathrm{s}(T_m(\mathbf{r}) - T_{m'}(\mathbf{r}))} \Psi_{X_m X_{m'}} \left|\Gamma_{mm'}(\Omega)\right|^2\,, \tag{3.42}$$

where $\Gamma_{mm'}(\Omega)$ indicates the coherence between channels m and m' at frequency Ω.

3.3.3 Practical aspects

In practice, source location estimates are computed on finite time records of the observed signals using the STFT domain representation of the discrete time signals $x_m(n)$. In this case, the expectation may either be dropped in favor of an instantaneous estimate for each frame b, or computed as a temporal average, for example by a first-order recursive smoothing along frames. For the former case, the cost function may be written in each frequency bin k as

$$\begin{aligned}\mathcal{J}(\mathbf{r},k,b) &= \mathbf{H}^H(\mathbf{r},k)\mathbf{X}(k,b)\mathbf{X}^H(k,b)\mathbf{H}(\mathbf{r},k)\\ &= \sum_{m,m'} H_m^*(\mathbf{r},k)H_{m'}(\mathbf{r},k)X_m(k,b)X_{m'}^*(k,b)\\ &= \sum_{m} \big|H_m(\mathbf{r},k)\big|^2\big|X_m(k,b)\big|^2\\ &\quad + \sum_{\substack{m,m'\\ m\neq m'}} H_m^*(\mathbf{r},k)H_{m'}(\mathbf{r},k)X_m(k,b)X_{m'}^*(k,b)\end{aligned} \tag{3.43}$$

from which point on, the procedure to localize the source is exactly the same as outlined in Sections 3.3.1 and 3.3.2, namely

$$\widehat{\mathbf{r}}_s(k,b) = \operatorname*{argmax}_{\mathbf{r}} \sum_{\substack{m,m'\\ m\neq m'}} H_m^*(\mathbf{r},k)H_{m'}(\mathbf{r},k)X_m(k,b)X_{m'}^*(k,b)\,. \tag{3.44}$$

Note that the cost function may also be weighted as in (3.39), yielding the source location estimate as

$$\widehat{\mathbf{r}}_s(k,b) = \operatorname*{argmax}_{\mathbf{r}} \sum_{\substack{m,m'\\ m\neq m'}} H_m^*(\mathbf{r},k)H_{m'}(\mathbf{r},k)\frac{X_m(k,b)X_{m'}^*(k,b)}{\big|X_m(k,b)\big|\big|X_{m'}(k,b)\big|}\,. \tag{3.45}$$

This normalization reduces the dependency of the cost function on the signal amplitudes, thus ameliorating the effects of sensor gain mismatch. We may also choose to weight the normalized components in a manner similar to that in (3.42), to provide further robustness against sensor defects.

When the signals to be localized are broadband, the computed cost function for each bin k as in (3.45) could additionally be averaged across all frequencies as in [33, 90]

$$\mathcal{J}(\mathbf{r},b) = \sum_{k}\sum_{\substack{m,m'\\ m\neq m'}} H_m^*(\mathbf{r},k)H_{m'}(\mathbf{r},k)\frac{X_m(k,b)X_{m'}^*(k,b)}{\big|X_m(k,b)\big|\big|X_{m'}(k,b)\big|} \tag{3.46}$$

yielding an estimate for the source location $\mathbf{r}_s(b)$ per frame b as

$$\widehat{\mathbf{r}}_s(b) = \operatorname*{argmax}_{\mathbf{r}} \mathcal{J}(\mathbf{r},b)\,. \tag{3.47}$$

3.3.4 Subspace based approaches

The MUltiple SIgnal Classification or MUSIC algorithm proposed in [107] is a subspace based approach. It works on the farfield model, but may be extended to the nearfield case, too. It was originally proposed as a solution to the problem of localization of Q narrowband,

uncorrelated sources, with $Q < M$. This approach can be extended to the wideband scenario in an incoherent [128] or a coherent [124, 136] manner. In our discussion, we shall restrict ourselves to the narrowband formulation.

We consider again the frequency domain model as in (2.8),

$$\mathbf{X} = \mathbf{AS} + \mathbf{V}\,, \tag{3.48}$$

where we have dropped the frequency variable for convenience. Computing the spectral covariance matrix $\mathbf{\Psi_{XX}}$, we obtain

$$\begin{aligned}\mathbf{\Psi_{XX}} &= \mathrm{E}\left\{\mathbf{X}\mathbf{X}^H\right\} \\ &= \mathbf{A}\mathrm{E}\left\{\mathbf{S}\mathbf{S}^H\right\}\mathbf{A}^H + \mathrm{E}\left\{\mathbf{V}\mathbf{V}^H\right\} \\ &= \mathbf{A}\mathbf{\Psi_{SS}}\mathbf{A}^H + \mathbf{\Psi_{VV}} \\ &= \mathbf{A}\mathbf{\Psi_{SS}}\mathbf{A}^H + \Psi_{VV}\mathbf{I}\,,\end{aligned} \tag{3.49}$$

where the source signals and the noise are assumed to be independent. The last step follows when the noise is spatially uncorrelated and with the same variance at each microphone[2]. The covariance matrix may be decomposed using the eigenvalue decomposition, yielding

$$\mathbf{\Psi_{XX}} = \mathbf{U}\left(\mathbf{D} + \Psi_{VV}\mathbf{I}\right)\mathbf{U}^H. \tag{3.50}$$

For $Q < M$ sources, the diagonal matrix $\mathbf{D}$ is singular and contains the Q dominant eigenvalues proportional to the spectral power of the sources. When the eigenvalues λ_q of $\mathbf{D}$ are arranged according to decreasing order of magnitude,

$$\lambda_1 > \lambda_2 > \ldots > \lambda_Q > \lambda_{Q+1} = \ldots = \lambda_M = 0\,, \tag{3.51}$$

the first Q eigenvectors $\mathbf{U}_q$ span the so-called signal-plus-noise subspace, whereas the $M-Q$ eigenvectors $\mathbf{U}_q$, $Q < q \leq M$ span the noise-only subspace.

From (3.49) and (3.50) it may be seen that the $(M-Q)$ eigenvectors of the noise-only subspace define the orthogonal complement of $\mathbf{A}$. Consequently, if we define a spatial spectrum $\mathcal{S}_{\mathrm{MUSIC}}(\mathbf{r})$ over all candidate source locations $\mathbf{r}$ as

$$\mathcal{S}_{\mathrm{MUSIC}}(\mathbf{r}) = \frac{1}{\mathbf{H}^H(\mathbf{r})\mathbf{U}_V\mathbf{U}_V^H\mathbf{H}(\mathbf{r})}\,, \tag{3.52}$$

where $\mathbf{H}(\mathbf{r}) = \left(e^{-\jmath\,\Omega f_s T_1(\mathbf{r})},\ \ldots,\ e^{-\jmath\,\Omega f_s T_M(\mathbf{r})}\right)^T$ is the steering vector towards candidate source location $\mathbf{r}$ and $\mathbf{U}_V$ is the $M \times (M-Q)$ matrix containing the eigenvectors corresponding to the noise-only subspace, the locations $\mathbf{r}$ corresponding to the Q peaks of the spectrum are the sought source positions

$$\widehat{\mathbf{r}}_{s_q} = \underset{\mathbf{r}}{\mathrm{argmax}}\ \mathcal{S}_{\mathrm{MUSIC}}(\mathbf{r})\,. \tag{3.53}$$

Alternatively, we may obtain the source position vectors as

$$\widehat{\mathbf{r}}_{s_q} = \underset{\mathbf{r}}{\mathrm{argmax}}\ \mathbf{H}^H(\mathbf{r})\mathbf{U}_S\mathbf{U}_S^H\mathbf{H}(\mathbf{r})\,, \tag{3.54}$$

where $\mathbf{U}_S$ is the $M \times Q$ matrix containing the eigenvectors corresponding to the Q dominant eigenvalues, spanning the signal-plus-noise subspace.

[2] If this is not the case, the signals may be prewhitened as in [38].

Single-frame MUSIC

The traditional MUSIC approach requires the estimation of the spectral covariance matrix to determine the number of dominant eigenvalues and corresponding eigenvectors. Assuming ergodicity of the signals, this estimate is computed by temporal averaging, making this approach batch-based. A simple modification of this approach leads to what we term the *single-frame* MUSIC approach, which bears a close relation to the SRP algorithm discussed previously.

The idea behind single-frame MUSIC is as follows. The matrix

$$\widehat{\mathbf{\Psi}}_{\mathbf{XX}} = \mathbf{X}\mathbf{X}^H \tag{3.55}$$

is of rank one. Thus, an eigenvalue decomposition of this matrix yields one dominant eigenvalue λ_1 with its corresponding eigenvector $\mathbf{U}_1$. The single-frame MUSIC spectrum $\mathcal{S}_{\text{MUSIC}}(\mathbf{r})$ is then computed from (3.52), where the matrix $\mathbf{U}_V\mathbf{U}_V^H$ is obtained as

$$\mathbf{U}_V\mathbf{U}_V^H = \mathbf{I} - \mathbf{U}_1\mathbf{U}_1^H. \tag{3.56}$$

The maxima of $\mathcal{S}_{\text{MUSIC}}(\mathbf{r})$ then indicate the position vectors.

For a single source, $\mathbf{A}(\mathbf{r}_s)$ simplifies to a column vector and (3.55) may be rewritten as

$$\widehat{\mathbf{\Psi}}_{\mathbf{XX}} = \left(\mathbf{A}(\mathbf{r}_s)S + \mathbf{V}\right)\left(\mathbf{A}(\mathbf{r}_s)S + \mathbf{V}\right)^H. \tag{3.57}$$

It is easily verified that the dominant eigenvector is

$$\begin{aligned}\mathbf{U}_1 &= \frac{\mathbf{A}(\mathbf{r}_s)S + \mathbf{V}}{\|\mathbf{A}(\mathbf{r}_s)S + \mathbf{V}\|} \\ &= \frac{\mathbf{X}}{\|\mathbf{X}\|},\end{aligned} \tag{3.58}$$

with the corresponding eigenvalue of $\lambda_1 = \|\mathbf{X}\|^2$. Maximizing the MUSIC spectrum of (3.52) as in (3.53), we have

$$\begin{aligned}\widehat{\mathbf{r}}_s &= \underset{\mathbf{r}}{\text{argmax}}\ \frac{1}{\mathbf{H}(\mathbf{r})^H\mathbf{U}_V\mathbf{U}_V^H\mathbf{H}(\mathbf{r})} \\ &= \underset{\mathbf{r}}{\text{argmax}}\ \frac{1}{\mathbf{H}^H(\mathbf{r})\left(\mathbf{I} - \mathbf{U}_1\mathbf{U}_1^H\right)\mathbf{H}(\mathbf{r})} \\ &= \underset{\mathbf{r}}{\text{argmin}}\ \frac{1}{\mathbf{H}^H(\mathbf{r})\mathbf{U}_1\mathbf{U}_1^H\mathbf{H}(\mathbf{r})} \\ &= \underset{\mathbf{r}}{\text{argmax}}\ \mathbf{H}^H(\mathbf{r})\mathbf{U}_1\mathbf{U}_1^H\mathbf{H}(\mathbf{r}),\end{aligned} \tag{3.59}$$

which, combined with (3.58), is closely related to the SRP cost function of (3.45).

3.3.5 Information theoretic criteria

Another interesting approach to the localization of sources in the frequency domain is through the application of information theoretic critera, specifically the principle of *minimum cross entropy* (MinxEnt) of Kullback [62].

Consider, again, the narrowband frequency domain signal model of (3.48):

$$\mathbf{X} = \mathbf{AS} + \mathbf{V}\,. \tag{3.60}$$

We shall now seek the optimal linear transformation that can represent the $Q < M$ sources by a more compact representation $\mathbf{Y}$, with $\mathbf{Y} \in \mathbb{C}^{Q\times 1}$. Note that as $\mathbf{Y}$ is of lower dimension than $\mathbf{X}$, this compression is *lossy*. The transformation $\mathbf{Y} = \mathbf{WX}$ should be such that we have the *minimum loss of information*, where the latter is defined in the sense proposed by Shannon [109]. Thus:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmax}}\ \mathfrak{I}\left(\mathbf{X};\mathbf{Y}\middle|\mathbf{W}\right), \tag{3.61}$$

where, as in Section 3.2.4, $\mathfrak{I}(a;b)$ denotes mutual information between a and b. Assuming that the signals and the noise are Gaussian distributed and uncorrelated, the respective entropies of $\mathbf{X}$ and $\mathbf{Y}$ may be expressed using the definition in [30] as:

$$\begin{aligned}
\mathfrak{H}\left(\mathbf{X}\middle|\mathbf{W}\right) &= \frac{1}{2}\ln\left(|\mathbf{\Psi_{XX}}|\right) + M\mathfrak{K} \\
\mathfrak{H}\left(\mathbf{Y}\middle|\mathbf{W}\right) &= \frac{1}{2}\ln\left(|\mathbf{W}\mathbf{\Psi_{XX}}\mathbf{W}^H|\right) + Q\mathfrak{K}, \quad \text{and} \\
\mathfrak{H}\left(\mathbf{X},\mathbf{Y}\middle|\mathbf{W}\right) &= \mathfrak{H}\left(\mathbf{X}\middle|\mathbf{W}\right) + \mathfrak{H}\left(\mathbf{Y}\middle|\mathbf{X},\mathbf{W}\right) \\
&= \mathfrak{H}\left(\mathbf{X}\middle|\mathbf{W}\right),
\end{aligned} \tag{3.62}$$

where $\mathfrak{K}$ is a constant. Using the eigenvalue decomposition and neglecting the constant terms, we may simplify the above to:

$$\begin{aligned}
\mathfrak{H}\left(\mathbf{X}\middle|\mathbf{W}\right) &= \frac{1}{2}\sum_{m=1}^{M}\ln\left(\lambda_m\right) \\
\mathfrak{H}\left(\mathbf{Y}\middle|\mathbf{W}\right) &= \frac{1}{2}\sum_{q=1}^{Q}\ln\left(\rho_q\right),
\end{aligned}$$

where the ρ_q represent the eigenvalues of $\mathbf{W}\mathbf{\Psi_{XX}}\mathbf{W}^H$.

The mutual information $\mathfrak{I}\left(\mathbf{X};\mathbf{Y}\middle|\mathbf{W}\right)$ is then (neglecting the additive constants):

$$\mathfrak{I}\left(\mathbf{X};\mathbf{Y}\middle|\mathbf{W}\right) \propto \sum_{q=1}^{Q}\ln\left(\rho_q\right). \tag{3.63}$$

It is maximized when $\sum_{q=1}^{Q}\ln\left(\rho_q^2\right)$ is maximum[3], i.e., the transformation $\mathbf{W}$ preserves the Q dominant eigenvectors of $\mathbf{\Psi_{XX}}$. Consequently, the columns of $\mathbf{W}$ span the signal+noise subspace, whence the propagation vectors may be found as in the MUSIC approach. Thus, MUSIC may also be derived from information theoretic criteria, assuming Gaussian distributions for the signal and the noise.

3.3.6 Maximum likelihood estimation (MLE)

The subspace and beamformer based approaches presented in the previous sections are computationally attractive. However, when multiple partially or fully coherent sources are present, the performance of these estimators is suboptimal. An alternative is to exploit

[3]For a given volume constraint on $\mathbf{W}$.

the underlying data model more completely. This leads to the development of the so-called *parametric* methods in the frequency domain, of which maximum likelihood (ML) estimators form an important class. Partial or complete signal coherence does not pose conceptual problems for the MLE approaches [67, 61, 15]. Moreover, estimates obtained using MLE approaches can be shown to be asymptotically consistent and attaining the Cramér–Rao lower bound.

ML approaches require models of the probability density function of the signals under consideration. We shall first consider the *deterministic* ML approach, where the source signals are modeled as deterministic and unknown and the noise is assumed to be stationary and Gaussian distributed. Next, we shall consider the more general *stochastic* ML approach, where both the source signals and the noise are assumed to be stationary and Gaussian distributed, with the source signals being independent of the noise. For both cases, we consider again, the signal model in the STFT domain where k and b denote the frequency bin index and the frame index respectively:

$$\mathbf{X}(k,b) = \mathbf{A}(k)\mathbf{S}(k,b) + \mathbf{V}(k,b)\,. \tag{3.64}$$

The noise is assumed to be spatially white. If this is not the case, the received signals at the microphones need to be prewhitened (for whitening approaches see, e.g., [38]). In what follows, we drop the frequency bin index k.

Deterministic maximum likelihood estimation

If we assume that the source signals are deterministic and unknown and that the noise is Gaussian distributed and spatially white, we have

$$\mathrm{E}\left\{\mathbf{V}(b)\mathbf{V}^H(b)\right\} = \Psi_{VV}\mathbf{I} \tag{3.65}$$

$$p\left(\mathbf{X}(b)\middle|\mathbf{A},\mathbf{S}(b)\right) = \frac{1}{\left(\pi\Psi_{VV}\right)^M}\exp\left(-\frac{\|\mathbf{X}(b)-\mathbf{A}\mathbf{S}(b)\|^2}{\Psi_{VV}}\right)\,, \tag{3.66}$$

where $p\left(\cdot\middle|\cdot\right)$ represents the conditional probability density function and $\|\cdot\|$ represents the Euclidean norm. Under the assumption that the measurement at each time frame is independent of the other measurements, we obtain, for B time records

$$p\left(\mathbf{X}^B\middle|\mathbf{A},\mathbf{S}^B\right) = \prod_{b=1}^{B} p\left(\mathbf{X}(b)\middle|\mathbf{A},\mathbf{S}(b)\right)\,, \tag{3.67}$$

where $\mathbf{X}^B = (\mathbf{X}(1),\ldots,\mathbf{X}(B))$ and $\mathbf{S}^B = (\mathbf{S}(1),\ldots,\mathbf{S}(B))$ are the $(M \times B)$ sensor and $(Q \times B)$ source signal *matrices*, respectively.

The aim of the deterministic ML approach is then to find the optimal parameter vector

$$\boldsymbol{\Theta} = \left(\mathbf{r}_{s_1}^T, \mathbf{r}_{s_2}^T, \ldots, \mathbf{r}_{s_Q}^T, \mathbf{S}^T(1),\ldots,\mathbf{S}^T(B), \Psi_{VV}\right)^T$$

such that we maximize the *likelihood* function $p\left(\mathbf{X}^B\middle|\boldsymbol{\Theta}\right)$.

For mathematical tractability often the log-likelihood function, defined as

$$\mathcal{L}(\mathbf{X}^B;\boldsymbol{\Theta}) = \ln\left(p\left(\mathbf{X}^B\middle|\boldsymbol{\Theta}\right)\right)\,,$$

is used. Owing to the monotonicity of the $\ln(\cdot)$ function, the $\boldsymbol{\Theta}$ that maximizes the log-likelihood function will also maximize the likelihood function. Thus

$$\mathcal{L}(\mathbf{X}^B;\boldsymbol{\Theta}) = \sum_{b=1}^{B} \ln\left(p\left(\mathbf{X}(b)\middle|\boldsymbol{\Theta}\right)\right) \tag{3.68}$$

$$= -MB\ln(\pi\Psi_{VV}) - \frac{1}{\Psi_{VV}}\sum_{b=1}^{B}\|\mathbf{X}(b)-\mathbf{A}\mathbf{S}(b)\|^2 \tag{3.69}$$

and

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\ \mathcal{L}(\mathbf{X}^B;\boldsymbol{\Theta}) \tag{3.70}$$

$$= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\ \left(MB\ln(\pi\Psi_{VV}) + \frac{1}{\Psi_{VV}}\sum_{b=1}^{B}\|\mathbf{X}(b)-\mathbf{A}\mathbf{S}(b)\|^2\right). \tag{3.71}$$

Minimizing (3.71) with respect to $\boldsymbol{\Theta}$ we obtain the well known solutions to the deterministic maximum likelihood (DML) problem [126, 29]:

$$\begin{aligned}\widehat{\mathbf{A}}_{\mathrm{DML}} &= \underset{\mathbf{H}}{\operatorname{argmin}}\ \sum_b \mathbf{X}^H(b)\mathbf{P}_{\mathbf{H}}^{\perp}\mathbf{X}(b)\\ &= \underset{\mathbf{H}}{\operatorname{argmax}}\ \sum_b \mathbf{X}^H(b)\mathbf{P}_{\mathbf{H}}\mathbf{X}(b)\end{aligned} \tag{3.72}$$

$$\widehat{\Psi}_{VV,\mathrm{DML}} = \frac{1}{MB}\sum_b \mathbf{X}^H(b)\mathbf{P}^{\perp}_{\widehat{\mathbf{A}}_{\mathrm{DML}}}\mathbf{X}(b) \tag{3.73}$$

$$\widehat{\mathbf{S}}_{\mathrm{DML}}(b) = \widehat{\mathbf{A}}^{\dagger}_{\mathrm{DML}}\mathbf{X}(b)\,, \tag{3.74}$$

where $\mathbf{P}_{\mathbf{H}} = \mathbf{H}\big(\mathbf{H}^H\mathbf{H}\big)^{-1}\mathbf{H}^H$ and $\mathbf{P}_{\mathbf{H}}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{H}}$ are, respectively, the projection matrices onto the column-space of $\mathbf{H}$ and onto the corresponding orthogonal complement, and

$$\widehat{\mathbf{A}}^{\dagger}_{\mathrm{DML}} = \left(\widehat{\mathbf{A}}^H_{\mathrm{DML}}\widehat{\mathbf{A}}_{\mathrm{DML}}\right)^{-1}\widehat{\mathbf{A}}^H_{\mathrm{DML}}$$

is the *Moore–Penrose pseudoinverse* [115] of $\widehat{\mathbf{A}}_{\mathrm{DML}}$. Note that $\mathbf{H}$ is the steering vector matrix that is parameterized by the hypothesized source locations. The optimal $\mathbf{H}$ should then provide a good estimate of $\mathbf{A}$ and of the source locations.

For the case of a single source S_1, $\mathbf{A} = \mathbf{A}_1$ and the optimal source location is obtained as the position that maximizes (from (3.72))

$$\begin{aligned}\widehat{\mathbf{r}}_s &= \underset{\mathbf{r}}{\operatorname{argmax}}\ \sum_b \frac{|\mathbf{H}^H(\mathbf{r})\mathbf{X}(b)|^2}{\|\mathbf{H}(\mathbf{r})\|^2}\\ &= \underset{\mathbf{r}}{\operatorname{argmax}}\ \sum_{\substack{m,m'\\ m\neq m'}} \widehat{\Psi}_{X_mX_{m'}}\frac{H^*_m(\mathbf{r})H_{m'}(\mathbf{r})}{\|\mathbf{H}(\mathbf{r})\|^2}\,,\end{aligned} \tag{3.75}$$

where

$$\widehat{\Psi}_{X_mX_{m'}} = \frac{1}{B}\sum_b X_m(b)X^*_{m'}(b) \tag{3.76}$$

is the estimate of the corresponding power spectral density and $\mathbf{H}(\mathbf{r})$ is the steering vector along $\mathbf{r}$. Recognize that (3.75) is fully equivalent to (3.37).

Stochastic maximum likelihood estimation

Stochastic ML assumes that the source and noise waveforms in (3.64) are realizations of a stochastic process with a parametric probability density function (pdf). Usually a zero-mean, complex Gaussian pdf is assumed, leading to the following likelihood function:

$$p\left(\mathbf{X}(b)\,\middle|\boldsymbol{\Theta}\right) = \frac{1}{\pi^M|\boldsymbol{\Psi}_{\mathbf{XX}}|}\exp\left(-\mathbf{X}^H(b)\,\boldsymbol{\Psi}_{\mathbf{XX}}^{-1}\mathbf{X}(b)\right), \tag{3.77}$$

where $\boldsymbol{\Psi}_{\mathbf{XX}} = \mathrm{E}\{\mathbf{X}(b)\,\mathbf{X}^H(b)\} = \mathbf{A}\boldsymbol{\Psi}_{\mathbf{SS}}\mathbf{A}^H + \Psi_{VV}\mathbf{I}$ is the covariance matrix of the microphone signals, and $\boldsymbol{\Psi}_{\mathbf{SS}} = \mathrm{E}\{\mathbf{S}(b)\,\mathbf{S}^H(b)\}$ is the covariance matrix of the source signals. Thus, the likelihood is a function of the source locations $\mathbf{R}_{\mathbf{S}} = (\mathbf{r}_{s_1}, \mathbf{r}_{s_2}, \ldots, \mathbf{r}_{s_Q})$, the source covariance matrix $\boldsymbol{\Psi}_{\mathbf{SS}}$ and the noise power Ψ_{VV}, all of which constitute $\boldsymbol{\Theta}$. Consequently, for B independent observations $\mathbf{X}(b)$, we have:

$$p\left(\mathbf{X}^B\middle|\boldsymbol{\Theta}\right) = \frac{1}{\pi^{MB}|\boldsymbol{\Psi}_{\mathbf{XX}}|^B}\prod_b \exp\left(-\mathbf{X}^H(b)\,\boldsymbol{\Psi}_{\mathbf{XX}}^{-1}\mathbf{X}(b)\right). \tag{3.78}$$

Optimizing the log-likelihood function $\mathcal{L}(\mathbf{X}^B;\boldsymbol{\Theta})$ for the unknown parameters $\boldsymbol{\Theta}$, we obtain the following solution to the stochastic maximum likelihood (SML) problem [61, 114]:

$$\widehat{\Psi}_{VV,\,\mathrm{SML}} = \frac{1}{B(M-Q)}\sum_{b=1}^{B}\mathbf{X}^H(b)\,\mathbf{P}_{\mathbf{H}}^{\perp}\mathbf{X}(b) \tag{3.79}$$

$$\widehat{\boldsymbol{\Psi}}_{\mathbf{SS},\,\mathrm{SML}} = \mathbf{H}^{\dagger}\left(\widehat{\boldsymbol{\Psi}}_{\mathbf{XX}} - \widehat{\Psi}_{VV}\mathbf{I}\right)\mathbf{H}^{\dagger H} \tag{3.80}$$

$$\widehat{\mathbf{A}}_{\mathrm{SML}} = \underset{\mathbf{H}}{\mathrm{argmin}}\ \ln\left(\left|\mathbf{P}_{\mathbf{H}}\widehat{\boldsymbol{\Psi}}_{\mathbf{XX}}\mathbf{P}_{\mathbf{H}} + \widehat{\Psi}_{VV}\mathbf{P}_{\mathbf{H}}^{\perp}\right|\right), \tag{3.81}$$

where $\widehat{\Psi}_{VV,\,\mathrm{SML}}$ is the estimate of the noise power and $\widehat{\boldsymbol{\Psi}}_{\mathbf{XX}}$ is the estimate of the microphone signal covariance matrix with elements $\widehat{\Psi}_{X_mX_{m'}}$ as defined in (3.76). $\mathbf{P}_{\mathbf{H}}$ and $\mathbf{P}_{\mathbf{H}}^{\perp}$ are as defined previously.

3.3.7 Relation between MLE and MUSIC

While at first glance the MUSIC approach seems to be removed from the algorithms studied so far, there exists a link between the deterministic ML and the MUSIC approach which we are now in a position to study. Notice first that equations (3.53) and (3.54) deliver the same results. In the first case, we look for candidate positions whose corresponding propagation vectors are orthogonal to the noise only subspace and in the second case we seek candidate positions that completely lie in the signal+noise subspace, and are thereby orthogonal to the noise only subspace. Also, MUSIC exploits the second order characteristic (as only the covariance matrix of the microphone signals is used) and, in practice, the signal covariance matrix is estimated by temporal averaging:

$$\widehat{\boldsymbol{\Psi}}_{\mathbf{XX}} = \frac{1}{B}\sum_b \mathbf{X}(b)\,\mathbf{X}^H(b)$$

The relation between the MUSIC and the ML approaches may now be obtained as in [138], by formulating (3.54) alternatively as:

$$\begin{aligned}\mathbf{r}_{s_q} &= \underset{\mathbf{r}}{\text{argmax}}\ \text{tr}\left(\mathbf{U}_S^H \mathbf{H}(\mathbf{r})\mathbf{H}^H(\mathbf{r})\mathbf{U}_S\right) \\ &= \text{tr}\left(\mathbf{U}_S^H \mathbf{P}_{\mathbf{H}(\mathbf{r})}\mathbf{U}_S\right) \\ &= \sum_{m=1}^{Q} \|\mathbf{P}_{\mathbf{H}(\mathbf{r})}\mathbf{U}_{m,S}\|^2, \quad m = 1 \ldots Q,\end{aligned} \tag{3.82}$$

where:

$$\mathbf{P}_{\mathbf{H}(\mathbf{r})} = \frac{\mathbf{H}(\mathbf{r})\mathbf{H}^H(\mathbf{r})}{\|\mathbf{H}(\mathbf{r})\|^2},$$

is the projector onto the space defined by $\mathbf{H}(\mathbf{r})$, $\mathbf{U}_{m,S}$ is the mth column vector of $\mathbf{U}_S$ and $\text{tr}(\cdot)$ is the trace operator. We may also write this for all the Q sources as:

$$\widehat{\mathbf{A}} = \underset{\mathbf{H}}{\text{argmax}} \sum_{m=1}^{Q} \|\mathbf{P}_{\mathbf{H}}\mathbf{U}_{m,S}\|^2, \quad m = 1 \ldots Q\,. \tag{3.83}$$

Thus, the matrix $\mathbf{H}$ onto the column space of which the projection of the signal+noise subspace is maximum, is the optimal solution. Note that only the eigenvectors are considered and the eigenvalues do not play a role, except for the purpose of demarcating the noise subspace.

Compare the result in (3.83) with that obtained by the deterministic ML approach in (3.72), which is repeated here for convenience:

$$\begin{aligned}\widehat{\mathbf{A}}_{\text{DML}} &= \underset{\mathbf{H}}{\text{argmax}} \sum_{b} \mathbf{X}^H(b)\, \mathbf{P}_{\mathbf{H}}\mathbf{X}(b) \\ &= \underset{\mathbf{H}}{\text{argmax}}\ \text{tr}\left(\mathbf{P}_{\mathbf{H}}\widehat{\mathbf{\Psi}}_{\mathbf{XX}}\right) \\ &= \sum_{m=1}^{M} \lambda_m \|\mathbf{P}_{\mathbf{H}}\mathbf{U}_m\|^2,\end{aligned} \tag{3.84}$$

where λ_m is the mth eigenvalue of $\widehat{\mathbf{\Psi}}_{\mathbf{XX}}$, with the corresponding eigenvector $\mathbf{U}_m$. Thus, we see that all eigenvectors contribute to the cost function, their contribution being proportional to their eigenvalue, in contrast to (3.83) where only the Q dominant eigenvectors are considered with equal weighting. Thus, the deterministic ML approach may be considered as a 'soft-decision' version of the MUSIC algorithm. Indeed it can been shown that the error variance of MUSIC approaches that of deterministic ML [113].

3.4 Evaluation of localization algorithms

As may be gleaned from the discussions above, most localization approaches utilize only the second-order statistics of the microphone signals and are closely related. This section presents the behavior of representative algorithms – for both indirect and direct methods – in reverberant and noisy situations. While the purpose of this section is not to

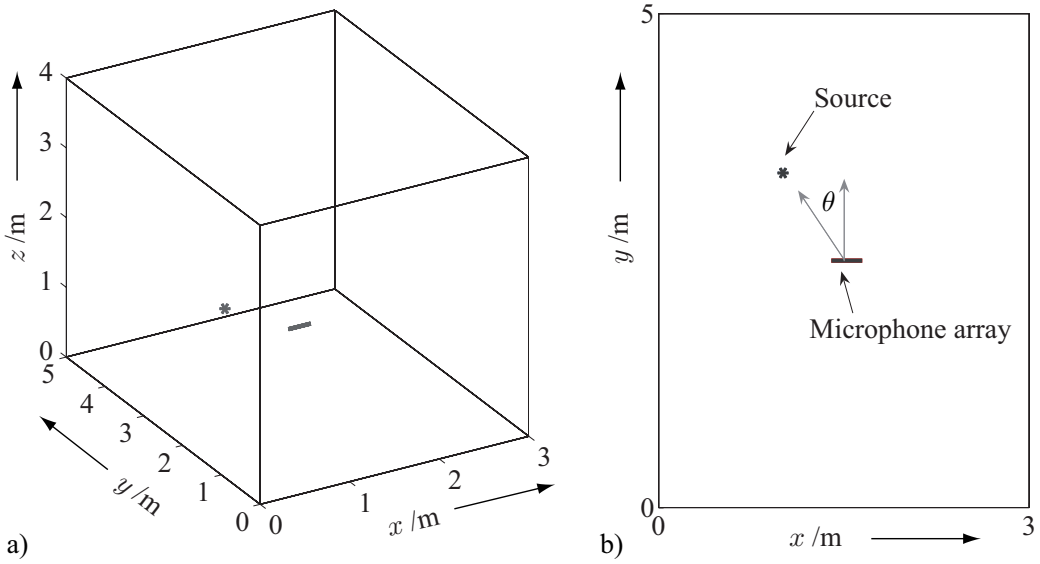


Figure 3.3: Overview of the simulation setup
a) 3-D image of the simulated room
b) Top-view of the simulated room ($\theta = \pi/6$)

perform an exhaustive comparison of the various methods, the advantages and disadvantages of the algorithms will be mentioned where appropriate. The algorithms considered are:

indirect methods:	GCC-PHAT, CC, LMS, AED
direct methods:	SRP-PHAT, MUSIC.

The localization experiments were carried out in a room simulated using the image method [3, 48]. The room dimensions were 3 m × 5 m × 4 m. A microphone array with $M = 5$ microphones was used. The microphones were placed linearly at distances of 3 cm, 8 cm, 15 cm, and 25 cm respectively from the first microphone and the array center was located at $\mathbf{r} = (1.5, 2.5, 1.0)^T$ m. The source was placed at $\mathbf{r}_s = (1.0, 3.366, 1.0)^T$ m, see Fig. 3.3.

Further, the simulations were carried out for three different reverberation times $\mathrm{T}_{60} \in \{0.12, 0.3, 0.5\}$ s, with corresponding critical radii of $d_{\mathrm{crit}} \in \{1.27, 0.81, 0.62\}$ m, and for each case, under three different signal-to-noise ratios $\mathrm{SNR} \in \{-5, 5, 15\}$ dB. The noise was white, spatially uncorrelated and independent of the source signal. The sampling frequency was $f_\mathrm{s} = 8000$ Hz. For the indirect methods, the two outermost microphones were selected with a resultant inter-microphone distance of $d_{\max} = 25$ cm.

For the GCC-PHAT and the simple cross-correlation approach, the required power spectral densities were estimated in the STFT domain by a first-order recursive, temporal smoothing with a smoothing constant η as

$$\Psi_{X_m X_{m'}}(k,b) = \eta\, \Psi_{X_m X_{m'}}(k,b-1) + (1-\eta)\, X_m(k,b) X^*_{m'}(k,b), \qquad m, m' \in \{1,2\}\,. \tag{3.85}$$

The generalized cross-correlation function is then obtained for each frame b by the inverse STFT of $\Psi_{X_m X_{m'}}(k,b)$.

For the direct methods considered, the search grid was defined in two ways: over the two-dimensional grid defined in the x–y plane and over a one-dimensional grid computed over the

Table 3.2: Algorithm parameters used in the simulations

Algorithm	Filter length (ms)	Window type/ length (ms)	DFT length (ms)	Frame shift (ms)	ς	η
CC / GCC-PHAT	N/A	Hamming/64	128	32	N/A	0.90
LMS	128	Rectangular/128	N/A	0.125	0.005	N/A
AED	128	Rectangular/64	N/A	0.125	0.005	N/A
SRP-PHAT	N/A	Hamming/128	128	64	N/A	N/A
MUSIC	N/A	Hamming/128	128	64	N/A	N/A

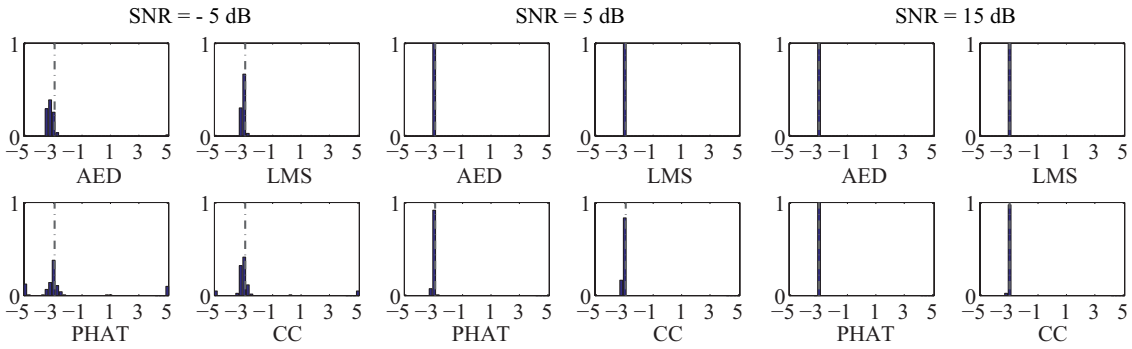
azimuth, measured with respect to the array axis from the array mid-point. For each case, the cost function was computed per frame as detailed in Section 3.3.3 for the SRP approach. For the MUSIC approach, the spectrum evaluated over the complete speech signal (5 s duration) was used to build the statistics upon which the estimate of the noise-only subspace $\mathbf{U}_{\mathcal{V}}(k)$ was computed for each frequency bin k. An estimate for the source location was then obtained over the respective 2-D or 1-D grid in each bin.

The parameters for the various methods are summarized in Table 3.2.

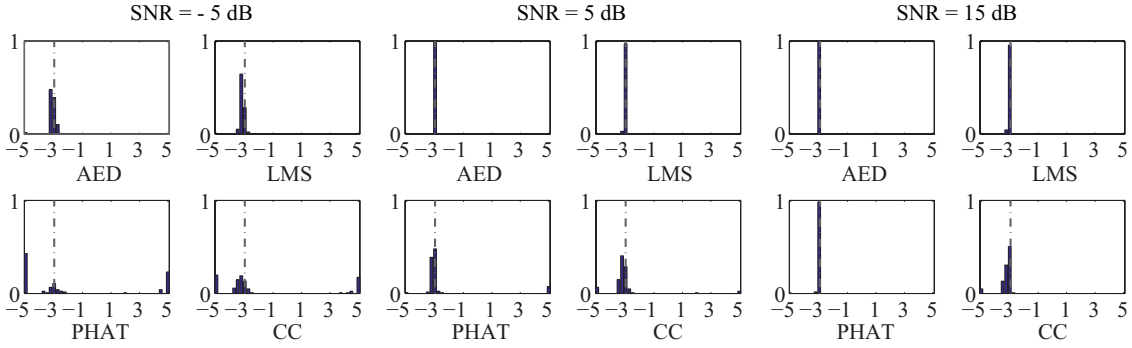
3.4.1 Performance of the indirect methods

With the chosen positions for the array and the source, the time difference of the direct path between the two outermost microphones corresponds to about three sampling intervals at 8000 Hz. Figure 3.4 depicts the histograms of estimated time delays of arrival between the microphones for the different simulation conditions. The histogram data was accumulated over 120 estimates, *after* the adaptive algorithms had converged. The delay axis is limited to the range of $[-5, 5]$ sampling intervals, as the maximum possible delay for the array configuration was about 5.8 sampling intervals. In all plots, the dotted line indicates the true time difference between the direct paths which was obtained from the room impulse response for each microphone. It may be seen that the performance of the algorithms improves with an increase in the SNR, which is to be expected. Under low reverberation conditions, the estimate of the time delay is almost perfect at high SNRs. However, as the reverberation increases, a spread may be observed about the true value.

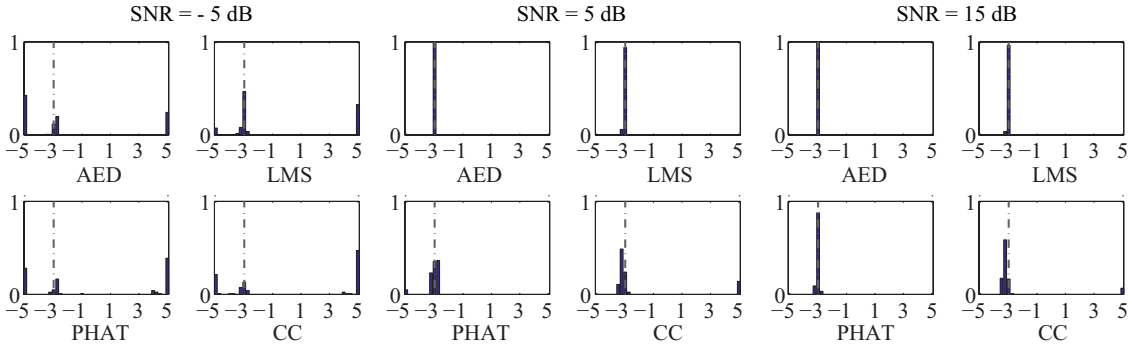
The simple cross-correlator (CC) has the worst performance amongst the four. This could be explained as follows: speech signals possess maximum energy at the lower frequencies. The CC algorithm applies no explicit weighting to the frequency bins and thus implicitly weights each frequency by the energy of the received signal in that frequency bin. Therefore, low frequencies are emphasized and higher frequency contributions are damped in this method. Conversely, time delay information is less accurate at low frequencies; more so in the presence of noise. Additionally, in reverberant conditions, it is the reflections at the higher frequencies that are damped to a greater extent than the lower frequencies. As GCC-PHAT, GCC-Roth, etc., on the other hand, remove this emphasis on the lower frequencies, they lead to an improvement in performance. This is noticeable especially in reverberant environments [47] where the higher frequencies – which are less affected by reverberation – contribute more to the delay estimate.



a) Reverberation time of $T_{60} = 0.12\,\mathrm{s}$.



b) Reverberation time of $T_{60} = 0.3\,\mathrm{s}$.



c) Reverberation time of $T_{60} = 0.5\,\mathrm{s}$.

Figure 3.4: Time delay estimates for different reverberation times and SNRs.

Consider now the performance of the functionally similar AED and LMS algorithms. Under low SNR conditions, the AED approach performs worse in comparison to the LMS approach. This could be the cumulative effect of errors on the individual filter estimates of the AED, as the noise is uncorrelated across the microphones. Under high SNR conditions, the respective performances of the AED and the LMS approaches are quite similar, with a slight advantage to the AED when the reverberation increases. We postulate that the added flexibility of the AED, in terms of two simultaneously adaptable filters (one for each microphone) helps it to converge better to the true delay difference under these conditions, as compared with the LMS approach. Consequently, in highly reverberant environments, it might be a good idea to use the AED when the SNR is high. When the SNR decreases however, it might be better to switch to the LMS algorithm.

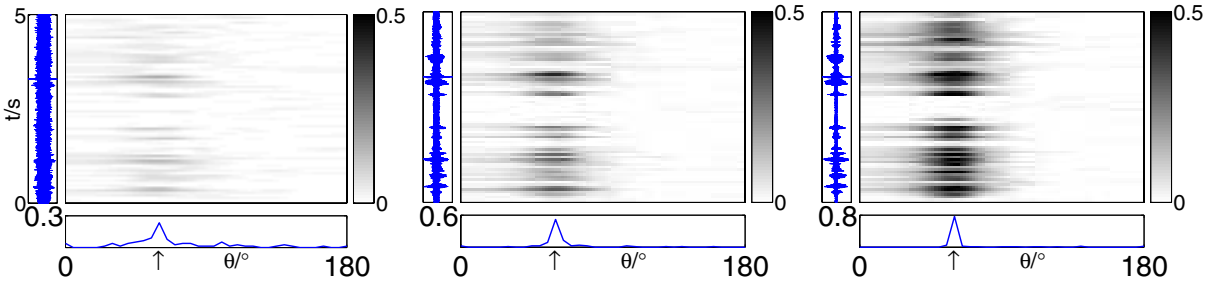
Another factor affecting the performance of these algorithms is the smoothing factor η for the GCC-PHAT/CC approaches and the step-size ς for the LMS/AED approaches. We find that increasing η improves the performance of the GCC-PHAT/CC approaches bringing

them – especially GCC-PHAT – close to the performance of the LMS/AED approaches. Similarly, a lower step-size ς improves the robustness of the LMS/AED approaches against noise, but the convergence is slower. In general, the step-size ς and the smoothing parameter η could be made adaptive: one could use a larger value of ς (and a lower value of η) in high SNR environments and a lower value for ς (and a larger smoothing constant η) when the noise level increases.

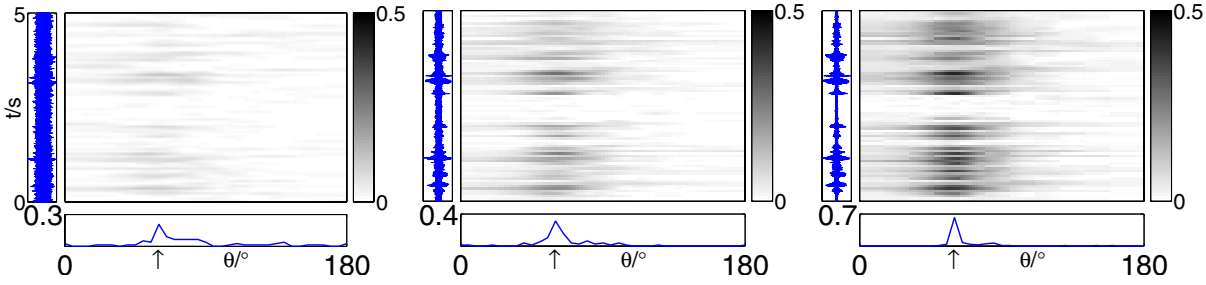
3.4.2 Performance of the direct methods

This section deals with the behavior of the SRP-PHAT and the MUSIC algorithms under the same conditions as for the indirect methods, except for the number of microphones. The direct approaches use all five microphones of the array. The performance of the SRP-PHAT algorithm will be discussed first, followed by the MUSIC approach.

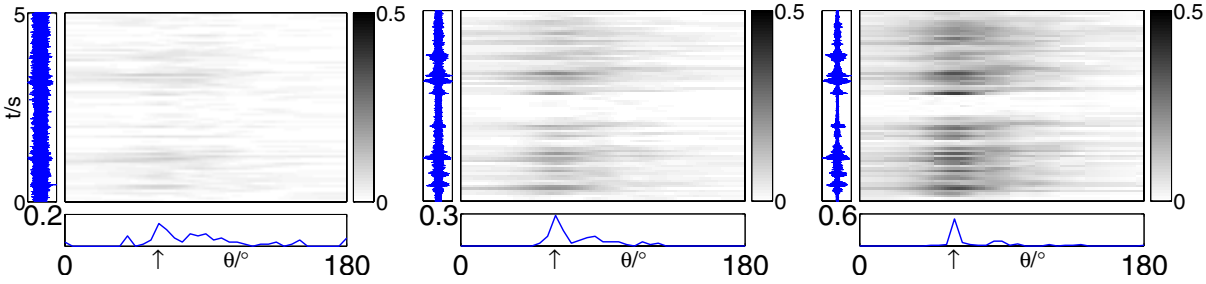
SRP approach



a) Reverberation time of $T_{60} = 0.12\,\mathrm{s}$.



b) Reverberation time of $T_{60} = 0.3\,\mathrm{s}$.



c) Reverberation time of $T_{60} = 0.5\,\mathrm{s}$.

Figure 3.5: $\mathcal{J}_{\mathrm{SRP}}(\theta, b)$ estimates for different reverberation times and SNRs.

Figure 3.5 depicts the cost function $\mathcal{J}_{\mathrm{SRP}}(\theta, b)$ computed according to (3.46) over the azimuth. The x-axis indicates the candidate azimuth angles, the y-axis indicates the time frame (in seconds) and the intensity of a point at any set of co-ordinates is a measure of

the cost function value for that time frame and that candidate location. For the given source-array constellation, the azimuth angle is $\pi/3$ (or 60°). This forms the 'ground-truth' for comparison and is indicated by an $\uparrow$ along the x-axis in all the plots. The time-domain signal at the first microphone is also plotted parallel to the y-axis. The plot beneath the cost function indicates the localization performance in terms of a histogram of the individual estimates computed for each frame b using (3.47). This illustrates both the spread in the estimates and the localization errors.

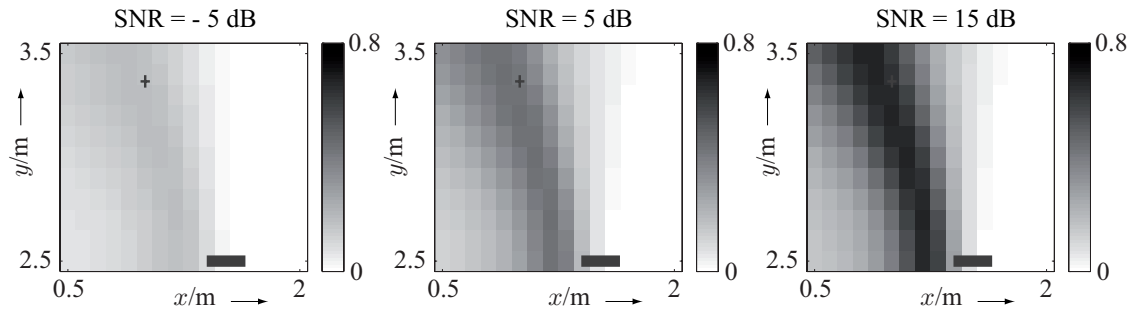
Three trends are clearly perceivable in all the plots: first, the range of the cost function values increases with the SNR – this is to be expected because, as the noise level decreases, the cost function depends increasingly upon the incident signal and yields a high value only at the true azimuth as perfect phase alignment is obtained among all microphone pairs. When the SNR is low, the SRP cost function gets smeared due to coincidental phase alignments at different azimuths, between different microphone pairs, reducing the range of values. Secondly, a good estimate to the azimuth is obtainable even at low SNRs due to the larger spatial diversity available as compared with the TDE approaches considered in the previous section. Thirdly, notice the broadening of the cost function peaks as reverberation increases – this is to be expected as, due to the multipath propagation, partial phase alignments are possible along a wider range of search locations. However, in terms of localizing the source, the performance does not seem to significantly deteriorate with increasing reverberation, as long as the direct path is dominant.

As an illustration of the capabilities of the SRP algorithm, the cost function is evaluated over a two dimensional grid along the x–y plane and is given in Fig. 3.6 for a sample time frame b. The true position of the source is indicated by a $+$. The grey bar at the bottom of each plot represents the microphone array.

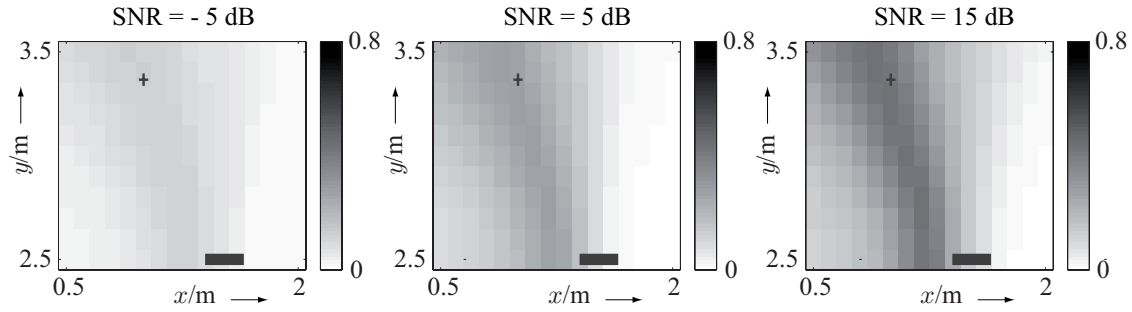
Physical considerations dictate that, for a linear array with a relatively small aperture, it is difficult to obtain both the range and the azimuth of the source when the source is in the farfield. This can be seen in the plots. Note that there is a broad range of co-ordinates with similar cost function values lying along a half-hyperbola centered about the array axis. This is sometimes termed the *cone of confusion* – sources lying anywhere on this cone would generate the same phase deviations at the microphones in the absence of reverberation and noise and it is difficult to pinpoint the location of the source on this cone without additional information, e.g., from a second array mounted perpendicular to the first. The trends with respect to SNR and reverberation observed in the previous case may be perceived here too.

MUSIC approach

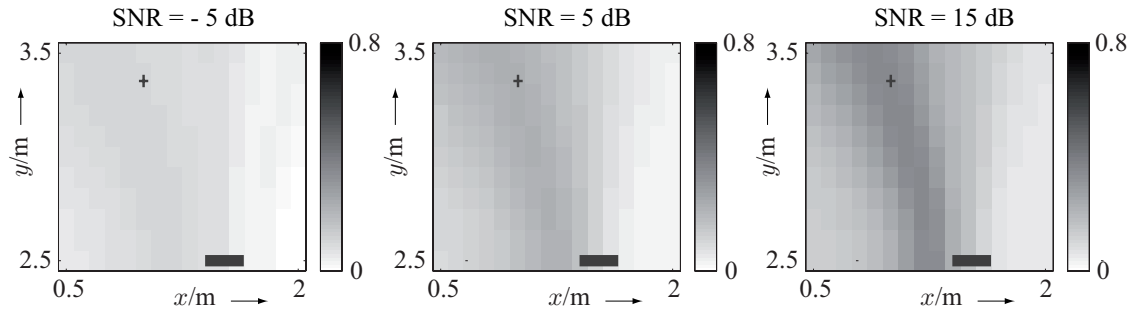
As mentioned before, the MUSIC approach first estimates the noise-only subspace in each frequency bin from the power spectral density matrix. The latter is obtained, in practice, from a temporal averaging of the signal spectrum. This averaging might be either recursive (in which case, MUSIC may be used to yield a source location estimate in each frame b) or block based (in which case the location estimates are computed on a batch basis). The first step lies in determining the number of signals present in the system. This is done either by comparing the values of the eigenvectors or by the application of information-theoretic criteria as described in [127]. Once this is done, the identification of the noise-only subspace is performed as described in Sec. 3.3.4. For the simulations, the $\mathbf{\Psi_{XX}}(k)$ matrix was obtained as a temporal average over the spectrum of the complete speech signal. Figure 3.7 showcases the performance of the MUSIC algorithm over the 1-D search grid.



a) Reverberation time of $T_{60} = 0.12\,\mathrm{s}$.



b) Reverberation time of $T_{60} = 0.3\,\mathrm{s}$.

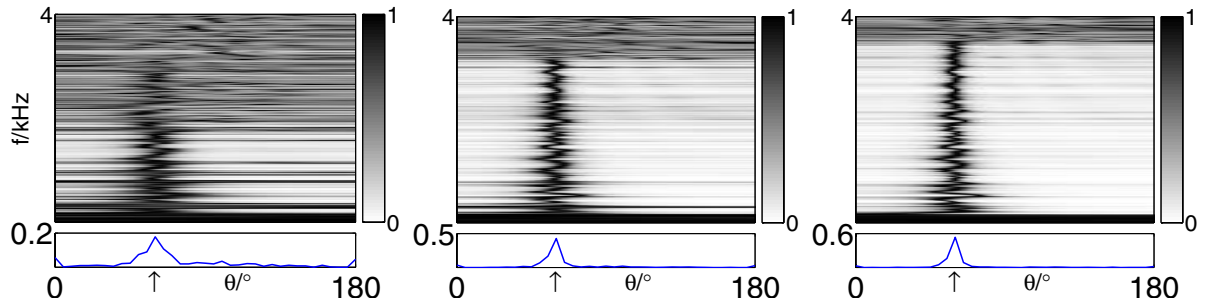


c) Reverberation time of $T_{60} = 0.5\,\mathrm{s}$.

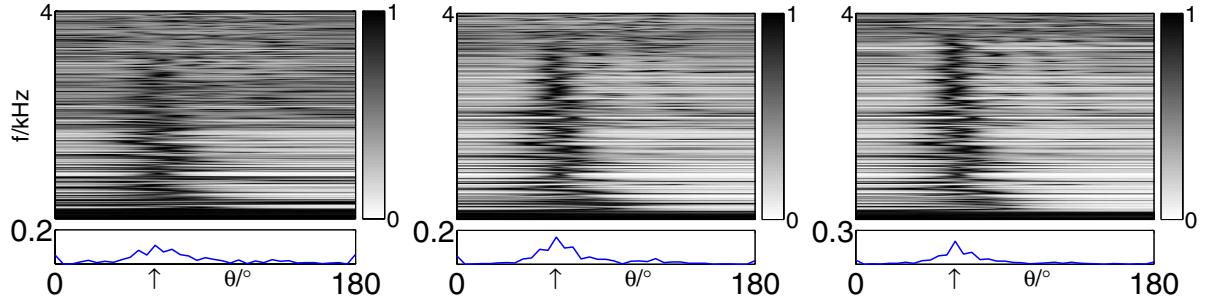
Figure 3.6: $\mathcal{J}_{\mathrm{SRP}}(\mathbf{r}, b)$ estimates for different reverberation times and SNRs.

The plots indicate the MUSIC spectrum values (normalized to a maximum of 1) for each frequency (plotted along the y-axis). Note the lack of directivity at frequencies below 200 Hz. In these bands, it is difficult to obtain an estimate of the source position due to the infinitesimal phase difference between the microphone signals. As the frequencies increase, the MUSIC spectrum spread narrows down – corresponding to increasing directivity at higher frequencies. As expected, the performance improves with an increase in the SNR. Further, as the room becomes more reverberant, the MUSIC spectrum begins to spread out across the azimuth – again, an effect that is to be expected due to the correlated multipath propagation. The plot below the cost function depicts the localization histogram similar to the previous section, only in this case the histogram is computed over the localization estimates for each *frequency bin*.

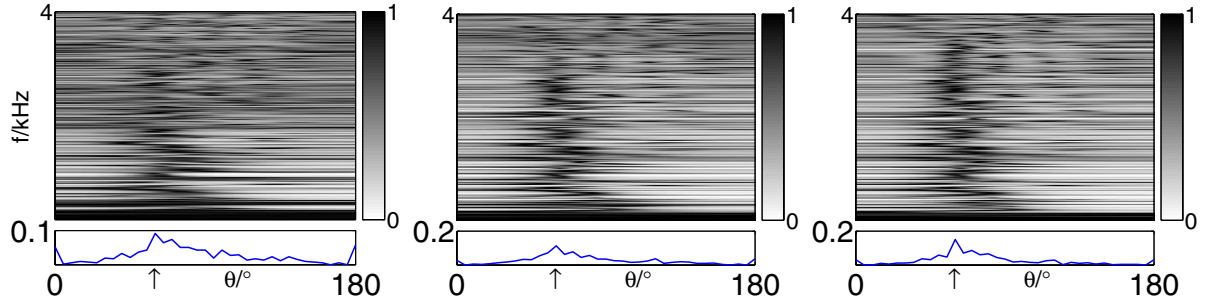
As in the case of the SRP, a more robust estimate of the source location may be found, for broadband sources, by averaging the normalized MUSIC spectrum along the source



a) Reverberation time of $T_{60} = 0.12\,\mathrm{s}$.



b) Reverberation time of $T_{60} = 0.3\,\mathrm{s}$.



c) Reverberation time of $T_{60} = 0.5\,\mathrm{s}$.

Figure 3.7: $\mathcal{J}_{\mathrm{MUSIC}}(\theta, b)$ estimates for different reverberation times and SNRs.

bandwidth and then finding the maximum as

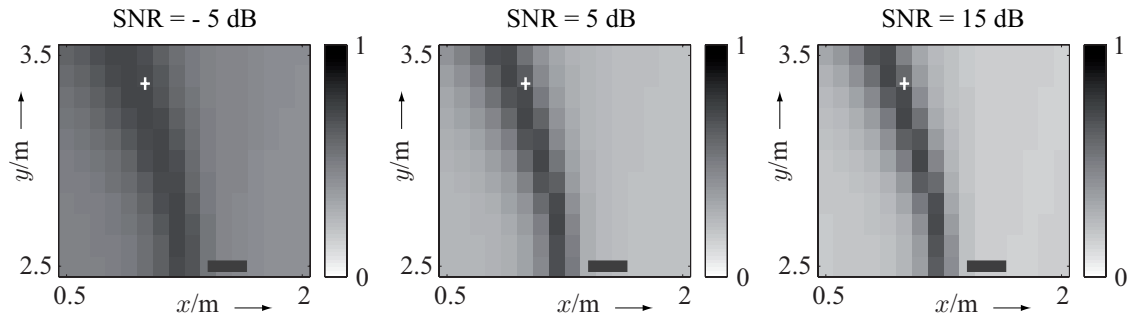
$$\begin{aligned} \mathcal{J}_{\mathrm{MUSIC}}(\theta) &= \sum_k \frac{\mathcal{J}_{\mathrm{MUSIC}}(\theta, k)}{\max\limits_{\theta} \left(\mathcal{J}_{\mathrm{MUSIC}}(\theta, k)\right)} \\ \widehat{\theta}_s &= \operatorname*{argmax}_{\theta} \ \mathcal{J}_{\mathrm{MUSIC}}(\theta) \end{aligned} \tag{3.86}$$

when evaluating over a one-dimensional grid, or

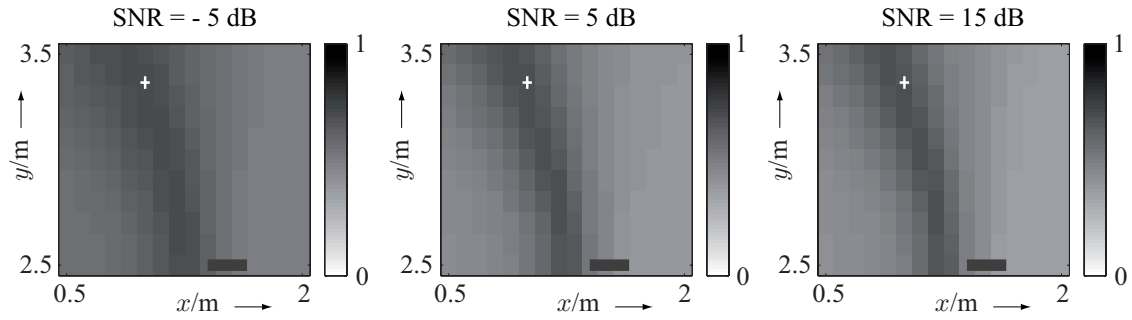
$$\begin{aligned} \mathcal{J}_{\mathrm{MUSIC}}(\mathbf{r}) &= \sum_k \frac{\mathcal{J}_{\mathrm{MUSIC}}(\mathbf{r}, k)}{\max\limits_{\mathbf{r}} \left(\mathcal{J}_{\mathrm{MUSIC}}(\mathbf{r}, k)\right)} \\ \widehat{\mathbf{r}}_s &= \operatorname*{argmax}_{\mathbf{r}} \ \mathcal{J}_{\mathrm{MUSIC}}(\mathbf{r}) \end{aligned} \tag{3.87}$$

over a two-dimensional grid. This leads to the incoherent MUSIC approach. Alternatively, one may choose to weight the estimates in each frequency band before averaging.

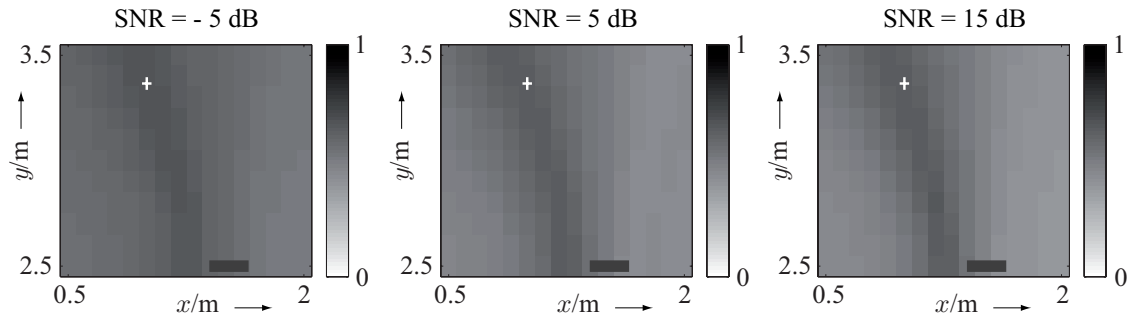
The MUSIC spectrum evaluated over a two-dimensional grid is illustrated in Fig. 3.8. Notice, again, the cone of confusion and the broadening of this cone with increasing reverberation.



a) Reverberation time of $T_{60} = 0.12\,\mathrm{s}$.



b) Reverberation time of $T_{60} = 0.3\,\mathrm{s}$.



c) Reverberation time of $T_{60} = 0.5\,\mathrm{s}$.

Figure 3.8: $\mathcal{J}_{\mathrm{MUSIC}}(\mathbf{r}, k)$ estimates for different reverberation times and SNRs.

As before, the grey bar at the bottom of each plot indicates the microphone array and the + the source position.

3.4.3 The two-source case

The last simulation presented here illustrates the performance of the localization algorithms for a multi-source case. In addition to the first source, we consider another source located at $\mathbf{r}_{s_2} = (2.0, 3.366, 1.0)^T$ m, with a corresponding time delay of -3 samples or an azimuth of $2\pi/3$. The two sources are simultaneously active. The room is simulated with a $T_{60} = 0.12\,\mathrm{s}$ and the SNR is 35 dB.

From Fig. 3.9 it is evident that the indirect methods as discussed here make a single source assumption and thus find it difficult to cope with the multi-speaker scenario. There are, however, methods that locate multiple speaker time delays by examining the secondary peaks of the cross-correlation function, e.g., [105, 106, 10]. SRP and MUSIC, on the other hand, can exploit the additional freedom of different frequency bins to localize multiple speakers even within a single time frame.

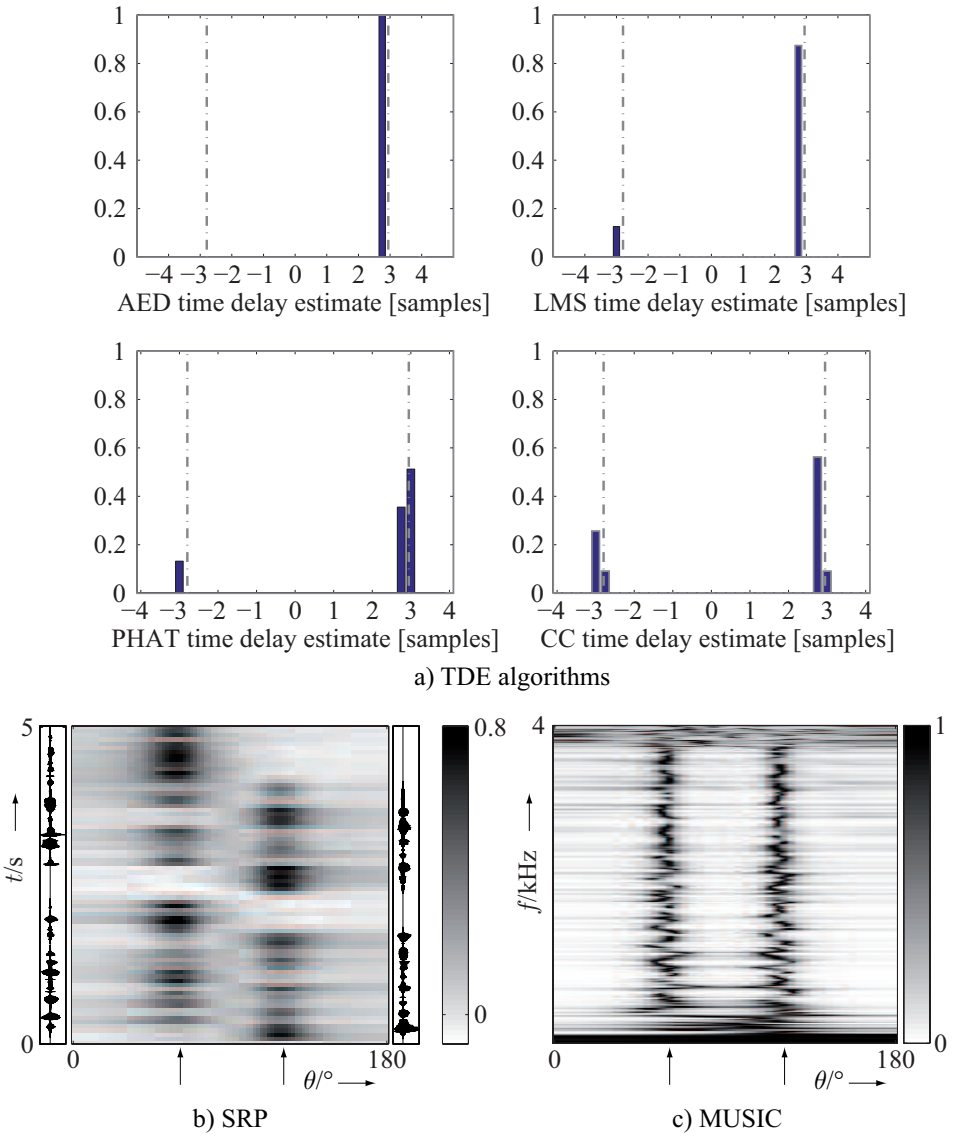


Figure 3.9: Performance in a multi-source scenario at 35 dB SNR.

3.5 Localization as a detection problem

In the previous sections we have examined localization approaches from the point of view of estimating the position of the acoustically active source(s). In some applications knowledge of the nominal source positions may be available *a priori*. In such a case, given the presence of an acoustic event (broadband or narrowband), the requirement is to attribute the event to the source(s) that generated it. In the following section we shall present solutions to this problem, that are based on the localization model.

While we shall restrict ourselves in the subsequent discussion to the case of only two principal sources for the purpose of brevity, the formulation may be generalized to any number of individual sources and combinations thereof.

Detection model

Consider an array of M microphones in the field of two principal sources whose approximate positions are known. We also assume that the sources emit signals that are narrowband, and at the same center frequency. Further, at a given time, any, both or neither of the

sources can be 'active'. Under these assumptions, when we consider a time record of B frames

$$\mathbf{X}^B(k) = \big(\mathbf{X}(k,1),\, \ldots,\, \mathbf{X}(k,B)\big)$$

of the signals received at the array, at frequency bin k, we may construct the following hypotheses along with their corresponding narrowband signal models at the sensors:

$\mathcal{H}_0$: neither source is active:

$$\mathbf{X}(k,b) = \mathbf{V}(k,b), \tag{3.88}$$

$\mathcal{H}_1$: source 1, with associated propagation vector $\mathbf{A}_1 \in \mathbb{C}^{M\times 1}$, is active:

$$\mathbf{X}(k,b) = \mathbf{A}_1(k)\, S_1(k,b) + \mathbf{V}(k,b), \tag{3.89}$$

$\mathcal{H}_2$: source 2, along $\mathbf{A}_2 \in \mathbb{C}^{M\times 1}$, is active:

$$\mathbf{X}(k,b) = \mathbf{A}_2(k)\, S_2(k,b) + \mathbf{V}(k,b), \tag{3.90}$$

$\mathcal{H}_3$: both sources are active:

$$\begin{aligned}\mathbf{X}(k,b) &= \sum_{q=1}^{2} \mathbf{A}_q(k)\, S_q(k,b) + \mathbf{V}(k,b),\\ &= \mathbf{A}_3\mathbf{S}(k,b) + \mathbf{V}(k,b),\end{aligned} \tag{3.91}$$

where $\mathbf{A}_3 = (\mathbf{A}_1\, \mathbf{A}_2)$ and $\mathbf{S}(k,b) = (S_1(k,b),\, S_2(k,b))^T$.

In the above equations, $\mathbf{V}(k,b)$ represents, as before, the vector of noise signals at the microphones and $S_q(k,b)$ the source. Further, if we consider that the sources are not point sources but distributed in space, the corresponding signal propagation vector for a source may be defined as:

$$\mathbf{A}_q(k) = \int_{\mathcal{R}_q} \mathcal{W}(\mathbf{r}_q)\mathbf{A}_q(\mathbf{r}_q,k)\, \mathrm{d}\mathbf{r}_q, \tag{3.92}$$

where $\mathcal{R}_q$ is the region associated with source q, $\mathbf{A}_q(\mathbf{r}_q,k)$ is the propagation vector associated with the position $\mathbf{r}_q$, and $\mathcal{W}(\mathbf{r}_q)$ is the spatial weight or *importance window* function associated with the position $\mathbf{r}_q$ for source q. For point sources, for example, this function may be expressed by a dirac impulse at the position associated with the source and zero elsewhere.

3.5.1 Hypothesis testing

The idea here is to decide between the four hypotheses described in Section 3.5, for which we adopt a sequential hypothesis testing approach (see, e.g., [66, 79], for more details). In short, given an appropriate score function definition,

1. Compute the score for the null hypothesis $\mathcal{H}_0$ (no source is present).
2. Compute the score for the alternative hypothesis $\mathcal{H}_1'$, namely at least one source is present, i.e., $\mathcal{H}_1$ or $\mathcal{H}_2$ holds. The source position is selected as the one that yields the best score. Denote the corresponding steering vector as $\mathbf{A}_q$ and the source as $q,\ q \in \{1,2\}$.

3. Compare $\mathcal{H}_0$ and $\mathcal{H}_1'$ based on the selected score function and decide if the null hypothesis is to be accepted or rejected.
4. If $\mathcal{H}_0$ is accepted, conclude that no sources are active $\boxed{\rightarrow \mathcal{H}_0 \text{ is } \texttt{True.}}$
5. If the null hypothesis is *rejected*, we conclude that *at least* source q is active. In this case,
   a) Set new null hypothesis $\mathcal{H}_0'$: only q is active (either $\mathcal{H}_1$ or $\mathcal{H}_2$),
   b) Set new alternative hypothesis $\mathcal{H}_1''$: both sources, S_1 and S_2, are active (i.e., $\mathcal{H}_3$ holds).
6. Compute the score functions for these new hypotheses and compare them as in step 3.
7. If $\mathcal{H}_0'$ is accepted, conclude that only source S_q is active $\boxed{\rightarrow \mathcal{H}_q \text{ is } \texttt{True.}}$
8. If the null hypothesis is *rejected*, we conclude that *both* S_1 and S_2 are active and have contributed to the acoustic event in the examined time record, at frequency k $\boxed{\rightarrow \mathcal{H}_3 \text{ is } \texttt{True.}}$

Note that such a sequential hypothesis test can be extended to the detection of any number of sources in general. Only, when defining the alternative hypothesis for more than one source being present, we select, each time, that new source as active whose *inclusion* in the existing set of active sources yields the best test score.

This approach is illustrated in the examples below, where we have considered the noise to be independently and identically distributed (i.i.d) with an underlying Gaussian pdf at all the sensors. We further assume the source signals are mutually independent of one another and of the noise, and also possess an underlying Gaussian distribution.

Hypothesis testing based on log-likelihood

The score function chosen in this case is the log-likelihood defined on the stochastic model of (3.78) as:

$$\mathcal{L}(\mathbf{X}^B; \boldsymbol{\Theta} | \mathcal{H}, \mathbf{A}) = \ln \left(p \left(\mathbf{X}^B | \boldsymbol{\Theta}, \mathcal{H}, \mathbf{A} \right) \right), \tag{3.93}$$

with the difference that the propagation matrix $\mathbf{A}$ is known for a given hypothesis. Further, under our assumptions on the signals, the parameter vector $\boldsymbol{\Theta}$ for the different hypotheses may be summarized as in Table 3.3, wherein $\mathcal{P}$ represents the number of free parameters (parameters to be estimated) for each hypothesis. The parameter estimates for each hypothesis are obtained using (3.79) and (3.80)

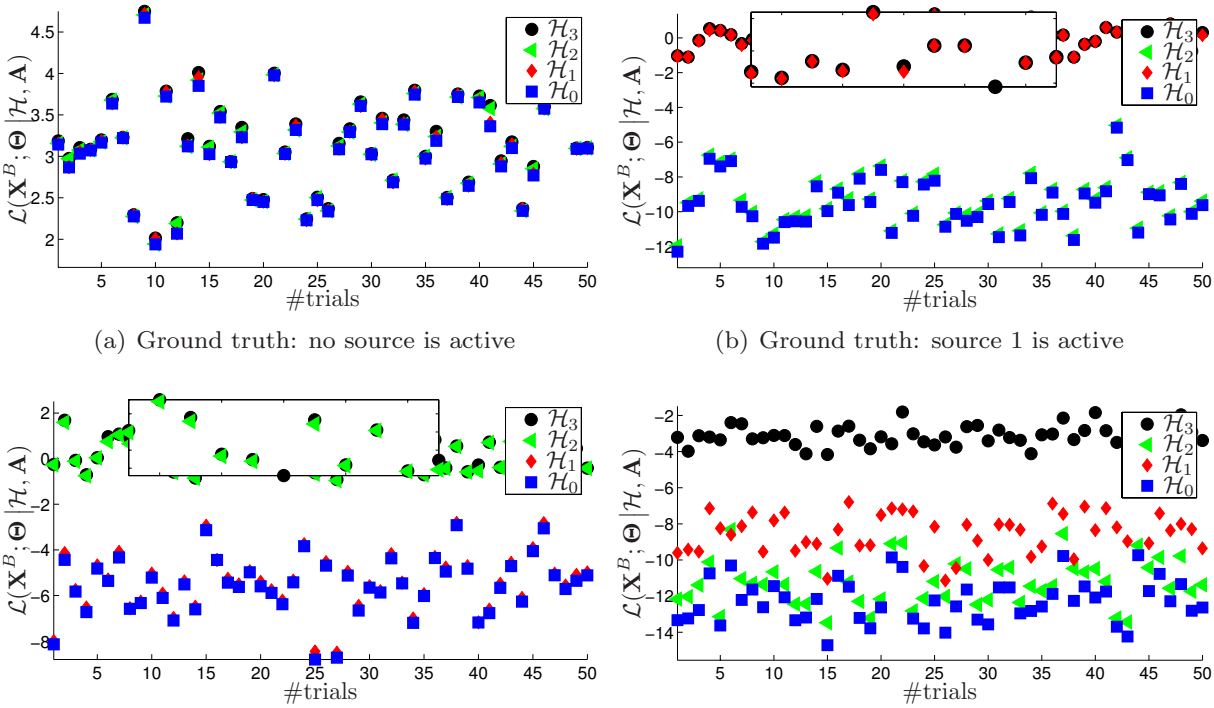
Table 3.3: Free parameters for each hypothesis.

$\mathcal{H}_i$	$\mathcal{P}$	$\boldsymbol{\Theta}_i$
$\mathcal{H}_0$	1	$\boldsymbol{\Theta}_0 = \Psi_{VV}$
$\mathcal{H}_1$	2	$\boldsymbol{\Theta}_1 = (\Psi_{VV},\, \Psi_{S_1S_1})^T$
$\mathcal{H}_2$	2	$\boldsymbol{\Theta}_2 = (\Psi_{VV},\, \Psi_{S_2S_2})^T$
$\mathcal{H}_3$	3	$\boldsymbol{\Theta}_3 = (\Psi_{VV},\, \Psi_{S_1S_1},\, \Psi_{S_2S_2})^T$

The performance of the likelihood-based approach is indicated in Figure 3.10, for each hypothesis, with the parameters for the simulation as in Table 3.4, where $\Delta\theta$ indicates the azimuthal spread of the source, i.e., $\mathcal{R}_q \in [\theta_q \pm \Delta\theta]$, with a spatial weighting provided by a normalized von Hann window centered around the source azimuth.

Table 3.4: Parameters for the hypothesis test simulations.

f_s (kHz)	B	k (kHz)	K	M	θ_1 (°)	θ_2 (°)	$\Delta\theta$ (°)	$\Psi_{S_1S_1}$	$\Psi_{S_2S_2}$	Ψ_{VV}
32	22	2	1024	8	45	135	5	1	0.5	0.25



(a) Ground truth: no source is active

(b) Ground truth: source 1 is active

(c) Ground truth: source 2 is active

(d) Ground truth: both sources are active

Figure 3.10: Log-likelihood values for the different hypotheses, based on the maximum likelihood estimate of the parameters. The results are presented for 50 Monte Carlo trials.

For the case where no source is active (Figure 3.10(a)), all hypotheses yield similar $\mathcal{L}$-scores. Thus, the null hypothesis will be accepted. For the case where only one source is active (Figures 3.10(b)–3.10(c)), the corresponding $\mathcal{L}$-score is significantly higher than for $\mathcal{H}_0$, which will consequently be rejected in step 3. When comparing the new null hypothesis ($\mathcal{H}_0' \triangleq \mathcal{H}_1$ or $\mathcal{H}_2$) and the new alternative hypothesis ($\mathcal{H}_1'' \triangleq \mathcal{H}_3$), we see that the scores for $\mathcal{H}_0'$ and $\mathcal{H}_1''$ are very similar. This is because the two source case can be thought of as being always true, with the second source having very low power. Here, the null hypothesis $\mathcal{H}_0'$ will be *accepted*. For the case where both sources are active (Figure 3.10(d)), $\mathcal{H}_0$ will be rejected in favor of $\mathcal{H}_1'$, which will, subsequently, be rejected in favour of $\mathcal{H}_1''$.

The issue now is to define the comparison criterion. One way is to compute the *distribution* of an appropriate function (difference or ratio) of the likelihood values for the null and alternative hypotheses, and make a decision based on this distribution and a predefined significance level. Alternatively, we may estimate the distributions of the *individual* likelihood values and base our decision on these distributions.

We shall examine the latter alternative, where the distributions $p\left(\mathcal{L}\middle|\mathcal{H}_i\right)$ are estimated for each hypothesis i by *bootstrapping* [139] the corresponding $\mathcal{L}$ values for *each* time frame. An

example of such a distribution is depicted in Figure 3.11.

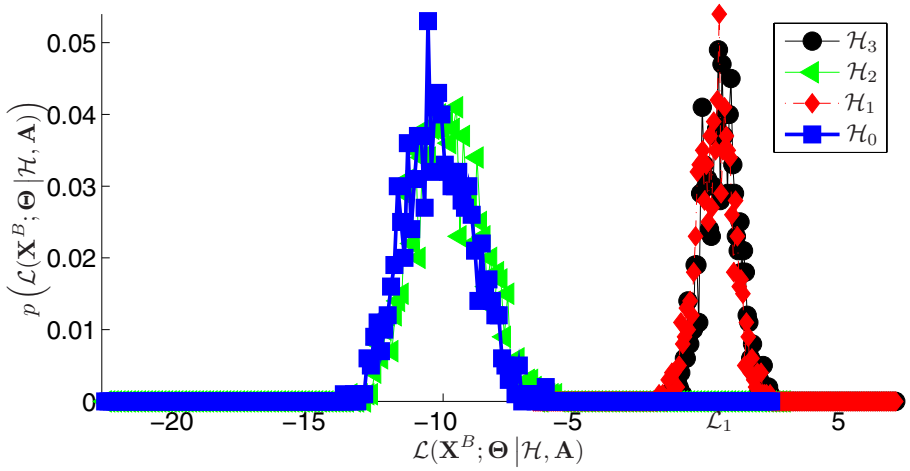


Figure 3.11: Sample bootstrap distribution on the $\mathcal{L}$ values for the case where only source 1 is active.

Then, with respect to this figure, to decide between $\mathcal{H}_1$ and $\mathcal{H}_0$, we compute

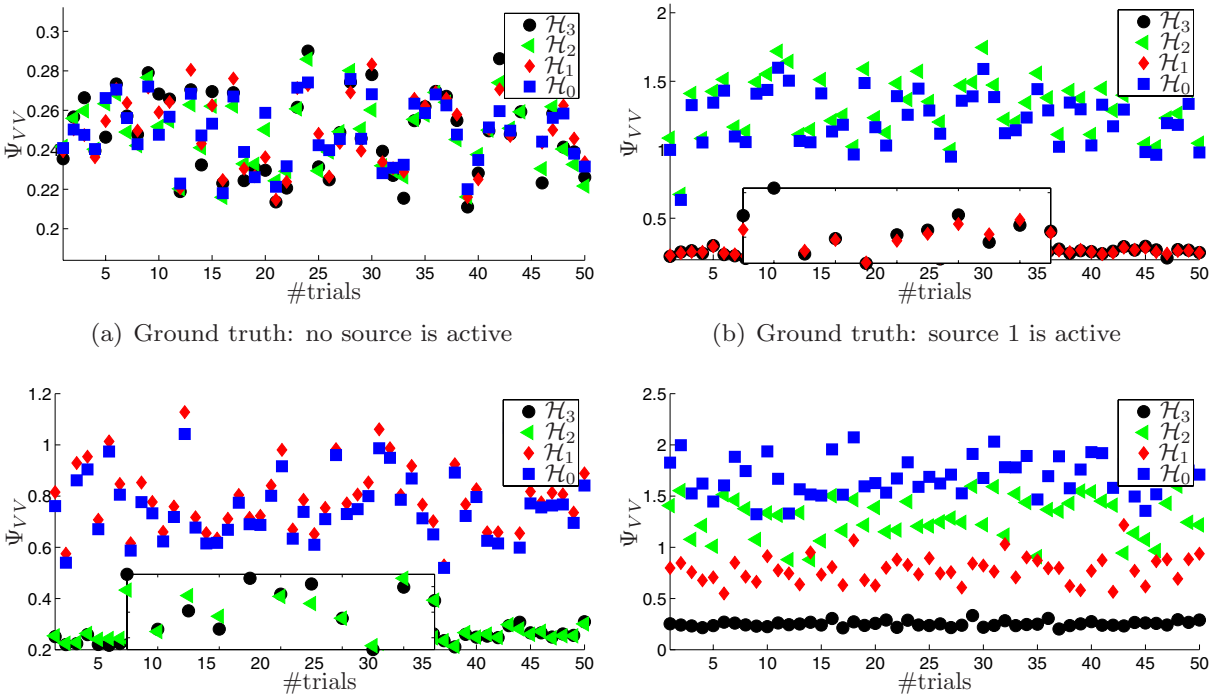
$$\int_{-\infty}^{\mathcal{L}_1} p\left(\mathcal{L}|\mathcal{H}_0\right) \mathrm{d}\mathcal{L} \gtrless \Upsilon_{\mathcal{L}} \begin{cases} \text{select } \mathcal{H}_1 \\ \text{select } \mathcal{H}_0 \end{cases} \tag{3.94}$$

where $\Upsilon_{\mathcal{L}}$ is the significance level of the test and $\mathcal{L}_1$ is the likelihood-score obtained for hypothesis 1. In essence, (3.94) measures the overlap of the bootstrapped distributions of the respective likelihood estimates and classifies them as being different or similar for a given threshold value $\Upsilon_{\mathcal{L}}$. Applying these observations to the examples presented in Figure 3.10, we attain convergence to the ground truth for appropriately chosen $\Upsilon_{\mathcal{L}}$.

Hypothesis testing based on the noise variance

The hypothesis test in this alternative case is based on the ML estimate of the noise power computed over the time record over the B frames. Since $\Psi_{VV} < \Psi_{S_qS_q}$, $\forall q$, the score function in this case is based on the minimum estimate of the noise variance, i.e., we opt for that hypothesis, which yields the lowest value of $\widehat{\Psi}_{VV}$. Results using this approach are presented in Figure 3.12 for the same data as in the previous approach.

Here too, the proposed sequential hypothesis test may be applied, with the decision being based on the bootstrap distribution: $p\left(\Psi_{VV}\middle|\mathcal{H}_i\right)$ of the noise variance estimates and a detection threshold Υ_V. A sample bootstrap distribution is presented in Figure 3.13. Again, for properly defined Υ_V, a convergence to the ground truth is attained.



(a) Ground truth: no source is active

(b) Ground truth: source 1 is active

(c) Ground truth: source 2 is active

(d) Ground truth: both sources are active

Figure 3.12: ML estimates of the noise variance for the different hypotheses. The results are presented for 50 Monte Carlo trials.

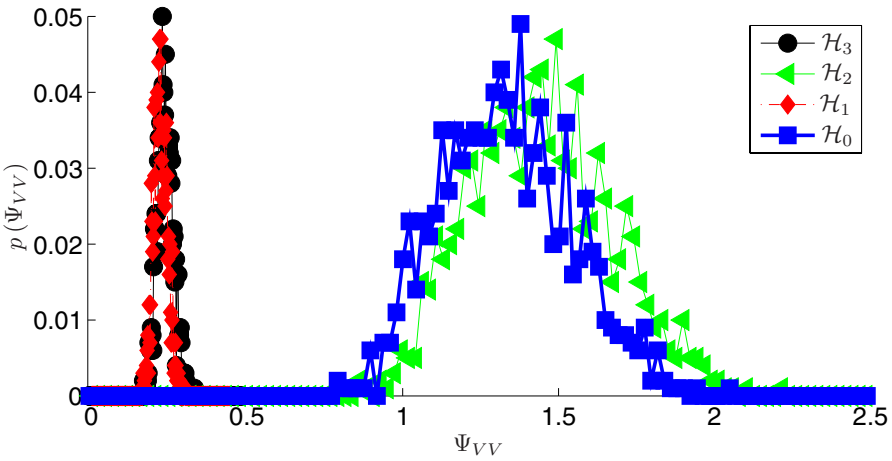


Figure 3.13: Sample bootstrap distribution on the ML estimates of Ψ_{VV} values for the case where only source 1 is active.

Hypothesis testing based on model order selection

Another alternative to test the hypotheses is on the basis of the model *order* corresponding to the observation. Model order in this context refers to the number of sources that contribute to the observed acoustic event. In other words we select the model that best describes the observations. The fit of a model to the observations is usually measured by the likelihood of that model, along with a second term that penalizes overfitting. Examples of such score functions are the Bayesian information criterion (BIC) [108] and the Akaike's information

criterion (AIC) [2].

The cost function for the Bayesian information criterion for a given hypothesis on a model is:

$$\mathrm{BIC}(\mathcal{H}) = \mathcal{P}\ln(B) - 2\ln(\mathcal{L}(\mathbf{X}^B; \boldsymbol{\Theta} \,|\mathcal{H}, \mathbf{A})) \tag{3.95}$$

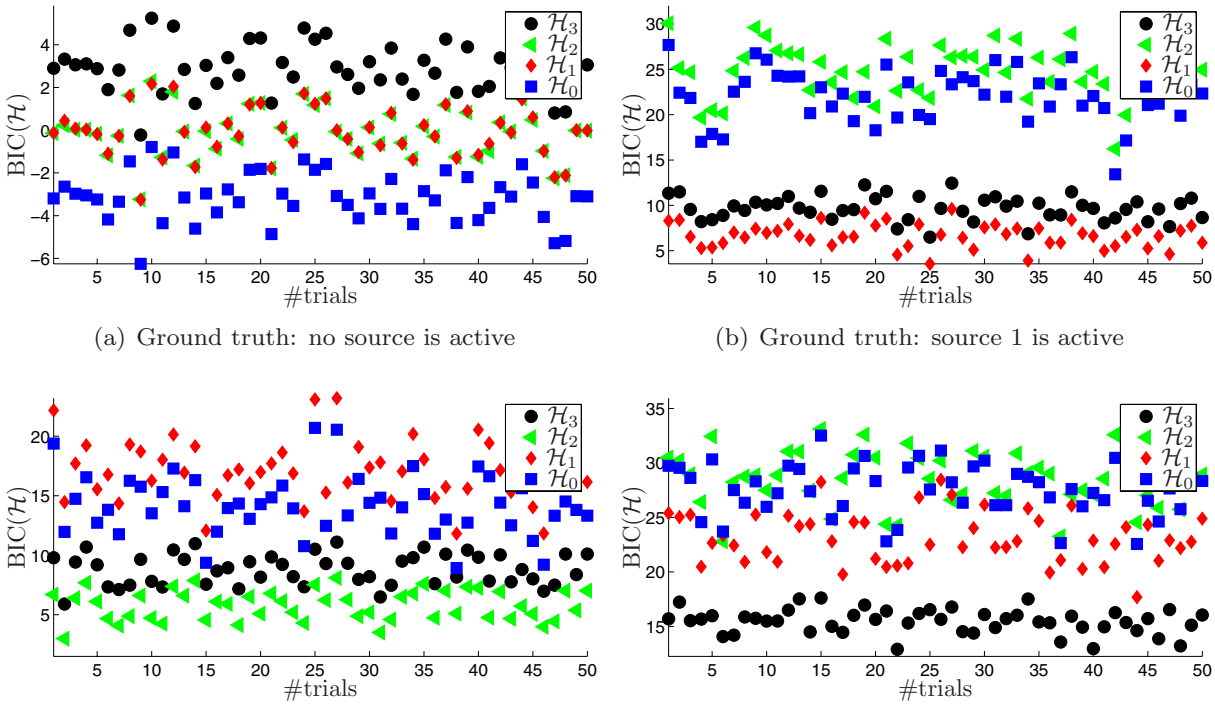
and that for the Akaike's information criterion is:

$$\mathrm{AIC}(\mathcal{H}) = 2\ln(\mathcal{P}) - 2\ln(\mathcal{L}(\mathbf{X}^B; \boldsymbol{\Theta} \,|\mathcal{H}, \mathbf{A})) \tag{3.96}$$

where $\mathcal{P}$ represents, as before, the number of free parameters.

To make a decision on the basis of these criteria, we select that hypothesis as true which has the lowest value of the BIC (or AIC). As the BIC imposes a larger penalty on increased model order as compared to the AIC, it may provide a conservative estimate when the number of sources is large.

The results obtained using this approach on the data of the previous sections are presented in Figures 3.14–3.15. In each case, the true hypothesis has the lowest value. Furthermore, there is a clear distinction between all the hypotheses, obviating the need for a comparison as described previously.



(a) Ground truth: no source is active

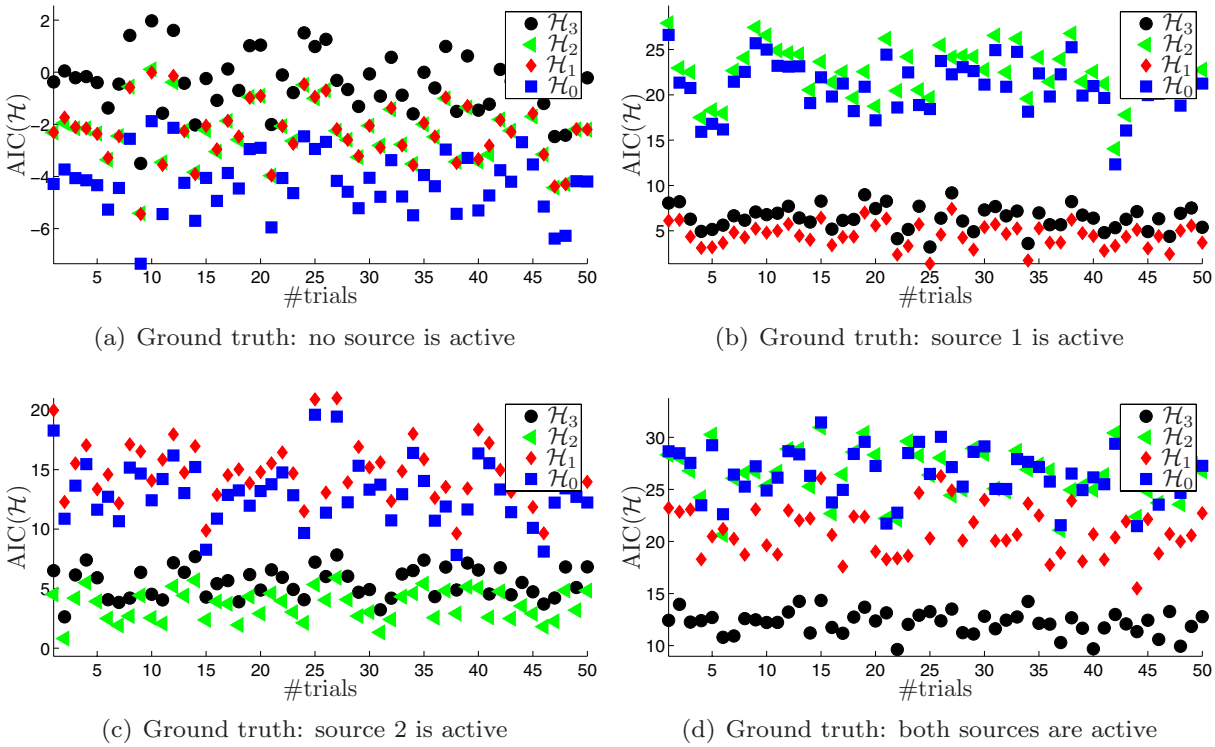
(b) Ground truth: source 1 is active

(c) Ground truth: source 2 is active

(d) Ground truth: both sources are active

Figure 3.14: Values of the BIC for the different hypotheses, based on the maximum likelihood estimate of the parameters. The results are presented for 50 Monte Carlo trials.

As a last point, note that when the nominal source positions are unknown, the detection approaches presented above may be further extended to *include* source localization in the framework, as done in [66, 79].



(a) Ground truth: no source is active

(b) Ground truth: source 1 is active

(c) Ground truth: source 2 is active

(d) Ground truth: both sources are active

Figure 3.15: AIC values for the different hypotheses, based on the maximum likelihood estimate of the parameters. The results are presented for 50 Monte Carlo trials.

3.6 Conclusions

This chapter has provided an overview of various contemporary source localization approaches. These have been classified into direct approaches and indirect approaches. Algorithms catalogued under indirect approaches first estimate the time delay of arrival (TDOA) between various microphone pairs and then, based on these values and the geometry of the array, estimate the source positions. Direct approaches, on the other hand, pick the most likely set of source positions from a given set of candidate locations.

Further, relations between the various approaches were derived and it was shown that most approaches exploit only the second-order statistics of the observed microphone signals. In general, this second-order dependence comes about as a result of making Gaussian assumptions regarding the signal and noise statistics. As these assumptions might not be realistic for speech, these are perhaps not optimal source location estimators. Consequently, future research could focus on incorporating *a priori* knowledge of the signal/noise statistics as done, for instance, in [1]. One could also retrieve the source location information as by-products of the algorithms for blind source separation.

The performance of representative direct and indirect algorithms was also illustrated for different noise and reverberation scenarios. Generally, the presence of noise and reverberation degrades the algorithm performance. However, localization is still possible under dominance of the direct path.

In general, having demonstrated the close relationship that exists between the various source localization approaches, we state that no algorithm may be termed as *the optimal*. Optimal-

ity is dependent upon the application-specific customizations that need to be made on the basic algorithm and each application requires its own, appropriately 'tuned' approach. We shall speak, instead, of optimal algorithms for the *specific* application under consideration. Such an algorithm takes into account all available *a priori* knowledge regarding the source and design constraints. As a brief example to illustrate our point, the time-domain broadband approaches are not applicable for the case where the sources are narrowband. Another example is the case where we do have broadband sources, but whose spectra are disjoint. Here too, it may be advisable to perform localization using the narrowband algorithms in the spectral regions corresponding to each source.

For the case of single broadband source localization using two microphones, the LMS/AED approaches might be a better choice than the GCC-PHAT. However, with more sensors or for narrowband sources or for the case of multi-source localization when the sources have disjoint spectra, the narrowband direct approaches such as the SRP, MUSIC or the ML based approaches allow for easier integration of application specific constraints and should be evaluated as the basic methods of choice. This point is explored in more detail in the next chapter.

As a last point, we have also presented a different view of localization – as a detection problem. Such a case arises when we have multiple acoustic sources, the nominal locations of which are known *a priori*. The goal here is to detect, when an acoustic event occurs, which source combination was the generator of that event. Two possible solutions based on hypothesis testing were also presented. The first approach utilizes the likelihood score functions and a comparison criterion based on bootstrapping, which does not require the knowledge of the distribution of the test statistic for decision making. The second approach is based on model order estimation using the AIC or the BIC, where the hypothesis representing the source combination that best fits the observations is selected as the true hypothesis. This latter approach is computationally less expensive than the bootstrap, but may yield more conservative results when the number of active sources increases.

If you're willing to restrict the flexibility of your approach, you can almost always do something better.

– John Carmack, Id Software

Chapter 4

Localization: Algorithm Design Principles

As demonstrated in the previous chapter, most contemporary algorithms for source localization are dependent only on the second order statistics of the sensor input signals – the differences between the algorithms being primarily in the amount of *a priori* knowledge available and on the prefiltering of the sensor signals before the computation of the test statistic. Both these aspects are related to the specific application under consideration and, consequently, it is not possible to find one algorithm or implementation that is successfully applicable to all classes of localization applications. Each application requires its own localization approach that takes into consideration all available *a priori* knowledge regarding the sources e.g., their spatial distributions, number, frequency ranges, any peculiarities in their time-frequency distributions, etc. Such knowledge alongwith any associated design constraints (such as the number of microphones, the tolerable latency, etc.) may then be put to use in formulating the optimal localization approach for *that* specific application, right from the design of an optimal microphone array to the localization algorithm itself. This chapter serves to illustrate this point using two contrasting application examples. The first is that of localization of brake squeal and the second, that of localizing multiple speech sources.

4.1 General array design considerations

Array design refers not only to the shape and dimensionality of the array, but also to the inter-element spacing in the selected geometry. This is an important step in array signal processing and one that is too often overlooked in algorithm descriptions; in some cases because the optimality of the selected array is obvious, e.g., arrays for the processing of narrowband signals with known center frequency typically have inter-element spacing of $\lambda/2$, with λ being the wavelength of the center frequency; and in some cases because the array design is dictated by constraints over which the algorithm designer may not have any influence, e.g., hearing-aid applications, where the array design is subject to power and space constraints.

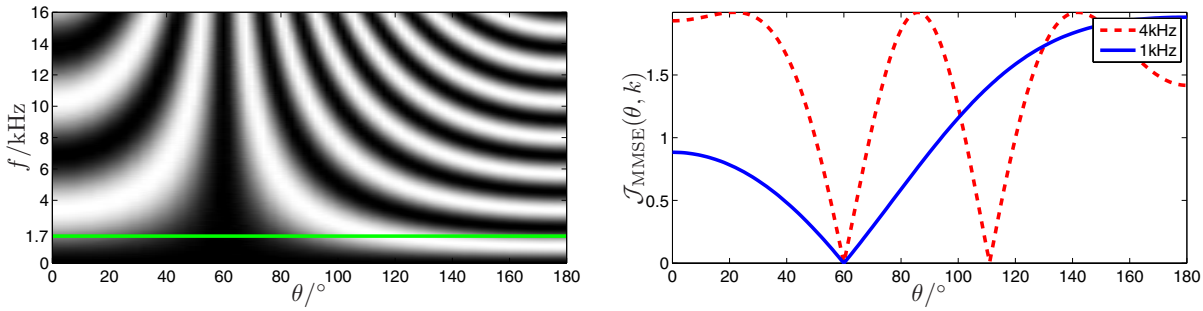
When the freedom of array design is available in conjunction with algorithm development for a specific application, it makes sense to tailor the design to the application-specific constraints. For example, in a video-conferencing scenario where the speakers are seated around a table, a good choice might be a planar array mounted on the table, allowing a

360° localization [75, 70]. If one were to consider an automotive application, such as hands-free telephony, one would design a linear array mounted in the vicinity of the driver. Where the area to be spanned is large, one finds the use of distributed sensor arrays (sensors spread out over a wide area) as for example in [16] and in the brake squeal localization presented in the next section. For multimedia applications or human-computer interaction, where usually the speakers are in front of the device, in the same half-plane, one could envisage a linear microphone array mounted on the device.

While partly dependent on the selected geometry, the selection of the inter-element spacing for source localization is mainly motivated by the issue of *spatial aliasing* – the existence of multiple modes (minima/maxima) in the narrowband localization cost function due to the 2π periodicity of the phase compensation factors. For a two sensor case with a given inter-element distance d, spatial aliasing occurs when the wavelength of the signals is smaller than $2d$, i.e.,

$$\lambda < 2d \quad \text{or} \quad f > \frac{c}{2d}.$$

This is illustrated in Figure 4.1 for $d = 0.1\,\text{m}$ (corresponding to a frequency of $f \approx 1.7\,\text{kHz}$) and a target source to be localized at 60°, using the MMSE approach to source localization (Chapter 3, Section 3.3.2). Note the emergence of multiple minima above the threshold frequency of $f > 1.7\,\text{kHz}$.



(a) $\mathcal{J}_{\text{MMSE}}$ over frequency and θ.

(b) Cost function for two specific frequencies.

Figure 4.1: The spatial aliasing issue in narrowband direct localization approaches, illustrated for a two microphone array with $d = 0.1\,\text{m}$. Figure 4.1(a) depicts the behaviour of the cost function over all frequencies. Spatial aliasing starts to occur for $f > 1.7\,\text{kHz}$. Figure 4.1(b) presents a closer look at the cost function for two specific frequencies. Note the ambiguity in the source location at 4kHz.

In the subsequent sections, we shall look more closely into the issue of array design on the basis of the application considered.

4.2 Brake squeal localization

Wikipedia [130] defines brake squeal as the high-pitched noise sometimes emitted when brakes are applied. Most brake squeal is produced by vibration of the brake components, especially the pads and discs, due to resonance. Thus, the source of the squeal is spread over part of the disc and pad surface, giving it a significant spatial spread. The squeal is narrow-band in nature, with the squeal frequencies typically ranging from 1 kHz to 16 kHz. Also, the squeal frequencies are usually different for each wheel and dependent upon various

factors such as speed of the automobile when braking, age of the components, environmental conditions (e.g., heat, moisture), brake pressure, etc.

From the acoustic signal processing point of view, we have:

1. an environment that contains several narrowband sources, the *a priori* knowledge of whose approximate positions are available (the four wheel positions), and
2. at any time none, one, or several of these sources could be 'active' (i.e., emit a squeal), occupying possibly overlapping frequencies. The presence of active sources within an observed time range constitutes what we shall define as an acoustic event.

Brake squeal localization defined in this context consists of detecting such acoustic events and locating them to their contributing sources (generator brakes). For each detected squeal frequency, given that there are 4 brakes, there are $(2^4 - 1)$ possible combinations[1] that could have contributed to the generation of the squeal. Thus, selecting the right combination constitutes the solution to the localization problem. In other words, brake squeal localization may be seen as a multiple-hypothesis testing or a *detection* problem based on the directional properties of the sound field measured by appropriately situated microphone arrays.

4.2.1 Array design for brake squeal localization

Given that brake-squeal is narrowband, we cannot use broadband approaches to localize the squeal to the generating brakes. Consequently, the algorithms that lend themselves to solving this problem should be based on narrowband, direct approaches for each active frequency.

In order to make an unambiguous decision at each frequency in the desired range, we need to ensure that the spatial aliasing problem will not occur. One solution to this is to set the smallest inter-element spacing of the microphone array to $d_{\min} \leq \lambda_{\min}/2$, where $\lambda_{\min}$ is the wavelength corresponding to the maximum frequency of interest. In our system, therefore, this was set to 0.5 cm, given that we wanted to be able to localize up to (and above) 16 kHz. However, as we have a broad range of frequencies to detect and localize, a single microphone pair will not do – especially one with $d = 0.5$ cm – as such small inter-sensor spacings do not provide sufficient spatial information for lower frequencies (the corresponding phase shift is too small to measure accurately and additionally would be swamped by the noise floor and the jitter from the windowed STFT). Thus, the next step is the decision regarding the additional number of sensors required, and their placement.

The decision regarding the number of sensors in an array is motivated by a number of factors:

- allowable size of the arrays: increasing the number of sensors would increase the array aperture. While this is of advantage in that we have increased spatial diversity, allowing for more degrees of freedom in the design of the localization algorithm, the size of the array is often governed by more practical issues such as available space, mounting constraints, etc. For the application under consideration, the requirement is to have as compact an array-length as possible, in the range of 15 cm–20 cm;
- hardware constraints: the signals from the microphones need to be sampled and digitized. This sampling and digitization must be synchronous across all the microphones of the array in order to prevent any errors in the localization. Small sample-time jitter can be catastrophic for localization at higher frequencies as such a jitter introduces an

[1] The presence of a squeal event precludes the null hypothesis that no brakes were active.

additional phase shift into the signals. Our survey of off-the-shelf A/D converters led us to constrain the maximum number of microphones in the array to 8, as this seemed to be the number of simultaneous A/D channels most widely available.

From the above considerations, we obtain the remaining inter-element spacings based on the same algorithmic constraints that governed the design of the first pair. Beginning with the second microphone of the first pair, the succeeding elements of the array are harmonically nested in order to obtain frequency invariant beamwidth [125] in the cost function over the frequency range of interest. Thus, we have inter-sensor spacings of 1 cm (16 kHz), 2 cm (8 kHz), and 4 cm (4 kHz), the distances being measured with respect to the last microphone in the previous pair. This setup is mirrored around the first microphone, yielding the 8-element microphone array with the largest aperture of 14.5 cm, approximately optimal for the 1 kHz lower band limit. The designed array is depicted in Figure 4.2.

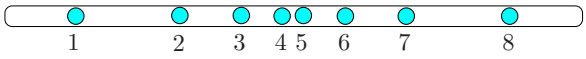


Figure 4.2: The designed array for brake squeal localization. The spacings are symmetric about the center, with $d_{12} = 4\,\text{cm}$, $d_{23} = 2\,\text{cm}$, $d_{34} = 1\,\text{cm}$, and $d_{45} = 0.5\,\text{cm}$

The next question is the issue of array *placement*. While we described above the configuration for a linear array, intuition suggests a planar array configuration mounted underneath the chassis, in the *middle*, having all the sources in its line of sight. However the acoustic environment underneath the automobile distorts the sound field, making localization impossible. This was verified not only by measurements in a controlled environment (brake disc excitation of a stationary automobile, using a shaker) but also on a simulation of the acoustic field using the finite-element method[2]. This (and exhaustive tests of other configurations) led to the use of a distributed microphone array system consisting of four linear sub-arrays, placed along the outside border of the auto-chassis. Each sub-array consists of 8 microphones and is configured as in Figure 4.2. The setup is shown in Figure 4.3. The numbers represent the global channel indices of the first and the last microphone of the respective array.

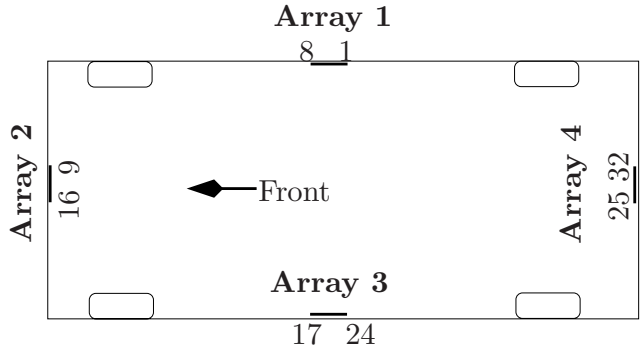


Figure 4.3: Array positions and microphone indices.

The requirement of four sub-arrays came about as the placement along the border of the chassis preserves the dominance of the direct path from a brake to the arrays to which it has a *direct* line of sight. To clarify the issue of line-of-sight: array 2 can distinguish

[2]We are thankful to Dr.-Ing. Henning Taschke and Dr.-Ing. Sebastian Schmidt for developing the finite-element model and performing the simulations.

accurately between events originating front right and front left, but an event at the rear right/left brakes is not accurately localized (construction of the chassis is such that line-of-sight is obstructed). The reverse holds true for array 4. Array 1, on the other hand, cannot distinguish between an event at *front* right and *front* left. In fact, due to the shadowing of the brake at front right by the wheel, and the presence of a parallel reflecting surface in the vicinity of the brake at front left, array 1 perceives a source emanating from front *right* as coming from front *left*. The same holds for squeals originating from the rear wheels, while the reverse argument holds for array 3. Thus, arrays 1 & 3 can distinguish only between events coming from the front and the rear. This discrimination capability for each array is illustrated in Figure 4.4. Accordingly, we require separate arrays for the front and rear wheels and two arrays on the sides. The latter redundancy will be justified in the subsequent sections.

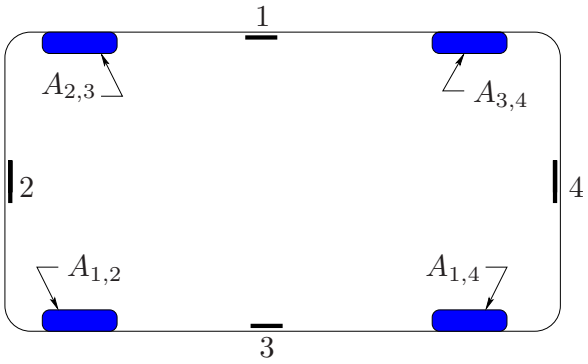


Figure 4.4: Spatial discrimination capability for each array. The subscripts for each wheel indicate the arrays capable of localizing a squeal from that wheel. Note that each array is principally capable of distinguishing only between two sources.

4.2.2 Further system considerations and constraints

A high precision placement of the arrays as depicted in Figure 4.3 is not guaranteed, given that the automobile silhouette is idealized and that there are usually constructional elements that one needs to take into account when mounting the arrays. Therefore provision should be made for *small deviations* in the microphone array positions.

Further, the detection of acoustic events must be accomplished in real-time, with low latency. For this reason, we analyze the signals in segments of 200 ms duration, which is also the latency of the system. This time frame is chosen as it provides the best trade-off between having enough time-records in the discrete frequency domain at sufficient frequency resolution and being close to the order of magnitude of signal length that is recognizable as a harmonic[3] (approximately 50 ms [119]) – allowing, on the one hand, for even short events to be detected and, on the other hand, the possibility of tracking the modulation of the squeal as it is often observed that even during a single squeal event, the squeal frequency for a single brake may vary with time.

4.2.3 Algorithm development

Assume the presence of an initial detection stage that can detect the occurrence of brake squeal and extract the corresponding squeal frequencies. Given these frequencies at which we

[3]Harmonic signals of shorter lengths are perceived more as a dull ‘thump’ than as sinusoidal excitations.

have an acoustic event, we are required to decide which of the $2^4 - 1$ combinations of brakes generated the event. In other words, we have a hypothesis testing problem based on the directional statistics available from the sound field observed by the arrays.

Hypothesis tests which consider the complete 4-array system are not optimal as no array is capable of localizing all the sources. Taking cognizance of the discrimination capabilities of the respective arrays, we arrive at the following two-stage hypothesis testing approach. In the first stage, we use arrays 1 and 3 to test

1. $\mathcal{H}_F$: squeal emanated from the front brakes
2. $\mathcal{H}_B$: squeal emanated from the back brakes
3. $\mathcal{H}_{FB}$: squeal emanated simultaneously from the front & back brakes

Next, depending upon the outcome of this stage, arrays 2 and 4 are considered to detect if the event originated from the right (R) or the left (L), i.e.,

1. Given the truth of $\mathcal{H}_F$, use array 2 to test the following hypotheses
   a) $\mathcal{H}_{L|F}$: squeal emanated from the brake on front left
   b) $\mathcal{H}_{R|F}$: squeal emanated from the brake on front right
   c) $\mathcal{H}_{LR|F}$: squeal emanated from front left and front right
2. Given the truth of $\mathcal{H}_B$, use array 4 to test the following hypotheses
   a) $\mathcal{H}_{L|B}$: squeal emanated from the brake on back left
   b) $\mathcal{H}_{R|B}$: squeal emanated from the brake on back right
   c) $\mathcal{H}_{LR|B}$: squeal emanated from back left and back right

Note that the accuracy of the front/back detection is critical as the next stage depends on the outcome of this test. It is to increase this accuracy and to provide robustness against sensor defects that two arrays are used for this stage.

It remains to decide the nature of the hypothesis tests, for which we present two approaches. The first one is based on heuristic considerations and is described in Section 4.2.4. The second one is based on a model that is robust against imperfections in the knowledge of the propagation vectors from the source to the array, and is described in Section 4.2.5.

4.2.4 Hierarchical approach to brake squeal localization

In this approach our aim is to assign a measure of *contribution* of each brake to each acoustic event, allowing for the possibility of arriving at a 'soft' decision. We assume, not unreasonably, that each brake squeal signal is independent of the others. For each detected squeal frequency, we shall compute a *likelihood* $\mathcal{L}$ of activity for each brake and assume that this may be factorized into a conditional form as in (4.1) below:

$$\begin{aligned} \mathcal{L}(F,R) &= \mathcal{F}\{\mathcal{L}(R|F), \mathcal{L}(F)\}, \\ \mathcal{L}(F,L) &= \mathcal{F}\{\mathcal{L}(L|F), \mathcal{L}(F)\}, \\ \mathcal{L}(B,R) &= \mathcal{F}\{\mathcal{L}(R|B), \mathcal{L}(B)\}, \\ \mathcal{L}(B,L) &= \mathcal{F}\{\mathcal{L}(L|B), \mathcal{L}(B)\}, \end{aligned} \tag{4.1}$$

where $\mathcal{F}$ is some, as yet unknown, function. Further, $\mathcal{L}(F,R)$ denotes the likelihood of activity of the front, right brake; $\mathcal{L}(F)$ denotes the likelihood of activity from the front brakes; $\mathcal{L}(R|F)$ denotes the likelihood of activity of the front right brake *conditional* upon the hypothesis that the event originated from the front, and so on.

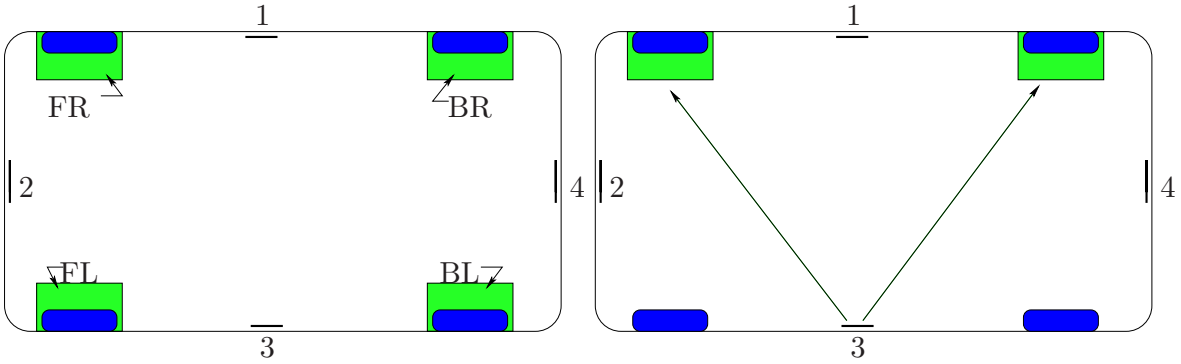
The likelihood is based on an appropriate choice of the localization cost function selected from the narrowband approaches discussed in Chapter 3 and is computed, for each active frequency k and each array i, for the two regions of discrimination defined for *that* array. The cost function chosen is a normalized version of (3.42). Let us denote this cost function by $\mathcal{J}(\mathbf{r}, k, i)$, noting its dependence on the frequency k, the candidate position $\mathbf{r}$ and the array i. For convenience, the cost function is reproduced for a generic array in (4.2) below:

$$\mathcal{J}(\mathbf{r}, k) = - \sum_{\substack{m,m' \\ m \neq m'}} e^{\jmath\, \Omega_k (T_m(\mathbf{r}) - T_{m'}(\mathbf{r}))} \frac{\Psi_{X_m X_{m'}}(k)}{|\Psi_{X_m X_{m'}}(k)|} |\Gamma_{mm'}(k)|^2 \ . \tag{4.2}$$

When computing these cost functions, we should address the concerns previously raised – namely the robustness of the system against small deviations in positioning of the microphone arrays and the spatially spread nature of the acoustic sources. To this end, the candidate positions are extended to encompass the possible spread of the source and the deviations in the microphone array, leading to a spatially averaged estimate of $\mathcal{J}$ over a candidate *region* $\mathcal{R}$.

$$\mathcal{J}(\mathcal{R}, k) = \int_{\mathbf{r} \in \mathcal{R}} \mathcal{J}(\mathbf{r}, k)\, \mathrm{d}\mathbf{r} \tag{4.3}$$

We assume that a squeal event would be concentrated in the region around the respective wheels. This allows us to use the wheel dimensions to define the extent of the candidate regions. These regions are illustrated by the shading in Figure 4.5(a) for all the wheels and in Figure 4.5(b) for array 3 in particular.



(a) Search regions for each wheel (b) Candidate positions for array 3

Figure 4.5: Candidate search regions. The shaded regions indicate the *a priori* knowledge of regions of source presence which is used to obtain the candidate search regions for the cost function computation as in (4.3).

With these cost functions, we now aim to define the likelihood measure in (4.1) and the function for the conditional representation. To begin with, we consider $\mathcal{J}(\mathcal{R}_{FL}, k, 1)$, $\mathcal{J}(\mathcal{R}_{BL}, k, 1)$, $\mathcal{J}(\mathcal{R}_{FR}, k, 3)$ and $\mathcal{J}(\mathcal{R}_{BR}, k, 3)$ – the spatially averaged cost functions of (4.3) for arrays 1 and 3 over their respective candidate regions. We next define:

$$\begin{aligned} \mathcal{J}_F(k) &= \mathcal{J}(\mathcal{R}_{FL}, k, 1) + \mathcal{J}(\mathcal{R}_{FR}, k, 3)\,, \\ \mathcal{J}_B(k) &= \mathcal{J}(\mathcal{R}_{BL}, k, 1) + \mathcal{J}(\mathcal{R}_{BR}, k, 3)\ . \end{aligned} \tag{4.4}$$

Now, to decide between the three possible hypotheses in the first stage – namely:

$$\boldsymbol{\mathcal{H}}(k) = \{\mathcal{H}_F(k)\,,\, \mathcal{H}_B(k)\,,\, \mathcal{H}_{FB}(k)\}\,, \tag{4.5}$$

where $\mathcal{H}_{\spadesuit}(k)$ indicates the hypothesis that the signal came from $\spadesuit$ at frequency k – we may proceed as:

$$\begin{aligned}
\mathcal{H}_F(k) &\triangleq \left(\frac{\mathcal{J}_F(k)}{\mathcal{J}_B(k)} > \Upsilon_1\right), \\
\mathcal{H}_B(k) &\triangleq \left(\frac{\mathcal{J}_F(k)}{\mathcal{J}_B(k)} < \Upsilon_2\right), \\
\mathcal{H}_{FB}(k) &\triangleq \left(\Upsilon_3 - \delta \leq \frac{\mathcal{J}_F(k)}{\mathcal{J}_B(k)} \leq \Upsilon_3 + \delta\right),
\end{aligned} \tag{4.6}$$

with Υ_1, Υ_2, Υ_3 and δ being thresholds to be decided. However, a better way to proceed, and one that also permits a more intuitive understanding, is to define the respective likelihoods as:

$$\begin{aligned}
\mathcal{L}(F,k) &\triangleq \frac{\mathcal{J}_F(k)}{\max\left(\mathcal{J}_F(k)\,,\mathcal{J}_B(k)\right)}\,, \\
\mathcal{L}(B,k) &\triangleq \frac{\mathcal{J}_B(k)}{\max\left(\mathcal{J}_F(k)\,,\mathcal{J}_B(k)\right)}\,.
\end{aligned} \tag{4.7}$$

This gives us a normalized value for the occurrence of either event - akin to the *probability* of that event. Note that the sum of the 'probabilities' in (4.7) need not be 1, as these events are *not* mutually exclusive. The definition of the probability as above nicely takes this into account, allowing for the possibility that both $\mathcal{L}(F,k)$ and $\mathcal{L}(B,k)$ may be 1, when events originate simultaneously from the front and the rear.

Based on similar considerations, in the second stage we may define the conditional likelihood for the individual brakes. For this, arrays 2 and 4 are used, and we now define

$$\begin{aligned}
\mathcal{L}\left(L|F,k\right) &\triangleq \frac{\mathcal{J}\left(\mathcal{R}_{FL},k,2\right)}{\max\left(\mathcal{J}\left(\mathcal{R}_{FR},k,2\right),\,\mathcal{J}\left(\mathcal{R}_{FL},k,2\right)\right)} \\
\mathcal{L}\left(R|F,k\right) &\triangleq \frac{\mathcal{J}\left(\mathcal{R}_{FR},k,2\right)}{\max\left(\mathcal{J}\left(\mathcal{R}_{FR},k,2\right),\,\mathcal{J}\left(\mathcal{R}_{FL},k,2\right)\right)} \\
\mathcal{L}\left(L|B,k\right) &\triangleq \frac{\mathcal{J}\left(\mathcal{R}_{BL},k,4\right)}{\max\left(\mathcal{J}\left(\mathcal{R}_{BR},k,4\right),\,\mathcal{J}\left(\mathcal{R}_{BL},k,4\right)\right)} \\
\mathcal{L}\left(R|B,k\right) &\triangleq \frac{\mathcal{J}\left(\mathcal{R}_{BR},k,4\right)}{\max\left(\mathcal{J}\left(\mathcal{R}_{BR},k,4\right),\,\mathcal{J}\left(\mathcal{R}_{BL},k,4\right)\right)}\,.
\end{aligned} \tag{4.8}$$

With the conditional probabilities defined as in (4.8) and the marginals defined as in (4.7), the probability of activity for each wheel can be calculated using (4.1) where, as the likelihood functions are now treated as probabilities, we may define $\mathcal{F}$ as:

$$\mathcal{F}(x,y) \triangleq xy\,. \tag{4.9}$$

This yields a value in the range of $[0,1]$. At this stage, we may take a hard decision regarding the activity of a particular brake, which is done by thresholding the likelihood values obtained from (4.1) by substituting the values of (4.7) and (4.8), and using the definition of $\mathcal{F}$ as in (4.9). We denote this threshold as $\Upsilon_{\mathcal{L}}$ and it is empirically set.

As a side note, the smaller the cost function, the better (more trustworthy) the localization estimate is. As the range of the cost function is limited to $[-1,1]$ by an appropriate normalization, it might happen that some cost functions attain a value greater than 0, making some 'probability' estimates negative in (4.7) and (4.8). This problem is avoided by setting the upper bound on the cost function values to 0.

4.2.5 MaxChoice approach to brake squeal localization

This approach is derived from the ML detection framework proposed in Section 3.5. While the ML detection approach is optimal when the assumptions concerning the signals and the propagation vector are met, such information is hard to gather for brake squeal localization. The reason lies in part in the insufficient knowledge regarding the source locations (the positions are only approximately known; the spatial extent during a squeal is unknown and time and frequency variant and must be estimated during the localization procedure). In addition, we require a model for the signal distortion caused by the acoustic environment under the automobile. This is further complicated by the fact that such a model is time variant and depends not only on the relative source–receiver positions, but also on the external environment, which is continually changing. Thus, given the short time frames in which the detection and localization must be done, applying the ML framework requires a high-dimensional optimization, which increases the computational complexity and makes a real-time implementation well-nigh impossible. This is the motivation to find an approach that is not only real-time capable (low computational complexity) but also more tolerant and robust against such imperfect knowledge. In this section, we shall outline one such method which also has ramifications for the ML-based hypothesis testing problem outlined previously.

Since each array principally evaluates two regions, consider one such generic array with the corresponding sources $S_1(k,b)$ and $S_2(k,b)$. In this case, we have a similar situation as in Section 3.5, i.e., we know the approximate positions of the two sources to be localized by the array. Furthermore, we shall assume that each brake squeals at frequencies *unique* to that brake. This is not an unreasonable assumption given the manufacturing tolerances of the components, assembly variations, possible age differences, different influence from the environment, etc. On the basis of this assumption, we shall consider the sources to be *spectrally disjoint*, i.e., each active frequency is dominated by just one source, which we require to localize[4]. Note the contrast to the previous algorithm which allows multiple wheels to contribute to a squeal event at a given frequency and time frame. The spectral disjointness assumption further implies that given an active frequency, it *must* be allocated to *one* source. To do this, we sequentially test each of the two regions for the presence of source activity whilst *blocking* out contributions from the other region (treated as a *clutter* region). The approximate knowledge of the source positions can be used to good effect for designing such blocking systems with sufficient tolerance to compensate for possible inaccuracies in the model. Assume for now that we have such a blocking system $\mathbf{P}_q^{\perp}(k)$ for region q at frequency bin k. Then,

$$\begin{aligned}\mathbf{Y}^{(q)}(k,b) &= \mathbf{P}_q^{\perp}(k)\,\mathbf{X}(k,b)\\ &= \mathbf{P}_q^{\perp}(k)\,\Big(\sum_{q'=1}^{2}\breve{\mathbf{A}}_{q'}(k)\,S_{q'}(k,b)+\mathbf{V}(k,b)\Big)\\ &\approx \mathbf{P}_q^{\perp}(k)\,\Big(\breve{\mathbf{A}}_{3-q}(k)\,S_{3-q}(k,b)+\mathbf{V}(k,b)\Big)\end{aligned} \tag{4.10}$$

where $\breve{\mathbf{A}}$ indicates the propagation vector from source to receiver, taking into account the unknown disturbances. Note that the approximation in the last step holds since, by designing the blocking system to have a suitable tolerance, we may negate the contributions from the blocked source q. The use of such a blocking system to obtain the resultant signals $\mathbf{Y}^{(q)}(k,b)$ is key to our approach – which we term the *Maximum Choice* (MaxChoice) algorithm.

[4]We shall test the validity of this assumption later.

We next compute the residual energy, over the B time records, after the application of the blocking system.

$$\begin{aligned}\breve{\Psi}_{qq}(k) &= \sum_{b=1}^{B} \mathrm{tr}\left(\mathbf{Y}^{(q)}(k,b)\mathbf{Y}^{(q),H}(k,b)\right) \\ &\approx \mathrm{tr}\left(\mathbf{P}_q^{\perp}(k)\left(\sum_{b=1}^{B}\left(\breve{\mathbf{A}}_{3-q}(k)\,\breve{\mathbf{A}}_{3-q}^{H}(k)\,|S_{3-q}(k,b)|^2 + \right.\right.\right. \\ &\qquad \left.\left.\left.\mathbf{V}(k,b)\mathbf{V}^H(k,b)\right)\right)\mathbf{P}_q^{\perp,H}(k)\right).\end{aligned} \tag{4.11}$$

Recognize that this is a measure of the energy received by the microphone from spatial regions *other* than the region associated with the presence of source q. Once this residual energy has been computed, a binary decision is made for that bin as follows:

$$\frac{\breve{\Psi}_{22}(k)}{\breve{\Psi}_{11}(k)} \underset{\mathcal{H}_2}{\overset{\mathcal{H}_1}{\gtrless}} 1, \tag{4.12}$$

with $\mathcal{H}_q$ representing the hypothesis that the squeal originated from region q. Note that in this approach we do not perform explicit source localization. Rather, by configuring the blocking system with suitable tolerance around the *nominal* source positions, we compensate for imperfections in the knowledge of the propagation model and source position.

When the propagation model is well-known and the source positions are available, it is straightforward to extend this approach to perform a hypothesis test similar to that described in Section 3.5 to test for the presence or absence of source activity *independently* in each region. Thereby it is also possible to handle the case of correlated or even *coherent* sources, and the assumption of spectral disjointness is not necessary. Indeed one may consider such a system to be a generalized form of the approach presented in [116] that better uses the *a priori* knowledge available regarding the nominal source positions and spatial extent.

A further extension to perform source *localization* when the source positions are unknown, but the propagation model is well defined, also follows in a rather straightforward manner.

Design of the blocking system

We now address the design of the blocking matrix for a generic frequency k, a general M channel array, and for a source q with its corresponding blocking region $\mathcal{R}_q$. We require to find $\mathbf{P}_q^{\perp}(k)$ such that:

$$\begin{aligned}&\int_{\mathcal{R}_q} \|\mathbf{P}_q^{\perp}(k)\,\mathbf{A}(\mathbf{r},k)\|^2\,\mathrm{d}\,\mathbf{r} \approx 0 \\ \Rightarrow\ &\mathrm{tr}\left(\mathbf{P}_q^{\perp}(k)\left(\int_{\mathcal{R}_q}\mathbf{A}(\mathbf{r},k)\mathbf{A}^H(\mathbf{r},k)\,\mathrm{d}\,\mathbf{r}\right)\mathbf{P}_q^{\perp,H}(k)\right) \approx 0\,.\end{aligned} \tag{4.13}$$

Defining:

$$\mathbf{R}_q(k) = \int_{\mathcal{R}_q}\mathbf{A}(\mathbf{r},k)\mathbf{A}^H(\mathbf{r},k)\;\mathrm{d}\,\mathbf{r}\,, \tag{4.14}$$

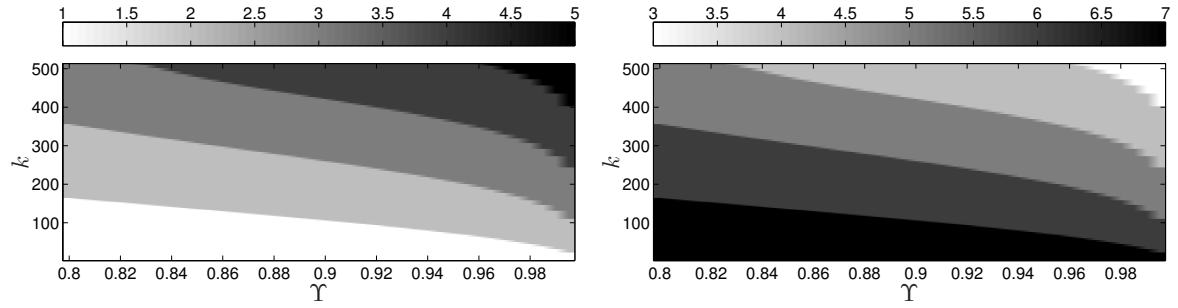
we see that (4.13) is minimized if we select $\mathbf{P}_q^{\perp}(k)$ to span the null space of $\mathbf{R}_q(k)$, i.e., if the eigenvectors of $\mathbf{R}_q(k)$ are denoted as $\mathbf{U}_i(k)$, then

$$\mathbf{P}_q^{\perp}(k) \triangleq \big(\mathbf{U}_{M-N_q(k)+1}(k)\,,\,\ldots\,,\,\mathbf{U}_M(k)\big)\big(\mathbf{U}_{M-N_q(k)+1}(k)\,,\,\ldots\,,\,\mathbf{U}_M(k)\big)^H. \tag{4.15}$$

The frequency-dependent parameter $N_q(k)\,(<M)$ is estimated, for a given threshold Υ $(0 \leq \Upsilon \leq 1)$, in the following manner:

$$N_q(k) = \underset{N}{\operatorname{argmin}} \left| \Upsilon - \frac{\sum_{i=1}^{M-N} \lambda_i(k)}{\sum_{i=1}^{M} \lambda_i(k)} \right| \tag{4.16}$$

where the $\lambda_i(k)$ are the eigenvalues of $\mathbf{R}_q(k)$, sorted in descending order of magnitude. In effect, (4.16) selects the dimension of the null space by setting a threshold on the energy that is present in the target region. The dimension of the target region subspace and the null space are illustrated below for an azimuthal blocking region $\mathcal{R}_\theta = [25°, 65°]$ for various values of Υ.



(a) Target region dimension

(b) Null space dimension

Figure 4.6: Subspace dimensions for the array shown in Figure 4.2. The sampling frequency is $f_s = 32\,\text{kHz}$ and a $K = 1024$ point DFT has been used. Note that the target region dimension and the null space dimension add up to M for each combination of (k, Υ).

As may be seen in Figure 4.6(a), the target region dimensions are low for the lower frequencies. This is because of the lack of directivity in these frequencies, making $\mathbf{A}(\mathbf{r}, k)$ in (4.13) almost independent of $\mathbf{r}$. As k increases, so does the directivity and the matrix $\mathbf{R}_q(k)$ increases in rank. The blocking capabilities (or performance) of such a designed system may be defined as the *pseudo*-beampattern:

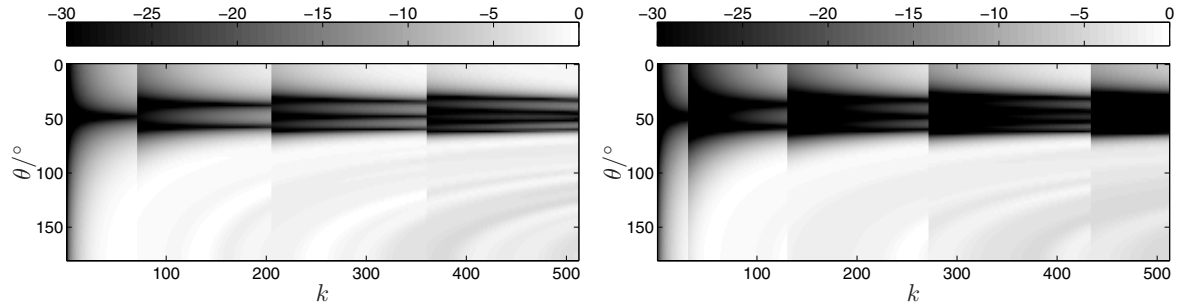
$$\chi(q, \mathbf{r}, k) = \|\mathbf{P}_q^{\perp}(k)\,\mathbf{A}(\mathbf{r}, k)\|^2 \tag{4.17}$$

and is illustrated in Figure 4.7 for two sample values of Υ.

Table 4.1: Blocking regions for the two sources corresponding to each array

$\mathcal{R}_{\theta_1}$	$\mathcal{R}_{\theta_2}$
$[25°, 65°]$	$[115°, 155°]$

The blocking matrix is computed according to (4.14)–(4.16), where the integration is either performed numerically or, under simplifying assumptions, may also be obtained in a closed form solution (see Appendix A). Note that the broad spread of the blocking region ($\pm 20°$



(a) $\Upsilon = 0.95$ (b) $\Upsilon = 0.99$

Figure 4.7: Performance of the blocking system (in dB) for the array of Figure 4.2 for a sampling frequency of $f_\mathrm{s} = 32\,\mathrm{kHz}$ and a $K = 1024$ point DFT. The discontinuities in the plot are the frequency bins at which the dimensions of the null space change. The blocking region was set to $\mathcal{R}_\theta = [25°, 65°]$.

about the mean source DOA) is chosen to account for the inaccuracies in array positioning and source spread and propagation model uncertainties. The target regions for the two sources corresponding to each array are defined as in Table 4.1. Note that the tolerance regions are the same for *all* arrays, requiring only a single computation of $\mathbf{P}_q^{\perp}(k)$ unlike the hierarchical approach where the cost function parameters for each array have to be computed separately. Further, this approach is easy to tailor to almost any automobile size and does not require a separate recalibration of the search regions as in the hierarchical approach.

4.2.6 Experimental evaluation

The presented approaches to brake squeal localization were evaluated on actual recordings made at the Volkswagen R&D testing grounds in Wolfsburg, Germany. The automobile used in the test was a VW Touran. The recordings were made under different driving conditions and surroundings. As mentioned previously, the signals were analyzed in 200 ms time segments for the presence of active frequencies and, if such frequencies were found, the hypothesis tests were carried out in the corresponding frequency bins. For more details regarding the measurement setup and the hardware, the interested reader is referred to Appendix D. The analysis parameters are presented in Table 4.2.

Table 4.2: Parameters for the experimental evaluations

f_s (kHz)	DFT length (ms)	Frame shift (ms)	Window type/ length (ms)	Likelihood threshold $\Upsilon_{\mathcal{L}}$
32	32	8	von Hann/32	0.7

A sample result is presented in Figure 4.8. Figure 4.8(a) depicts the signals recorded by the microphones[5] – here one may easily discern the squeal regions. The results obtained from

[5] As the amplitudes of the signals at each microphone vary considerably depending upon the position of the squeal event and other environmental conditions, it is difficult to select a 'good' reference microphone for all squeal conditions for plotting the reference spectrogram. To overcome this issue, Figure 4.8(a) is plotted as a *maximum* spectrogram, where the amplitude of each depicted time-frequency point is selected as the maximum across all the 32 channels. The localization results, however, are superimposed only on the spectrogram of the signal from microphone 1.

the localization approaches are presented in the next two figures, superimposed on the signal spectrogram in a color coded manner consistent with Table 4.3.

Table 4.3: Color-coding for the different wheels

FR	FL	BL	BR

We assume we have a squeal frequency detection algorithm that can detect the presence of harmonic components in the observed signal segment and extract the corresponding frequencies. Appendix C presents one algorithm for the detection of these acoustic events.

Relative comparison of the approaches

We now present a relative comparison of the hierarchical and the MaxChoice approach, evaluated on the available data. The comparison is according to four different criteria:

a.) detection and localization rate

b.) validity of the spectral disjointness assumption in MaxChoice

c.) agreement between the two approaches

d.) consistency of the two approaches during disagreement in the localization result

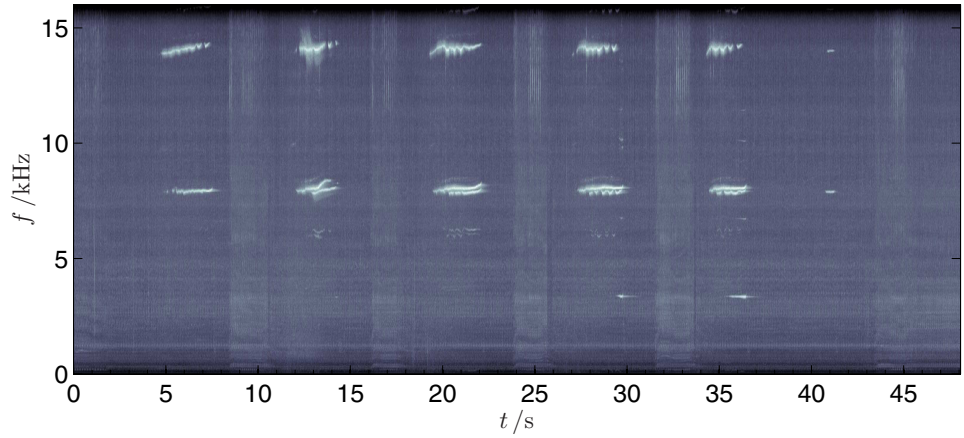
a.) Detection and localization rate This criterion measures the number of instances where each algorithm detects and localizes a squeal. This is indicated in Table 4.4, where $\mathcal{Z}$ denotes the number of localization decisions and $\overline{\mathcal{Z}}$ represents those segments where a squeal was detected but no localization decision was taken. Additionally indicated is the number of soft decisions taken by the hierarchical approach ($\mathcal{Z}_{\mathrm{s}}$). We see that the

Table 4.4: Performance comparison of hierarchical and MaxChoice – detection and localization capability.

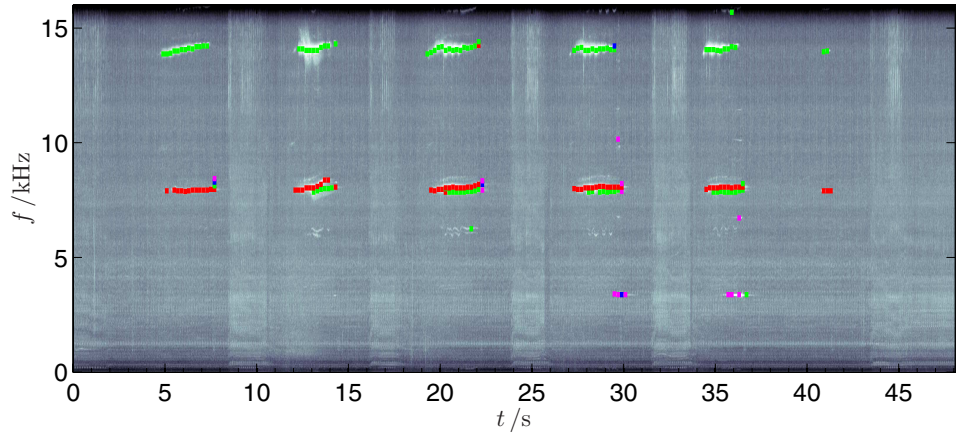
	$\mathcal{Z}$	$\overline{\mathcal{Z}}$	$\mathcal{Z}_{\mathrm{s}}$
Hier.	861	456	27
MaxCh.	895	395	N/A

data contains 1290 instances of *detected* and 895 instances (under MaxChoice) of *localized* squeal events. For any particular time frame and frequency, we say we have *detected* a squeal event if at least one array marks this frequency as active in the observed time frame. The discrepancy in the number of detected and localized events arises due to the following reasons:

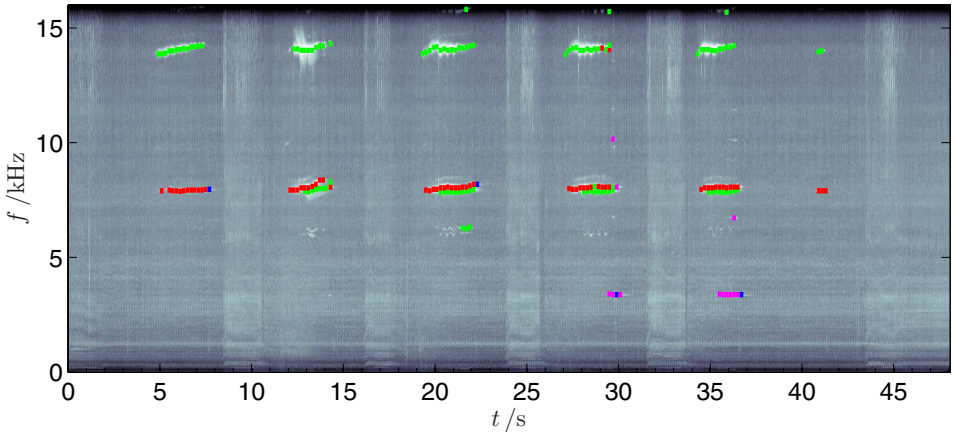
C1 When the SNR is low, there are squeal events that are detected only by the one or the other array, and therefore do not contribute to the localization result. Note that this aspect of detection does not, in any way, detract from our previous argument where we have precluded the possibility of $\mathcal{H}_0$, the hypothesis that no signal is present, in the hypothesis testing stage.



(a) Sample signal spectrum



(b) Localization results, hierarchical approach



(c) Localization results, MaxChoice approach

Figure 4.8: Sample results obtained on data recorded in real driving conditions. Note the finely spaced squeal frequencies in the spectrum. Both the proposed approaches detect them as being from different brakes and localize them correspondingly.

C2 In the *hierarchical* approach, given the zero upper threshold value of the cost function values, some frequencies where a squeal is present (i.e., detected by all the arrays) may

not be localized as the cost function assumes positive values over the corresponding search regions. Manual examination of such instances indicate that in these cases the minimum of the cost function is shifted to lie *outside* the demarcated search region. This could happen due to temporary divergence of the direct path as, for example, when turning. The hierarchical approach is more sensitive to such misalignments as compared to the MaxChoice approach due to the stronger influence of the cost function in the decision chain in the former case, as compared to the binary decision propagation based on relative differences in the latter case.

C3 In very rare cases, the MaxChoice approach may also fail to make a decision. This occurs when, for example, in the first stage the squeal is localized to the front of the automobile, but array 2, responsible for the subsequent FL/FR decision, does not detect the squeal at that frequency and time frame.

Since C3 is very rare (based on manual examination of the results), we take the MaxChoice approach as the baseline and assume that the lack of decision here is principally due to C1. As this condition must also hold for the hierarchical approach, the difference in the $\mathcal{Z}$ and $\overline{\mathcal{Z}}$ values between the two algorithms should indicate the frequency of occurrence of C2 – which is principally a drawback of the hierarchical approach. In this context, we see that the hierarchical approach fails to make a decision in an additional 61 instances as compared to the MaxChoice approach.

b.) Validity of the spectral disjointness assumption of MaxChoice We analyze here the soft decisions taken by the *hierarchical* approach, indicating the number of times multiple wheels have been allocated to a specific event. This metric should give us an idea of the validity of our assumption regarding the temporal and spectral disjointness of the squeal signals originating from different brakes. We see from Table 4.5, where $\mathcal{Z}_{(i)}$ indicates the number of instances where i wheels were localized, that only in 27 cases was a soft decision taken, a proportion of 3.4%, indicating the validity of our spectral disjointness assumption for the MaxChoice approach. Furthermore, we see that in 23 of these cases one of the decisions of the hierarchical approach correspond to the decision taken by the MaxChoice approach. The 4 instances where no agreement is reached with MaxChoice are engendered by the occurrence of condition C3, which forces MaxChoice to a 'no-decision' state.

Table 4.5: Performance comparison of hierarchical and MaxChoice – analysis of soft decisions by the hierarchical approach

$\mathcal{Z}_{\mathrm{s}}$	$\mathcal{Z}_{(2)}$	$\mathcal{Z}_{(3)}$	$\mathcal{Z}_{(4)}$	$\mathcal{Z}_{\mathrm{agree}}$ MaxChoice
27(3.4%)	15	1	1	23

c.) Agreement between the two approaches This is presented in Table 4.6 for two cases. In the first case, we consider the instances where the squeal has been detected by all arrays and therefore must be localized at least by the MaxChoice approach. We measure here how often both approaches localize such events to the same wheel. In the second case, we consider the agreement between the approaches when no decision is taken for the time segment. The discrepancy between $\overline{\mathcal{Z}}$ and $\overline{\mathcal{Z}}_{\mathrm{agree}}$ is because of the rare occurrence of C3 for the MaxChoice approach. In general we see a very good agreement between the two proposed approaches.

Table 4.6: Performance comparison of hierarchical and MaxChoice – agreement.

$\mathcal{Z}_{\text{agree}}$ (decision segments)	$\overline{\mathcal{Z}}_{\text{agree}}$ (no decision segments)
800 (89.4%)	391 (99.0%)

d.) Consistency of the results during disagreement This criterion presents the temporal consistency of the localization results for the cases where both approaches are not in agreement for a squeal event. In effect, we treat the localization result for a frame $B(l)$ and frequency k as *consistent* if the same wheel has been localized at frame $B(l \pm 1)$, for that same frequency (assuming, of course, that the squeal event spans more than one time frame – consistency in this context cannot be defined for events lasting only one time frame). We see from the presented results in Table 4.7 that the MaxChoice approach seems to be consistent for a larger proportion of the cases where the results of the approaches were different, for detected squeal events. Note, however, that the consistency measure is heuristic in the absence of ground truth and it is difficult to ascertain if a consistent result is indeed the true result. Therefore these results should be treated with some caution. $\mathcal{Z}_{\text{div}}$ indicates the total number of detected squeal events for which the two approaches were not in agreement and $\mathcal{Z}_{\text{consistent}}$ indicates the number of instances from $\mathcal{Z}_{\text{div}}$, where each approach is consistent.

Table 4.7: Performance comparison of hierarchical and MaxChoice – analysis during divergent behaviour

$\mathcal{Z}_{\text{div}}$	$\mathcal{Z}_{\text{consistent}}$		
	MaxChoice	Hierarchical	Neither
99	65	21	13

Comparison summary

Both approaches are capable of detecting and localizing closely spaced harmonics that originate from different wheels. This is illustrated in Figure 4.8, around frequencies of 7.9–8.1 kHz. Furthermore, both show a high degree of agreement in the localization performance, in general making the same decision when the squeal component has a high SNR. The time frames where the approaches diverge correspond mainly to segments that contain the squeal only in a small fraction of their total length or where the SNR is low. In such cases, the temporal consistency of the results obtained by MaxChoice is higher than that of the hierarchical approach. We reiterate that in the absence of a ground-truth, such a comparative evaluation cannot completely predict which algorithm is the better choice. However, the MaxChoice approach seems to be less sensitive to positioning errors and errors in the propagation model.

The hierarchical approach makes a soft decision in only a small percentage of the cases. In all other cases, only a single decision is made, leading us to conclude the validity of the assumption that an active frequency stems from a single source.

As a last point of note, the finite spectral resolution associated with the discrete Fourier transform coupled with the frequency jitter of the sources smears each narrowband acoustic event across more than one frequency bin. The accretion of such adjacent frequency bins

builds what we denote as a frequency *band* and the presence of an active band characterizes an acoustic event. Both the hierarchical approach and the MaxChoice approach are modified to take this spectral smearing into account. In the former case, the probability of activity is computed using the spectrally-averaged cost functions for each array and in the latter case, the decision is based on a weighted average of the decisions in the individual bins of a band. This consideration is a practical issue and does not detract from the discussion presented above.

4.3 Multiple talker localization and tracking – A parametric approach

The aim of multiple talker localization is to detect and localize a number of overlapping or competing speakers under varying levels of background noise using the spatial diversity afforded by microphone arrays. Active speaker localization is an important part of state-of-the-art communication systems and finds its use, for example, in steering the video camera towards the active speaker in video conferencing scenarios, for interference cancellation and noise suppression in hands-free systems, in automated speech diarization, and so on.

When more than two microphones are available for this purpose, the localization approach of choice is frequently the steered response power (SRP) algorithm and its variants, due to its ease of implementation and scalability in terms of addition of new sensors and selection of the candidate locations (see e.g., [110]). These algorithms are usually implemented in the narrowband STFT domain.

The imperative question handled in this application example is deciding *when* and *where* (in an appropriately defined reference system) a speaker was active. This requires

- determination of the number of active sources in any time frame,
- deciding when a new source has become active ('birth' of a source), and
- deciding when a source has fallen silent ('death' of a source).

4.3.1 Array design for speaker localization

Speech signals are broadband in nature, with most of the information in the range from 200 Hz to about 6 kHz. Often, in speech signal processing, we speak of narrowband ('telephone-quality') or wideband speech, where the former refers to signals sampled at 8 kHz and the latter to signals sampled at 16 kHz. Arrays designed for speech processing should take this frequency range into account. For the work presented here, we shall design the array to be used with both wideband and telephone-quality speech and, as an example application, we continue with the automotive scenario, this time for localizing speakers in the interior of the car. Thus we aim to design a linear array to be mounted on the rear-view mirror, which additionally constrains the maximum available length to about 25 cm. With the maximum length fixed and the lower length (about 3 cm) dictated by spatial aliasing considerations at the highest frequency of interest, the remaining sensors are distributed in this range. Setting a maximum sensor limit of 5, we come up with the staggered logarithmic spacing as in Figure 4.9, where the staggering is done in a manner so as to avoid regularity in the inter-sensor spacing. This consideration hampers the formation of grating lobes, leading to dominant minima in the localization cost function for the frequency range considered – which is a desired property.

Figure 4.9: Designed array for speaker localization. The spacings are as follows: $d_{12} = 3\,\text{cm}$, $d_{23} = 5\,\text{cm}$, $d_{34} = 7\,\text{cm}$, and $d_{45} = 10\,\text{cm}$

4.3.2 Further constraints and assumptions

Despite the broadband nature of speech signals, they demonstrate considerable sparsity in the time and frequency domain representations. Additionally, if we consider the STFT representations (obtained by the DFT on windowed and overlapped time samples) of two speaker signals, we see that they are approximately disjoint [100]. This means that if the corresponding spectra $S_1(k,b)$ and $S_2(k,b)$ are overlaid, there are very few time-frequency (T-F) points (k,b) at which the spectra overlap. Mathematically this may be expressed as:

$$S_q(k,b)S_{q'}(k,b) \approx 0 \quad \forall q' \neq q\,. \tag{4.18}$$

This is a key property we shall exploit for the multi-source localization. The disjointness also depends upon the resolution of the DFT. This has been evaluated in [135] for a sampling frequency of 16 kHz, in which case it attains its maximum for $K \in \{512, 1024, 2048\}$. Correspondingly, we fix our DFT resolution to lie in this range.

Another important aspect of multiple speaker localization is the adaptive detection of the number of active speakers. While there exist methods for this purpose, e.g., Akaike's Information Criterion (AIC), Rissanen's Minimum Description Length (MDL), or the Bayesian Information Criterion (BIC) (see, e.g., [127] and references therein), these are principally formulated for the case of stationary signals with a fixed bandwidth, and the formulation of these criteria is difficult for the case of non-stationary and spectrally disjoint signals such as speech and music. The non-stationarity in these signals is compounded by the fact that speech signals in a natural scenario are dynamic: a speaker may start, be active for a while, fall silent, and then start again. Even *within* active speaker segments, we may have speech pauses (which may be interposed with activity from other sources). Moreover, the detection using the above-mentioned approaches is coupled with the localization, requiring a multidimensional, non-linear, maximum likelihood optimization, which adds to the complexity. For these reasons, most applications either assume the number of concurrently active speakers to be known, or implicitly assume a single, dominant speaker.

We aim to design our algorithm such that we need impose *neither* the constraint of constant multi-speaker activity (competing situation) *nor* that of single source dominance.

4.3.3 Algorithm development

Consider a generic M sensor array, with the localization being done along the direction of arrival $\theta\,(\in [0, 180^\circ])$, in the azimuthal plane, where θ is measured with respect to the array axis[6]. Further, we shall consider the signal model in the STFT domain and select the subset $\{k : k_{\text{low}} < k \leq k_{\text{max}}\}$ of $K'(< K/2)$ bins for the localization. The upper bound is due to the symmetric nature of the DFT and array geometry considerations, and the lower

[6] We hasten to add that the restriction to azimuth localization in this section is only for the sake of illustration; the framework presented here will also work on a multidimensional localization grid.

bound is because very low frequencies do not yield good directional estimates for small array apertures.

For each selected discrete frequency bin k and each frame b, we compute the SRP function $\mathcal{J}_{\text{SRP}}(\theta, k, b)$ as in (3.43) over the preselected grid of search locations. Given the sparsity of speech in the STFT domain and the assumption of approximate speech disjointness, each time-frequency point (k,b) can be attributed to the dominant speaker at that T-F point, i.e.,

$$\begin{aligned} X_m(k,b) &= \sum_q A_{mq}(k)\, S_q(k,b) + V_m(k,b) \\ &\approx A_{mq'}(k)\, S_{q'}(k,b) + V_m(k,b)\,. \end{aligned} \tag{4.19}$$

Therefore, the maximum of the SRP cost function would yield an estimate of the azimuth of that speaker:

$$\widehat{\theta}_{q'}(k,b) = \underset{\theta}{\text{argmax}}\ \mathcal{J}_{\text{SRP}}(\theta, k, b). \tag{4.20}$$

Despite having computed only a single source location estimate per frequency bin, we hypothesize that the broadband nature of speech would yield enough data over these bins to localize multiple speakers. This is done by a scalar clustering of the $\widehat{\theta}(k,b)$ estimates obtained at that frame[7].

A pertinent question to be answered at this stage is the issue of multi-source localization by clustering the individual results at each frequency bin versus multi-source localization by modelling the peaks on the spectrally-averaged cost function

$$\mathcal{J}_{\text{SRP}}(\theta, b) = \sum_k \mathcal{J}_{\text{SRP}}(\theta, k, b)\,.$$

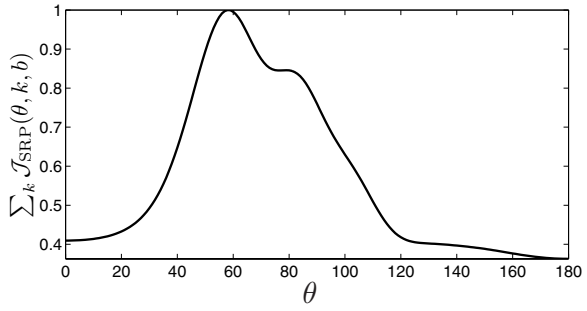
For this, we present in Figure 4.10, a sample localization result for a two speaker competing situation. Note that due to the non-linearity of the argmax in (4.20), there are two clear peaks visible in the histogram of the localization results from each bin, whereas the spectrally averaged cost function barely hints at the presence of the second source. This is examined further in Section 4.3.6. Further, as shall be shown in Chapter 6, the clustering model provides additional useful information that can be used in a wide variety of ways.

Another point to note is that the clustering should allow for overlapping clusters. Approaches such as k-means do a 'hard' partitioning of the sample space, yielding biased estimates of the cluster centriods when the data histograms are multi-modal, with the modes being in the immediate vicinity of one another.

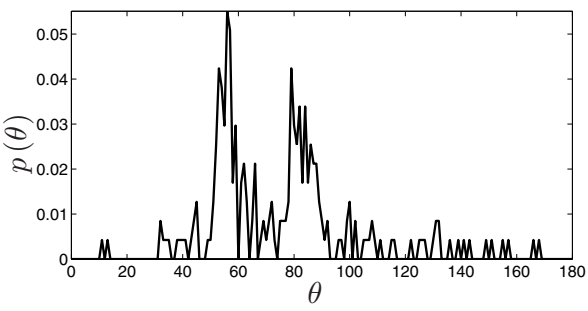
[7] An alternative approach is to use the vector of frequency normalized TDOA estimates, obtained over all microphone pairs (m, m') as:

$$\tau_{mm'}(k,b) = \frac{c}{\Omega_k}\,\text{imag}\left(\log\left(\frac{X_m(k,b)}{X_{m'}(k,b)}\right)\right)\,,$$

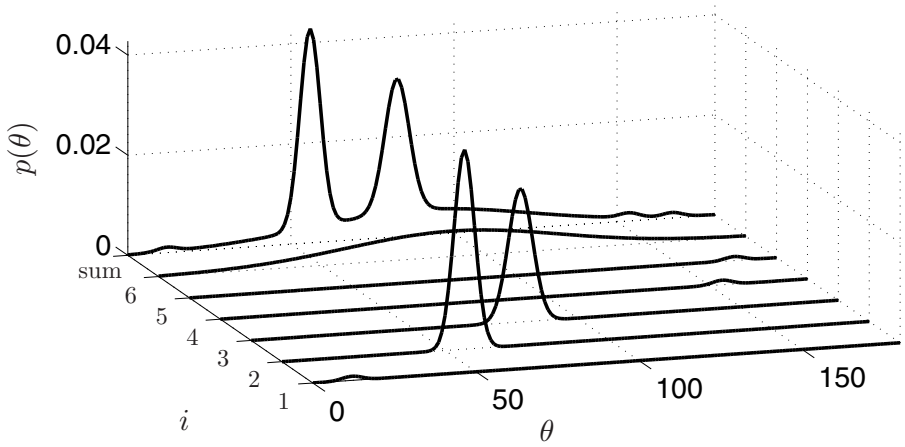
to obtain the source location from these $\binom{M}{2}$ values by non-linear optimization/parameter fitting, assuming the sparsity to hold. A similar approach is adopted in [23, 7]. The subsequent clustering is a *vector* clustering on the length $\binom{M}{2}$ vector $\boldsymbol{\tau}$ of pair-wise TDOA. The approach proposed is applicable to both clustering models, although we note that the clustering models over $\widehat{\theta}$ are more robust than those obtained from the vector clustering on $\boldsymbol{\tau}$ due to the implicit spatial averaging in the SRP which reduces the effect of grating lobes at higher frequencies in properly designed arrays. However, the SRP approach requires knowledge of the array geometry. In the *rare* cases where this is unavailable, we can still use the TDOA based clustering. However, localization is only an abstract concept in this case, having no spatial equivalent.



(a) Cost function averaging



(b) Histogram of $\widehat{\theta}(k, b)$



(c) Individual Gaussians and complete MoG fit

Figure 4.10: Justifying the 'hard-decision first' approach, symbolized by the clustering of the per-bin localization results against the 'hard-decision last' counterpart obtained by a spectral average of the SRP cost function. Two sources were active and they are clearly visible as the largest components in the clustering.

4.3.4 Cluster modelling

'Soft' clustering of the azimuth estimates is possible when one considers parametric models for the histogram, e.g, a mixture of distributions (MoD) form with the type of the individual distribution being chosen according to mathematical tractability and sample observations. Accordingly, we treat the elements $\widehat{\theta}(k, b)$ of the vector

$$\widehat{\boldsymbol{\theta}}(b) = \left(\widehat{\theta}(k_{\mathrm{low}}, b),\ \ldots,\ \widehat{\theta}(k_{\mathrm{max}}, b)\right)^T \tag{4.21}$$

as a sequence of K' realizations of a mixture of distributions process, and estimate the parameters of this process using the Expectation Maximization (EM) [13] approach. The distribution chosen here is the Gaussian distribution:

$$\widehat{\theta} \sim \sum_{i=1}^{\mathcal{I}} P_i \mathcal{N}\left(\theta_i, \sigma_i^2\right) , \tag{4.22}$$

where $\mathcal{I}$ is the model order, P_i is the weight (*a priori* evidence) and θ_i is the centroid of the ith element, with the corresponding variance σ_i^2 indicating the spatial spread of that component.

As this estimation is done on a per-frame basis, we shall subsequently drop the frame index for convenience and reintroduce it when necessary. Further, as the number of sources is not

known *a priori*, we start with a predefined model order $\mathcal{I}$. The EM clustering on the K' values $\widehat{\theta}(k)$ then yields

the means : $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_{\mathcal{I}}\right)^T$,
the variances: $\boldsymbol{\Xi} = \left(\sigma_1^2, \ldots, \sigma_{\mathcal{I}}^2\right)^T$, and
the weights/probabilities: $\mathbf{P} = \left(P_1, \ldots, P_{\mathcal{I}}\right)^T$
of the $\mathcal{I}$ components.

As the initial value of $\mathcal{I}$ is chosen to overestimate the underlying process, we next *shrink* our model if necessary. To this end we define a shrink threshold Υ_θ and:

$$\begin{aligned}
&\text{if } \exists\, i, i' \text{ such that } |\theta_i - \theta_{i'}| \leq \Upsilon_\theta, \\
&\qquad \theta_i \leftarrow \frac{P_i\theta_i + P_{i'}\theta_{i'}}{P_i + P_{i'}} \\
&\qquad \sigma_i^2 \leftarrow \frac{P_i\sigma_i^2 + P_{i'}\sigma_{i'}^2}{P_i + P_{i'}} \\
&\qquad P_i \leftarrow P_i + P_{i'} \\
&\qquad \mathcal{I} \leftarrow \mathcal{I} - 1
\end{aligned} \tag{4.23}$$

The rationale behind the selected shrinkage operation is to remove the θ_i that are very close together (in practical situations, we do not have point sources and thus the sources always have a minimum separation. Υ_θ implements this constraint). Following the sequence of steps in (4.23), the parameters for the reduced model of order $\mathcal{I} - 1$ are re-estimated *using the newly averaged values* as initial seeding for the EM:

$$\begin{aligned}
\boldsymbol{\theta}_{\text{init}} &= \left(\theta_1, \ldots, \theta_i, \ldots, \theta_{i'-1}, \theta_{i'+1}, \ldots, \theta_{\mathcal{I}}\right)^T \\
\boldsymbol{\Xi}_{\text{init}} &= \left(\sigma_1^2, \ldots, \sigma_i^2, \sigma_{i'-1}^2, \sigma_{i'+1}^2, \ldots, \sigma_{\mathcal{I}}^2\right)^T \\
\mathbf{P}_{\text{init}} &= \left(P_1, \ldots, P_i, P_{i'-1}, P_{i'+1}, \ldots, P_{\mathcal{I}}\right)^T.
\end{aligned} \tag{4.24}$$

This process is repeated until all cluster centroids have a minimum separation of Υ_θ and we are left with a model order of $\mathcal{I}_1 \leq \mathcal{I}$.

Source number estimation

The model obtained after the iterative shrinkage might still contain some ghost clusters – clusters not belonging to any source. Such a situation occurs typically when the model utilizes its extra degrees of freedom to model outliers. Such ghost components may be reduced by the following consideration [63]. The weights P_i define a discrete probability mass function (pmf). Denote by $\mathfrak{H}$ the entropy of this distribution:

$$\mathfrak{H} = \sum_{i=1}^{\mathcal{I}_1} P_i \log_2(P_i). \tag{4.25}$$

Using this value, we may estimate the number of *significant* components in the model as:

$$\mathcal{I}_2 = \underset{\mathcal{I}'}{\operatorname{argmin}} \; |\mathfrak{H} - \log_2(\mathcal{I}')|, \quad \mathcal{I}' \in \{0, 1, \ldots, \mathcal{I}_1\}. \tag{4.26}$$

If $\mathcal{I}_2 < \mathcal{I}_1$,

- select the $\mathcal{I}_1 - 1$ components with the highest weights,

- normalize weights to yield a well defined but reduced pmf, i.e.,

$$P_i \leftarrow \frac{P_i}{\sum_{i=1}^{\mathcal{I}_1-1} P_i} \quad \forall i \leq (\mathcal{I}_1 - 1) \tag{4.27}$$

- $\mathcal{I}_1 \leftarrow \mathcal{I}_1 - 1$
- repeat (4.25)–(4.27) until $\mathcal{I}_2 = \mathcal{I}_1$.

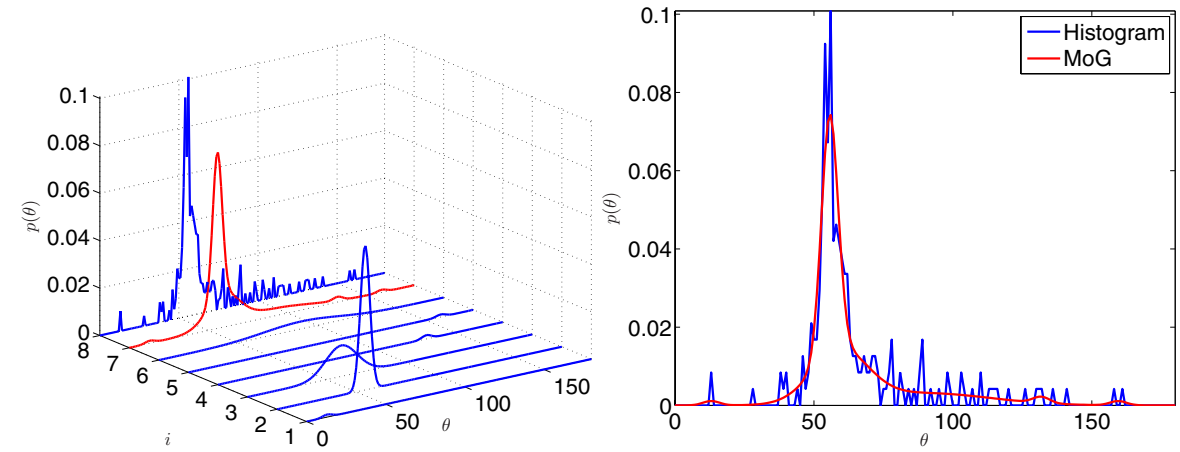
Note that, irrespective of the difference between $\mathcal{I}_1$ and $\mathcal{I}_2$, the shrinkage in the first step only reduces the number of elements by 1. This conservative shrinkage strategy was chosen as it yielded the best results as compared to a direct shrinkage by $\mathcal{I}_1 - \mathcal{I}_2$ elements.

The number of clusters thus obtained indicates the estimated number of sources in that frame, the means correspond to the respective azimuth estimates, the weights to the respective probability of activity and the variance to the spatial spreads.

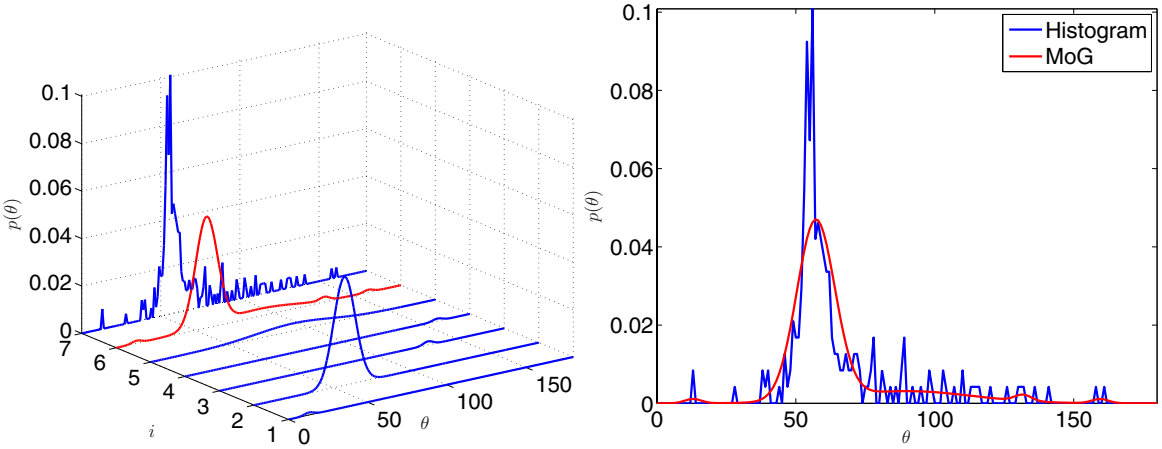
The rationale for the computation of $\mathcal{I}_2$ as in (4.26) derives from the information theoretic relationship between entropy and the optimal average symbol bitlength in source coding. Only, in this case, we aim to find the most significant components in the estimated mixture of Gaussians (MoG) representation. If all the components are equally significant, we would have a uniform distribution on the P_i, and $\mathfrak{H} = \log_2(\mathcal{I}_1)$. If, however, this distribution is 'peaky', some components are more significant, and this information can be used to reduce the redundancy (the ghost values) by the proposed iterative procedure.

Why re-estimate?

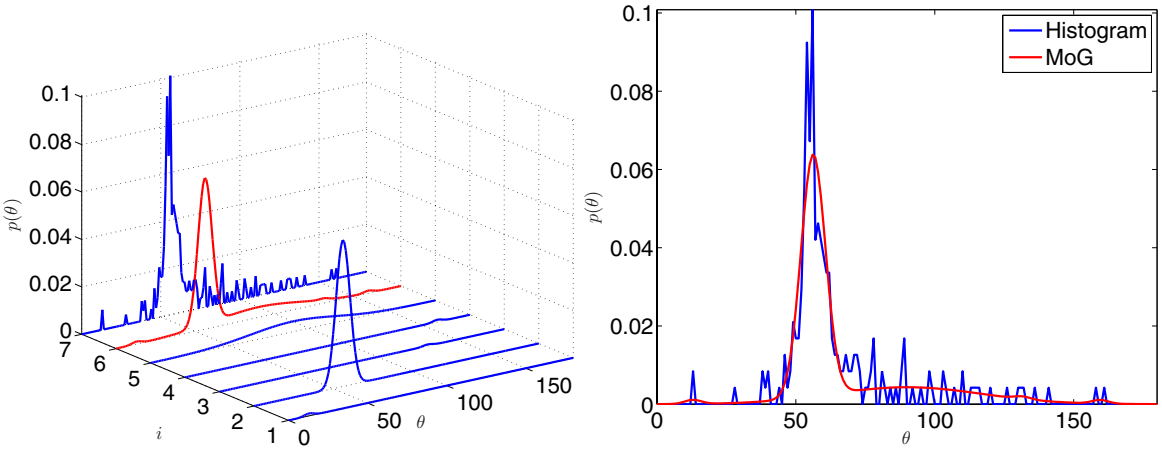
To justify the necessity of the re-estimation step after updating the parameters as in (4.23), consider Fig. 4.11. This presents the histograms and the estimated MoG fit (the penultimate curve) for the three cases: (a) initial estimation (overdetermined), (b) shrunk model (from (4.23)) and (c) re-estimated, compacter model ($\mathcal{I} - 1$ elements) with initial seeding from (4.24). The histogram obtained from the SRP function is the last component. We see that the MoG fit to the histogram improves after re-estimation, as compared to simply shrinking. This is also indicated by the corresponding lower Kullback-Leibler divergence (KLD).



(a) Overdetermined model ($\mathcal{I} = 5$), KLD = 0.3803.



(b) Shrunk model ($\mathcal{I} = 4$), KLD = 0.4966.



(c) Shrunk and re-estimated model ($\mathcal{I} = 4$), KLD = 0.4510.

Figure 4.11: MoG fit illustrating the effect of shrinkage. (a) indicates the initial (overdetermined) estimate; (b) the estimate after shrinkage according to (4.23) and (c) the model obtained after shrinkage and re-estimation with initial seeding as in (4.24). The azimuth is plotted along the x-axis, the respective MoG components on the y-axis. The z-axis denotes the pdf of the MoG components over θ. One source was active.

4.3.5 Source Tracking

In general, the number of detected sources $\mathcal{I}_2$ changes from frame to frame. This time variance of $\mathcal{I}_2$ is due to the following reasons:

- the shrinkage- and the entropy-based model order selection cannot completely eliminate the formation of ghost clusters without further knowledge. Proper intialization and denoising of the $\widehat{\boldsymbol{\theta}}$ help, but only marginally. Thus, some means might be spurious, caused by errors in the cost function, especially at lower frequencies and frequencies with low SNR,
- some means might indicate the onset of a new source,
- additionally, it could also happen that a source detected in the previous frame(s) is not present in the current frame (within the threshold window $[-\Upsilon_\theta, \Upsilon_\theta]$) – indicating either a speech pause for that source or a sign that the source has fallen silent.

Note that only a constantly active source would contribute to a MoG component i whose mean (θ_i) does not change significantly from frame to frame. The MoG model of Section 4.3.4 cannot account for this time variant nature of $\mathcal{I}_2$, and cannot distinguish between transient and active sources. Therefore, we shall extend our localization framework to include a non-linear, token-based temporal smoothing stage to preserve the sources of interest and eliminate the transient sources. This is denoted hereafter as the tracking stage.

For source tracking, we maintain a *frame-independent* record of *averaged* means, variances and probabilities, denoted as:

$$\begin{aligned}\overline{\boldsymbol{\theta}} &= \left(\overline{\theta_1}, \ldots, \overline{\theta_{\mathcal{I}_T}}\right)^T \\ \overline{\boldsymbol{\Xi}} &= \left(\overline{\sigma_1^2}, \ldots, \overline{\sigma_{\mathcal{I}_T}^2}\right)^T \\ \overline{\mathbf{P}} &= \left(\overline{P_1}, \ldots, \overline{P_{\mathcal{I}_T}}\right)^T\end{aligned} \tag{4.28}$$

where $\mathcal{I}_T$ indicates the number of elements currently present in the averaged mixture. Further, borrowing an idea from packet-switched networks, we associate with each source $\bar{i}$ in the averaged set a 'time to live', $\mathrm{TTL}_{\bar{i}}$ token. Now consider the $\mathcal{I}_2(b)$ elements of the current frame b obtained as detailed in the previous section.

If $\exists\, i(b)\,, \bar{i}$ such that $|\theta_{i(b)} - \overline{\theta_{\bar{i}}}| \leq \Upsilon_\theta$

$$\begin{aligned}\overline{\theta_{\bar{i}}} &\leftarrow \frac{\overline{\theta_{\bar{i}}}\,\overline{P_{\bar{i}}} + \theta_{i(b)} P_{i(b)}}{\overline{P_{\bar{i}}} + P_{i(b)}} \\ \overline{\sigma_{\bar{i}}^2} &\leftarrow \frac{\overline{\sigma_{\bar{i}}^2}\,\overline{P_{\bar{i}}} + \sigma_{i(b)}^2 P_{i(b)}}{\overline{P_{\bar{i}}} + P_{i(b)}} \\ \overline{P_{\bar{i}}} &\leftarrow \overline{P_{\bar{i}}} + P_{i(b)} \\ \mathrm{TTL}_{\bar{i}} &\leftarrow \min\left(\mathrm{TTL}_{\max}, \mathrm{TTL}_{\bar{i}} + 1\right)\end{aligned} \tag{4.29a}$$

and for each $i(b)$ such that $\nexists\, \bar{i}$, with $|\theta_{i(b)} - \overline{\theta_{\bar{i}}}| \leq \Upsilon_\theta$,

$$\begin{aligned}\overline{\boldsymbol{\theta}} &\leftarrow \overline{\boldsymbol{\theta}} \cup \theta_{i(b)} \\ \overline{\boldsymbol{\Xi}} &\leftarrow \overline{\boldsymbol{\Xi}} \cup \sigma_{i(b)}^2 \\ \overline{\mathbf{P}} &\leftarrow \overline{\mathbf{P}} \cup P_{i(b)} \\ \mathcal{I}_T &\leftarrow \mathcal{I}_T + 1 \\ \mathrm{TTL}_{\mathcal{I}_T} &\leftarrow \mathrm{TTL}_{\min}\end{aligned} \tag{4.29b}$$

and for each $\bar{i}$ such that $\nexists\, i(b)$, with $|\theta_{i(b)} - \overline{\theta_{\bar{i}}}| \leq \Upsilon_\theta$,

$$\mathrm{TTL}_{\bar{i}} \leftarrow \mathrm{TTL}_{\bar{i}} - 1\,. \tag{4.29c}$$

Following this, $\overline{\mathbf{P}}$ is renormalized to guarantee $\sum_{\bar{i}} \overline{P_{\bar{i}}} = 1$.

Equation (4.29a) indicates the update for a source that was already present in previous frames. Equation (4.29b) handles the situation where a new source has possibly entered the system and (4.29c) indicates the case where an existing source was absent in the current frame. Within this framework, sources with a TTL ≤ 0 are considered to have 'died' and are removed from the averaged set. Note that the limitation $\mathrm{TTL}_{\max}$ is required in order to limit the source lifetime, giving us a reasonable period of aging and death for a source that is not active anymore.

4.3.6 Experimental evaluation

A preliminary version of this approach presented in [76] showcased many of the algorithm features – specifically: the capability of using coarser search grids due to the implicit interpolation provided by the model, the capability of maintaining a speaker location estimate in the presence of short speech pauses and primitive speaker tracking for slowly moving sources (small movements of the head or gradual position change).

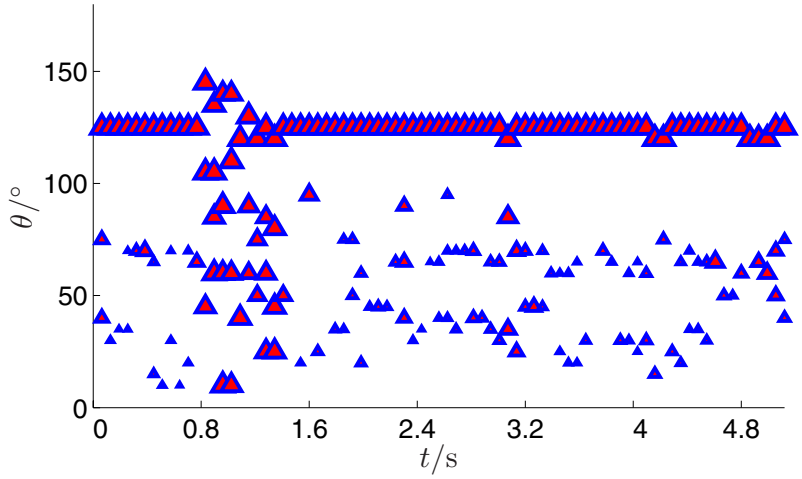
In the evaluation here we are concerned with a more rigorous testing of the approach vis-à-vis a modified version of the traditional SRP-PHAT approach that can localize multiple sources. Principally this modification consists of selecting a *maximum* of $\mathcal{I}_{\mathrm{SRP}}$ local maxima from the spectrally-averaged SRP-PHAT cost function of (3.46). The corresponding algorithm is denoted in the following as the multi-source SRP (M-SRP).

Throughout the simulations and the comparisons, the system parameters were set as in Table 4.8. The selfsame parameter settings are used in the speech enhancement algorithms developed in Chapter 6.

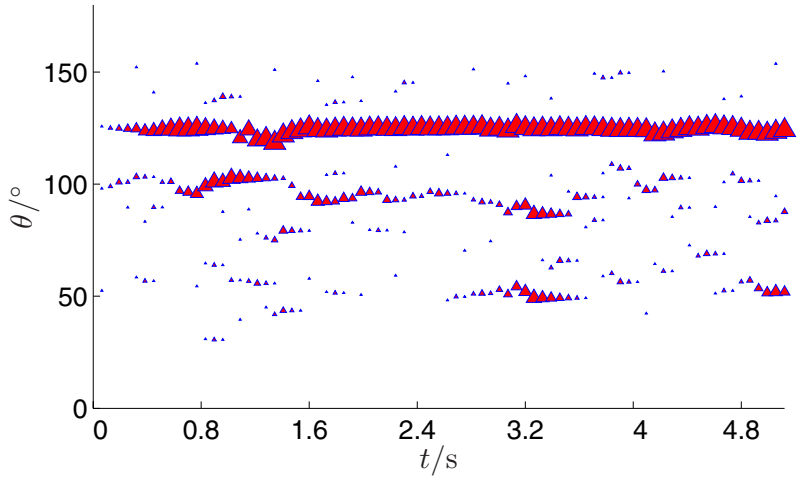
Table 4.8: Parameters for the experimental evaluations

f_{s} (kHz)	DFT length K (ms)	Frame shift (ms)	Window type/ length (ms)	$\mathcal{I}$	$\mathcal{I}_{\mathrm{SRP}}$	Υ_θ (°)
8	128	16	von Hann/128	5	5	10

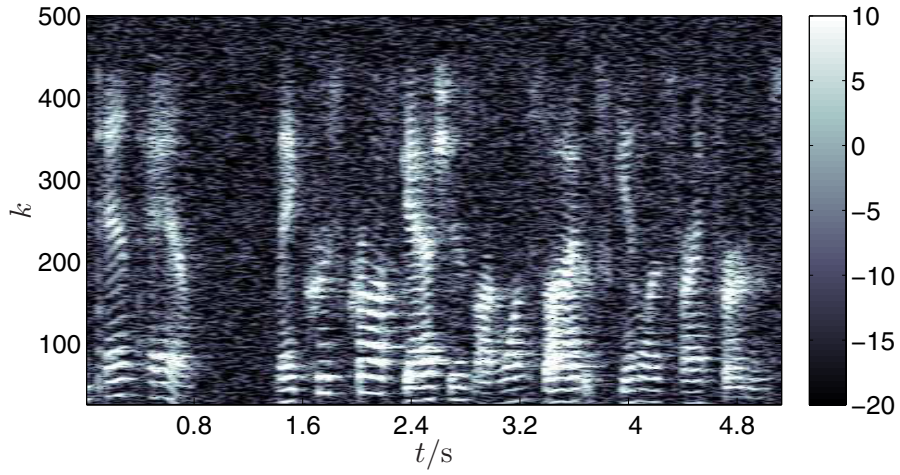
Before we proceed with the rigorous comparisons however, we shall illustrate the performance of the algorithm in specific scenarios so as to give the reader unacquainted with [76] a feel for the algorithm behaviour. To begin with, we present the localization of a single source in a reverberant room.



(a) M-SRP, single source



(b) MoG, single source

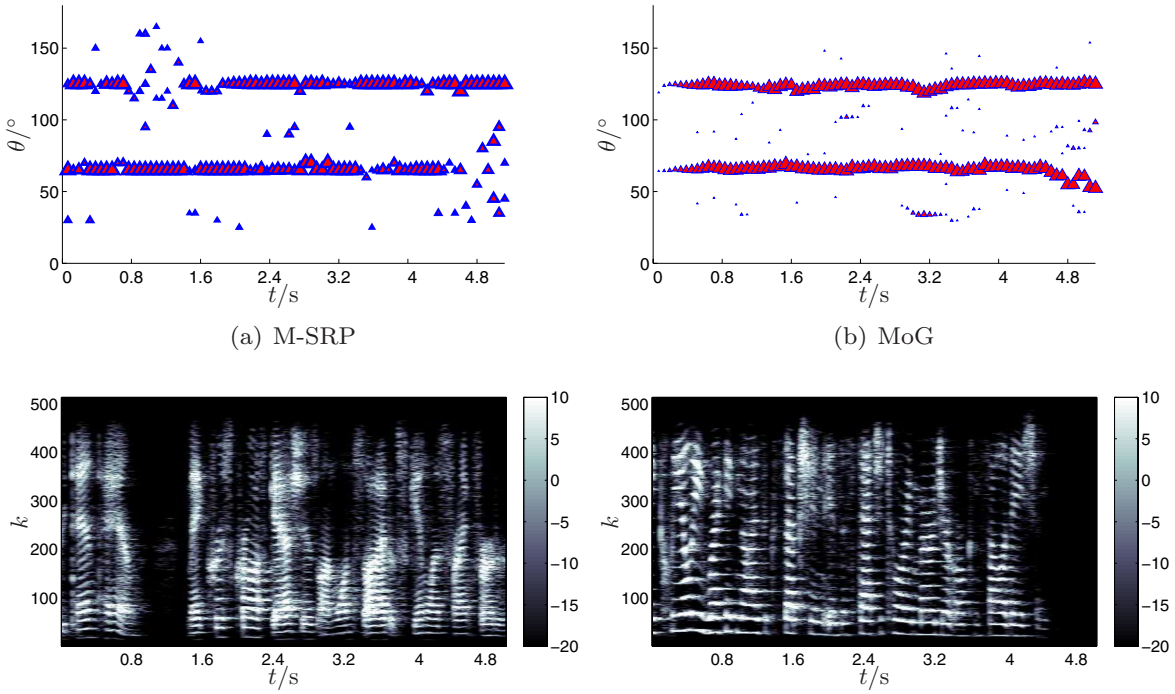


(c) Source signal spectrum

Figure 4.12: The first figure shows the estimated azimuth from the M-SRP approach and the second plot indicates the performance of the MoG based approach. The signal spectrum is presented in the third plot. The size of the marker in the SRP plot is proportional to the relative height of the corresponding local maximum and that in the MoG plot is proportional to the TTL of the corresponding MoG element. The search grid quantization is $5°$ and the true azimuth is $120°$. The SNR was 15 dB.

With reference to Figure 4.12, we see that both approaches are able to obtain a good estimate of the source azimuth. Note, however, that the MoG approach, with the implicit smoothing provided by the tracking mechanism, is capable of maintaining the source position estimate even in the presence of small speech pauses, as seen in the time frame 0.8 s–1.6 s. The M-SRP produces random results in this period.

Next, consider the case of two simultaneously active sources at $\theta_1 = 60°$ and $\theta_2 = 120°$ presented in Figure 4.13.

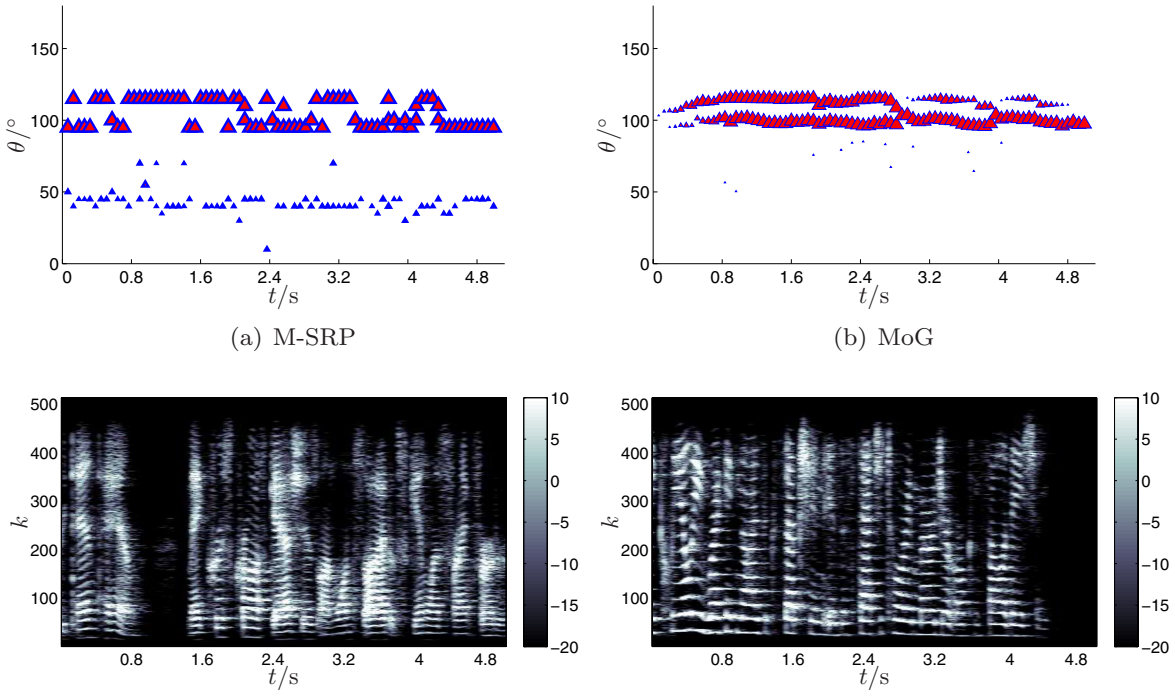


(a) M-SRP (b) MoG

(c) Source 1 signal spectrum (d) Source 2 signal spectrum

Figure 4.13: The left figure shows the estimated azimuth from the spectrally averaged SRP-PHAT approach and the plot on the right indicates the performance of the MoG based approach.

We see that both approaches detect two sources and with comparable accuracy. The advantage of the tracking mechanism is again evident during speech pauses. But as this framework could just as easily be integrated into the M-SRP approach, it seems thus far that there is no clear advantage to the MoG-based model for localization. However, a look at the comparative performance with closely spaced sources immediately disabuses us of this notion (Figure 4.14). Here, the M-SRP either localizes the one source or the other, as the spectrally averaged cost function tends to merge the peaks. The MoG-based algorithm, on the other hand, shows two clear 'tracks', corresponding to the two sources. We note that in some frames the two tracks merge – this is a consequence of the tracking mechanism and occurs when one speaker dominates the localization histogram. When this dominance fades, the second speaker is again allocated a dedicated track (this is evidenced by the small TTL values which gradually increase around $t = 3\,\mathrm{s}$ and $t = 4\,\mathrm{s}$).



(a) M-SRP (b) MoG

(c) Source 1 signal spectrum (d) Source 2 signal spectrum

Figure 4.14: The left figure shows the estimated azimuth from the M-SRP approach and the plot on the right indicates the performance of the MoG based approach.

In short, when the sources to be localized are relatively widely spaced, the M-SRP and the MoG based approach are able to localize both sources with similar accuracy. But, as the azimuthal separation between the sources reduces, the M-SRP approach begins to fail. We expect a similar effect when the number of concurrent sources increases, causing individual peaks in the M-SRP function to merge.

Now we come to the rigorous testing of the algorithms. As mentioned previously, a comparison of the M-SRP approach with the MoG approach incorporating source tracking is slightly unfair, as such a framework may also be incorporated for the SRP approach. Therefore, we shall compare the performance of the M-SRP approach with the results obtained using the *raw* MoG model, MoG_r, which is the intermediate result obtained by the MoG clustering at each time frame, and *without* tracking. This comparison of M-SRP with MoG_r should provide an indication of the performance enhancement arising from using the clustering approach over the frequency-averaged maxima search of the M-SRP. We shall also separately compare the results of MoG_r with the MoG approach incorporating source tracking, which should indicate the additional gain in performance obtained by the implicit temporal smoothing.

The experimental setup and data used are detailed in Appendix E. The performance of the respective algorithms are tested according to their (a) hit percentage and (b) localization accuracy. The hit percentage $\mathcal{Z}$ is defined separately for the single source case and the multi-source case. In the former, it is the percentage of time frames in which the algorithm estimates a source position in the *vicinity* of the true source, with the vicinity threshold being set to $\Upsilon_\text{hit} = \pm 10°$. This criterion measures how often the localization algorithm localizes a source around the true source position and does not, in any way, indicate the accuracy of the localization estimate. For the multi-source case, we define two kinds of hit percentages,

1. $\mathcal{Z}_1$, which measures the percentage of frames where the algorithm localizes *at least* one source within its vicinity, and
2. $\mathcal{Z}_2$, which measures the percentage of frames where the algorithm localizes both sources within their respective vicinities.

In all cases the speakers were always active with no significant speech pauses, thereby obviating the need for a separate voice activity detector for the performance evaluations.

The localization accuracy is measured by the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\text{E}\{(\theta - \widehat{\theta})^2\}}, \tag{4.30}$$

where the expectation is replaced by a temporal average in practice. For this averaging, we only consider the $\widehat{\theta}$ in frames where the sources have been localized to within the hit threshold.

The results for each combination of spatial environment, background noise, SNR and position are averaged over all the speakers for the single speaker case. For the concurrent speaker situation, the results corresponding to the same azimuthal *difference* in the speaker location are averaged over all such speaker combinations, e.g., results for $\{\theta_1 = 60°, \theta_2 = 90°\}$ and $\{\theta_1 = 90°, \theta_2 = 120°\}$ are averaged over all the speaker combinations in these positions.

Single source localization in noise

Table 4.9: Performance for single source localization in the synthetic room

Noise type	SNR (dB)	$\mathcal{Z}$ (%)			RMSE (°)		
		SRP	MoG_r	MoG	SRP	MoG_r	MoG
Clean	N/A	99.86	100	100	0.71	1.42	1.22
White	0	64.74	76.84	96.18	3.68	4.81	4.20
	10	85.54	89.36	99.33	2.89	3.50	2.65
	20	96.01	97.89	100	1.91	2.42	1.65
White, diffuse	0	77.18	90.75	99.23	3.25	4.39	3.77
	10	90.01	95.74	100	2.64	3.20	2.59
	20	97.20	99.39	100	1.81	2.29	1.75
Babble, diffuse	0	88.76	95.95	99.95	2.31	3.05	2.42
	10	97.52	99.32	100	1.59	2.10	1.77
	20	99.49	100	100	0.97	1.62	1.45

Table 4.10: Performance for single source localization in the office room (Room 2)

Noise type	SNR (dB)	$\mathcal{Z}$ (%)			RMSE (°)		
		SRP	MoG_r	MoG	SRP	MoG_r	MoG
Clean	N/A	100	100	100	0	1.31	1.00
White	0	81.03	83.92	96.03	3.39	4.09	3.00
	10	95.00	95.68	99.59	2.24	2.74	1.97
	20	99.15	99.00	99.95	1.31	1.86	1.46
White, diffuse	0	86.81	92.10	98.77	3.07	3.67	3.09
	10	96.71	98.21	99.73	1.93	2.38	1.83
	20	99.36	99.54	100	1.01	1.76	1.64
Babble, diffuse	0	95.33	98.09	99.87	1.75	2.28	1.89
	10	99.17	99.76	100	0.82	1.74	1.58
	20	99.87	99.93	100	0.44	1.56	1.38

Multi-source localization in noise, $\Delta\theta = 60°$

Table 4.11: Localization of concurrent speakers, Room 1, azimuthal separation = 60°.

Noise type	SNR (dB)	$\mathcal{Z}_1$ (%)			$\mathcal{Z}_2$ (%)			RMSE (°)		
		SRP	MoG_r	MoG	SRP	MoG_r	MoG	SRP	MoG_r	MoG
Clean	N/A	100	100	100	92.07	97.99	99.93	2.22	2.06	1.41
White	0	91.96	95.30	99.95	66.94	74.57	95.71	3.82	4.61	4.15
	10	98.46	99.41	100	82.11	90.53	99.57	3.05	3.50	2.72
	20	99.73	99.95	100	88.83	97.20	100	2.56	2.52	1.64
White, diffuse	0	98.85	99.86	100	84.00	92.01	99.73	2.97	4.37	3.80
	10	99.85	100	100	90.48	97.07	99.87	2.63	3.07	2.19
	20	100	100	100	91.64	98.43	100	2.31	2.40	1.57
Babble, diffuse	0	99.46	100	100	86.07	96.77	99.98	2.53	2.77	1.88
	10	99.95	100	100	90.95	98.08	100	2.32	2.23	1.46
	20	100	100	100	92.50	98.43	99.98	2.23	2.15	1.46

Table 4.12: Localization of concurrent speakers, Room 2, azimuthal separation = 60°.

Noise type	SNR (dB)	$\mathcal{Z}_1$ (%)			$\mathcal{Z}_2$ (%)			RMSE (°)		
		SRP	MoG_r	MoG	SRP	MoG_r	MoG	SRP	MoG_r	MoG
Clean	N/A	100	100	100	94.92	93.66	99.02	1.39	1.66	1.28
White	0	97.68	99.15	100	82.12	83.60	97.84	3.10	3.95	3.04
	10	99.90	99.94	100	95.55	93.87	99.95	2.42	2.89	2.18
	20	100	100	100	97.91	95.50	100	1.94	2.10	1.53
White, diffuse	0	99.25	100	100	89.11	93.06	99.14	2.94	3.53	2.89
	10	100	100	100	96.97	96.29	100	2.02	2.35	1.62
	20	100	100	100	98.11	94.94	100	1.64	1.75	1.16
Babble, diffuse	0	99.96	100	100	95.19	94.61	99.96	1.93	2.52	1.95
	10	100	100	100	96.69	94.38	99.98	1.59	2.03	1.49
	20	100	100	100	96.13	93.55	100	1.45	1.75	1.33

Multi-source localization in noise, $\Delta\theta = 30°$

Table 4.13: Localization of concurrent speakers, Room 1, azimuthal separation = 30°.

Noise type	SNR (dB)	$\mathcal{Z}_1$ (%)			$\mathcal{Z}_2$ (%)			RMSE (°)		
		SRP	MoG_r	MoG	SRP	MoG_r	MoG	SRP	MoG_r	MoG
Clean	N/A	100	99.90	100	85.22	93.69	99.32	2.52	2.54	2.18
White	0	87.89	96.49	99.97	62.08	82.03	98.07	3.59	4.54	3.76
	10	96.65	99.46	100	71.77	93.12	99.95	2.94	3.50	2.47
	20	99.55	100	100	79.29	95.17	99.74	2.57	2.70	2.17
White, diffuse	0	93.39	98.30	99.97	72.09	91.99	99.82	3.35	4.40	3.60
	10	98.41	99.85	100	77.78	95.59	99.85	2.86	3.31	2.52
	20	99.82	99.98	100	81.82	95.14	99.81	2.59	2.70	2.18
Babble, diffuse	0	99.33	99.92	100	83.79	95.58	99.81	2.66	2.99	2.35
	10	99.85	100	100	85.29	94.42	99.75	2.50	2.62	2.25
	20	99.95	99.81	100	85.28	93.55	99.68	2.48	2.49	2.25

Table 4.14: Localization of concurrent speakers, Room 2, azimuthal separation = 30°.

Noise type	SNR (dB)	$\mathcal{Z}_1$ (%)			$\mathcal{Z}_2$ (%)			RMSE (°)		
		SRP	MoG_r	MoG	SRP	MoG_r	MoG	SRP	MoG_r	MoG
Clean	N/A	100	100	100	82.54	91.74	97.96	2.14	2.47	2.35
White	0	91.18	97.66	100	58.42	87.95	99.30	3.18	3.79	2.86
	10	98.62	99.82	100	66.59	89.70	98.97	2.75	3.00	2.81
	20	99.98	99.90	100	74.89	89.55	98.24	2.32	2.68	2.85
White, diffuse	0	97.09	99.34	100	67.73	93.25	99.69	2.86	3.33	2.52
	10	99.62	99.98	100	71.82	92.30	99.74	2.32	2.62	2.22
	20	100	99.88	100	77.16	90.87	98.78	2.02	2.36	2.25
Babble, diffuse	0	99.82	99.95	100	79.93	91.34	99.14	1.88	2.62	2.40
	10	100	99.98	100	80.81	91.01	98.86	1.83	2.53	2.48
	20	100	99.98	100	81.90	90.88	98.85	1.99	2.41	2.50

Discussion

The following points are evident from the tables. Firstly, we see that the M-SRP and MoG approaches localize the sources with comparable accuracy. This is partly to be expected as the MoG algorithm utilizes the SRP function to obtain the individual source estimates. Nevertheless, it is a sanity check and is an indicator of the performance of the MoG modelling stage. We reiterate that this performance measure is computed on localization estimates that lie within a neighbourhood of the true azimuth and is *not* influenced by the detection performance.

With respect to the hit percentage for single source localization, the performance of M-SRP and MoG_r are very similar at high SNRs. In low SNR conditions, we have a flattening of the spectrally averaged SRP function along with the introduction of spurious maxima, which leads to false estimates of the source positions using this approach. On the other hand, as the MoG approach performs a localization in each frequency bin, enough information is available (due to the correct localization in bins with high SNR) to localize the sources. Consequently MoG_r performs better than M-SRP under low SNR conditions. The introduction of the tracking mechanism further improves the performance of the MoG-based approach, as this framework can preserve the source location estimate over frames where the source signal is swamped by the noise floor or during speech pauses.

With respect to the competing speaker situation, we note that when the sources are widely-spaced and the SNR is high, the M-SRP and MoG_r yield comparable results. Again, at low SNRs, MoG_r performs better for the reasons mentioned previously.

When the sources are closely-spaced, the performance degradation of the M-SRP is dramatic, especially at low SNRs. This behaviour is expected as the spectral averaging of the M-SRP tends to coalesce the cost function peaks corresponding to the individual sources. The MoG_r approach, however, optimally utilizes the sparsity and disjointness of speech to effect a good localization even under such conditions. The corresponding disparity between the M-SRP and MoG_r results are clearly evident (Tables 4.13–4.14), pointing towards the wisdom of using a 'hard-decision first' approach in this case.

As expected, introducing the tracking framework consistently boosts performance.

A slightly surprising result may be seen in the two-source localization case, where the performance of the algorithms in the presence of noise is marginally *improved* at high SNRs, as compared to the clean mixture. We postulate that the reason for this in M-SRP is the flattening of the cost functions in noise-only bins, which has the effect of enhancing the peaks of weak signal components, distinguishing them as local maxima; and in the MoG approach the generation of a noise floor, which assists the convergence of the EM algorithm to peaks centered around the true source azimuths.

4.4 Conclusions

This chapter presented the design of localization algorithms, from the ground-up, for two radically different scenarios. Although the basic statistic required for each approach is the same (namely the multi-channel cross correlation), the algorithms are drastically different because of the different source characteristics, design constraints and other *a priori knowledge* available.

The first application was that of brake squeal localization, the aim being to locate a squeal event to the brakes that generated it. Since the *a priori* knowledge of the approximate source

positions is available and these remain fixed, we treat this application from the point of view of a detection problem. Given the complex time-varying acoustic environment underneath the chassis, a model of the propagation paths is not easy to construct. Additionally, a model that is robust against errors in sensor placement and defects is even more complex. Thus, standard approaches for hypothesis testing as in [66, 79], or those developed in the previous chapter, fail. Consequently, we construct approaches that implicitly consider such variations. We have presented two such approaches, the first arising from heuristic considerations and the second based on robustness against model imperfections and the further simplifying assumption of spectral disjointness of the sources.

Both approaches perform well on the test data. A comparison of the two approaches shows that their performance is very similar, highlighting, on the one hand, the good choice of the heuristic and, on the other hand, the validity of the simplifying assumptions of the second approach.

The next application considered was the localization of multiple talkers. When the number of sources is not known, or when it is time-variant, traditional model-order estimation algorithms are difficult to formulate. This problem is compounded for speech signals given their sparsity, non-stationarity, disjointness of spectra, and the real-time requirements. In this regard we have proposed a simple and elegant framework, based on a mixture of Gaussians clustering, that exploits the sparsity and disjointness of speech. We have shown that our framework functions at least as well as the generic SRP in the case of a single talker, in different background noise scenarios, and outperforms it in the case of multiple concurrent sources – more so when the spatial separation between the sources to be localized decreases. Additionally, our approach incorporates a simple tracking framework for moving speakers. The tracking will work, as long as the source does not move more than Υ_θ from one frame to the next, and may be enhanced by suitable state-of-the-art algorithms (e.g., [44]).

We have also presented some design considerations for the array geometry and layout, a topic not often considered in conjunction with algorithm development. The array design is coupled to the algorithm being used and the application at hand. Under given constraints of geometry and maximum permissible number of sensors, the main motivating considerations for determining the inter-sensor spacing is the spatial-aliasing over the effective frequency range of the application. In general, use of logarithmic or staggered logarithmic inter-sensor spacing hampers the formation of grating lobes and allows for approximately constant-width beamforming over frequency, which is of benefit for source localization.

In summary, this chapter along with the previous one should present the reader with some 'rule-of-thumb' ideas for developing application-specific source localization algorithms. Further extensions to the speaker localization framework will be considered in the context of source separation in Chapter 6.

II

Source Separation

...'So two and one more is . . . ?' Detritus looked panicky. This was calculus territory.

Men at arms
– Terry Pratchett

Chapter 5

Multi-Channel Source Separation and Enhancement – Overview

5.1 Introduction

Central to the idea of source separation is the concept of adaptive 'nulling', where all sources other than the target source are suppressed. Multi-channel approaches in this context usually function by exploiting the spatial diversity. When such a system is implemented as a *linear* spatial filter, the process is known as 'beamforming' and the corresponding spatial filters are termed as beamformers. In such a case, we often speak of interference *cancellation* or source *enhancement*.

The suppression of the interferences may also be done by varying the system gain: setting a low gain when the interference is dominant and a high gain when the source of interest predominates. We shall term such an operation 'masking' and the corresponding gain functions masks. The spatial diversity is used here to obtain information regarding source or interference presence. Such algorithms are implemented as *non-linear* filters. In this case, we refer to the improvement obtained as interference *suppression*.

We note here that the term non-linear filter, as used in this context, can engender some confusion. A short discussion on this issue is therefore presented in Appendix F.

Note that while the term *interferer* is used to represent a source (generally directive) other than the source of interest, in signal separation where the goal is to resynthesize individual source signals from a mixture, this term may be misleading as, depending upon the context, a source may function as an interferer or a target. As an example, in the two source scenario with signals $s_1(n)$ and $s_2(n)$, the latter is the interferer when $s_1(n)$ is to be extracted and vice-versa. The role of a source will be explicitly disambiguated where required in the following discussion.

Beamforming algorithms may be divided into two broad categories: algorithms based on second-order statistics (SOS) and those based on higher-order statistics (HOS). Examples of the latter are the independent component analysis (ICA) based algorithms [57] whereas the generalized sidelobe canceller (GSC) [45] and the PCA based algorithms of [32] are examples from the former category.

Mask-based separation algorithms [134, 101, 123] are realized based on specific properties of the source signals. In a very general sense, the aim of such algorithms is to partition a

set of data points (sensor signals in their time or frequency domain representations) into clusters, with the data points in each cluster belonging to a specific source. The clustering may be 'hard' (non-overlapping clusters – each data point can belong only to one cluster) or 'soft' (overlapping clusters – many sources may share a data point, albeit to different degrees).

A *practical* realization of both categories of separation approaches requires some kind of localization information, e.g., for resolving the permutation ambiguity in the case of the ICA approaches, for parameterizing the blocking matrix in the GSC structure or as a feature for the clustering in the mask-based approaches. Further, as in the case of localization, where the custom alterations are required for optimal performance in each application scenario, source separation algorithms too must be customized for the application in question. This requires taking into consideration, when designing the source separation system, all available *a priori* knowledge such as array dimensions, source characteristics, etc.

Our work in this part focusses on the application of multi-channel separation and enhancement algorithms to acoustic sources, specifically speech, in a noisy environment with competing (multiple sources simultaneously active) and non-competing source scenarios. It is to this end that we devote this and the following chapter.

5.2 Signal model

The signal model we shall assume for the sources that need to be extracted is the STFT domain formulation introduced in (2.20) in Chapter 2, reproduced below with the first sensor taken as the reference. We again assume the dominance of the direct path as, for diffuse target sources spatial diversity does not yield much of an advantage in terms of beamforming or masking, and algorithms for the extraction of the corresponding source signals in such scenarios is out of the scope of this work.

$$\begin{aligned}\mathbf{X}(k,b) \approx & \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ \left|\frac{A'_{M1}(k)}{A'_{11}(k)}\right| e^{\jmath\,\Omega_k \Delta\tau_{M1}} & \cdots & \left|\frac{A'_{MQ}(k)}{A'_{1Q}(k)}\right| e^{\jmath\,\Omega_k \Delta\tau_{MQ}} \end{pmatrix} \mathbf{S}(k,b) \\ & + \begin{pmatrix} \breve{A}_{11}(k) & \cdots & \breve{A}_{1Q}(k) \\ \vdots & \ddots & \vdots \\ \breve{A}_{M1}(k) & \cdots & \breve{A}_{MQ}(k) \end{pmatrix} \mathbf{S}(k,b) + \mathbf{V}(k,b) \\ & \triangleq \mathbf{A}(k)\,\mathbf{S}(k,b) + \mathbf{V}(k,b), \end{aligned} \tag{5.1}$$

where $\mathbf{A}(k)$ represents the matrix of *relative* transfer functions from each source signal component at the reference microphone to all the other microphones. Note, again, that the relation is only approximate in the STFT domain as the DFT is computed on windowed, finite-length signals, whereas the relative transfer functions are usually much longer than the window length. The assumption of equality, however, does not detract from the ideas propounded below.

In addition to the spatial diversity, a property of speech signals used frequently in enhancement algorithms is their sparseness[1], both in their time and their frequency representations. As a result of this sparseness, speech signals show evidence of a strongly super-Gaussian distribution of their amplitudes in the time domain and of the real and imaginary parts of their corresponding DFT coefficients in the STFT domain, on a long-term basis [92, 95, 80]. The *approximate* disjointness (see Section 4.3.3) of speech signals in the STFT domain is another property that is exploited in conjunction with sparsity, especially by masking approaches.

5.3 Mask-based, non-linear approaches

Masks are usually generated in the STFT domain at each time-frequency (T-F) point. Suppression of interferers is obtained by masking T–F points which do not belong to the target source. The prerequisite for such masking is the availability of a classifier that allows a distinction between the various sources. For speech sources, this may be obtained, in the multi-channel case, by clustering the T–F points according to two features: the inter-sensor signal *delay* and the inter-sensor signal *attenuation*, both of which are functions of the source location. Consider, for an $M = 2$ case, the following estimates of attenuation and delay:

$$\begin{aligned} \widehat{|A_2(k,b)|} &= \left| \frac{X_2(k,b)}{X_1(k,b)} \right| \\ \widehat{\tau_2(k,b)} &= -\Omega_k^{-1} \arg\left(\frac{X_2(k,b)}{X_1(k,b)} \right). \end{aligned} \tag{5.2}$$

Then, if source disjointness holds, the estimated values for attenuation and delay for each (k,b) should correspond to only one source. Subsequently, if the features thus estimated are grouped into Q clusters in the two dimensional plane defined by $|A_2(k)|$ and $\tau_2(k)$, the centroid $(\tau_{2q}(k)\,,\ |A_{2q}(k)|)^T$ of each cluster q may be taken as an estimate for the true parameters for each source. Note that these parameters may, in general, be frequency variant, which fact is indicated by expressing them as a function of the respective frequency bin. Masks $\mathcal{M}$ may then be defined for a source q as:

$$\mathcal{M}_q(k,b) = \begin{cases} 1 & q = \underset{q'}{\operatorname{argmin}} \left\| \left(\widehat{\tau_2(k,b)},\ \widehat{|A_2(k,b)|}\right)^T - \left(\tau_{2q'}(k)\,,\ |A_{2q'}(k)|\right)^T \right\| \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

and the target source spectrum obtained by applying the mask to the spectrum of any microphone signal m:

$$\widehat{S}(k,b) = \mathcal{M}_q(k,b) X_m(k,b)\,. \tag{5.4}$$

The corresponding time domain signal may then be resynthesized using standard techniques.

The advantage of mask-based approaches is that they may be applied in underdetermined scenarios. Further, the interference suppression obtained using such approaches is quite high. The drawback, however, is the distortion of the target signal. Also, mask based approaches usually evince artefacts in the form of spurious spikes in the signal spectrum leading to what has been termed as *musical noise* in the reconstructed signals.

[1] Note that while sparseness in the strict sense implies that most signal components are zero, in the practical sense we take it to indicate that most signal components have a very small value.

Early approaches to alleviating these problems considered the usage of 'soft' masks $\mathcal{M} \in [0, 1]$ in (5.3). However such improvement was usually at the cost of lower interference suppression [46]. Recently however, it was demonstrated in [18] that a temporal smoothing of gain functions in their *cepstral* representation preserves target signal quality and interference suppression whilst successfully reducing musical noise. The result is a more natural sounding signal. While the approach was presented in the context of single channel speech enhancement, it has also been successfully applied to source separation in [74].

5.4 Linear approaches – ICA

The disjointness assumption in the previous section places a rather restrictive constraint on the separation performance. Where more than one source is active, we either have a *winner takes all* strategy when using binary masks, leading to the degradation of the other signal, or increased cross-talk when using soft masks. The linear spatial filtering approaches, on the other hand, allow for the fact that more than one source may be active at any T-F point. Enhancement (or, equivalently, interference cancellation) is achieved by steering a null in the direction of the interferer. If the direction of all the sources are unknown, they must first be estimated before beamforming may be performed.

Alternatively, when the sources are concurrently active, approaches exploiting the non-stationarity of the signals [21, 20, 94, 60] or the higher order statistics [57, 103, 8, 104, 102] may be used to directly separate the signals, the positions of the sources being implicitly estimated during the process. In this section we shall consider approaches based on the ICA model, which usually operate on the STFT domain representation of the signal model as in (5.1).

The linear separation problem is then posed as the search for a so-called demixing *matrix* $\mathbf{W}(k)$ for each frequency bin k such that:

$$\mathbf{Y}(k, b) = \mathbf{W}(k)\,\mathbf{X}(k, b) \approx \mathbf{S}(k, b), \tag{5.5}$$

where, following our convention, $\mathbf{S}(k, b)$ is the vector of individual source signals as received at the *reference* microphone. In order to cast our contribution to this class of approaches in the proper light, we shall first briefly describe the ICA based approach. For a more detailed explanation the interested reader is referred to [57], or to [59] for an excellent introduction.

ICA based separation approaches assume a generative mixing model of the kind in equation (5.1) and treat the source signals as realizations of Q independent (but not necessarily identical) random processes with an underlying *non-Gaussian* distribution, and basically aim to find a demixing system such that the elements Y_q of $\mathbf{Y}$ are 'as statistically independent of each other as possible'. The logic is intuitive: if the underlying signals are statistically independent, obtaining independent variables at the output of the demixing system can only mean that the sources have been separated.

The search for any such $\mathbf{W}$ lies in the generation and optimization of a $\mathbf{W}$–parameterized cost function that penalizes the dependence between the variables Y_q. Examples of popular cost functions are:

- the absolute value of the kurtosis,
- the Kullback-Leibler divergence,
- non-polynomial approximations to the entropy, and

- mutual information.

Further, use is often made of the FastICA [58, 56] algorithm, which is a fixed-point algorithm operating on *pre-processed* data, and which restricts the search for the optimal separation matrix to the space of orthonormal matrices $\overset{\vee}{\mathbf{W}}$. For a brief introduction to FastICA, the reader is referred to Appendix B. The pre-processing consists of two stages:

- centering, consisting of mean removal (the means may be re-introduced post-separation); and
- sphereing, which mutually decorrelates the sources and, additionally, normalizes each source to unit variance. Dimension reduction for the case of overdetermined systems ($M > Q$) may also be done here.

The latter stage may be recognized as a standard PCA. It may be shown [57, 73] that such pre-processing reduces the search to that over the space of orthonormal matrices which, in turn, allows for the implementation of the FastICA principle. Thus, we may write:

$$\mathbf{W}(k) = \overset{\vee}{\mathbf{W}}(k)\,\overset{\circ}{\mathbf{W}}(k)\,, \tag{5.6}$$

with $\overset{\circ}{\mathbf{W}}(k)$ being the sphereing stage estimated from the correlation matrix of the input signals and $\overset{\vee}{\mathbf{W}}(k)$ being the orthonormal matrix to be estimated using FastICA.

In the absence of any *a priori* knowledge regarding the elements $A_{mq}(k)$ of $\mathbf{A}(k)$ or the underlying processes, the solution for $\mathbf{W}(k)$ contains two inherent ambiguities: *scale* and *permutation*. Any $\mathbf{W}(k)$ that fulfills:

$$\mathbf{W}(k) = \mathbf{J}(k)\,\mathbf{D}(k)\,\mathbf{A}^{-1}(k)\,, \tag{5.7}$$

with $\mathbf{J}(k)$ being a permutation matrix and $\mathbf{D}(k)$ being a diagonal scaling matrix, is a valid solution to the optimization problem. Such ambiguities are, in themselves, not critical. The problem arises because $\mathbf{J}(k)$ and $\mathbf{D}(k)$ are, in general, different in different frequency bins. We shall term a permutation and scaling as *local* if it is for a particular frequency bin, and as *global* if it is common to all bins. Obviously source reconstruction from the spectral domain is only successful if the permutation and scaling matrices are global. In other words if

$$\begin{aligned} &\mathbf{J}(k_a) = \mathbf{J}(k_b) \quad \text{and} \\ &\mathbf{D}(k_a) = \mathbf{D}(k_b) \quad \forall\ a, b \in \{0, \ldots, K\}\,. \end{aligned} \tag{5.8}$$

We note here that the key issue with ICA-based approaches is *not* in the selection of the optimal *core* algorithm. Rather, it lies in finding a practicable solution to the permutation and scaling problem. While the scaling problem is rather reliably solved by the minimal distortion principle [82], inter-frequency permutation remains a rather formidable issue, especially due to its combinatorial nature. It is to this aspect of the ICA solution that much research has been devoted in the recent past.

Work done in this area includes the inter-frequency amplitude envelope correlation (AmDeCor) approach of [4], where the spectral amplitude envelopes for a given source are assumed to be similar over frequencies, and this similarity is exploited for permutation resolution, and the inter-frequency transfer function correlation approaches of [89], which assume that the transfer functions across neighbouring bins are correlated and use the estimate of the demixing system of a previous frequency bin to initialize the search for the current bin.

These methods are based, however, on assumptions that might not hold in general. Further, the danger with such serial approaches is the propagation of errors made at one frequency through successive bins.

Approaches based on localization cues obtained either explicitly or implicitly (estimated within the ICA framework) may be applied *independently* at each bin, preventing error propagation across frequencies, and are therefore quite popular [60, 102, 85]. In [60], the 'beam-pattern' of the demixing matrix is used to generate direction of arrival (DOA) information, which is subsequently used for permutation resolution; in [102] the demixing matrices are selected for each frequency bin as the best option from a null-beam solution and an ICA solution, using source localization estimates and a quality criterion that is based upon the coherence function; [8] uses an anechoic approximation to the mixing model and forms a cumulant based cost function for optimizing the null-beamformers for the considered model, again using the DOA estimates for permutation resolution, whereas [94] argues that proper initialization of the demixing matrices – using geometric constraints based on beamformers – obviates the need for permutation correction in the proposed algorithm. The DOA based approach of [85] or the clustering approach of [7] for permutation solution are similar to the beam-pattern approach of [60] in that each cluster in the pattern is allocated to a source.

A recent algorithm that iteratively applies the amplitude correlation approach and the DOA based approach was proposed in [104]. However, while this algorithm is indeed rather robust, it is also computationally very expensive. Also, none of the beamformer based approaches utilize the beamformed signals themselves and while a proper initialization of the demixing matrices as in [94] decreases the reconstruction error, the problem is not completely solved. In Section 6.1 of the following chapter, we shall propose our alternative, based on an appropriately spatially-filtered input, which allows for a single-step resolution of the permutation problem.

5.5 Linear approaches – GSC

The ICA based HOS approaches work well, but have a few drawbacks:

1. they require $M \geq Q$ and
2. they are implemented as batch realizations, which implies large latency in the output signals.

Consequently, in this section we shall explore other alternatives based on the second order statistics (SOS) – namely the linearly constrained minimum variance (LCMV) beamformer [42] and its adaptive variant the generalized sidelobe canceller (GSC) [45].

5.5.1 The LCMV beamformer and the GSC

The LCMV is designed primarily for a single target source, in the presence of directive interferers and noise. Given the *a priori* knowledge of the steering vector[2] $\mathbf{A}_q(k)$ for the

[2] The LCMV and related algorithms were originially formulated for a system containing only a delay, implying that the elements A_{mq} of $\mathbf{A}$ consist simply of a phase shift. However, an extension to a more general propagation matrix is straightforward and it is this model we imply when we refer to the LCMV and the other algorithms in this category.

desired source q, the signal model may be reformulated as:

$$\begin{aligned}\mathbf{X}(k,b) &= \mathbf{A}_q(k)\, S_q(k,b) + \sum_{\substack{q'=1\\ q'\neq q}}^{Q} \mathbf{A}_{q'}(k)\, S_{q'}(k,b) + \mathbf{V}(k,b) \\ &= \mathbf{A}_q(k)\, S_q(k,b) + \breve{\mathbf{V}}(k,b),\end{aligned} \tag{5.9}$$

where $\breve{\mathbf{V}}(k,b)$ now represents the combined effect of interference and noise. The desired signal may then be enhanced by finding an optimal vector $\mathbf{W}_q(k)$ such that:

$$\mathbf{W}_q(k) = \underset{\mathbf{W}(k)}{\operatorname{argmin}}\ \mathbf{W}^H(k)\, \mathbf{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}(k)\, \mathbf{W}(k) \tag{5.10}$$

subject to

$$\mathbf{W}^H(k)\, \mathbf{A}_q(k) = \mathcal{G}(k)\,, \tag{5.11}$$

where $\mathcal{G}(k) \in \mathbb{C}^{1\times 1}$ is the desired linear constraint. The solution to this constrained optimization problem is:

$$\mathbf{W}_q(k) = \frac{\mathcal{G}^* \mathbf{\Psi}^{-1}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}\, \mathbf{A}_q(k)}{\mathbf{A}_q^H(k)\, \mathbf{\Psi}^{-1}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}(k)\, \mathbf{A}_q(k)}\,. \tag{5.12}$$

For the case $\mathcal{G} = 1$, the solution is also known as the minimum variance distortionless response (MVDR) beamformer [31].

In general, however, the statistics of the interference and noise signal $\breve{\mathbf{V}}(k,b)$ are not known. An alternative is then to obtain $\mathbf{W}_q(k)$ as:

$$\mathbf{W}_q(k) = \underset{\mathbf{W}(k)}{\operatorname{argmin}}\ \mathbf{W}^H(k)\, \mathbf{\Psi}_{\mathbf{X}\mathbf{X}}(k)\, \mathbf{W}(k)\,, \tag{5.13}$$

subject to the same constraint. This is known as the linearly constrained minimum power (LCMP) solution. It is easy to show that the estimate of $\mathbf{W}_q$ using (5.10) or (5.13) yields the same result if $\mathbf{A}_q(k)$ is accurately known (see [121], for example).

The GSC is based on the LCMP principle. However, the system as proposed in [45] optimizes the elements of $\mathbf{W}(k)$ in an *unconstrained* manner, making the algorithm computationally more efficient. The GSC for a target source q consists of three basic building blocks: the fixed beamformer $\mathbf{W}_{q,\mathrm{d}} \in \mathbb{C}^{M\times 1}$, the blocking matrix $\mathbf{B}_q \in \mathbb{C}^{M\times(M-1)}$ and the adaptive multi-channel noise canceller $\mathbf{W}_{q,\breve{\mathbf{V}}} \in \mathbb{C}^{(M-1)\times 1}$, with:

$$\mathbf{W}_q(k) = \mathbf{W}_{q,\mathrm{d}}(k) - \mathbf{B}_q(k)\, \mathbf{W}_{q,\breve{\mathbf{V}}}(k)\,. \tag{5.14}$$

The fixed beamformer is responsible for imposing the linear constraint and is designed so that:

$$\mathbf{W}^H_{q,\mathrm{d}}(k)\, \mathbf{A}_q(k) = \mathcal{G}(k)\,. \tag{5.15}$$

The columns of the blocking matrix are designed to span the *orthogonal complement* of the desired signal subspace. Consequently, it does not affect the desired signal component and generates the reference for the spatially correlated noise, which is subsequently cancelled from the output of the fixed beamformer by the adaptive component $\mathbf{W}_{q,\breve{\mathbf{V}}}$. In other words, if the output is denoted as

$$Y_q(k,b) = \mathbf{W}_q^H(k)\, \mathbf{X}(k,b)\,,$$

we have:

$$Y_q(k,b) = \mathbf{W}_{q,\mathrm{d}}^H(k)\,\mathbf{A}_q S_q(k,b) \; + \big(\mathbf{W}_{q,\mathrm{d}}(k) - \mathbf{B}_q(k)\,\mathbf{W}_{q,\breve{\mathbf{V}}}(k)\big)^H \breve{\mathbf{V}}(k,b) \tag{5.16}$$

whence the optimal noise canceller is found as

$$\begin{aligned}\mathbf{W}_{q,\breve{\mathbf{V}},\mathrm{opt}}(k) &= \underset{\mathbf{W}_{q,\breve{\mathbf{V}}}(k)}{\operatorname{argmin}} \; \mathrm{E}\left\{|Y_q(k,b)|^2\right\} \\ &= \big(\mathbf{B}_q^H(k)\,\mathbf{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}(k)\,\mathbf{B}_q(k)\big)^{-1}\mathbf{B}_q^H(k)\,\mathbf{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}(k)\,\mathbf{W}_{q,\mathrm{d}}(k)\;.\end{aligned} \tag{5.17}$$

Note that, in the above, the direction of the desired source is assumed known, allowing us to compute the fixed beamformer and the blocking matrix in advance. Most current applications of GSC to speech are designed for a single target source under such assumptions. However, these assumptions are idealistic, leading to target signal cancellation and corresponding output degradation when they are violated as, e.g, when $\mathbf{A}_q(k)$ is imperfectly known. To counter this, several safeguards have been proposed in the recent past. These include the robust GSC solution of [51] – where the noise canceller is constrained to adapt only within the range *outside* a tolerance region for the target source, preventing target signal cancellation – and the TF-GSC approach of [43] – where the signal model considered is similar to that in (5.1) and the $\mathbf{A}_q$ is *estimated* assuming stationary background noise. Such approaches belong to the class of 'data-driven' MVDR beamformers. A more recent, but essentially similar, alternative to the TF-GSC has been proposed in [28] and is summarized below.

5.5.2 Data-driven MVDR beamforming

We shall again consider the narrow band formulation for the multi-channel signal enhancement case as in (5.9), namely, for microphone m:

$$X_m(k,b) = A_{mq}(k)\,S_q(k,b) + \breve{V}_m(k,b)\,. \tag{5.18}$$

We further assume the target source signal to be uncorrelated with the noise $\breve{V}_m(k,b)$. In the following, the source index and the frequency bin index are dropped for convenience, as the processing is considered to be done for a generic source and independently at each bin. Further, defining $Z_m \triangleq A_m S$, we may also write (5.18) compactly as:

$$\mathbf{X} = \mathbf{Z} + \breve{\mathbf{V}}\,. \tag{5.19}$$

Now, assuming that there exists a mapping of the target signal component at all microphones m' to that at a specific *reference* microphone m, i.e., that there exist filters $H_{m'} \in \mathbb{C}^{1\times 1}$ such that:

$$Z_{m'} = H_{m'} Z_m\,, \quad \forall m' \in \{1,\ldots,M\}\,, \tag{5.20}$$

then, for signal enhancement, we seek a filter $\mathbf{W}$ such that:

$$\mathbf{W}_{\mathrm{opt}} = \underset{\mathbf{W}}{\operatorname{argmin}} \; \mathrm{E}\big\{|\mathbf{W}^H\mathbf{X}|^2\big\} \tag{5.21}$$

subject to

$$\mathbf{W}^H\mathbf{H} = 1\,, \tag{5.22}$$

where $\mathbf{H} = (H_1\,,\ldots\,,H_M)^T$.

This constrained optimization leads to the solution:

$$\mathbf{W}_{\text{opt}} = \frac{\mathbf{\Psi}_{\mathbf{XX}}^{-1}\mathbf{H}}{\mathbf{H}^H\mathbf{\Psi}_{\mathbf{XX}}^{-1}\mathbf{H}}\,. \tag{5.23}$$

Consider our signal model of (5.9), modified using the definition of $\mathbf{Z}$ as previously. If we assume that (5.20) holds, we may write an alternative signal model for the system as:

$$\mathbf{X} = \mathbf{H}Z_m + \check{\mathbf{V}}\,, \tag{5.24}$$

with

$$\mathbf{\Psi}_{\mathbf{XX}} = \mathbf{H}\mathbf{H}^H\Psi_{Z_mZ_m} + \mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}\,, \tag{5.25}$$

in which case, the solution for the optimal $\mathbf{W}$ from (5.23) simplifies to

$$\mathbf{W}_{\text{opt}} = \frac{\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}^{-1}\mathbf{H}}{\mathbf{H}^H\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}^{-1}\mathbf{H}}\,. \tag{5.26}$$

This form may be recognized as the standard MVDR beamformer, with the target direction vector being estimated by $\mathbf{H}$. The vector $\mathbf{H}$ may also be seen as the *relative transfer function* between the microphones and, in this sense, the solution is *equivalent* to the TF-GSC approach.

Estimation of H

The coefficients $H_{m'}$ indicate the mapping of the target signal component at all microphones m' to that at a specific *reference* microphone m. These coefficients may be estimated in the mean square error sense as:

$$H_{m',\,\text{opt}} = \underset{H_{m'}}{\text{argmin}}\ \mathrm{E}\big\{|Z_{m'} - H_{m'}Z_m|^2\big\}, \tag{5.27}$$

which leads to the solution

$$H_{m',\,\text{opt}} = \frac{\Psi_{Z_{m'}Z_m}}{\Psi_{Z_mZ_m}}. \tag{5.28}$$

Stacking the results for each $H_{m'}$ into a vectorial form, we obtain:

$$\mathbf{H} = \Psi_{Z_mZ_m}^{-1}\mathbf{\Psi}_{\mathbf{Z}Z_m}\,. \tag{5.29}$$

While this approach compensates for irregularities in the knowledge of the true propagation vector, it still requires estimates of $\mathbf{\Psi}_{\mathbf{ZZ}}$ and $\widehat{\mathbf{\Psi}}_{\check{\mathbf{V}}\check{\mathbf{V}}}$. In [43, 28], the case of a single desired target source enhancement in a *stationary* noise field that is not fully coherent across the array is considered. Here, the power spectral density of the noise $\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}$ is estimated by temporal averaging in the noise-only periods. Next, during periods of target signal activity, the power spectral density matrix of the microphone signals $\mathbf{\Psi}_{\mathbf{XX}}$ is estimated. From this, and noting that target signal and noise are uncorrelated, we obtain the matrices required for the estimation of $\mathbf{H}$ as:

$$\mathbf{\Psi}_{\mathbf{ZZ}} = \mathbf{\Psi}_{\mathbf{XX}} - \mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}} \tag{5.30}$$

$$\begin{aligned}\mathbf{\Psi}_{\mathbf{Z}Z_m} &= \mathrm{E}\big\{\mathbf{Z}Z_m^*\big\}\\ &= \mathbf{\Psi}_{\mathbf{ZZ}}\mathbf{e}_m\end{aligned} \tag{5.31}$$

where $\mathbf{e}_m = [0, \ldots, 0, 1_m, 0, \ldots, 0]^T \in \mathbb{R}^{M\times 1}$ is the column selection vector, with only element $e_m = 1$.

Note, however, that additional algorithms are required for detecting the noise-only periods. For the case of spatially uncorrelated noise, for example, such detection is accomplished by setting a threshold on the spatial coherence function of the microphone signals.

5.6 Optimum multi-channel filtering

We shall again consider the narrowband formulation for the multi-channel signal enhancement case, as in (5.19), for a generic source. We further assume the target source signal $A_m(k)\, S(k,b)$ at any microphone m to be independent of the noise $\breve{V}_m(k,b)$ at that microphone.

5.6.1 Multi-channel Wiener filter (MWF)

The traditional multi-channel Wiener filter can be formulated in the following manner: given the microphone signals and knowledge of the target source, the aim is to design a spatial filter $\mathbf{W}_{\text{opt}}$ such that:

$$\mathbf{W}_{\text{opt}} = \underset{\mathbf{W}}{\text{argmin}}\ \mathrm{E}\{|\mathcal{G}S - \mathbf{W}^H\mathbf{X}|^2\}, \tag{5.32}$$

where $\mathcal{G} \in \mathbb{C}^{1\times 1}$ is an arbitrary, desired response. Optimizing (5.32) yields:

$$\mathbf{W}_{\text{opt}} = \mathcal{G}^*\boldsymbol{\Psi}_{\mathbf{XX}}^{-1}\boldsymbol{\Psi}_{\mathbf{X}S}\,, \tag{5.33}$$

where $\boldsymbol{\Psi}_{\mathbf{XX}} = \mathrm{E}\{\mathbf{X}\mathbf{X}^H\}$, $\boldsymbol{\Psi}_{\mathbf{X}S} = \mathrm{E}\{\mathbf{X}S^*\}$.

An analysis of this solution is instructive. We begin with writing:

$$\boldsymbol{\Psi}_{\mathbf{XX}} = \mathbf{A}\mathbf{A}^H\Psi_{SS} + \boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}.$$

Then, applying Woodbury's identity, we compute the inverse of $\boldsymbol{\Psi}_{\mathbf{XX}}$ as:

$$\boldsymbol{\Psi}_{\mathbf{XX}}^{-1} = \boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1} - \Psi_{SS}\frac{\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}}{\left(1 + \Psi_{SS}\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}\right)}. \tag{5.34}$$

Substituting (5.34) in (5.33), and using $\boldsymbol{\Psi}_{\mathbf{X}S} = \Psi_{SS}\mathbf{A}$, we obtain:

$$\begin{aligned}\mathbf{W}_{\text{opt}} &= \mathcal{G}^*\Psi_{SS}\frac{\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}}{\left(1 + \Psi_{SS}\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}\right)} \\ &= \underbrace{\frac{\mathcal{G}^*\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}}{\left(\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}\right)}}_{\text{MVDR}}\underbrace{\frac{\Psi_{SS}\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}}{\left(1 + \Psi_{SS}\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}\right)}}_{\text{post-filter}}.\end{aligned} \tag{5.35}$$

Note that the post-filter may be manipulated into a more recognizable form as:

$$\frac{\Psi_{SS}\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}}{\left(1 + \Psi_{SS}\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}\right)} = \frac{\Psi_{SS}}{\left(\Psi_{SS} + \left(\mathbf{A}^H\boldsymbol{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1}\mathbf{A}\right)^{-1}\right)}, \tag{5.36}$$

which is readily recognizable as an analogue of the single-channel noise reduction Wiener filter. Thus, we see that the optimum filter for the MMSE estimation in the multi-channel case consists of an MVDR beamformer and a single channel Wiener post-processor. Note that while the MVDR beamformer preserves the signal in the target direction, leading to undistorted response, the post-filtering introduces target signal distortion. Thus, a multi-channel Wiener filter (MWF), by default, is *not* a distortionless system and *always* perturbs the target signal when $\left|\left(\mathbf{A}^H \mathbf{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}^{-1} \mathbf{A}\right)^{-1}\right| > 0$. This is evident here and has also been pointed out in [53].

Alternative implementation

In general, the filter developed above is impractical as the source signal is usually not available. For practical purposes, the desired signal is chosen to be the *target signal component* at any particular microphone m. In other words, the optimal filter of (5.32) is modified such that:

$$\mathbf{W}_{\text{opt}} = \underset{\mathbf{W}}{\operatorname{argmin}}\ \mathrm{E}\{|Z_m - \mathbf{W}^H \mathbf{X}|^2\}, \tag{5.37}$$

where $Z_m \triangleq A_m S$ is the target signal component of channel m. We may also pose the cost function for the above optimization in the form:

$$\mathcal{J}(\mathbf{W}) = \mathrm{E}\{|\mathbf{e}_m^T \mathbf{Z} - \mathbf{W}^H \mathbf{X}|^2\}, \tag{5.38}$$

where $\mathbf{e}_m = [0, \ldots, 0, 1_m, 0, \ldots, 0]^T \in \mathbb{R}^{M\times 1}$ is the selection vector as before, and $\mathbf{Z} = \mathbf{A}S$. Solving the above, we obtain the optimal filter $\mathbf{W}_{\text{opt}}$ as:

$$\begin{aligned} \mathbf{W}_{\text{opt}} &= \mathbf{\Psi}_{\mathbf{XX}}^{-1} \mathbf{\Psi}_{\mathbf{XZ}} \mathbf{e}_m \\ &= \mathbf{\Psi}_{\mathbf{XX}}^{-1} \mathbf{\Psi}_{\mathbf{ZZ}} \mathbf{e}_m \,. \end{aligned} \tag{5.39}$$

Under similar assumptions as in Section 5.5.2 for (5.30), this now allows a more practical way to estimate the required quantities: in the noise-only periods, the power spectral density of the noise $\widehat{\mathbf{\Psi}}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}$ is estimated. Next, during periods of target signal activity, the power spectral density matrix of the microphone signals $\widehat{\mathbf{\Psi}}_{\mathbf{XX}}$ is estimated. From this, and noting that target signal and noise are uncorrelated, we obtain an estimate for $\mathbf{\Psi}_{\mathbf{ZZ}}$ as in (5.30):

$$\widehat{\mathbf{\Psi}}_{\mathbf{ZZ}} = \widehat{\mathbf{\Psi}}_{\mathbf{XX}} - \widehat{\mathbf{\Psi}}_{\breve{\mathbf{V}}\breve{\mathbf{V}}}.$$

5.6.2 Weighted Wiener filter

Here we look at another implementation of the multi-channel Wiener filter. Consider the signal model as in the section above:

$$X_m = Z_m + \breve{V}_m, \tag{5.40}$$

with $Z_m \triangleq A_m S$ as in the previous section. Now we aim to find the vector $\mathbf{W}$ such that, given the following cost function $\mathcal{J}(\mathbf{W})$:

$$\begin{aligned} \mathcal{J}(\mathbf{W}) &= \mathrm{E}\{|\mathbf{W}^H \mathbf{X}|^2\}, \\ &= \mathrm{E}\{|\mathbf{W}^H (\mathbf{Z} + \mathbf{V})|^2\} \\ &= \Psi_{SS} + \underbrace{\Psi_{SS}(\mathbf{W}^H \mathbf{A}\mathbf{A}^H \mathbf{W} - 1)}_{\text{speech distortion}} + \underbrace{\mathbf{W}^H \mathbf{\Psi}_{\breve{\mathbf{V}}\breve{\mathbf{V}}} \mathbf{W}}_{\text{noise reduction}}, \end{aligned} \tag{5.41}$$

the vector $\mathbf{W}$ should minimize the noise and *simultaneously* keep speech distortion to a minimum. Choosing this minimization criterion leads to the following:

$$\begin{aligned} \frac{\partial \mathcal{J}(\mathbf{W})}{\partial \mathbf{W}^*} &= 0, \\ \Rightarrow \big(\mathbf{A}\mathbf{A}^H \Psi_{SS} + \mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}\big)\mathbf{W} &= 0, \end{aligned} \tag{5.42}$$

the solution to which is the eigenvector belonging to the null eigenvalue of the microphone array covariance matrix $\mathbf{\Psi}_{\mathbf{XX}}$. If this matrix is full rank, then $\mathbf{W}$ may be selected as the eigenvector corresponding to the smallest eigenvalue. However, this approach is not practical, as such a vector $\mathbf{W}$ minimizes not only the noise power but also the target signal.

A variation of this approach is to consider a *weighted* minimization of the cost function:

$$\mathcal{J}(\mathbf{W}) = \frac{1}{\mu}\Psi_{SS}(\mathbf{W}^H\mathbf{A}\mathbf{A}^H\mathbf{W} - 1) + \mathbf{W}^H\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}\mathbf{W}\,. \tag{5.43}$$

Then

$$\begin{aligned} \frac{\partial \mathcal{J}(\mathbf{W})}{\partial \mathbf{W}^*} &= 0, \\ \Rightarrow \big(\frac{1}{\mu}\mathbf{A}\mathbf{A}^H \Psi_{SS} + \mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}\big)\mathbf{W} &= 0, \end{aligned} \tag{5.44}$$

where the factor $1/\mu$ indicates the importance given to speech distortion. $\mu \to 0$ constrains the solution for $\mathbf{W}$ to be orthogonal to $\mathbf{A}$, with no specific emphasis on noise reduction. Setting $\mu \to \infty$ places full emphasis on noise reduction, but speech need not necessarily be preserved. In this case, $\mathbf{W}$ either lies in the null space of $\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}$, if it is rank deficient, or is given by the eigenvector corresponding to the smallest eigenvalue of $\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}$. Any other value of μ balances speech distortion against noise suppression. Again, while this approach is not very practical, it lays the foundations for the speech distortion weighted multi-channel Wiener filtering proposed in [35] and explained below.

5.6.3 Speech distortion weighted MWF

From the model in (5.40), the cost function is formulated and expanded as:

$$\begin{aligned} \mathcal{J}(\mathbf{W}) &= \mathrm{E}\big\{|Z_m - \mathbf{W}^H\mathbf{X}|^2\big\} \\ &= \mathrm{E}\big\{|\mathbf{e}_m^T\mathbf{Z} - \mathbf{W}^H(\mathbf{Z}+\mathbf{V})|^2\big\} \\ &= \mathrm{E}\big\{|(\mathbf{e}_m - \mathbf{W})^H\mathbf{Z} - \mathbf{W}^H\mathbf{V}|^2\big\} \\ &= \mathbf{e}_m^T\mathbf{\Psi}_{\mathbf{ZZ}}\mathbf{e}_m + \mathbf{W}^H\mathbf{\Psi}_{\mathbf{XX}}\mathbf{W} - \mathbf{W}^H\mathbf{\Psi}_{\mathbf{XZ}}\mathbf{e}_m - \mathbf{e}_m^T\mathbf{\Psi}_{\mathbf{ZX}}\mathbf{W} \end{aligned} \tag{5.45}$$

$$= \underbrace{|A_m|^2\Psi_{SS}}_{\text{desired output}} + \underbrace{\big(|\mathbf{W}^H\mathbf{A}|^2 - \mathbf{W}^H\mathbf{A}\mathbf{A}^H\mathbf{e}_m - \mathbf{e}_m^T\mathbf{A}\mathbf{A}^H\mathbf{W}\big)\Psi_{SS}}_{\text{speech distortion}} + \underbrace{\mathbf{W}^H\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}\mathbf{W}}_{\text{noise reduction}}. \tag{5.46}$$

The aim now is to optimize the noise reduction, subject to a specified amount of signal distortion. Thus, similar to (5.43), we have an optimization of the form:

$$\begin{aligned} \mathcal{J}(\mathbf{W},\mu) &= \frac{1}{\mu}\Psi_{SS}\big|(\mathbf{W}-\mathbf{e}_m)^H\mathbf{A}\big|^2 + \mathbf{W}^H\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}\mathbf{W} \\ \frac{\partial \mathcal{J}(\mathbf{W},\mu)}{\partial \mathbf{W}^*} &= 0 \\ \Rightarrow \mathbf{W}_{\text{opt}} &= \big(\mu\mathbf{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}} + \mathbf{A}\mathbf{A}^H\Psi_{SS}\big)^{-1}\mathbf{A}\mathbf{A}^H\mathbf{e}_m\Psi_{SS}\,. \end{aligned} \tag{5.47}$$

In contrast to the MVDR optimization, which is constrained to preserve the target signal in magnitude and phase, the weighted Wiener filter only constrains the *power* of the *distortion*.

Further, expanding (5.47) into its constituent MVDR and post-filter using Woodbury's identity, we obtain:

$$\mathbf{W}_{\text{opt}} = \mathbf{W}_{\text{MVDR}} \frac{\Psi_{SS}}{\left(\Psi_{SS} + \mu(\mathbf{A}^H \boldsymbol{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}^{-1} \mathbf{A})^{-1}\right)} \tag{5.48}$$

with

$$\mathbf{W}_{\text{MVDR}} = A_m^* \frac{\boldsymbol{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}^{-1} \mathbf{A}}{\mathbf{A}^H \boldsymbol{\Psi}_{\check{\mathbf{V}}\check{\mathbf{V}}}^{-1} \mathbf{A}}. \tag{5.49}$$

For $\mu \to 0$, (5.48) converges to the MVDR solution. For $\mu = 1$, (5.48) yields the traditional MWF. In general, $\mu < 1$ preserves the target signal at the cost of lower noise reduction whereas $\mu > 1$ places more emphasis on the noise reduction as compared to signal preservation. An adaptive value for μ, i.e., a higher μ in noise-dominant sections and a lower value in speech-dominant sections, can be obtained for example by using the speech presence probability as in [86].

5.7 Conclusions

In this chapter we have presented an overview of the state-of-the-art in multi-channel source separation/enhancement techniques. These were broadly classified into linear and non-linear approaches. Non-linear approaches are gain-based and are implemented in the STFT domain. These approaches function by lowering the output signal gain at STFT points where the interference is dominant or holding the gain close to unity at STFT points where the desired or target signal is dominant. Such approaches perform interference *suppression* but not, strictly speaking, target enhancement.

The linear algorithms for the multi-channel source separation/enhancement problem, on the other hand, utilize the spatial diversity to steer a spatial null in the direction of the interference, thus performing interference *cancellation*. Such algorithms may also, depending upon the implementation, enhance the target signal. In the case where the spatial null is data-dependent, the linear algorithms perform what is termed as 'adaptive' beamforming.

The linear algorithms are further classified into SOS and HOS based algorithms. The MVDR and its adaptive counterpart, the GSC, were introduced as examples of the SOS based algorithms. Conceptually, these algorithms are independent of source signal statistics and are applicable to a wide range of applications. A representative class of HOS algorithms based on ICA was also briefly discussed. These algorithms rely on specific signal characteristics (e.g., non-Gaussianity) and generative models (e.g., instantaneous mixing in the time or frequency domain) for their functionality, which might not always be present. Thus, these approaches are rather restricted in their applicability. In defence of such algorithms, when these assumptions are fulfilled, they are theoretically capable of source separation in scenarios with almost no *a priori* information.

Further, an optimal (in the MMSE sense) multi-channel solution was detailed as a combination of the MVDR beamformer with a Wiener post-filter, which is a gain based approach. This solution may be seen as a convergence of the linear and the non-linear

approaches, and provides a guideline to the path we must take in designing such algorithms.

You say you don't know the microphone distances, I say 'Here is a ruler, measure it!'.

A pragmatic take on blind source separation
– Ivan Tashev, Microsoft Research

Chapter 6

Localization Based Source Separation – Algorithms and Examples

We shall apply the theory of localization and separation developed in the previous chapters to perform speech signal enhancement under more general conditions. We begin with a simple two-microphone, two-source case and estimate the individual source signals using ICA. We shall show here how a localization stage is beneficial to robustly solving the permutation problem. Next, we move to the general case of M microphones and Q sources using the MoG speaker localization model of Chapter 4. It shall be shown how this localization model may be used to implement a wide variety of enhancement algorithms, from a soft-mask based approach to a generalized GSC, capable of extracting more than one source from a mixture, in the presence of multiple competing sources and background noise.

6.1 Beam-initialized ICA (BI-ICA)

The beam-initialized ICA (BI-ICA) approach proposed in this section explicitly considers a 2×2 source-microphone model and makes use of the *null* and *direct path* characteristics of beamformers to generate *reference sources* for permutation resolution. This approach is computationally less expensive as compared to the method proposed in [104] while, at the same time, yielding comparable results. Another advantage of the proposed approach is that it is tolerant to position estimation errors and, further, the position of only one source need be estimated for the 2×2 system presented here.

BI-ICA consists of three stages: an initial spatial prefiltering stage, traditional ICA (preprocessing+FastICA) and, lastly, the permutation and scaling resolution stage.

6.1.1 Spatial prefiltering

Assume that, of the two sources, we know the approximate azimuthal location θ_t of one source – the 'target' source – with the DOA being measured with respect to the array axis. This information may be obtained by any of the methods suggested in Chapter 3. Then,

the steering vector corresponding to the *direct* path from this position to the array, under the assumption of equal microphone gains, is:

$$\mathbf{A}_{\mathrm{t}}(k) = [1,\ \exp(-\jmath\,\Omega_k d\cos(\theta_{\mathrm{t}})/c)]^T\,, \tag{6.1}$$

where Ω_k represents the k th discrete frequency, d the inter-microphone distance and c the speed of sound in air. Correspondingly we define two signals:

- the null-beam signal: $Z_{0,\mathrm{t}}(k,b) = [1,\ -\exp(\jmath\,\Omega_k d\cos(\theta_{\mathrm{t}})/c)]\mathbf{X}(k,b)$
- the direct-path signal: $Z_{\mathrm{d,t}}(k) = \mathbf{A}_{\mathrm{t}}^H(k)\,\mathbf{X}(k,b)$.

Note that both signals still contain a mixture of the sources. However, the interferer would be predominant in $Z_{0,\mathrm{t}}(k,b)$ (in the absence of grating lobes). This null-beam signal generates our 'reference-source' for permutation resolution, as will be shown in Section 6.1.4.

6.1.2 Sphereing

The input to this pre-processing stage is the composite vector

$$\mathbf{Z}(k,b) = \left[Z_{0,\mathrm{t}}(k,b)\ Z_{\mathrm{d,t}}(k,b)\right]^T \tag{6.2}$$

of the previous step. We use the standard PCA to decorrelate the elements of the vector $\mathbf{Z}(k,b)$ to obtain the whitened vector $\overset{\circ}{\mathbf{Z}}(k,b)$:

$$\overset{\circ}{\mathbf{Z}}(k,b) = \mathbf{\Psi}_{\mathbf{ZZ}}^{-\frac{1}{2}}(k)\ \mathbf{Z}(k,b), \tag{6.3}$$

with $\mathbf{\Psi}_{\mathbf{ZZ}}(k)$ being estimated by a temporal averaging over the B time records of the signals per frequency bin:

$$\mathbf{\Psi}_{\mathbf{ZZ}}(k) = \frac{1}{B}\sum_b \mathbf{Z}(k,b)\mathbf{Z}^H(k,b).$$

6.1.3 ICA

The vector $\overset{\circ}{\mathbf{Z}}(k,b)$ from (6.3) is input to the ICA stage. Due to the pre-processing, we now search for an orthogonal matrix $\overset{\vee}{\mathbf{W}}(k)$ that decomposes the vector $\overset{\circ}{\mathbf{Z}}(k,b)$ into mutually independent components $\overset{\circ}{\mathbf{Y}}(k,b) = \overset{\vee}{\mathbf{W}}(k)\,\overset{\circ}{\mathbf{Z}}(k,b)$. Choosing the Kullback-Leibler divergence as the cost function measure and using the polar co-ordinate non-linearity of [103] to approximate the derivative of the probability density functions, we arrive at the following FASTICA update rule:

$$\begin{aligned}\nabla_{\overset{\vee}{\mathbf{W}}(k)}\mathcal{J}(\overset{\vee}{\mathbf{W}}(k)) &= \left(\mathbf{I} - \mathrm{E}\left\{\mathcal{K}\big(\overset{\circ}{\mathbf{Y}}(k,b)\big)\overset{\circ}{\mathbf{Y}}^H(k,b)\right\}\right)\overset{\vee}{\mathbf{W}}(k)\,,\\ \overset{\vee}{\mathbf{W}}(k) &\leftarrow \nabla_{\overset{\vee}{\mathbf{W}}(k)}\mathcal{J}(\overset{\vee}{\mathbf{W}}(k))\,,\end{aligned} \tag{6.4}$$

where $\mathcal{K}(x) = \tanh(|x|)e^{\jmath\arg(x)}$ and $E\{.\}$ stands for the expectation operator, replaced in practice by the temporal average. $\nabla_{\overset{\vee}{\mathbf{W}}(k)}$ represents the complex matrix gradient operator with respect to $\overset{\vee}{\mathbf{W}} = [\overset{\vee}{\mathbf{W}}_1, \ldots, \overset{\vee}{\mathbf{W}}_Q]^H$.

This particular choice of cost function and non-linearity has the fastest convergence [46, 103] among the non-linearities proposed in [57, 103, 111]. The update is followed by the orthonormalization of the updated matrix $\overset{\vee}{\mathbf{W}}(k)$:

$$\overset{\vee}{\mathbf{W}}(k) \leftarrow \big(\overset{\vee}{\mathbf{W}}(k)\,\overset{\vee}{\mathbf{W}}{}^{H}(k)\,\big)^{-\frac{1}{2}}\,\overset{\vee}{\mathbf{W}}(k)\ . \tag{6.5}$$

6.1.4 Permutation and scaling resolution

The result of the sphereing and ICA steps is a scaled and permuted estimate of the underlying source components:

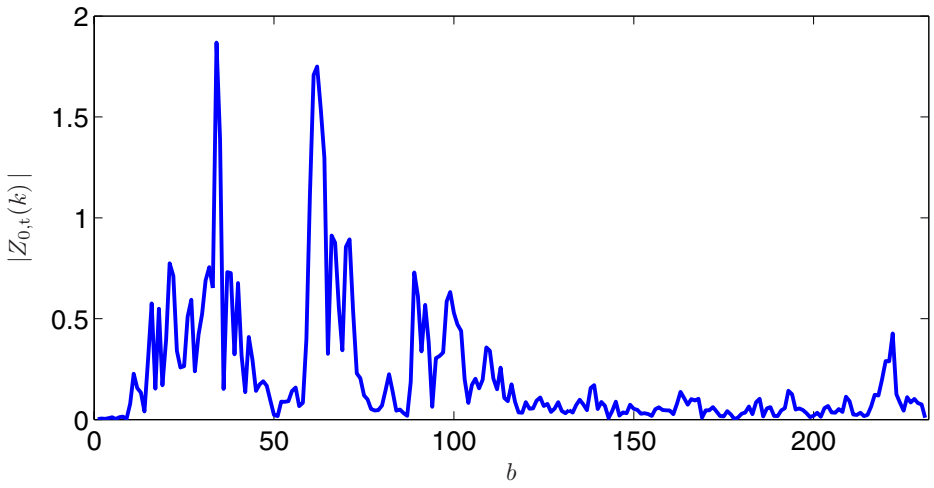
$$\overset{\circ}{\mathbf{Y}}(k,b) \approx \mathbf{J}(k)\,\mathbf{D}(k)\,\mathbf{S}(k,b).$$

To resolve the permutation, we shall consistently assign the signal from the known location to channel 1 and the interferer to channel 2. To make this distinction, we use the null-beam signal as the reference. The required permutation is then such, that:

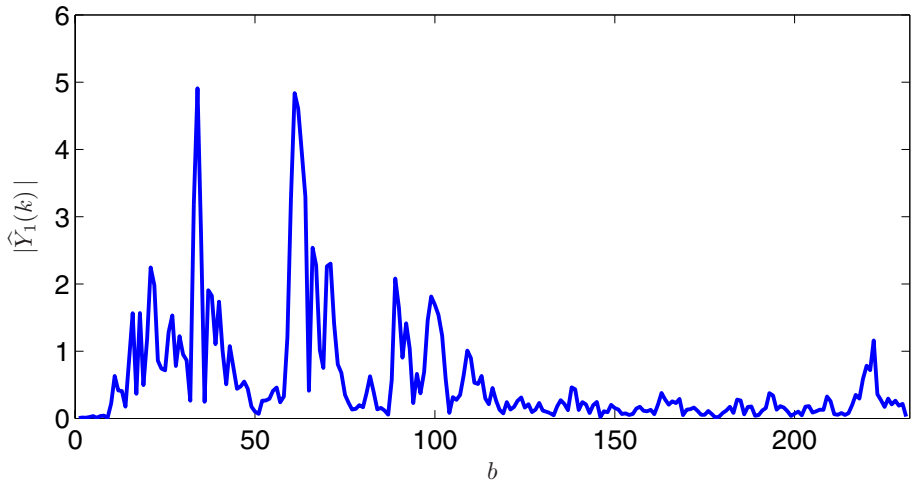
$$\mathbf{J}_{\Pi}(k) = \underset{\Pi(i)}{\operatorname{argmax}}\ \operatorname{cor}\Big(\Big|\big(\mathbf{J}_{\Pi(i)}\overset{\circ}{\mathbf{Y}}(k)\,\big)_1\Big|, \big|Z_{0,\mathrm{t}}(k)\,\big|\Big), \tag{6.6}$$

where $(\mathbf{J}_{\Pi(i)}\overset{\circ}{\mathbf{Y}})_1$ is the first element of the permuted output signal vector under the permutation $\Pi(i)$. Thus (6.6) seeks to align the recovered interference signal component of $\overset{\circ}{\mathbf{Y}}$ on the first output channel. This consideration arises from the fact that a null-beam should remove much of the direct path energy from the target signal, S_{t}, resulting in amplitude spectra that are least correlated with the corresponding recovered versions of the target (Figure 6.1). However, as the interferer is still present in the null-beam signal, the correlation between the amplitude spectra of the recovered interferer and the null-beam signal should be higher. The correlation is a normalized value, similar to the correlation coefficient defined in [93, Eq. 7-8].

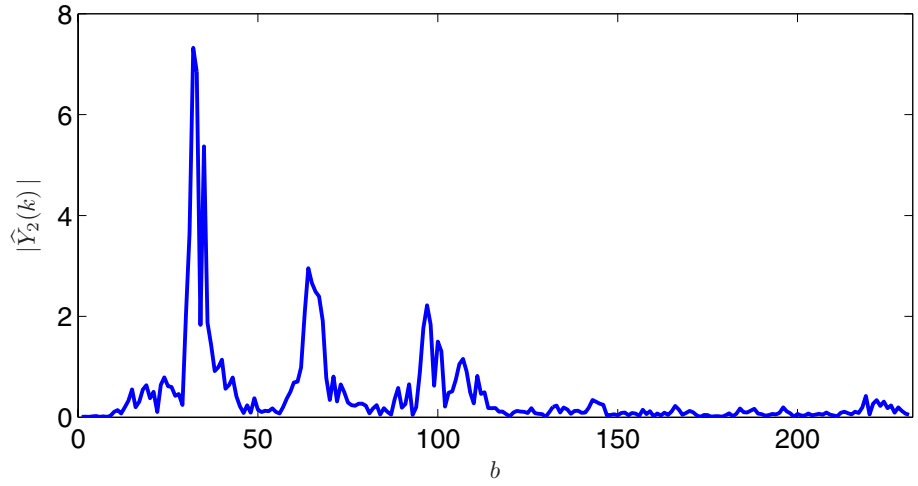
Note that this approach may be seen as a combination of the DOA based approaches and the correlation based approaches to permutation resolution. Further, as the permutation resolution is done independently in each frequency bin, it prevents the error propagation peculiar to the amplitude correlation based approaches. The system schematic is presented in Figure 6.2.



(a) Null beam on target



(b) Recovered interferer



(c) Recovered target

Figure 6.1: Amplitude envelope evolution at 1.0 kHz. The x-axis indicates the time frame. Note the similarity between the envelopes of the interferer and the null beam.

Once the permutation has been resolved, the minimum distortion principle is used to obtain an estimate of the scaling matrix $\widehat{\mathbf{D}}(k)$, from which we may obtain:

$$\mathbf{Y}(k,b) = \widehat{\mathbf{D}}^{-1}(k)\,\mathbf{J}_{\Pi}(k)\,\overset{\circ}{\mathbf{Y}}(k,b) \approx \mathbf{S}(k,b), \tag{6.7}$$

where $\mathbf{S}(k,b)$ represents the source signal as observed at a reference microphone (usually microphone 1). The time-domain signal can then be reconstructed using standard overlap-add techniques, yielding the separated sources.

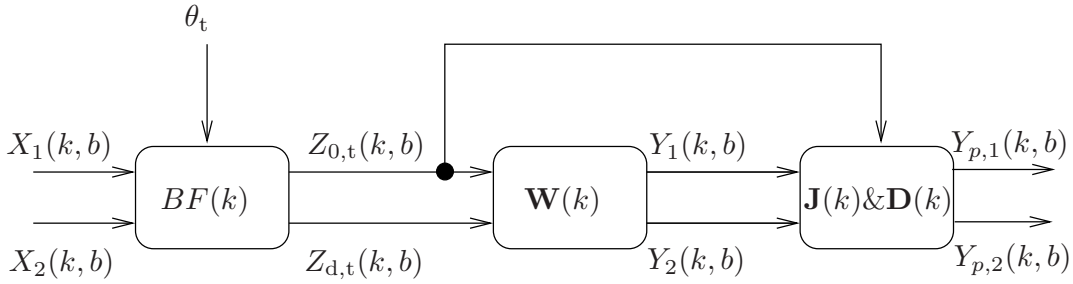


Figure 6.2: BI-ICA system schematic illustrated for a particular frequency bin.

6.1.5 Experimental setup and results

BI-ICA was compared vis-à-vis the *robust and precise method* (ICARP) of [104] and an 'ideal' method (ICAID) where the permutation ambiguity is resolved using the amplitude envelopes of the original sources. The ICAID approach represents the upper bound, in terms of performance, for the ICARP and BI-ICA algorithms. To keep the comparison fair, the core *ICA* algorithm of each approach is set to the one described previously. For the proposed approach, and the robust and precise method, the DOAs were given to the algorithms and correspond to the values from the setup.

Table 6.1: Algorithm parameters used for BI-ICA

Sample rate, f_{s} (kHz)	T_{60} (s)	Window type/ length (ms)	DFT length (ms)	Frame shift (ms)	d (cm)	Source distance (m)
8	0.6	von Hann/128	128	32	3	1.0

The experiments were conducted on data measured in Room 2 (Appendix E). The algorithm parameters were set as in Table 6.1. The microphones used were the Sennheiser KE-4 omni-directional capsules. The sources were selected from the Timit database and consisted of three male and three female speakers, each uttering a single sentence. The algorithms were run for different combinations of speakers, over different azimuthal spacings (from 30° to 120°) between the sources, and over different positions of these *pairs* in the azimuthal plane. The results presented have been averaged over all the experiments.

The two instrumental measurement criteria selected for evaluating the algorithms were the Signal to Interference Ratio (SIR) *improvement* (ΔSIR) and the Signal to Distortion ratio (SDR). ΔSIR is computed as the difference between the Signal to Interference Ratio (SIR)

after separation and the average SIR before separation (i.e., in the input signals). The SDR, a measure of the quality of the extracted signal, is defined as:

$$\text{SDR} = 10\log_{10}\frac{\text{Energy}_{\text{clean source}}}{\text{Energy}_{\text{clean source}-\text{filtered source}}} \tag{6.8}$$

In each case, the SIR and the SDR are computed in the time domain as proposed in [6]. The results are as shown in Figure 6.3.

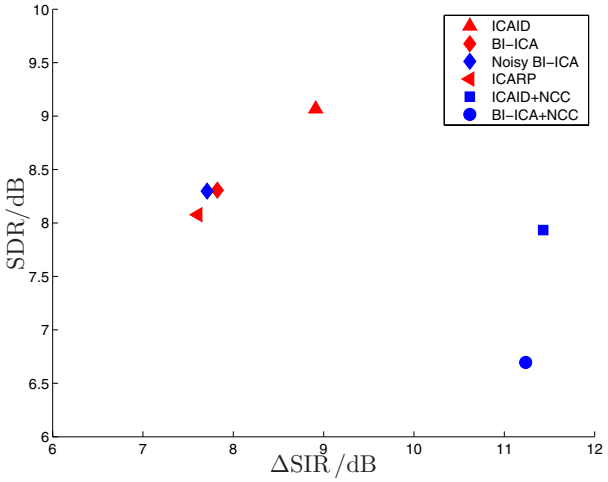


Figure 6.3: Experimental results in the ΔSIR-SDR plane.

Both BI-ICA and the reference approach (ICARP) perform comparably. The ideal approach (ICAID) has the best performance, as expected. Another interesting aspect of the proposed approach is its relative tolerance to DOA estimation errors in the localization stage. This was tested on the measured data, where the beamforming was done with corrupted DOA values (randomly perturbed to within $\pm 10^\circ$ of the true azimuth). The results (Noisy BI-ICA) clearly corroborate the tolerance of our approach.

However, this approach still has some cross-talk (confirmed by listening tests and by the improvement in SIR due to post-processing). This is an indication of the larger approximation error of the STFT model in reverberant and noisy environments and the corresponding inability of the ICA algorithms to steer sharp nulls towards the interfererence in such a case. The cross-talk may be reduced by post-filtering the extracted signals. One way is to use a simple binary-mask post processor based on the assumption of disjointness of speech (similar to [65]). For an extracted source q, such a binary-mask post-processor may be defined as:

$$\mathcal{M}_{q,\text{post}}(k,b) = \begin{cases} 1 & \Upsilon_{\text{post}}\,|Y_q(k,b)| > \max\limits_{\forall q' \neq q}\left(|Y_{q'}(k,b)|\right) \\ 0 & \text{otherwise,} \end{cases} \tag{6.9}$$

where $0 < \Upsilon_{\text{post}} \leq 1$ is used to prevent spurious triggering of the masks. Results obtained using such a post-processor (denoted as the non-linear cross-talk canceller (NCC)) show improved performance as far as cross-talk suppression is concerned (ΔSIR increases).

However, such masks also introduce audible artefacts, termed as musical noise, into the resulting signal. More sophisticated masking procedures (e.g. [74]) allow for interference suppression to the same level as that afforded by binary-masks, while reducing the effect of artefacts, yielding a more natural-sounding signal.

6.2 Generalized approach for speech enhancement

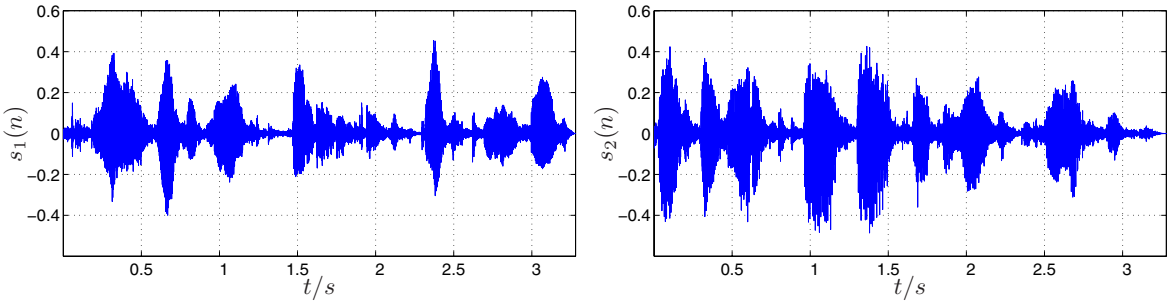
The ICA-based algorithm presented above serves only to illustrate that even so-called 'blind' algorithms require additional information, such as spatial cues, to allow a practical realization. In general, ICA algorithms are characterized by large latency and relative inflexibility: they do not perform well in the presence of high levels of background noise and in underdetermined cases.

In this section, we shall capitalize on the speaker localization system developed in Chapter 4 and develop a versatile framework that allows for the implementation of a wide range of alternatives for target speaker extraction in the presence of competing sources and background noise, and nicely unifies the goal of localization with source separation.

For the separation problem, the first stage is identifying the MoG components that correspond to the active sources. Through the unsupervised clustering of the per-bin localization estimates using an underlying MoG model, we obtain, for each frame, the number of sources estimated in *that* frame (the number of elements in the MoG model), an estimate of the source positions (given by the MoG means) and some information regarding their spatial extent and relative importance in that frame (the variances and weights of the elements respectively). Recognize that the per-frame estimates may contain elements that do not correspond to any source. We termed such elements ghost-clusters. Under the assumption that the speaker signals are stationary in space or move only very slowly, we postulated that one of the characteristics of a source was consistent excitation of the MoG model around the source position, leading to the propagation of the estimated means across multiple time frames. This behaviour was captured by the non-linear, token-based temporal smoothing framework and used to reduce the influence of ghost clusters. Subsequently we obtained localization 'tracks' in the vicinity of the positions containing actual sources and a small amount of clutter. In the separation framework, such dominant tracks (a track is deemed dominant if its TTL value exceeds a threshold) indicate the positions at which active sources are present and are to be extracted.

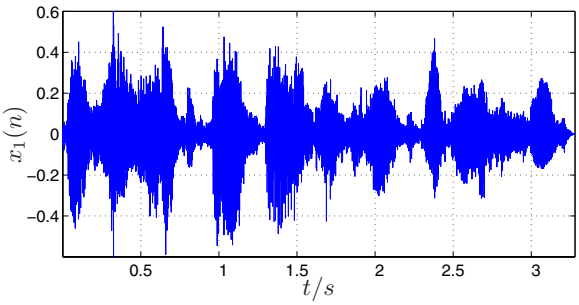
We shall start with the simplest of all spatial approaches, the delay-and-sum beamformer (DSB), for target source enhancement and progress through a mask-based approach before considering the linear, adaptive beamformers. This development is instructive in that it provides an indication of the fusion of the linear and non-linear approaches to source separation and serves to better illustrate the links between these methods. As in Section 4.3, we consider the azimuth-based localization model.

The framework development shall be illustrated throughout on a mixture of two sources recorded in Room 2 (Appendix E) using the 5 channel array introduced in Chapter 3. The original source signals at microphone 1 and the mixture at that microphone are presented in Figure 6.4.



(a) Source 1 at mic. 1: $s_1(n)$

(b) Source 2 at mic. 1: $s_2(n)$



(c) Mixture observed at mic. 1: $x_1(n) = s_1(n) + s_2(n)$

Figure 6.4: Clean input signal at microphone 1 and the mixture at the same microphone.

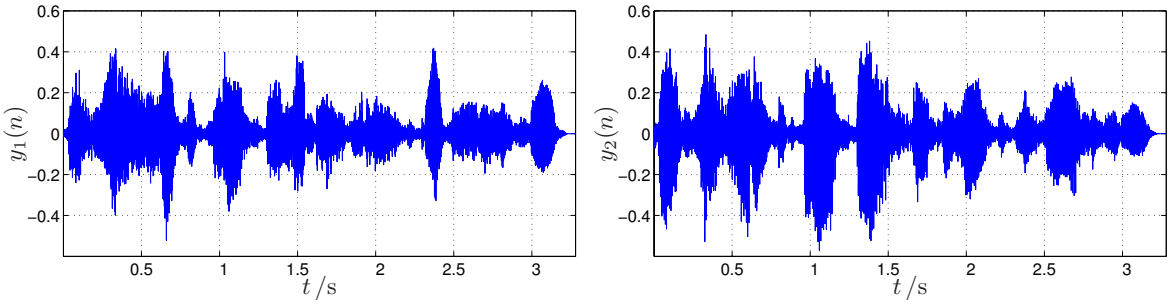
6.2.1 DSB-based extraction

Given the number of dominant tracks and the corresponding means, we may easily formulate the DSB to obtain these sources. Assume that Q corresponds to the number of dominant tracks and M represents the number of microphones in the array. Then, for any frame b and frequency bin k, the DSB estimate of the source q, with a corresponding MoG element around θ_q, is:

$$Y_{q,\mathrm{DSB}}(k,b) = \mathbf{H}^H(\theta_q,k,b)\mathbf{X}(k,b), \tag{6.10}$$

where $\mathbf{H}(\theta_q,k,b)$ is the delay-and-sum beamformer along θ_q which is obtained either as the estimated value from the MoG model in the current frame, if the source is found in that frame, or as the time-averaged mean $\overline{\theta}_q$ if the source is not detected by the MoG model in that frame. The DSB is formulated in a manner similar to (3.38).

The DSB enhanced signals may then be resynthesized from their STFT estimate (see (5.4)) using standard techniques. Applying this procedure to the mixture of Figure 6.4, we obtain the signals in Figure 6.5. With reference to this figure, we see that the DSB provides only limited interference suppression. This is to be expected as the DSB does not actively cancel interferers and, as shall be shown, is best suited for uncorrelated noise at the sensors.



(a) $y_1(n)$ (b) $y_2(n)$

Figure 6.5: Separated signals obtained using the DSB on the input signals.

6.2.2 Mask-based extraction

For the DSB approach of the previous section, we have only utilized the knowledge of the source positions for source extraction. If we consider the other estimated parameters of the MoG and examine the MoG model from the point of view of a signal presence detector, then for each bin k and frame b, we may now compute a mask for the MoG element q based on the generalized likelihood [120] as follows (when using the azimuth-based localization):

$$\mathcal{M}_q(k,b) = \frac{\frac{1}{\sigma_q} P_q \exp\left(-\frac{\left(\widehat{\theta}(k,b) - \theta_q\right)^2}{2\sigma_q^2}\right)}{\sum_{q'=1}^{\mathcal{I}} \frac{1}{\sigma_{q'}} P_{q'} \exp\left(-\frac{\left(\widehat{\theta}(k,b) - \theta_{q'}\right)^2}{2\sigma_{q'}^2}\right)}. \tag{6.11}$$

This is guaranteed to be in the range $[0,1]$. From this, we may form an estimate of the underlying source signal spectrum:

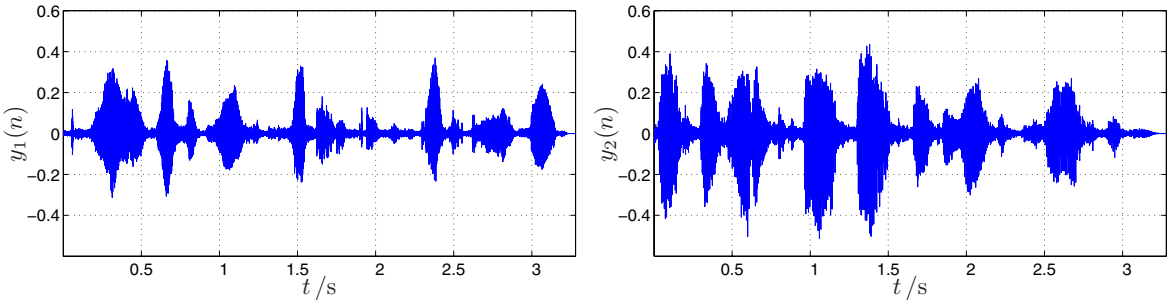
$$Y_{q,\mathrm{Msk}}(k,b) = \mathcal{M}_q(k,b) X_m(k,b) \tag{6.12}$$

or

$$Y_{q,\mathrm{Msk}}(k,b) = \mathcal{M}_q(k,b) Y_{q,\mathrm{DSB}}(k,b). \tag{6.13}$$

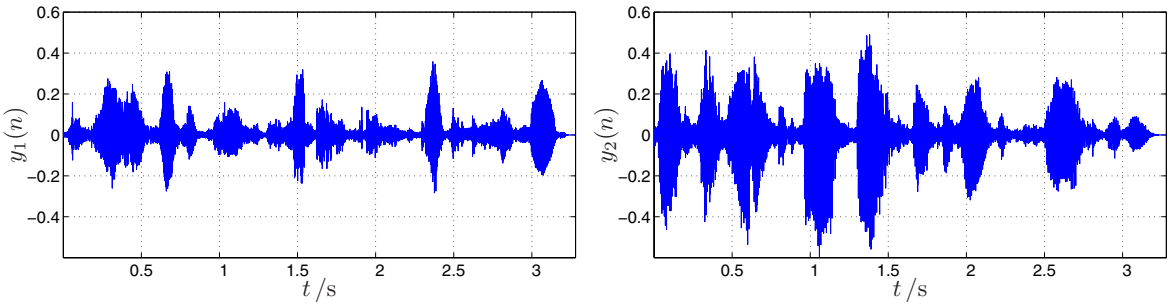
In practice, the latter estimate is a better choice due to the additional gain afforded by the DSB before the mask. Also, as in the DSB approach, the parameters used to estimate the mask stem either from the estimated MoG model for the current frame – when the target source is found in that frame – or from the time-averaged model when the source is not found within that frame.

We have observed that, especially for noisy conditions, further improvement in the separated signal quality may be obtained by smoothing the masks in their cepstral representation, as proposed in [74]. Applying this smoothing to the example under consideration, we obtain the results shown in Figure 6.7. Note that such a smoothing principally improves the target signal quality. However, as the interferer is not nulled, there exists some cross-talk in the masked signals, which is slightly exacerbated by the smoothing of the masks. This represents the classic trade-off between intereference suppression and target-signal degradation. Such a trade-off is less evident in the experiments conducted in [74] as the smoothing was applied in the post-processing stage, *after* linear separation. We reiterate, however, that the cepstral-temporal smoothing of the masks preserves target speech onsets and reduces musical noise.



(a) $y_1(n)$ (b) $y_2(n)$

Figure 6.6: Separated signals obtained by time-frequency masking of the DSB signal.



(a) $y_1(n)$ (b) $y_2(n)$

Figure 6.7: Separation performance using the cepstro-temporal mask smoothing proposed in [74].

Consequently, even though the separated signals obtained using smoothed masks contain more cross-talk, the net result sounds more natural.

6.2.3 Data-driven parsimonious excitation-based GSC (PEG)

The data-driven GSC approaches of Section 5.5.2 require an estimate of the transfer function ratios or, equivalently, the target source covariance matrix. As was shown in Chapter 5, under certain assumptions regarding the noise characteristics, this can be done by estimating the noise covariance during target signal pauses and subtracting the value from the covariance measured during target signal activity, (see (5.30)). However, this requires an estimate of target silence detection, which is not trivial when the noise contains non-stationary sources (like competing speakers). In such cases, the algorithms have to be modified to take into account the number of sources present, the nature of the sources and the background noise, etc. Such an approach is the dual TF-GSC algorithm [99] which is developed for a single, non-stationary, *directive* interference and *directive* background noise. However such additional information may not be easily available, necessitating alternative implementations.

In the following, we propose one such approach. We shall show how the required information is implicitly available when using the MoG based localization approach, allowing us to formulate a fully adaptive beamformer based on the GSC structure for each source to be extracted. Such an adaptive beamformer consists of two main components: the steering vector in the direction of the target source (this stage is similar to the fixed beamformer in

the traditional GSC) and a blocking matrix with a corresponding adaptive canceller. Unlike the traditional GSC implementations, however, the steering vector and blocking matrix computations are data-driven. The canceller is estimated as in the traditional GSC, albeit with the adaptively estimated blocking matrix.

For output channel q, frame b and frequency bin k, we begin with the estimation of the target signal component as in the DSB approach of (6.10). Next, we compute the probability that the target source is present in that frame and bin as:

$$P_{q|\widehat{\theta}}(k,b) = \frac{p\left(\widehat{\theta}(k,b)\middle|q\right) P_q}{\sum_{q'} p\left(\widehat{\theta}(k,b)\middle|q'\right) P_{q'}}, \tag{6.14}$$

which, when expanded from the MoG model, is recognizable as the mask from (6.11).

Using this source presence probability, we estimate the parameters required in the GSC approach – namely the blocking matrix $\mathbf{B}_q(k,b) \in \mathbb{C}^{M\times(M-1)}$ and the noise canceller $\mathbf{W}_{q,V}(k,b) \in \mathbb{C}^{(M-1)\times 1}$. For this, we begin with an estimation of the target signal subspace as:

$$\mathbf{P}_q(k,b) = \left(1 - P_{q|\widehat{\theta}}(k,b)\right) \mathbf{P}_q(k,b-1) + P_{q|\widehat{\theta}}(k,b) \frac{\mathbf{X}(k,b)\mathbf{X}^H(k,b)}{\|\mathbf{X}(k,b)\|^2}. \tag{6.15}$$

Next, for generating the noise reference signals, we *approximate* the projector onto the orthogonal complement space of the target signal as:

$$\mathbf{P}_q^{\perp}(k,b) = \left(\mathbf{I} - \mathbf{P}_q(k,b)\right), \tag{6.16}$$

from which the blocking matrix $\mathbf{B}_q$ is obtained as:

$$\mathbf{B}_q(k,b) = \mathcal{D}_{M,(M-1)}\{\mathbf{P}_q^{\perp}(k,b)\} \tag{6.17}$$

where $\mathcal{D}_{a,b}\{\cdot\}$ is a selection operator that selects the first a rows and b columns of the (matrix) argument. Next, we update the cancelling vector $\mathbf{W}_{q,V}$ as:

$$\begin{aligned}\mathbf{W}_{q,V}(k,b+1) = \mathbf{W}_{q,V}(k,b) \\ + \widehat{\varsigma}(k,b) \frac{\left(Y_{q,\mathrm{DSB}}(k,b) - \mathbf{W}_{q,V}^H(k,b)\mathbf{B}_q^H(k,b)\mathbf{X}(k,b)\right)^* \mathbf{B}_q^H(k,b)\mathbf{X}(k,b)}{\|\mathbf{B}_q(k,b)\mathbf{X}(k,b)\|^2},\end{aligned} \tag{6.18}$$

where $Y_{q,\mathrm{DSB}}(k,b)$ is the delay-and-sum beamformed signal in the direction of source q. The GSC enhanced output is then obtained as:

$$Y_{q,\mathrm{GSC}}(k,b) = Y_{q,\mathrm{DSB}}(k,b) - \mathbf{W}_{q,V}^H(k,b)\mathbf{B}_q^H(k,b)\mathbf{X}(k,b)\,. \tag{6.19}$$

Note that the stepsize in (6.18) is adaptively set for each time-frequency point, and is estimated as:

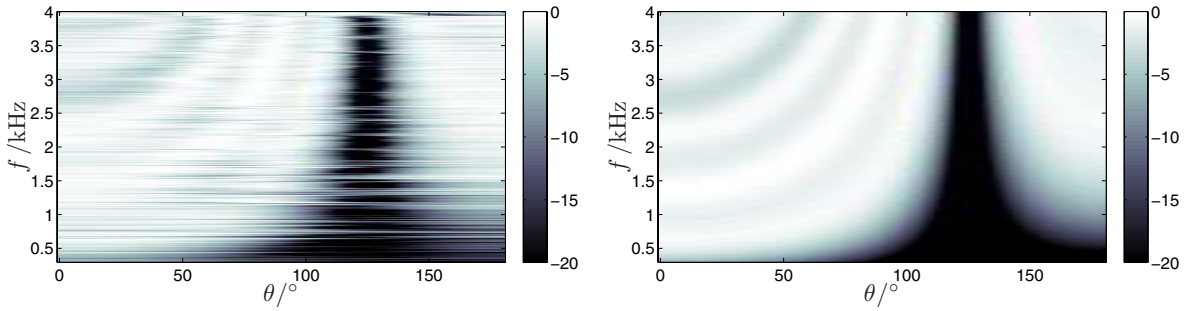
$$\widehat{\varsigma}(k,b) = \left(1 - P_{q|\widehat{\theta}}(k,b)\right)\varsigma_{\mathrm{c}}\,, \tag{6.20}$$

with $\varsigma_{\mathrm{c}} \leq 2$ being a fixed constant. In other words, when the interference is dominant, the adaptation of the canceller is more rapid, as the stepsize is larger than when the target signal is dominant. Here the canceller is adapted with a very small stepsize thereby providing an additional guard against any possible target signal leakage through the blocking matrix.

As may be seen, key to the proposed GSC implementation is the mask-steered adaptation of the beamformer for each source. Given the sparse nature of the source signal, the

masks and, correspondingly, the excitation of the frequencies for adaptation also reflect this characteristic. Hence we term this approach the *parsimonious-excitation based GSC* or PEG.

Note that the blocking matrix should accurately reject the source signal component and maybe seen, in some measure, as implementing the distortionless constraint of the MVDR beamformer. If the target signal component is successfully rejected, the noise canceller can be adapted in an unconstrained manner, utilizing the spatial diversity to cancel out the correlated interference in the output signal as represented in (6.19). Thus, the design of the $\mathbf{B}_q$ is a critical aspect. Figure 6.8 illustrates the rejection capability of the blocking matrix estimated using (6.15)–(6.17) and the subsequent performance of the adaptive canceller for a sample frame of the example under test. Note the formation of the null along the direction of the interferer (around $\theta = 60°$), coupled with low target direction degradation in the resultant signal.



(a) Target rejection capability of the estimated $\mathbf{B}$ (target around $\theta = 120°$)

(b) Target rejection using oracle-determined $\mathbf{B}$

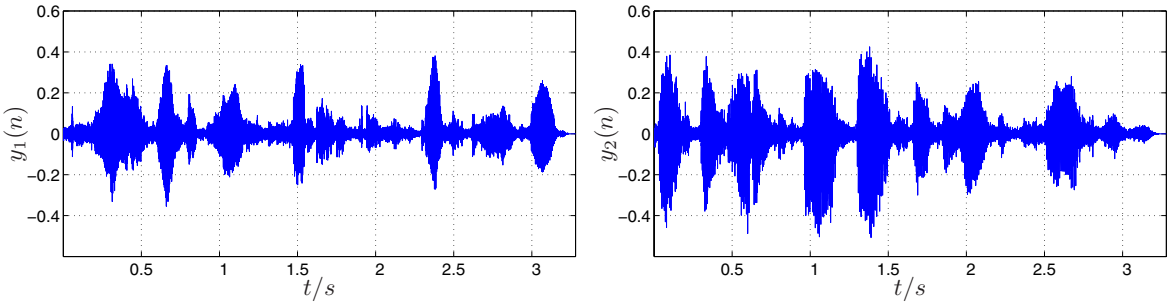
(c) PEG beampattern using the estimated $\mathbf{B}$. Interference angle is around $\theta = 60°$. Target signal is preserved

(d) The DSB beampattern for the same situation

Figure 6.8: Illustrating the effect of PEG in the spectral domain. Figures 6.8(a) and 6.8(b) depict the corresponding $\mathbf{B}$-weighted inner-products of the steering vectors for each azimuth and frequency. The effect of the blocking matrix is visible as a notch along the target direction. Note that the estimated blocking matrix closely approximates the oracle-determined blocking matrix in performance. Note, further, that the PEG steers a null along the *interference*, whilst the target signal is preserved (compare to the DSB beampattern).

The separated signals obtained using this completely data-driven structure are illustrated in Figure 6.9. Based on Figure 6.8 and Figure 6.9, we make the following observations: a) the linear structure preserves the target signal as well as the DSB (indicating that the orthogonal complement space is well estimated), and b) it also has good interference suppression capability (comparable to the mask based approaches), as long as the interference is correlated across the array. This latter condition is a prerequisite since the GSC basically uses

spatial diversity in the adaptive canceller to steer a null in the direction of the interferer.



(a) $y_1(n)$ (b) $y_2(n)$

Figure 6.9: Separated signals obtained using PEG.

6.2.4 Analysis of the PEG approach

Now that we have empirically established the performance of the PEG approach, we shall devote the following sections to a more detailed analysis of this algorithm. Specifically, we shall justify the selected update rule for the blocking matrix instead of an *oracle* approach and illustrate the performance of the proposed approach in cases where *a priori* knowledge regarding the sensors may be imperfect. We begin with our description of the oracle approach, before comparing it to the proposed PEG.

The oracle estimation of B_q

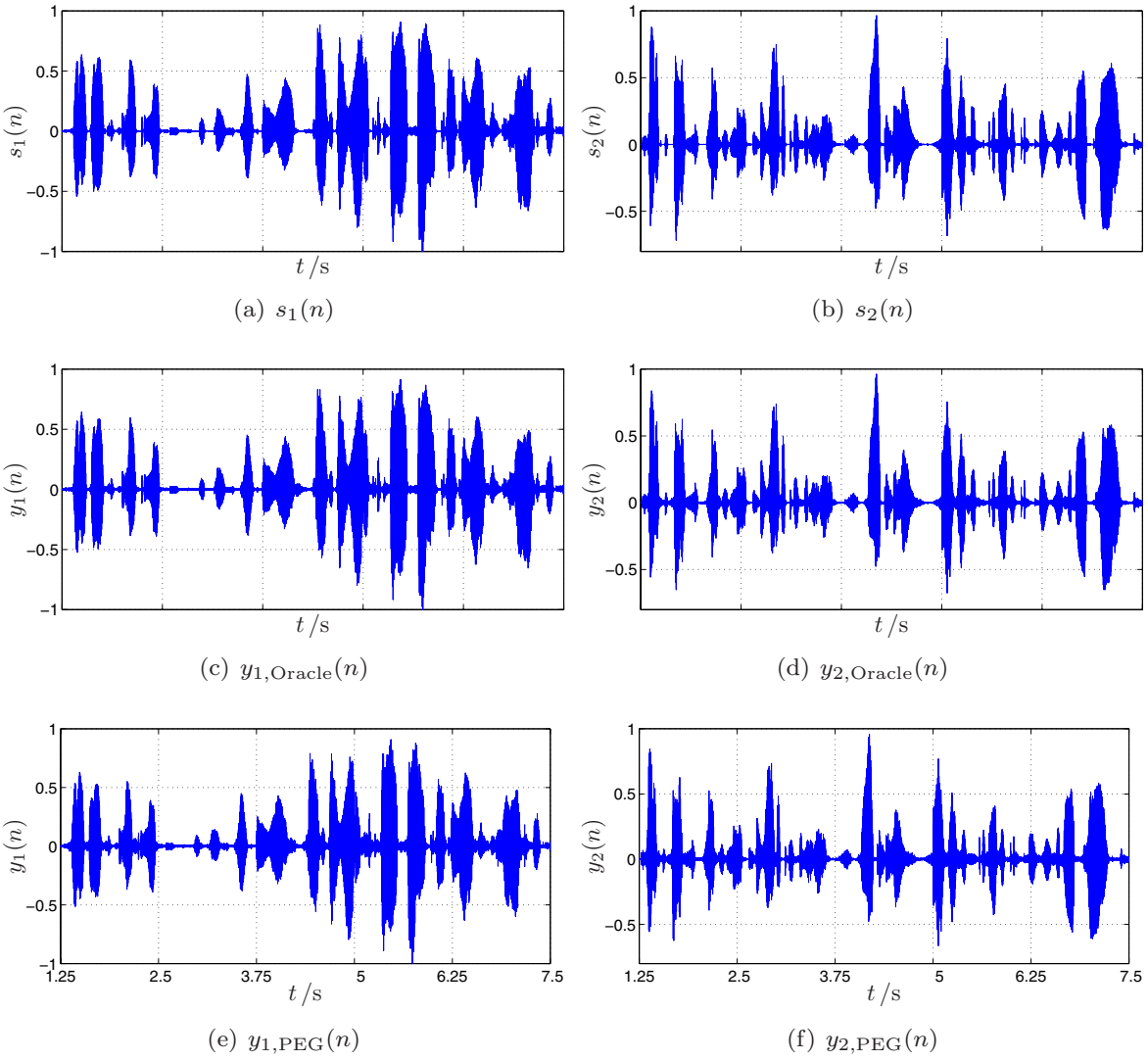
The blocking matrix has, at its heart, the projector into the orthogonal complement of the target signal subspace. If the propagation vector from the source q to the microphone array, $\mathbf{A}_q(k)$, is known accurately, we obtain the best estimate of the corresponding projector as

$$\mathbf{P}^{\perp}_{\mathbf{A}_q}(k) = \mathbf{I} - \mathbf{A}_q(k)\left(\mathbf{A}_q^H(k)\,\mathbf{A}_q(k)\right)^{-1}\mathbf{A}_q^H(k)\ . \tag{6.21}$$

Such an estimate of the projector onto the orthogonal complement and the corresponding estimate of the blocking matrix $\mathbf{B}_q(k)$ is what we shall refer to as the *oracle* estimate. We shall examine, subsequently, the performance of this estimate against the adaptive estimate in an anechoic situation.

Oracle vs adaptive $\mathbf{B}_q$ estimate – matched sensors

With respect to Figure 6.10, we see that the performance of the GSC using the adaptive estimate of the blocking matrix is close to the oracle approach. This is also verified on the instrumental performance measures computed for the two algorithms, where PEG is within a dB of the performance attained by the oracle-GSC.



(a) $s_1(n)$ (b) $s_2(n)$

(c) $y_{1,\mathrm{Oracle}}(n)$ (d) $y_{2,\mathrm{Oracle}}(n)$

(e) $y_{1,\mathrm{PEG}}(n)$ (f) $y_{2,\mathrm{PEG}}(n)$

Figure 6.10: Separation performance under anechoic conditions with matched sensors.

Oracle vs adaptive B_q estimate – mismatched sensors

When the sensor gains are not matched and the mismatch is unknown, the oracle estimate of the orthogonal complement is no longer perfect. Using the oracle estimate of $\mathbf{B}_q$ now leads to target signal cancellation. The adaptive estimate, however, continues to perform well, and is robust against such mismatches.

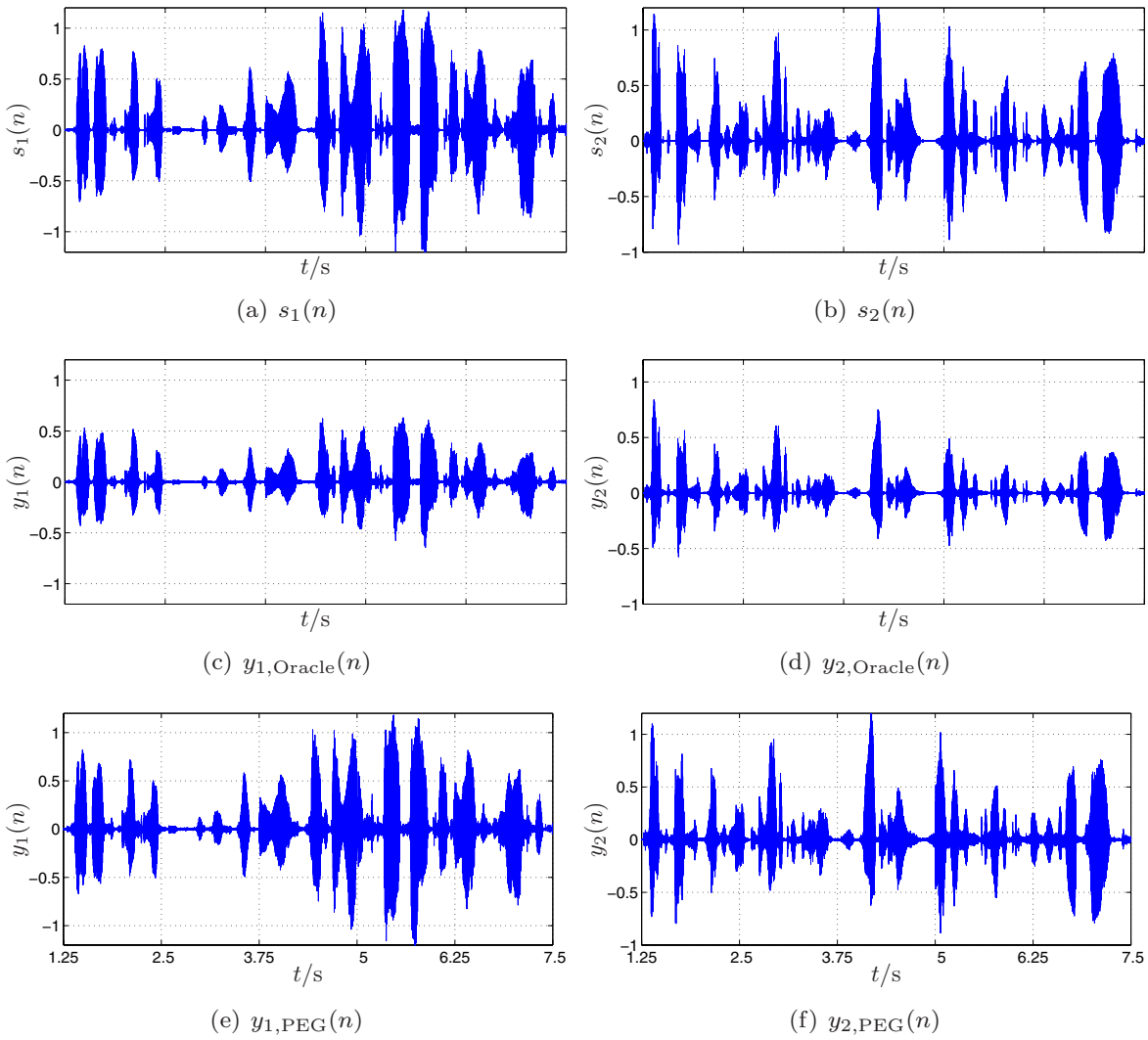
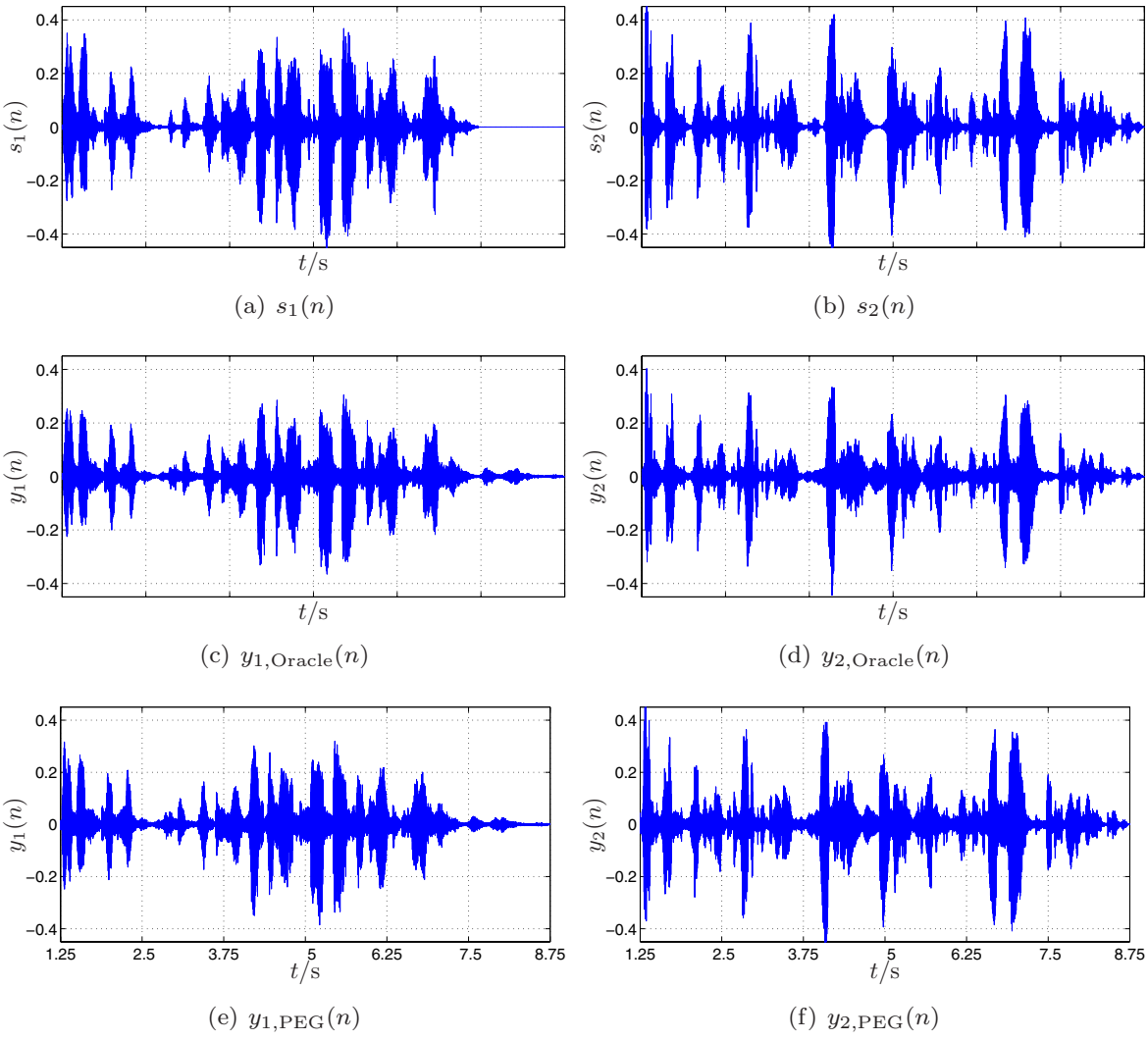


Figure 6.11: Separation performance under anechoic conditions with mismatched sensors.

Oracle vs adaptive $\mathbf{B}_q$ estimate – reverberant environments

Even if the sensor gains could be matched (for example, by array calibration approaches proposed in Chapter 7), oracle based methods suffer from another drawback when the environment is not anechoic: in reverberant environments the propagation matrix from the source to the array also includes the reflections. An accurate estimate of $\mathbf{A}_q$ is therefore not available, even if the direction of arrival is accurately known, and in this case too, using the oracle estimate of $\mathbf{B}_q$ leads to target signal cancellation (with well calibrated sensors, this was about 1 dB for the oracle-steered GSC in the above example).



(a) $s_1(n)$ (b) $s_2(n)$

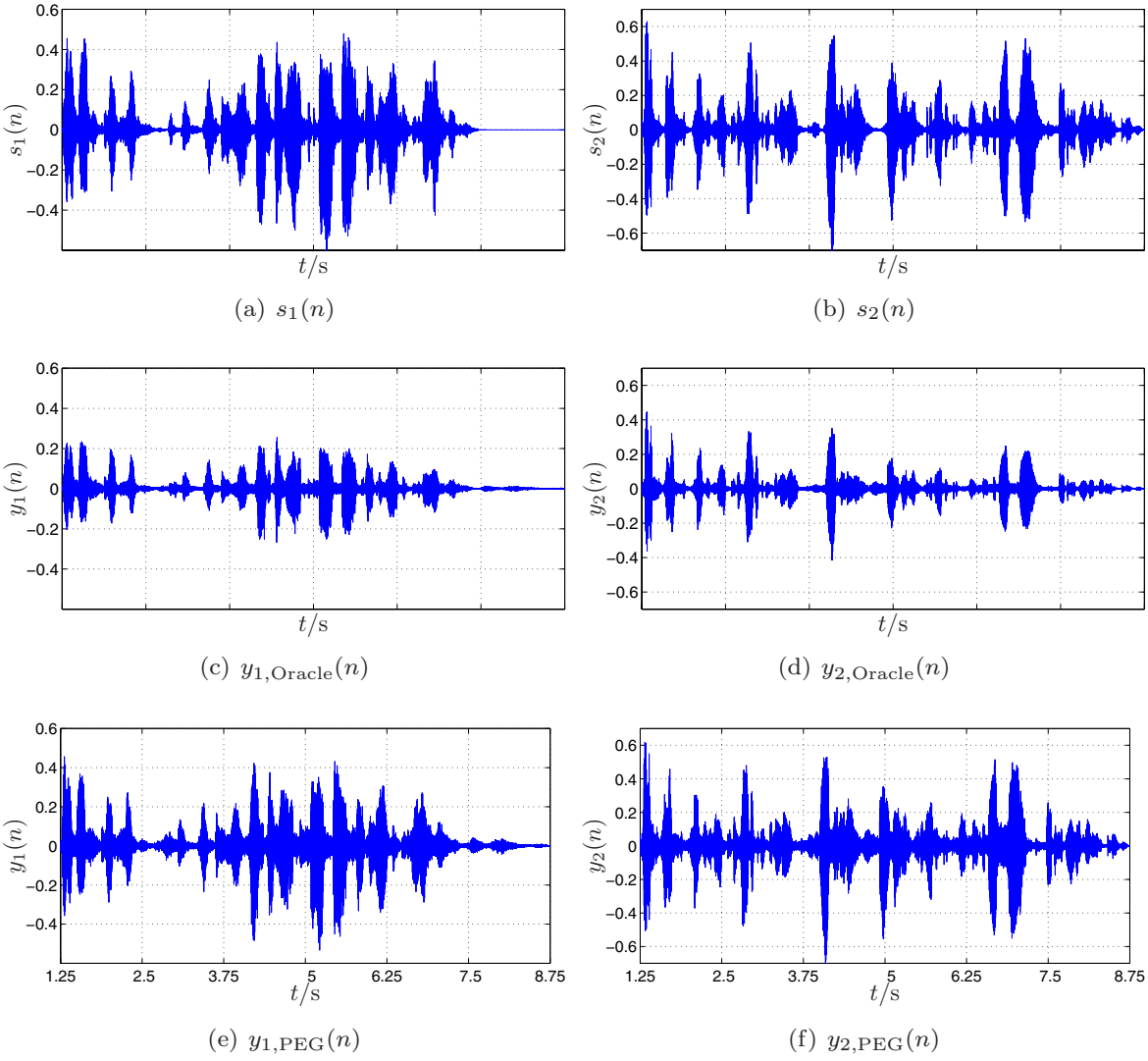
(c) $y_{1,\mathrm{Oracle}}(n)$ (d) $y_{2,\mathrm{Oracle}}(n)$

(e) $y_{1,\mathrm{PEG}}(n)$ (f) $y_{2,\mathrm{PEG}}(n)$

Figure 6.12: Separated signals obtained in reverberant environments with matched sensors.

Oracle vs adaptive $\mathbf{B}_q$ estimate – reverberant environments, mismatched sensors

This last comparison considers mismatched sensors in reverberant environments. The performance degradation of the oracle-estimate is worse than in the anechoic case due to the combined effect of reverberation and sensor mismatch.



(a) $s_1(n)$ (b) $s_2(n)$

(c) $y_{1,\mathrm{Oracle}}(n)$ (d) $y_{2,\mathrm{Oracle}}(n)$

(e) $y_{1,\mathrm{PEG}}(n)$ (f) $y_{2,\mathrm{PEG}}(n)$

Figure 6.13: Separated signals obtained in reverberant environments using mismatched sensors.

Oracle comparison summary

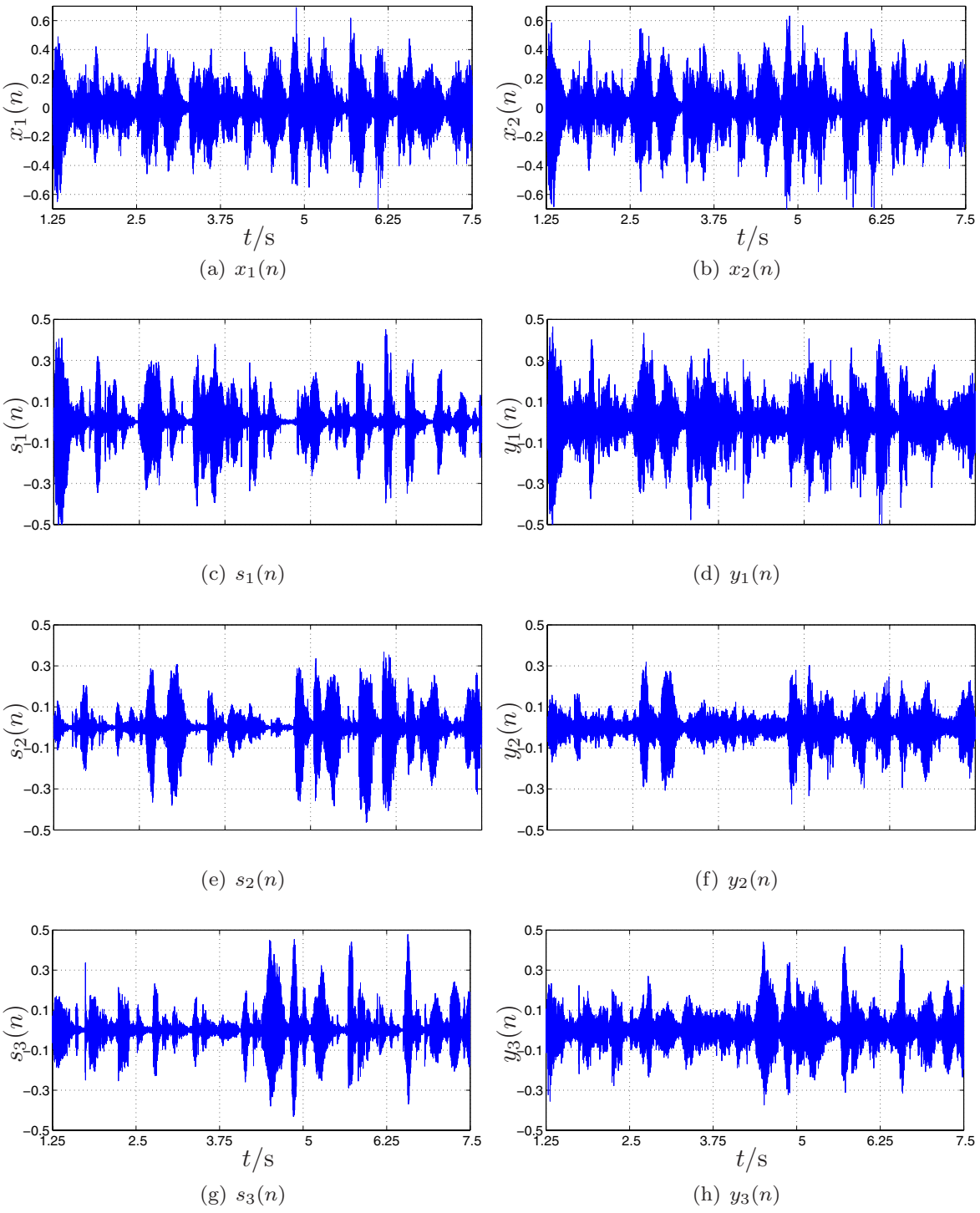
We see that the proposed method to estimate the blocking matrix performs as well as the oracle approach when the sensors are well calibrated and the source positions are accurately known. The degradation of the oracle approach is evident when the sensor gains are mismatched. The adaptive system, however, remains robust. As an aside, even if the sensor gains are well known, the oracle based blocking matrix estimate is not viable in practice. This is because the true source positions are rarely known accurately and must be estimated. As the oracle approach is then based on these estimates, the degradation of the

target signal is more severe than in Figure 6.12, even when the sensors are well matched in gain.

Underdetermined PEG

A last example presented here to show the versatility of our approach is that of underdetermined source separation. We consider the separation of three speaker signals (at $\theta \in \{60°, 90°, 120°\}$) from a mixture observed by two microphones, spaced 8 cm apart (corresponds to mic. 1 & 3 in our 5 sensor array). Observe that it is still possible to approximate the underlying source signals. One may also view PEG in this example as a generalized version of the approach followed in [71], which makes rather primitive assumptions on the mixing system. The batch approaches of ICA and the SOS approaches of [21, 20] fail in such a situation.

Note that $y_2(n)$ and $y_3(n)$, which correspond to sources at 60° and 120° respectively, are recovered better than the source at broadside. This is because for each of these sources, the interferers are in the same direction (either all to the left or all to the right). Thus, given the single degree of freedom to steer a null, it is possible when extracting these sources, to suppress both the interference signals. For the source at broadside, this is difficult, as there is one interference to the left and another to the right. Additionally, this spatial distribution makes it difficult to quickly steer the null towards the dominant interferer given their wide azimuthal separation. However, all three extracted sources are clearer (the individual sentences can be better focussed upon) than in the mixtures.



(a) $x_1(n)$ (b) $x_2(n)$

(c) $s_1(n)$ (d) $y_1(n)$

(e) $s_2(n)$ (f) $y_2(n)$

(g) $s_3(n)$ (h) $y_3(n)$

Figure 6.14: PEG performance in the underdetermined case ($M = 2$, $Q = 3$). The left column depicts the clean source signal and the right column, the signals obtained from the mixture.

6.2.5 Experimental evaluation of the generalized separation algorithms

In the previous sections, we have demonstrated the performance of the various data-driven separation algorithms on the basis of a single example. This was primarily to illustrate the behaviour of the various algorithms. We shall now attempt to quantify their performance

using instrumental measures. For this, we assume the presence of two concurrent speakers, whose signals are picked up by the 5 channel microphone array, in the presence of various kinds of background noise and at varying SNR. The speakers are located at azimuths of 60° and 120° with respect to the array axis. For the competing speaker signals, we use a subset of the speakers considered in Chapter 4. Each speaker at an azimuth is tested against all the other speakers at the interference position. The results obtained are averaged over all such speaker-pair combinations. The background noise and the simulation scenarios are, otherwise, the same as in Chapter 4. The performance of the algorithms are evaluated according to the following metrics:

- ΔSIR, the signal to interference ratio *improvement*,
- ΔSegSIR, the segmental SIR *improvement*,
- ΔIW-SegSIR, the segmental intelligibility-weighted SIR *improvement*,
- ΔSNR, the signal to noise ratio *improvement*,
- ΔSegSNR, the segmental SNR ratio *improvement*,
- ΔIW-SegSNR, the segmental intelligibility-weighted SNR *improvement*,
- ΔSINR, the signal to interference + noise ratio improvement,
- ΔSegSINR, the segmental SINR *improvement*,
- ΔIW-SegSINR, the segmental intelligibility-weighted SINR *improvement*, and
- $\log_{\mathrm{SD}}$, the log spectral distortion.

For details of the instrumental measures, the reader is referred to Appendix G and to [40].

The table below summarizes the performance of the various algorithms for the envisaged scenario.

Separation results in Room 1

Clean mixing conditions (no background noise present)

Table 6.2: Performance under noise-less mixing conditions. Only a target source and an interferer are present.

Metric (db)	DSB	PEG	Mask	Smth. Mask
ΔSIR	1.69	4.60	6.17	4.66
ΔSegSIR	1.66	3.40	5.88	3.96
ΔIW-SegSIR	2.43	3.04	5.20	3.57
ΔSNR	N/A	N/A	N/A	N/A
ΔSegSNR	N/A	N/A	N/A	N/A
ΔIW-SegSNR	N/A	N/A	N/A	N/A
ΔSINR	1.69	4.60	6.17	4.66
ΔSegSINR	1.66	3.40	5.88	3.96
ΔIW-SegSINR	2.43	3.04	5.20	3.57
$\log_{SD}$	4.86	6.61	7.78	6.54
long-$\log_{SD}$	1.05	1.66	2.27	2.06

White noise

Table 6.3: Performance of separation algorithms in room 1, with spatially uncorrelated, white background noise at the sensors

Metric (db)	DSB			PEG			Mask			Smth. Mask		
	0db	10db	20db	0db	10db	20db	0db	10db	20db	0db	10db	20db
ΔSIR	1.77	1.76	1.71	5.17	5.02	4.70	6.75	6.65	6.10	4.52	4.70	4.49
ΔSegSIR	1.67	1.68	1.67	2.96	3.44	3.52	4.86	5.64	5.92	3.04	3.77	3.89
ΔIW-SegSIR	2.36	2.39	2.41	1.76	2.45	2.85	4.31	4.87	5.15	2.99	3.38	3.52
ΔSNR	5.62	5.64	5.64	4.09	1.62	-1.08	7.63	7.99	8.04	7.42	7.46	6.92
ΔSegSNR	5.17	5.44	5.50	3.40	1.97	-0.19	5.55	6.84	7.45	5.59	6.24	6.14
ΔIW-SegSNR	3.94	4.71	4.78	1.86	1.16	-0.23	4.79	6.71	7.56	4.31	5.48	5.71
ΔSINR	3.36	2.02	1.74	4.81	4.73	4.63	6.82	6.64	6.10	5.71	4.90	4.51
ΔSegSINR	3.73	2.78	2.06	3.91	3.93	3.70	5.06	6.05	6.21	4.52	4.56	4.21
ΔIW-SegSINR	2.95	3.04	2.77	2.57	3.07	3.17	3.99	5.19	5.58	3.44	4.03	3.99
$\log_{SD}$	4.84	4.87	4.87	7.12	6.84	6.63	8.46	7.94	7.77	6.52	6.54	6.54
long-$\log_{SD}$	1.10	1.06	1.05	1.93	1.74	1.67	2.41	2.35	2.31	1.86	2.00	2.06

Diffuse white noise

Table 6.4: Performance in room 1, with diffuse white background noise at the sensors. The noise is spatially *correlated*.

Metric (db)	DSB			PEG			Mask			Smth. Mask		
	0db	10db	20db	0db	10db	20db	0db	10db	20db	0db	10db	20db
ΔSIR	1.79	1.75	1.72	4.20	4.70	4.42	5.57	6.19	6.02	3.54	4.64	4.58
ΔSegSIR	1.69	1.68	1.67	2.60	3.29	3.39	4.49	5.45	5.78	2.87	3.72	3.96
ΔIW-SegSIR	2.33	2.39	2.41	1.68	2.44	2.84	3.91	4.70	5.04	2.88	3.35	3.53
ΔSNR	1.63	1.63	1.62	1.42	0.16	-1.29	3.14	4.07	4.43	2.54	3.12	3.09
ΔSegSNR	1.43	1.52	1.55	1.75	0.84	-0.32	2.01	3.39	4.18	1.67	2.31	2.52
ΔIW-SegSNR	0.94	1.27	1.38	1.47	1.19	0.36	1.34	2.92	4.10	1.07	1.88	2.30
ΔSINR	1.70	1.74	1.72	2.90	4.20	4.34	3.92	5.77	5.98	2.96	4.44	4.56
ΔSegSINR	1.43	1.58	1.64	2.62	3.31	3.50	2.46	4.60	5.69	1.93	3.20	3.82
ΔIW-SegSINR	1.08	1.65	2.06	2.02	2.81	3.06	1.57	3.45	4.75	1.28	2.46	3.22
$\log_{SD}$	4.84	4.86	4.86	6.63	6.59	6.57	7.72	7.75	7.74	6.16	6.36	6.47
long-$\log_{SD}$	1.08	1.05	1.05	1.71	1.65	1.65	2.22	2.24	2.32	1.81	1.94	2.06

Diffuse babble noise

Table 6.5: Performance in room 1, with diffuse babble background noise at the sensors. The noise is spatially *correlated* and non-stationary.

Metric (db)	DSB			PEG			Mask			Smth. Mask		
	0db	10db	20db	0db	10db	20db	0db	10db	20db	0db	10db	20db
ΔSIR	1.72	1.71	1.69	4.06	4.51	4.50	5.61	6.08	6.02	3.45	4.02	4.64
ΔSegSIR	1.68	1.67	1.66	2.86	3.32	3.38	5.10	5.73	5.86	3.16	3.69	4.00
ΔIW-SegSIR	2.41	2.42	2.42	2.27	2.81	2.96	4.51	5.01	5.15	3.19	3.45	3.59
ΔSNR	1.13	1.11	1.11	1.23	-0.06	-1.19	2.73	3.69	4.28	2.20	2.77	3.03
ΔSegSNR	0.79	0.85	0.86	1.51	0.47	-0.58	2.48	3.76	4.49	1.74	2.38	2.62
ΔIW-SegSIR	1.45	1.63	1.59	2.04	1.20	0.40	2.73	4.04	4.60	1.91	2.48	2.61
ΔSINR	1.37	1.63	1.69	2.54	3.85	4.39	3.46	5.52	5.95	2.66	3.82	4.61
ΔSeg-SINR	0.97	1.29	1.52	2.48	3.15	3.47	2.91	4.90	5.82	2.01	3.15	3.88
ΔIW-SegSINR	1.60	2.00	2.25	2.80	3.15	3.21	2.85	4.40	5.16	2.10	2.98	3.49
$\log_{SD}$	4.82	4.86	4.86	6.44	6.51	6.54	7.69	7.77	7.77	6.16	6.35	6.47
long-$\log_{SD}$	1.05	1.05	1.05	1.54	1.66	1.69	2.33	2.37	2.31	2.09	2.16	2.04

Separation results in Room 2

Separation performance in the seminar room - clean mixing conditions

Table 6.6: Algorithm performance in room 2, under clean mixing conditions, with no background noise.

Metric (db)	DSB	PEG	Mask	Smth. Mask
ΔSIR	4.24	10.26	15.09	9.96
ΔSegSIR	3.43	7.65	11.79	7.30
ΔIW-SegSIR	4.48	6.77	10.46	6.15
ΔSNR	N/A	N/A	N/A	N/A
ΔSegSNR	N/A	N/A	N/A	N/A
ΔIW-SegSNR	N/A	N/A	N/A	N/A
ΔSINR	4.24	10.26	15.09	9.96
ΔSegSINR	3.43	7.65	11.79	7.30
ΔIW-SegSINR	4.48	6.77	10.46	6.15
$\log_{SD}$	3.89	5.10	7.40	6.42
long-$\log_{SD}$	0.85	1.46	1.67	1.61

Separation performance in the seminar room - uncorrelated white noise conditions

Table 6.7: Performance of separation algorithms with real-room recordings and uncorrelated white noise at the sensors.

Metric (db)	DSB			PEG			Mask			Smth. Mask		
	0db	10db	20db	0db	10db	20db	0db	10db	20db	0db	10db	20db
ΔSIR	4.23	4.19	4.21	10.67	10.63	10.49	15.42	15.15	14.71	9.54	9.83	9.76
ΔSegSIR	3.40	3.40	3.42	6.56	7.41	7.71	11.13	11.60	11.33	6.31	7.00	7.14
ΔIW-SegSIR	4.39	4.45	4.47	4.13	5.60	6.45	9.05	9.83	10.03	5.64	6.01	6.12
ΔSNR	6.77	6.72	6.70	5.99	4.49	3.23	10.91	10.70	10.38	9.99	9.50	9.01
ΔSegSNR	6.11	6.19	6.12	4.80	3.92	2.98	8.26	8.87	8.67	7.34	7.57	7.23
ΔIW-SegSNR	4.90	5.36	5.18	2.84	2.60	2.21	6.83	8.07	7.96	5.56	6.17	5.86
ΔSINR	5.17	4.34	4.22	8.45	10.01	10.41	12.65	14.46	14.62	9.68	9.78	9.75
ΔSegSINR	4.91	4.11	3.61	6.39	7.37	7.72	9.23	10.97	11.27	6.96	7.29	7.23
ΔIW-SegSINR	4.02	4.11	4.13	4.19	5.57	6.51	6.84	8.51	9.37	4.93	5.50	5.73
$\log_{SD}$	3.83	3.90	3.90	6.52	5.77	5.28	8.39	7.85	7.33	6.74	6.67	6.41
long-$\log_{SD}$	0.88	0.85	0.85	1.99	1.74	1.58	2.40	1.97	1.72	2.15	1.81	1.64

Separation performance in the seminar room - diffuse white noise conditions

Table 6.8: Performance in room 2, with diffuse white noise at the sensors. Note that the noise signals show spatial *correlation*

Metric (db)	DSB			PEG			Mask			Smth. Mask		
	0db	10db	20db	0db	10db	20db	0db	10db	20db	0db	10db	20db
ΔSIR	4.28	4.23	4.23	9.35	10.18	10.29	14.95	14.96	14.69	9.15	9.69	9.84
ΔSegSIR	3.44	3.42	3.43	5.90	7.15	7.63	10.76	11.40	11.30	6.33	7.01	7.19
ΔIW-SegSIR	4.41	4.46	4.47	3.92	5.51	6.41	8.74	9.70	10.04	5.67	6.08	6.15
ΔSNR	2.74	2.72	2.72	3.38	2.62	2.00	6.70	7.00	6.85	5.49	5.56	5.26
ΔSegSNR	2.18	2.23	2.23	2.92	2.30	1.82	4.46	5.19	5.40	3.27	3.62	3.66
ΔIW-SegSNR	1.65	1.89	1.96	2.28	2.25	2.08	3.05	4.30	4.89	2.05	2.57	2.67
ΔSINR	3.50	4.08	4.21	6.34	9.24	10.17	9.50	13.41	14.48	7.13	9.16	9.76
ΔSegSINR	2.57	3.03	3.32	4.89	6.56	7.48	6.23	9.30	10.81	4.26	5.97	6.92
ΔIW-SegSINR	2.04	2.91	3.68	3.52	5.17	6.37	4.01	6.78	8.84	2.62	4.09	5.22
$\log_{\mathrm{SD}}$	3.81	3.88	3.89	5.92	5.49	5.20	8.04	7.72	7.30	6.46	6.53	6.36
long-$\log_{\mathrm{SD}}$	0.86	0.84	0.85	1.83	1.65	1.56	2.33	1.91	1.73	2.07	1.81	1.63

Separation performance in the seminar room - diffuse babble conditions

Table 6.9: Performance in room 2, with non-stationary diffuse babble noise at the sensors.

Metric (db)	DSB			PEG			Mask			Smth. Mask		
	0db	10db	20db	0db	10db	20db	0db	10db	20db	0db	10db	20db
ΔSIR	4.22	4.23	4.23	9.36	10.25	10.28	14.24	14.78	14.87	9.06	9.67	9.90
ΔSegSIR	3.42	3.42	3.43	6.35	7.36	7.58	10.65	11.25	11.55	6.61	7.08	7.30
ΔIW-SegSIR	4.46	4.47	4.47	5.07	6.24	6.63	9.20	9.95	10.27	5.80	6.07	6.16
ΔSNR	2.18	2.18	2.18	3.46	2.64	1.98	6.20	6.19	6.53	5.32	5.08	5.10
ΔSegSNR	1.48	1.50	1.50	2.91	2.21	1.53	4.44	4.96	5.46	3.25	3.51	3.64
ΔIW-SegSNR	2.11	2.25	2.06	3.11	2.87	2.34	4.12	5.03	5.22	2.67	3.00	2.86
ΔSINR	3.05	3.97	4.20	5.90	8.95	10.09	8.49	12.59	14.55	6.73	8.91	9.80
ΔSegSINR	2.09	2.77	3.21	4.78	6.51	7.37	6.10	8.99	10.96	4.28	5.93	6.97
ΔIW-SegSINR	2.83	3.63	4.15	4.88	6.23	6.77	5.71	8.29	9.82	3.64	4.96	5.78
$\log_{SD}$	3.77	3.88	3.89	5.44	5.24	5.15	7.77	7.48	7.39	6.37	6.37	6.41
long-$\log_{SD}$	0.84	0.84	0.84	1.61	1.51	1.49	1.96	1.70	1.67	1.84	1.65	1.61

6.2.6 Discussion

With respect to the tables above, we observe some rather interesting trends which, upon reflection, are in keeping with the corresponding algorithm design. Firstly, we see that the linear algorithms exceed the performance of the mask-based algorithms in terms of target signal quality, as seen by the lower log spectral distortion values.

Among the linear approaches, the noise suppression using PEG is worse than the DSB for spatially uncorrelated white noise, especially at lower input SNRs. This is because the noise suppression capability of PEG depends upon the spatial correlation of the interference signals (noise + competing speaker) – which prerequisite is not completely met in the case of white noise. Subsequently, from the point of view of *noise suppression*, the DSB is optimal here. In the presence of diffuse noise (white/babble), which evinces some spatial correlation, PEG performs better than the DSB, in terms of ΔSNR, at low SNRs. With increasing SNR, however, PEG tends to concentrate its resources on interference cancellation, indicated by the increasing ΔSIR and decreasing ΔSNR values as the input SNR increases. Note, also, that PEG is always better than the DSB in terms of SINR – which is to be expected in the presence of a directive interference. The more or less constant value of the DSB in terms of ΔSIR is also to be expected as the DSB does not actively cancel interferences.

In terms of the target signal distortion, PEG has a slightly higher log spectral distortion value than the DSB, when measured with respect to the reverberant input signal at reference microphone 1. In general, under such a comparison, any linear beamforming algorithm would suffer from some distortion as the beamformer reduces the reverberant component – which is present in the reference signal. This explains why the DSB also shows some target signal 'deterioration'. The higher values of the log spectral distortion for PEG could be due to the formation of a sharper beam as compared to the DSB, leading to a further reduction in the reverberant component and increased distortion. Note, also, that the spectral distortion values remain more or less constant for a given simulation environment under all the different noise types and SNR conditions. This is an indication of the robustness of the localization stage (for the DSB) and the blocking matrix and interference canceller adaptation stage (for PEG) against noise, and is a very welcome attribute.

The mask based approaches have a higher over all SINR improvement, as expected, but at the cost of higher target signal spectrum degradation. Smoothing the masks yields lower values for the $\log_{\mathrm{SD}}$, at the cost of correspondingly higher cross-talk. We note that the current implementation of the mask-based approaches are best at suppressing background noise. In the presence of a directive interference, they do not optimally utilize the spatial diversity afforded them by the microphone array. Conversely, linear algorithms do not optimally utilize the signal sparsity and disjointness, and aim to suppress the *average* power of the interference and noise.

The separation performance may be increased by combining the linear and the non-linear approaches in a single framework. As a point of note, simply piggy-backing a single channel post-processor onto the output of the linear algorithms, *without* considering the other output signals, does not perform well. An optimal post-processor should take extracted competing source signals into account, along with any spectral disjointness that might be present, and we believe that such a post-processor would be a hybrid between that used in the multi-channel Wiener filter and the mask-based approaches of [65, 74].

The rather dismal performance of the algorithms in Room 1 is due to the fact that the simulation software imposes a flat absorption spectrum on the room across the frequency. This is clearly unrealistic – the absorption coefficients increase with frequency, leading to

less reverberation (and the possibility of steering sharper nulls) in the higher frequencies. In the absence of such behaviour, the spatial selectivity is hindered and the beamforming algorithms are only able to steer rather shallow nulls, even in the higher frequencies. This leads to a larger amount of residual interference and noise and subsequently lowers the performance indices. Note however the drastic improvement in performance under more realistic conditions (Room 2), despite the larger T_{60}.

6.3 Conclusions

In this chapter we have studied the speaker separation problem using two categories of separation algorithms. The aim was to highlight the use of localization information in the practical realizations of such algorithms.

We began with the so-called 'blind' linear algorithms based on ICA, and showed how localization information can be used to generate practical realizations at lower computational cost as compared to the state-of-the-art. We also demonstrated the resilience of the proposed approach to errors in the localization information. In short, a rough estimate of any *one* target source position suffices for the 2×2 problem handled in this chapter.

Next we studied the more general case of M microphones and Q sources and developed an algorithm based explicitly on the localization approaches developed in Chapter 4. We also showed how the usage of this model allows for the development of a wide variety of source separation algorithms, from the mask-based to the linear GSC based approach.

In comparing the MoG based separation algorithm variants to the ICA approaches, note that the ICA algorithms are batch-based and rather inflexible in that they require, strictly, $M \geq Q$, that the number of sources must be known *a priori* and that the signals satisfy non-Gaussianity constraints. The MoG based approaches are, principally, immune to such limitations. Of course, the only limitations here are of a physical rather than algorithmic nature.

The mask-based variants using the MoG model offer very good interference suppression at the cost of larger target signal degradation and undesired artefacts. These disturbing effects may be mitigated to some extent by smoothing the masks. The result obtained from the GSC variant sounds most natural as the signal gain in the target direction is preserved and the spatial diversity is used to cancel the directive interference and suppress the diffuse components. Further, we also demonstrated that the design of the blocking matrix in the proposed GSC approach is robust against sensor mismatch.

The good suppression capabilities of the mask-oriented methods and the natural-ness of the GSC output leads us to believe that a weighted combination of these approaches could combine the best of both. Note that we do not mean a post-filter as developed in [83, 137] or the MSE-optimal approach of the previous chapter as such post-filter systems are still based on signal statistics. Rather, we would like to exploit the signal disjointness at the individual time frames. However, further considerations along this direction are deferred to a later work.

III

Array Calibration

The wizards were civilized men of considerable education and culture. When faced with being inadvertently marooned on a desert island they understood immediately that the first thing to do was to place the blame.

The last continent
– Terry Pratchett

Chapter 7

Microphone Fault Detection and Calibration

7.1 Introduction

The previous chapters on localization and separation implicitly assume the proper functioning of each microphone in the array. However, in applications where the arrays are deployed in adverse environments and for long periods of time, this assumption needs to be verified periodically. Furthermore, the underlying matched-microphone characteristics assumption of most multi-channel signal processing algorithms is not guaranteed due to the manufacturing tolerances of the microphones and that of the associated amplifier and sampling components. Depending upon the implementation, such mismatches may lead to a degradation in algorithm performance. Therefore, calibration of the microphone array becomes necessary.

The problem of calibration has been well studied in the literature. One approach is to calibrate the microphone arrays using known calibration sources, either in anechoic environments or *in situ* [87]. The calibration filters are then held constant during the operational phase. Periodic recalibration may be done to correct degradations due to changes in environment or age-related changes. If done properly, such approaches are able to correct both gain and phase mismatches and can considerably improve the performance of the multi-channel localization and enhancement algorithms.

Alternatively, in [88], the calibration filters are computed *online*, in a generalized sidelobe canceller (GSC) structure, using the enhanced signal after the fixed beamformer stage for better SNR conditions. The direction of the desired signal is assumed known and the calibration is done in the presence of this target signal. Similar structures are proposed in [22]. These approaches are better suited to dynamic enviroments where a periodic recalibration using known signals is impractical.

In [118] it is argued that the problem arising from mismatch is more due to the gain imbalance than the phase. Thus, to improve performance, it is sufficient to compute only the gain correction factor and neglect the phase. This is also the approach followed in [52]. The difference in the two approaches lies mainly in the adaptation of the gain correction factor. Whereas [118] adapts the gain factors in the presence of a *single* dominant source with a strong direct path, [52] models the gains over a long-term average of the signal powers, with

the assumption that, over a long observation period, the average power received by all the microphones must be the same in the absence of mismatch.

However, none of the above approaches explicitly test the microphone channels to see if they are functional. While detection of sensor failure can be done by manually disassembling and re-calibrating the arrays with the corresponding advantage of laboratory condition calibration and detection, the procedure is not only time-consuming, but also impractical. Online algorithms for sensor degradation detection are thus necessary and such algorithms are the focus of this chapter.

We start with a description of the system model and enumerate our key assumptions and simplifications. Based on this model, Sections 7.3 and 7.4 present a detailed discussion of two online approaches for the detection of sensor anomalies: the spectral correlation (SCORE) approach and the minimum mean square error approach. The described approaches are then tested under various error conditions and their performance is summarized.

7.2 System model

As in the previous chapters, the methods described subsequently operate on the STFT representation of the microphone signals. Furthermore, we consider the general case of an M channel array placed in an environment containing Q sources distributed within it. Note that, in contrast to the previous chapters, Q is any unknown number *inessential* to the proposed approaches. The general signal model we shall assume subsequently is:

$$\mathbf{X}(k,b) = \mathbf{A} \odot \left(\sum_{q=1}^{Q} \mathbf{S}_q(k,b) \right) + \mathbf{V}(k,b), \tag{7.1}$$

where $\odot$ represents the element-wise (Hadamard) product,

$$\mathbf{X}(k,b) = (X_1(k,b),\ X_2(k,b),\ \ldots,\ X_M(k,b))^T \in \mathbb{C}^{M\times 1}$$

is the M-dimensional vector of microphone signals at frequency bin k and frame b, and $\mathbf{A} = (A_1,\ \ldots,\ A_M)^T$ indicates the vector of calibration factors. Following [118], the A_m are assumed to be purely real and *independent* of frequency. The vector

$$\mathbf{S}_q(k,b) = (S_{1q}(k,b),\ S_{2q}(k,b),\ \ldots,\ S_{Mq}(k,b))^T \in \mathbb{C}^{M\times 1}$$

represents the contribution of the qth source to the signals at the array, with the propagation effects being implicitly included in the definition of $S_{mq}(k,b)$. The term $\sum_{q=1}^{Q} \mathbf{S}_q(k,b)$ thus represents the part of the array input that is dependent upon the signals in the environment and received at all the channels and the term

$$\mathbf{V}(k,b) = (V_1(k,b),\ \ldots,\ V_M(k,b))^T \in \mathbb{C}^{M\times 1}$$

represents the part of the input that contains self-induced noise. Further, as in the previous chapters, the microphone signals are assumed to be zero-mean and the Q sources statistically independent of one another and of the self-induced noise.

The key assumption we make for our detection approaches, namely that each microphone receives the same power on average, is similar to that in [52]. For arrays with closely spaced elements, as is often the case, this is a justifiable assumption. As a note, under some acoustic

conditions, e.g., when the array is in a standing wave field, the assumption of equal average power may be grossly violated, leading to strong power variations across the microphones of the array. These variations do not reflect the mismatch being modelled. Consequently, adapting the gain functions and/or detecting degradation etc., leads to biased results in such conditions. Such wave fields are usually generated by strongly directive sources and, in such cases, it might be advisable to allow the algorithm to be active only during periods where these sources are *absent*. The detection of such directive sources may be accomplished, for example, as in [78, 96, 77].

7.3 The SCORE approach

For a healthy array, a sufficient condition for equal average power at each sensor is that the average spectral power of each sensor should be similar. For a microphone m, the power spectrum may be computed by a temporal averaging of the instantaneous power: $\mathcal{E}_m(k,b) = |X_m(k,b)|^2$, of the DFT coefficients over multiple frames B. We define this value as:

$$\widehat{\Psi}_{X_m X_m}(k) = \frac{1}{B}\sum_b \mathcal{E}_m(k,b)\,. \tag{7.2}$$

To detect sensor degradations, we treat the power spectrum of each channel as a realization of a random process and compute the *correlations* between the spectral powers of the channels. Thus, for each channel pair (m, m'), we obtain the corresponding correlation coefficient [93, Eq. 7-8] $\Gamma_{mm'}$ as follows:

$$\widehat{\mu}_m = \frac{1}{K}\sum_{k=1}^{K} \widehat{\Psi}_{X_m X_m}(k)\,, \tag{7.3}$$

$$\Gamma_{mm'} = \frac{\sum_k \left(\widehat{\Psi}_{X_m X_m}(k) - \widehat{\mu}_m\right)\left(\widehat{\Psi}_{X_{m'} X_{m'}}(k) - \widehat{\mu}_{m'}\right)}{\sqrt{\sum_k \left(\widehat{\Psi}_{X_m X_m}(k) - \widehat{\mu}_m\right)^2 \sum_k \left(\widehat{\Psi}_{X_{m'} X_{m'}}(k) - \widehat{\mu}_{m'}\right)^2}}, \tag{7.4}$$

where K is the length of the (discrete) Fourier transform. Stacking the $\Gamma_{mm'}$ according to the indices, we obtain the corresponding correlation matrix $\mathbf{\Gamma}$ for the array under consideration. The distribution of the elements of such a correlation matrix is the key to detecting corrupt sensors. When the array is healthy, the elements $\Gamma_{mm'} \approx 1$. If the array contains defective sensors, the corresponding $\Gamma_{mm'}$ deviate from this value. Thus, detection of corrupt sensors may be done by setting a threshold on this statistic. To obtain a suitable detection threshold Υ_Γ, the distribution $p(\Gamma)$ of the Γ is estimated on the basis of the histogram of observed $\Gamma_{mm'}$ over the sensors of a healthy array (see Figure 7.1). On the basis of this distribution, and for a given probability of missed detections α, we select Υ_Γ such that $\int_{\Gamma<\Upsilon_\Gamma} p(\Gamma)\,\mathrm{d}\Gamma \leq \alpha$.

7.3.1 The advantage of centering

The advantage of using the correlation coefficients as defined in (7.4) becomes evident when we consider a noise floor that is spectrally flat (which is a realistic assumption for the sensor

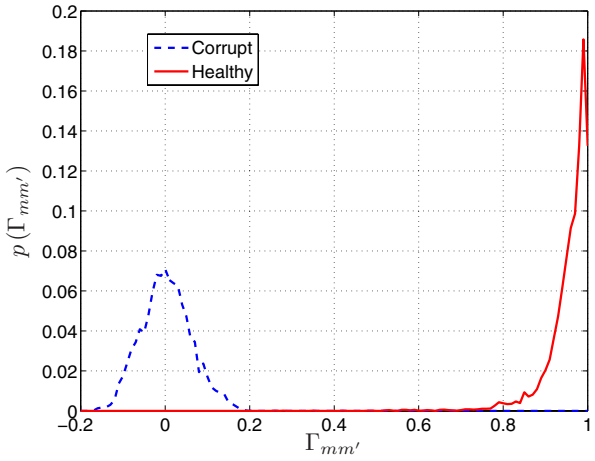


Figure 7.1: Estimated $p(\Gamma_{mm'})$ for a healthy array and for a sample, single sensor degradation.

noise). For this case, taking the statistical expectation of the instantaneous power, we obtain from (7.1) and the definition of $\mathcal{E}_m(k,b)$:

$$\begin{aligned} \mathrm{E}\left\{\mathcal{E}_m(k,b)\right\} &= \mathrm{E}\left\{|X_m(k,b)|^2\right\} \\ &= |A_m|^2 \sum_{q=1}^{Q} \Psi_{S_{mq}S_{mq}}(k) + \Psi_{V_mV_m}(k) \qquad (7.5) \\ &= |A_m|^2 \Psi_{SS}(k) + \Psi_{V_mV_m}, \qquad (7.6) \end{aligned}$$

where $\Psi_{S_{mq}S_{mq}}(k)$ is the power spectral density of the qth source at sensor m, for bin k, and $\Psi_{SS}(k) = \sum_{q=1}^{Q} \Psi_{S_{mq}S_{mq}}(k)$. Note that the equal average power assumption at each channel implies that $\Psi_{SS}(k)$ is the same for all the channels – indicated by dropping the channel index m in (7.6). Note, also, that the frequency index has been dropped for the noise because of the assumption of flat spectrum characteristics. Consequently,

$$\begin{aligned} \mu_m &= \frac{1}{K} \sum_{k=1}^{K} \mathrm{E}\left\{\mathcal{E}_m(k)\right\} \\ &= |A_m|^2 \frac{1}{K} \sum_{k=1}^{K} \Psi_{SS}(k) + \Psi_{V_mV_m}. \end{aligned} \qquad (7.7)$$

We now see that computing the correlation as in (7.4) yields a value that is not sensitive to white noise and is dependent only upon the incident signals only, thus avoiding any bias in the correlation function.

7.3.2 Determining the corrupted channels from $\mathbf{\Gamma}$

As shown in Fig. 7.1, the correlation between the power spectra is high when the channels function properly – illustrated below for an $M = 8$ sensor array:

$$\mathbf{\Gamma} = \begin{pmatrix} 1.000 & 0.970 & 0.963 & 0.964 & 0.964 & 0.962 & 0.953 & 0.943 \\ 0.970 & 1.000 & 0.992 & 0.986 & 0.982 & 0.969 & 0.942 & 0.914 \\ 0.963 & 0.992 & 1.000 & 0.997 & 0.994 & 0.983 & 0.960 & 0.929 \\ 0.964 & 0.986 & 0.997 & 1.000 & 0.999 & 0.992 & 0.972 & 0.943 \\ 0.964 & 0.982 & 0.994 & 0.999 & 1.000 & 0.996 & 0.979 & 0.950 \\ 0.962 & 0.969 & 0.983 & 0.992 & 0.996 & 1.000 & 0.989 & 0.964 \\ 0.953 & 0.942 & 0.960 & 0.972 & 0.979 & 0.989 & 1.000 & 0.983 \\ 0.943 & 0.914 & 0.929 & 0.943 & 0.950 & 0.964 & 0.983 & 1.000 \end{pmatrix}$$

Summing the $\Gamma_{mm'}$ values along the columns then yields values close to M:

$$\sum_{m'=1}^{8} \Gamma_{mm'} = \big(7.722,\ 7.758,\ 7.820,\ 7.856,\ 7.865,\ 7.858,\ 7.782,\ 7.628\big)^T.$$

However, when one or more channels are degraded, this correlation goes down as illustrated below for the case where sensor 1 is degraded.

$$|\mathbf{\Gamma}| = \begin{pmatrix} 1.000 & 0.026 & 0.029 & 0.033 & 0.032 & 0.033 & 0.026 & 0.012 \\ 0.026 & 1.000 & 0.992 & 0.984 & 0.981 & 0.975 & 0.962 & 0.946 \\ 0.029 & 0.992 & 1.000 & 0.997 & 0.995 & 0.988 & 0.973 & 0.957 \\ 0.033 & 0.984 & 0.997 & 1.000 & 0.999 & 0.994 & 0.981 & 0.963 \\ 0.032 & 0.981 & 0.995 & 0.999 & 1.000 & 0.997 & 0.985 & 0.966 \\ 0.033 & 0.975 & 0.988 & 0.994 & 0.997 & 1.000 & 0.992 & 0.972 \\ 0.026 & 0.962 & 0.973 & 0.981 & 0.985 & 0.992 & 1.000 & 0.983 \\ 0.012 & 0.946 & 0.957 & 0.963 & 0.966 & 0.972 & 0.983 & 1.000 \end{pmatrix}.$$

The column sum yields, in this case,

$$\sum_{m'=1}^{8} \Gamma_{mm'} = \big(0.805,\ 6.817,\ 6.873,\ 6.886,\ 6.893,\ 6.888,\ 6.852,\ 6.777\big)^T;$$

i.e, values significantly lower than M, with the first element having the lowest score. Using this observation, we arrive at our procedure, described in Fig. 7.2. Note that the approach is *iterative*: we discard the defective sensors in the descending order of their magnitude of degradation. Such a procedure allows us to set a uniform threshold for selection. If we choose to select *all* degraded sensors in a *single* step based on the column sum, we would not only be required to set a different threshold to accommodate the different combinations of degraded sensors, but also the number of hypotheses to be evaluated would increase combinatorially, leading to a rather unwieldy structure for the hypothesis tests.

Further, the limitiation of $M \geq 3$ in step 7 is required in order to have a majority vote. Obviously, given only two sensors with low correlation, it is difficult to decide which sensor is defective without any further information.

7.3.3 Determining the calibration factor $(A_m^{(m')})$

Once the degraded channels have been weeded out, the gain calibration can be done for the remaining elements of the array. For this, we first compute the net power at each sensor by

1. DegradedChannels = $\{\}$

2. $\forall$ channels m of M, do:
   Calculate the $\widehat{\Psi}_{X_m X_m}(k)$ (using (7.2)).
   End

3. Calculate $\mathbf{\Gamma}$ as in (7.3) & (7.4).

4. Let Υ_Γ := detection threshold.

5. Compute the average Γ_m for each sensor m with respect to the other sensors m': $\Gamma_m = \frac{1}{M}\sum_{m'=1}^{M}\Gamma_{mm'}$.

6. If ($\exists\ \Gamma_m < \Upsilon_\Gamma$) & ($M \geq 3$), do:
   Find $m_{\text{bad}} = \underset{m}{\text{argmin}}\ \Gamma_m$
   DegradedChannels = DegradedChannels $\cup\ m_{\text{bad}}$
   Remove the row and column for channel m_{bad} from $\mathbf{\Gamma}$
   $M \leftarrow M - 1$
   Go to Step 5.
   End

Figure 7.2: Algorithm to determine degraded channels using the SCORE approach.

summing the power spectrum:

$$\widehat{\Psi}_{X_m X_m} = \sum_{k=1}^{K} \widehat{\Psi}_{X_m X_m}(k) \quad \forall m, m \notin \text{DegradedChannels}. \tag{7.8}$$

Next, to set the reference gain, we select, without loss of generality, the sensor with the maximum power:

$$m_{\text{r}} = \underset{m}{\text{argmax}}\ \widehat{\Psi}_{X_m X_m}\,, m \notin \text{DegradedChannels}\,. \tag{7.9}$$

The calibration factors are then computed for the remaining channels with respect to this sensor as:

$$A_m^{(m_{\text{r}})} = \sqrt{\frac{\widehat{\Psi}_{X_{m_{\text{r}}} X_{m_{\text{r}}}}}{\widehat{\Psi}_{X_m X_m}(k)}}\,, \quad \forall m,\ m \notin \text{DegradedChannels}. \tag{7.10}$$

7.3.4 Additional considerations for the SCORE approach

The above development has assumed the existence of a stationary power floor. However, in practical situations, we need to consider the eventuality that the average signal power could change with time – necessitating some kind of moving average operator for the computation of the scaling factors and the detection of corrupt sensors. We also need to consider the eventuality that there may be signal segments which cannot be used (as discussed previously), leading to breaks in the averaging periods. Consequently, we parse the signal in blocks of length T seconds and compute the long-term average in the following manner: for any signal block, we first decide whether the block is to be selected for the averaging or not. If the

block does *not* contain any strongly directive sources, it is selected. Otherwise, it is dropped. Next, for each *selected* block o of length T seconds (corresponds to $B \approx (T f_\mathrm{s} - K)/O + 1$ frames[1]), the short-term temporal average is computed:

$$\widehat{\Psi}^{o}_{X_m X_m}(k) = \frac{1}{B} \sum_{b} \mathcal{E}_m(k, b)\,. \tag{7.11}$$

Lastly, the long-term average is computed on the selected blocks in a recursive manner – in order to conserve memory – as:

$$\widehat{\Psi}_{X_m X_m} \leftarrow \eta\, \widehat{\Psi}_{X_m X_m} + (1 - \eta)\widehat{\Psi}^{o}_{X_m X_m}(k)\ , \tag{7.12}$$

where η is the smoothing factor for the update. Equation (7.11) provides a robust estimate of the signal power *within* the block o, whereas (7.12) adapts the power spectrum to changes in the average signal power with time. As the detection of the anomalous sensors and the gain calibration is based on long-term averages, it suffices to compute the necessary parameters (g_m and $\mathbf{\Gamma}$) once every few *seconds*, reducing the computational load on the processing engine.

Note that the recursive averaging of the power spectrum also implicitly averages the estimate of the gain g_m. This has the beneficial effect of preventing any particular block from exerting undue influence on the gain estimates. For similar reasons, a recursive averaging of the gain estimates (albeit in a different manner) has also been proposed in [118].

7.3.5 Evaluation of SCORE

The proposed approach was tested on an extract from the brake squeal database. As a rigorous testing of all possible degradations is impossible, only the following representative scenarios are presented:

a. Complete failure of multiple sensors (such sensors record only noise)
b. One (or more) microphones are nonlinearly corrupted (clipping/rectification)
c. Multiple microphones pick up parasitic sinusoids (corresponding, e.g., to cross-talk with power supplies).

The system parameters used in the experiments are:

Table 7.1: System parameters for the calibration approaches.

DFT length (ms)	Window type/ length (ms)	Frame shift (ms)	f_s (kHz)	T (ms)	Υ_Γ	η
32	von Hann/32	8	32	200	0.7	0.5

The signal segment considered was 5 s long.

a. Complete sensor failure

In this evaluation, a degraded channel is represented by replacing the time domain signal by random uncorrelated white noise of unit variance. This would correspond to the case where a sensor is defective to the point where it does not respond to acoustic input.

[1] With a frame shift of O samples, sampling rate f_s and a K point DFT

Single channel corruption (channel 1) The obtained correlation matrix is:

$$\mathbf{\Gamma} = \begin{pmatrix} 1.000 & 0.141 & 0.153 & 0.152 & 0.150 & 0.139 & 0.128 & 0.119 \\ 0.141 & 1.000 & 0.990 & 0.982 & 0.978 & 0.972 & 0.963 & 0.950 \\ 0.153 & 0.990 & 1.000 & 0.996 & 0.994 & 0.987 & 0.970 & 0.947 \\ 0.152 & 0.982 & 0.996 & 1.000 & 0.999 & 0.994 & 0.976 & 0.948 \\ 0.150 & 0.978 & 0.994 & 0.999 & 1.000 & 0.997 & 0.982 & 0.951 \\ 0.139 & 0.972 & 0.987 & 0.994 & 0.997 & 1.000 & 0.990 & 0.957 \\ 0.128 & 0.963 & 0.970 & 0.976 & 0.982 & 0.990 & 1.000 & 0.970 \\ 0.119 & 0.950 & 0.947 & 0.948 & 0.951 & 0.957 & 0.970 & 1.000 \end{pmatrix}.$$

Note the low values for the correlation in the first row and column. Further, the average power at each sensor (computed using (7.8)) is as follows:

$$\widehat{\Psi}_{X_m X_m} = \left(1.966 \times 10^5,\ 1.155,\ 1.269,\ 1.359,\ 1.282,\ 1.321,\ 1.181,\ 1.161\right),$$

where, again, it may be seen that while channels 2–8 have approximately the same power, channel 1 differs in value. However, this alone is not *always* indicative of the improper functioning of a sensor.

Multiple channel degradation (channels 1,2 and 5) For the correlation matrix we have:

$$|\mathbf{\Gamma}| = \begin{pmatrix} 1.000 & 0.094 & 0.030 & 0.036 & 0.016 & 0.042 & 0.050 & 0.034 \\ 0.094 & 1.000 & 0.030 & 0.034 & 0.043 & 0.041 & 0.042 & 0.022 \\ 0.030 & 0.030 & 1.000 & 0.996 & 0.098 & 0.987 & 0.970 & 0.947 \\ 0.036 & 0.034 & 0.996 & 1.000 & 0.090 & 0.994 & 0.976 & 0.948 \\ 0.016 & 0.043 & 0.098 & 0.090 & 1.000 & 0.090 & 0.096 & 0.080 \\ 0.042 & 0.041 & 0.987 & 0.994 & 0.090 & 1.000 & 0.990 & 0.957 \\ 0.050 & 0.042 & 0.970 & 0.976 & 0.096 & 0.990 & 1.000 & 0.970 \\ 0.034 & 0.022 & 0.947 & 0.948 & 0.080 & 0.957 & 0.970 & 1.000 \end{pmatrix}.$$

Note the low correlation values for channels 1,2 and 5 clearly delineating them as corrupted channels.

b. Non-linear degradation

The non-linearity considered here is full wave rectification: $\breve{x}_m(n) = |x_m(n)|$, i.e., the absolute value of the input signal is taken. For this case, again channels 1,2 and 5 were considered to be degraded. The correlation matrix obtained is:

$$\mathbf{\Gamma} = \begin{pmatrix} 1.000 & 0.885 & 0.376 & 0.384 & 0.902 & 0.401 & 0.443 & 0.499 \\ 0.885 & 1.000 & 0.361 & 0.363 & 0.948 & 0.384 & 0.428 & 0.498 \\ 0.376 & 0.361 & 1.000 & 0.997 & 0.371 & 0.993 & 0.979 & 0.895 \\ 0.384 & 0.363 & 0.997 & 1.000 & 0.376 & 0.996 & 0.981 & 0.895 \\ 0.902 & 0.948 & 0.371 & 0.376 & 1.000 & 0.397 & 0.438 & 0.513 \\ 0.401 & 0.384 & 0.993 & 0.996 & 0.397 & 1.000 & 0.991 & 0.910 \\ 0.443 & 0.428 & 0.979 & 0.981 & 0.438 & 0.991 & 1.000 & 0.941 \\ 0.499 & 0.498 & 0.895 & 0.895 & 0.513 & 0.910 & 0.941 & 1.000 \end{pmatrix}.$$

In this case, too, we see that the power spectra of degraded channels show a lower correlation with the power spectra of healthy channels, and the iterative procedure described in Fig. 7.2 is able to correctly pick out the degraded microphones. The high correlations between pairs (1,2), (1,5) and (2,5) are to be expected, as they suffer from the same degradation.

c. Random parasitic sinusoids

For this test, a sinusoid of 100 Hz was added, at 0 dB SNR, to the signals of microphones considered to be degraded. We present here, only the case of a single harmonic as it is more difficult to detect than the case of multiple harmonics. Additionally, for the case of multiple sensor degradation, the parasitic sinusoid at each microphone is given a random phase offset.

Single channel (channel 1) corruption The correlation matrix is:

$$\mathbf{\Gamma} = \begin{pmatrix} 1.000 & 0.108 & 0.096 & 0.102 & 0.094 & 0.097 & 0.106 & 0.101 \\ 0.108 & 1.000 & 0.990 & 0.982 & 0.978 & 0.972 & 0.963 & 0.950 \\ 0.096 & 0.990 & 1.000 & 0.996 & 0.994 & 0.987 & 0.970 & 0.947 \\ 0.102 & 0.982 & 0.996 & 1.000 & 0.999 & 0.994 & 0.976 & 0.948 \\ 0.094 & 0.978 & 0.994 & 0.999 & 1.000 & 0.997 & 0.982 & 0.951 \\ 0.097 & 0.972 & 0.987 & 0.994 & 0.997 & 1.000 & 0.990 & 0.957 \\ 0.106 & 0.963 & 0.970 & 0.976 & 0.982 & 0.990 & 1.000 & 0.970 \\ 0.101 & 0.950 & 0.947 & 0.948 & 0.951 & 0.957 & 0.970 & 1.000 \end{pmatrix}.$$

Note again the low correlation between channel 1 and the rest.

Multiple channel degradation (channels 1,2 and 5)

$$\mathbf{\Gamma} = \begin{pmatrix} 1.000 & 0.999 & 0.096 & 0.102 & 0.999 & 0.097 & 0.106 & 0.101 \\ 0.999 & 1.000 & 0.085 & 0.091 & 0.999 & 0.083 & 0.091 & 0.086 \\ 0.096 & 0.085 & 1.000 & 0.996 & 0.080 & 0.987 & 0.970 & 0.947 \\ 0.102 & 0.091 & 0.996 & 1.000 & 0.086 & 0.994 & 0.976 & 0.948 \\ 0.999 & 0.999 & 0.080 & 0.086 & 1.000 & 0.080 & 0.087 & 0.081 \\ 0.097 & 0.083 & 0.987 & 0.994 & 0.080 & 1.000 & 0.990 & 0.957 \\ 0.106 & 0.091 & 0.970 & 0.976 & 0.087 & 0.990 & 1.000 & 0.970 \\ 0.101 & 0.086 & 0.947 & 0.948 & 0.081 & 0.957 & 0.970 & 1.000 \end{pmatrix}.$$

Note that the correlations between channel-pairs (1,2), (1,5), and (2,5) are high – which is not surprising since they suffer the same kind of degradation. However, their respective correlations with the other channels are low.

7.4 The minimum mean square (MMSE) approach

In this alternative approach for the detection of malfunctioning channels, we try to find the scalar factor that minimizes the difference between the power of the channels of an array with respect to a *reference* channel, also belonging to the same array. In other words, we scale the net power of a channel in order to bring it to the same level as the reference channel. The efficacy of the scaling is indicated by the magnitude of the *residual* power, obtained as the difference between the power of the reference channel and the (scaled) power of the particular channel under consideration.

7.4.1 Determining the scaling factor $(g_m^{(m')})$

Consider, for the sake of discussion, that we wish to calibrate the power of sensor 2 with respect to that of sensor 1. Thus, with respect to the system model of (7.1), we seek the factor $\widehat{g}_2$ such that:

$$\widehat{g}_2 = \underset{g_2}{\operatorname{argmin}}\ \mathcal{J}(g_2)\,, \tag{7.13}$$

with

$$\mathcal{J}(g_2) = \sum_{k=1}^{K} \left(\mathcal{E}_1(k,b) - g_2\mathcal{E}_2(k,b)\right)^2\,, \tag{7.14}$$

where $\widehat{g}_2$ represents the optimal estimate of g_2. Differentiating (7.14) and equating it to zero, we obtain:

$$\frac{d\mathcal{J}}{dg_2} = -2\sum_{k=1}^{K} \left(\mathcal{E}_1(k,b) - g_2\mathcal{E}_2(k,b)\right)\mathcal{E}_2(k,b) \stackrel{!}{=} 0, \tag{7.15}$$

whence,

$$\widehat{g}_2 = \frac{\displaystyle\sum_{k=1}^{K} \left(\mathcal{E}_1(k,b)\mathcal{E}_2(k,b)\right)}{\displaystyle\sum_{k=1}^{K} \mathcal{E}_2^2(k,b)}. \tag{7.16}$$

This may be seen as a kind of *Wiener* gain, which usually results when using a mean square error approach.

Similar factors g_m may be computed for the remaining channel pairs $(1,m),\ m = 3,\ldots,M$, giving us an M dimensional vector $\mathbf{g}$ of compensation parameters with channel 1 as the reference. We can generalize this, taking each channel iteratively as reference, obtaining an $M \times M$ matrix:

$$\mathbf{G} = \left[\mathbf{g}^{(1)}\ \mathbf{g}^{(2)},\ \ldots,\ \mathbf{g}^{(M)}\right], \tag{7.17}$$

where the superscript indicates the particular channel taken as the reference. Note that the diagonal elements of $\mathbf{G}$ are 1 and the non-diagonal elements $g_m^{(m')}$ indicate the compensation factor for channel m when channel m' is the reference[2].

7.4.2 Determining the residual power $(\mathfrak{E})$

In general, the compensation factors alone are not sufficient to detect the presence of sensor anomalies. A more significant indicator is the magnitude of the normalized *residual* power $(\mathfrak{E})$, obtained by substituting the optimal g from (7.16) in (7.14):

$$\mathfrak{E}_m^{(m')} = \frac{\displaystyle\sum_{k=1}^{K} \left(\widehat{\Psi}_{X_{m'}X_{m'}}(k) - \widehat{g}_m^{(m')}\widehat{\Psi}_{X_mX_m}(k)\right)^2}{\displaystyle\sum_{k=1}^{K} \widehat{\Psi}^2_{X_{m'}X_{m'}}(k)}. \tag{7.18}$$

[2]The scaling factor $g_m^{(m')}$ is related to the calibration factor $A_m^{(m')}$ as $A_m^{(m')} = \sqrt{g_m^{(m')}}$

The normalization in (7.18) is required in order to measure the effectiveness of the calibration independently of any microphone scaling. Obviously, if the normalized residual power is high, the corresponding sensor pair could not be properly calibrated, indicating a possible anomaly in *either* sensor. On the other hand, if the residual power is low, the channels have been well-matched in gain.

A rather interesting result is obtained if we expand (7.18) by substituting the value of the optimal g from (7.16):

$$\mathfrak{E}_m^{(m')} = \frac{1}{\sum_k \widehat{\Psi}^2_{X_{m'}X_{m'}}(k)} \sum_k \left(\widehat{\Psi}_{X_{m'}X_{m'}}(k) - \frac{\sum_k \widehat{\Psi}_{X_{m'}X_{m'}}(k)\, \widehat{\Psi}_{X_mX_m}(k)}{\sum_k \widehat{\Psi}^2_{X_mX_m}(k)}\, \widehat{\Psi}_{X_mX_m}(k) \right)^2 \tag{7.19}$$

$$= 1 - \frac{\left(\sum_k \widehat{\Psi}_{X_mX_m}(k)\, \widehat{\Psi}_{X_{m'}X_{m'}}(k)\right)^2}{\left(\sum_k \widehat{\Psi}^2_{X_{m'}X_{m'}}(k)\right)\left(\sum_k \widehat{\Psi}^2_{X_mX_m}(k)\right)}. \tag{7.20}$$

It may be seen that the second term in (7.20) is simply the square of the non-centered version of (7.4), indicating the overall match (over frequency) between the power spectra of the microphones. Note, however, that such a non-centralized correlation measure is sensitive to the presence of self-induced noise, which usually has a flat power spectrum. Therefore, we modify (7.16) according to:

$$\widehat{\mu}_m = \frac{1}{K} \sum_{k=1}^{K} \widehat{\Psi}_{X_mX_m} \tag{7.21}$$

$$\widehat{g}_m^{(m')} = \frac{\sum_{k=1}^{K} (\widehat{\Psi}_{X_{m'}X_{m'}}(k) - \widehat{\mu}_{m'})(\widehat{\Psi}_{X_mX_m}(k) - \widehat{\mu}_m)}{\sum_{k=1}^{K} (\widehat{\Psi}_{X_mX_m}(k) - \widehat{\mu}_m)^2}. \tag{7.22}$$

7.4.3 Detecting degradations using $g_m^{(m')}$ and $\mathfrak{E}_m^{(m')}$

Calibrating a fully functional array results in a low value of $\mathfrak{E}_m^{(m')}$ for all pairs (m, m'). Gain mismatch, if any, would be indicated by the respective divergence of the $g_m^{(m')}$ from 1. This is evident from the cluster plot of Fig. 7.3 for a 'healthy' 8 channel array.

On the other hand, the presence of sensor anomalies would increase the residual power. This is indicated in Fig. 7.4, where microphone 1 has failed (delivers only noise). Note, in this case, that the corresponding value of g is also significantly different from the value of g for the other microphones.

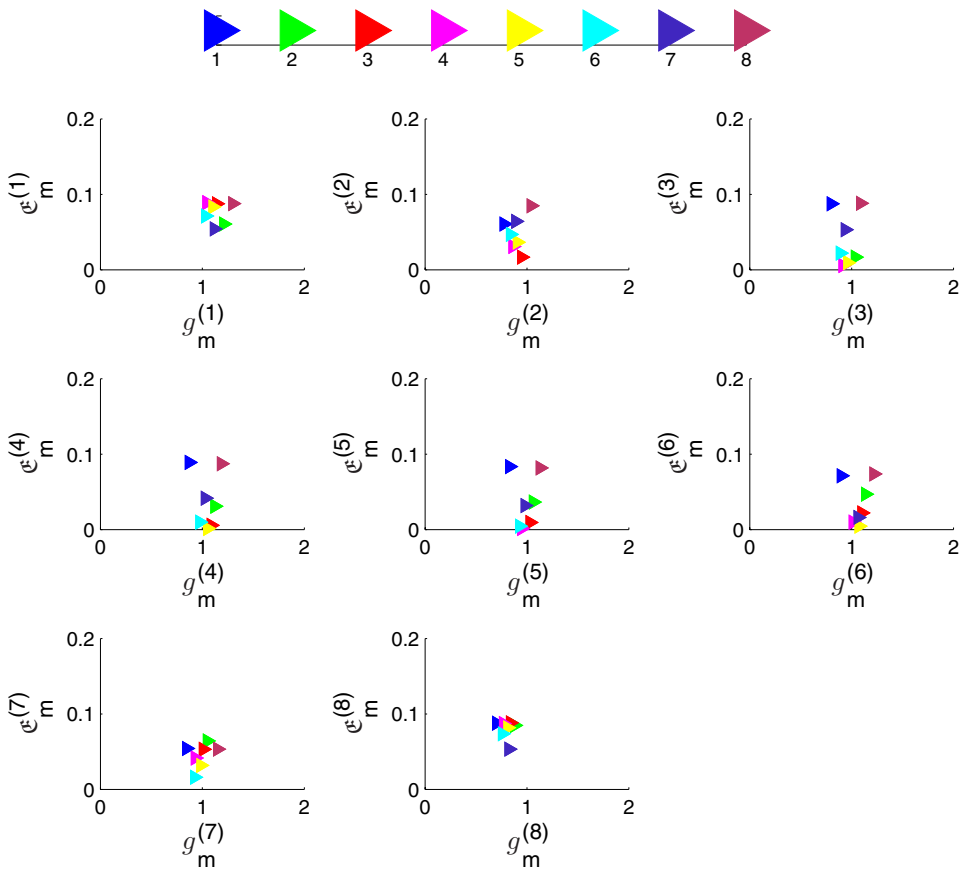


Figure 7.3: Cluster plot of g and $\mathfrak{E}$ for a 'healthy' array. The color-coding is used to differentiate the channels.

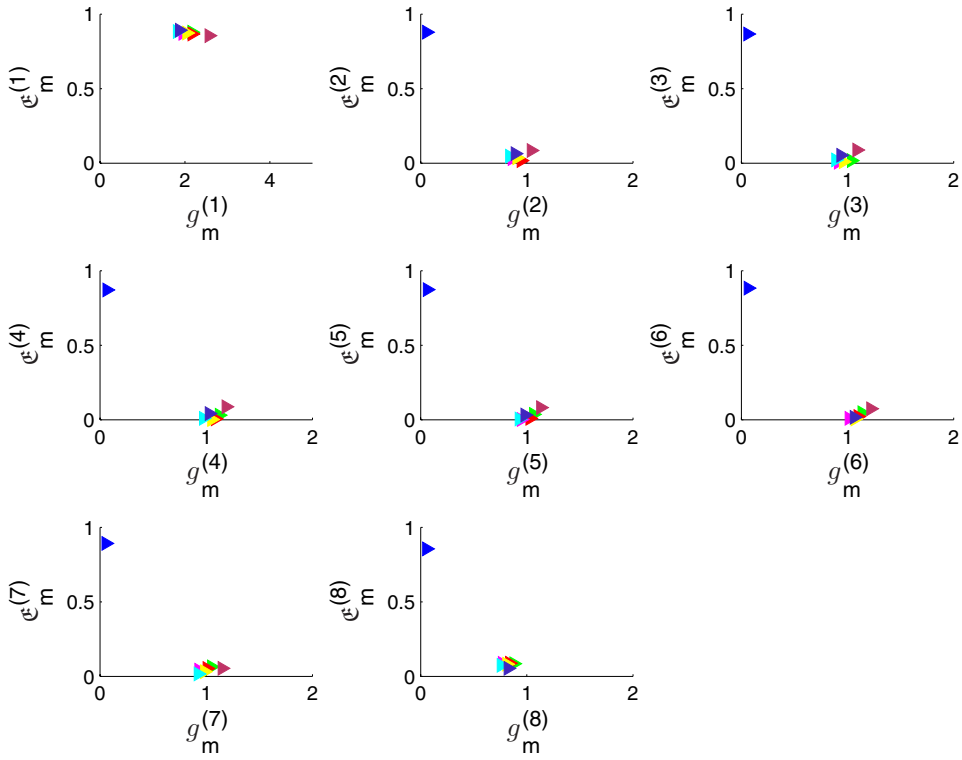


Figure 7.4: Cluster plot of g and $\mathfrak{E}$. Channel 1 has failed. Note the high values for $\mathfrak{E}_1^{(m')}$ and $\mathfrak{E}_m^{(1)}$, as compared to the residual power for the other microphones.

Thus, the residual power is a measure of the degradation suffered by a sensor.

This observation is used for detecting malfunctioning sensors using the $M \times M$ residual power matrix defined as:

$$\mathfrak{E} = \left(\mathfrak{E}^{(1)}, \ldots, \mathfrak{E}^{(M)}\right), \tag{7.23}$$

where $\mathfrak{E}^{(m)} = (\mathfrak{E}_1^{(m)}, \ldots, \mathfrak{E}_M^{(m)})^T$. The algorithm is similar to SCORE and is described in Fig. 7.5. The limitation of $M \geq 3$ in step 6 is required here too, and for the same reasons as in the SCORE approach. The detection threshold may be obtained in a manner similar to that in Section 7.3.

1. DegradedChannels = $\{\}$
2. $\forall$ channels m of M, do:
   Calculate $\mathfrak{E}^{(m)} = \left(\mathfrak{E}_1^{(m)}, \ldots, \mathfrak{E}_M^{(m)}\right)^T$ from (7.18).
   End
3. Calculate $\mathfrak{E}$ from the $\mathfrak{E}^{(m)}$ using (7.23).
4. Let $\Upsilon_{\mathfrak{E}}$:= allowed degradation in the channels.
5. Calculate the average residual error obtained for a sensor m when it is matched to all other sensors m':
   $\mathfrak{e} = \frac{1}{M} \sum_{m'=1}^{M} \mathfrak{E}_m^{(m')}$.
6. If ($\exists$ $\mathfrak{e}_m > \Upsilon_{\mathfrak{E}}$) and ($M \geq 3$), do:
   Find $m_{\text{bad}} = \underset{m}{\text{argmax}}\ \mathfrak{e}_m$
   DegradedChannels = DegradedChannels $\cup$ m_{bad}
   Remove the row and column for channel m_{bad} from $\mathfrak{E}$
   $M \leftarrow M - 1$
   Go to Step 5.
   End

Figure 7.5: Detecting the degraded channels using the residual power matrix.

7.4.4 Evaluation of the MMSE approach

The test setup is the same as the one used for the SCORE approach. Again, only the results for the representative degradation scenarios are presented. Further, only the case of multiple, simultaneously degraded sensors is considered. The parameters are the same as that used in the SCORE approach in Table 7.1.

a. Complete sensor failure

For the purpose of the tests, microphones 1, 2 and 5 are considered to be defective. The residual power matrix $\mathfrak{E}$ for this case is

$$\mathfrak{E} = \begin{pmatrix} 0.0000 & 0.9252 & 1.3232 & 1.3397 & 0.9192 & 1.3560 & 1.4070 & 1.8050 \\ 0.9186 & 0.0000 & 1.0238 & 1.0326 & 0.9525 & 1.0099 & 1.0108 & 0.9863 \\ 1.0066 & 1.0007 & 0.0000 & 0.0026 & 0.9864 & 0.0076 & 0.0256 & 0.1288 \\ 1.0073 & 1.0009 & 0.0026 & 0.0000 & 0.9856 & 0.0045 & 0.0208 & 0.1232 \\ 0.9241 & 0.9591 & 1.0120 & 0.9798 & 0.0000 & 0.9758 & 0.9201 & 0.8413 \\ 1.0073 & 1.0003 & 0.0076 & 0.0045 & 0.9868 & 0.0000 & 0.0094 & 0.1014 \\ 1.0088 & 1.0003 & 0.0256 & 0.0207 & 0.9858 & 0.0094 & 0.0000 & 0.0606 \\ 1.0163 & 0.9995 & 0.1285 & 0.1226 & 0.9856 & 0.1012 & 0.0603 & 0.0000 \end{pmatrix}$$

Note the high residual power when either channel 1, 2 or 5 is taken as the reference. Our detection algorithm is able to successfully demarcate these channels as degraded, for an appropriately set $\Upsilon_{\mathfrak{E}}$.

b. Non-linear degradation

The non-linearity considered here is full wave rectification: $\breve{x}_m(n) = |x_m(n)|$, i.e., the absolute value of the input signal is taken. For this case, again channels 1, 2 and 5 were considered to be degraded. The residual power matrix for this case is:

$$\mathfrak{E} = \begin{pmatrix} 0.0000 & 0.3829 & 1.0008 & 1.0009 & 0.1341 & 1.0011 & 1.0006 & 1.0009 \\ 0.3814 & 0.0000 & 1.0018 & 1.0019 & 0.4091 & 1.0024 & 1.0014 & 1.0020 \\ 1.0008 & 1.0015 & 0.0000 & 0.0399 & 1.0007 & 0.0254 & 0.1328 & 0.0947 \\ 1.0008 & 1.0016 & 0.0399 & 0.0000 & 1.0008 & 0.0395 & 0.0592 & 0.0177 \\ 0.1341 & 0.4108 & 1.0008 & 1.0008 & 0.0000 & 1.0010 & 1.0006 & 1.0009 \\ 1.0010 & 1.0020 & 0.0254 & 0.0395 & 1.0010 & 0.0000 & 0.1701 & 0.0842 \\ 1.0006 & 1.0011 & 0.1328 & 0.0592 & 1.0005 & 0.1701 & 0.0000 & 0.0368 \\ 1.0008 & 1.0017 & 0.0947 & 0.0177 & 1.0008 & 0.0842 & 0.0368 & 0.0000 \end{pmatrix}$$

Applying the detection algorithm of Fig. 7.5, we correctly localize microphones 1, 2 and 5 to be faulty. Note that, with respect to the residual power, these microphones seem to be well matched *between* themselves (relatively low values of the residual power for the elements $(1,2)$, $(1,5)$ and $(2,5)$ and their conjugate positions). This is not surprising as they suffer from the same kind of degradation. However, note the high values of the residual power when these microphones are matched to the other microphones.

c. Random parasitic sinusoids

For this purpose, a sinusoid of 100 Hz was added to the signals of microphone 1,2 and 5, at 0 dB SNR. We present, again, only the case of a single harmonic. The parasitic sinusoid at each microphone has a random phase offset. The residual power matrix is as shown below. Here, too, the degraded channels are evident.

$$\mathfrak{E} = \begin{pmatrix} 0.0000 & 0.0038 & 0.9936 & 0.9937 & 0.0052 & 0.9937 & 0.9938 & 0.9942 \\ 0.0038 & 0.0000 & 0.9804 & 0.9807 & 0.0003 & 0.9807 & 0.9810 & 0.9826 \\ 0.9934 & 0.9800 & 0.0000 & 0.0026 & 0.9775 & 0.0076 & 0.0256 & 0.1288 \\ 0.9935 & 0.9803 & 0.0026 & 0.0000 & 0.9775 & 0.0045 & 0.0208 & 0.1232 \\ 0.0052 & 0.0003 & 0.9779 & 0.9780 & 0.0000 & 0.9779 & 0.9782 & 0.9799 \\ 0.9935 & 0.9803 & 0.0076 & 0.0045 & 0.9774 & 0.0000 & 0.0094 & 0.1014 \\ 0.9935 & 0.9804 & 0.0256 & 0.0207 & 0.9775 & 0.0094 & 0.0000 & 0.0606 \\ 0.9938 & 0.9817 & 0.1285 & 0.1226 & 0.9790 & 0.1012 & 0.0603 & 0.0000 \end{pmatrix}$$

Note, again, the low residual power for the elements $(1,2)$, $(1,5)$, $(2,5)$ and their conjugates. As in the previous section, this is because the degradation suffered by the microphones is similar.

7.5 Conclusions

Current algorithms for online self-calibration of microphone arrays do not consider the eventuality of sensor degradation. Accordingly, in this chapter, we have introduced two simple online approaches for the *in situ* detection of corrupt sensors. The merits of the approaches were tested under three possible sensor degradation scenarios: complete sensor failure (only noise at the degraded sensor), full-wave rectification of the sensor signals (non-linear distortion) and parasitic harmonics in the channels. The approaches perform well for all these cases of sensor degradation. The proposed methods are computationally inexpensive: the computation of the averages and the detection of degradation is done only once every few *seconds*.

In addition, the methods can also be used for the gain calibration of the healthy microphones of an array. The computed gain compensation factors $A_m^{(m')}$ are similar to that of [52]. In contrast to other state of the art calibration algorithms [88, 22], these algorithms do not require phase equalization or time-delay compensation prior to calibration.

The MMSE-based approach is further flexible in that the gain compensation factors may be determined *separately* for each frequency *bin*. Clustering the obtained $g_m^{(m')}(k)$ over the frequency bins provides us with an additional feature for detecting sensor malfunctions: the cluster radius. The cluster radii for healthy sensors would be small. In the presence of degradations, however, the radii would grow. Despite the additional cost involved, we believe that a combination of the two features – the residual power and the cluster radii – would yield more robust results for a wider range of scenarios and an examination of this alternative implementation is worthwhile.

But engineering is not about finding perfect solutions. It is about doing the best you can with limited resources.

A concise summary
– Randy Pausch, 1960–2008

Chapter 8

Conclusions

This work was concerned with the development of algorithms for acoustic source localization and enhancement using sensor arrays. It was underpinned throughout by localization both as an end in itself and as a means to an end.

We began with an overview of contemporary source localization algorithms and introduced a taxonomy for their classification. Furthermore we showed the close link that exists between these algorithms and consistently highlighted the fact that the statistic necessary for all the presented approaches is of the second order. What distinguishes the algorithms is the way in which this statistic is utilized – this depends upon the application under consideration. We stress that, rather than try and find a single algorithm that may be used to localize sources *independently* of the available *a priori* knowledge and application, it is better to design algorithms specifically tailored to the application at hand. We have attempted to drive this point home with the help of two cases that are drastically different in the available *a priori* knowledge and the design constraints. We see here that despite the different nature of the algorithms, the statistic used is the same – namely, the multi-channel cross-correlation. However, generic algorithms that do not take cognizance of the application-specific properties and constraints are either useless (e.g., no off-the-shelf localization algorithm would work in the brake-squeal localization case) or do not perform as well (compare the performance of the generic M-SRP against the MoG model that utilizes the underlying source properties).

The next section dealt with the application of source localization to the problem of active speaker separation and enhancement. We first presented an overview of contemporary separation approaches and their taxonomy in Chapter 5 and followed this up with an examination of sample algorithms from each category in Chapter 6, showcasing the need for localization in practical realizations. For the sample algorithms, we began with a simple 2×2 case and showed how the HOS-based algorithms benefit from the localization information in practical realizations. Next, we extended the MoG based localization approach discussed in Chapter 4 to a flexible framework for source separation. The flexibility arises from the fact that, beginning with this framework, we are able to realize a wide class of algorithms for target speaker enhancement.

The separation algorithms developed include binary- and soft-mask based approaches and a linear algorithm based on the generalized sidelobe canceller (GSC) structure, which we termed the parsimonious excitation based GSC (PEG). In each case, the estimate of the source location is essential, either as the classification feature for the mask based approaches or for designing the blocking matrix and the fixed beamformer for PEG. We further demonstrated that PEG is robust against target signal cancellation in the presence of sensor gain mismatch – a very desirable property.

Lastly, we dealt with the issue of array health. The localization and separation algorithms presented in the previous chapters implicitly assume that the array functions perfectly. However, this assumption needs to be periodically tested, especially when the arrays are deployed in hostile environments. For this purpose we presented two closely related approaches based on the assumption of equal average received signal power at all the sensors of an array.

The theory and algorithms presented in this work pose many new questions which could serve as starting points for further investigations. For example, the blocking system based detection approach for brake squeal localization could be extended to a more general case for sources with *unknown* positions. This model might lend itself to the problem of localizing *correlated* sources, where the spatial blocking may permit each source to be detected *independently* of the others.

Another interesting point of research is the adaptation of the MoG based localization system to planar or 3-D array geometries, necessitating a *vector* clustering. It would also be interesting to consider other features (such as amplitude differences) for the clustering, as it might lead to more robust results.

Central to the source separation algorithms developed in this work is the MoG based generation of masks for the required target source. The further usage of these masks forms a starting point for further research, such as integration of the source location as a prior in the ICA algorithms, combining the HOS and mask based approaches, and so on.

We aim to address some of these issues in the near future.

IV

Appendices

Appendix A

Closed-form Solution for R

For a linear array and a blocking region demarcated in terms of the azimuth as $\mathcal{R}_\theta = [\theta_1, \theta_2]$, and under the assumption of well calibrated arrays, the propagation vector from the sources in this region to the array is defined as in Chapter 2 and we recall this here:

$$\mathbf{A}(\theta, k) = (e^{\jmath\, d_1 \Omega_k \cos(\theta)/c}, \ldots, e^{\jmath\, d_M \Omega_k \cos(\theta)/c})^T, \tag{A.1}$$

with d_m being measured with respect to the center of mass of the array. The covariance matrix $\mathbf{R}(k)$ for $\mathcal{R}_\theta$ is then:

$$\mathcal{R}(k) = \int_{\theta_1}^{\theta_2} \mathbf{A}(\theta, k)\mathbf{A}^H(\theta, k)\, \mathrm{d}\,\theta\,, \tag{A.2}$$

which may be evaluated numerically since a closed form solution to this is difficult. If we however make the simplifying assumption that the sources lie on the surface of a sphere, with the array at the center, and assume the blocking region to span the surface of the sphere between the specified angles, we may obtain a closed form solution (see e.g., [39]) as:

$$\begin{aligned} \mathcal{R}(k) &= \int_0^{2\pi} \int_{\theta_1}^{\theta_2} \mathbf{A}(\theta, \varpi, k)\mathbf{A}^H(\theta, \varpi, k) \sin(\theta)\, \mathrm{d}\,\theta\, \mathrm{d}\,\varpi \\ &= 2\pi \int_{\theta_1}^{\theta_2} \mathbf{A}(\theta, k)\mathbf{A}^H(\theta, k) \sin(\theta)\, \mathrm{d}\,\theta\,, \end{aligned} \tag{A.3}$$

with elements:

$$R_{m,m'} = \begin{cases} 2\pi(\cos(\theta_1) - \cos(\theta_2)), & m = m' \\ 2\pi \dfrac{e^{\jmath\, \Omega_k d_{mm'} \cos(\theta_1)/c} - e^{\jmath\, \Omega_k d_{mm'} \cos(\theta_2)/c}}{\jmath\, \Omega_k d_{mm'}/c}, & m \neq m'\,, \end{cases} \tag{A.4}$$

where $d_{mm'} = d_m - d_{m'}$.

Appendix B

The FastICA Update Rule

Central to the idea of FastICA is the theory of *iterated* functions and their *attractive fixed-points*. Specifically, a function that is composed with itself *ad infinitum* by iteration is known as an iterated function [132]. If $\mathcal{X}$ be the set on which a function $\mathcal{f}\colon \mathcal{X} \to \mathcal{X}$ is defined, the i th iterate $\mathcal{f}^{(i)}$ of the function is defined as:

$$\mathcal{f}^{(i)} = \mathcal{f} \circ \mathcal{f}^{(i-1)}\,, \tag{B.1}$$

where $(\mathcal{f} \circ \mathcal{g})(x) = \mathcal{f}(\mathcal{g}(x))$. Such an iterated function is said to have an *attractive fixed-point* if there exists an $x \in \mathcal{X}$ such that:

$$\mathcal{f}(x) = x\,. \tag{B.2}$$

In numerical analysis, fixed-point iteration [131] is a method of computing such fixed-points of iterated functions. In essence, given an initial starting point $x^{(0)}$ in $\mathcal{X}$, the fixed-point iteration is:

$$x^{(i+1)} = \mathcal{f}(x^{(i)})\,. \tag{B.3}$$

If $\mathcal{f}$ satisfies the conditions set by the Banach fixed-point theorem [129], then (B.3) is guaranteed to converge to a fixed-point.

Consider now the simple case of estimating a single independent component after the preprocessing stage. This requires the minimization/maximization of an appropriately chosen score function $\mathcal{J}(\mathbf{w})$, under the constraint $\|\mathbf{w}\|^2 = 1$, where $\|\cdot\|$ represents, as before, the Euclidean norm. Formulating this optimization using Lagrange multipliers, we have:

$$\mathbf{w}_{\text{opt}} = \underset{\mathbf{w}}{\text{argmin}(\text{argmax})}\, \mathcal{J}(\mathbf{w}) + \beta\big(\mathbf{w}^H\mathbf{w} - 1\big)\,. \tag{B.4}$$

At the optimal solution we must have, therefore,

$$\frac{\partial \mathcal{J}(\mathbf{w}_{\text{opt}})}{\partial \mathbf{w}^*_{\text{opt}}} + \beta\mathbf{w}_{\text{opt}} = 0\,, \tag{B.5}$$

or,

$$\frac{\partial \mathcal{J}(\mathbf{w}_{\text{opt}})}{\partial \mathbf{w}^*_{\text{opt}}} \triangleq -\beta\mathbf{w}_{\text{opt}} \tag{B.6}$$

where the complex gradient operator $\partial/\partial\mathbf{w}^*$ is used in the sense defined in [17].

From (B.6) we see that at an optimum solution the gradient $\partial\mathcal{J}(\mathbf{w})/\partial\mathbf{w}^*$ must point in the direction of $\mathbf{w}$. The scalar β is irrelevant given the subsequent normalization of $\mathbf{w}$ to unit norm. Thus we see that the optimal solution $\mathbf{w}_{\text{opt}}$ may be obtained as the fixed-point of $\partial\mathcal{J}(\mathbf{w})/\partial\mathbf{w}^*$, leading to the update rule:

$$\begin{aligned} \mathbf{w}^+ &= \left.\frac{\partial\mathcal{J}(\mathbf{w})}{\partial\mathbf{w}^*}\right|_{\mathbf{w}=\mathbf{w}^{(i)}} \\ \mathbf{w}^{(i+1)} &\leftarrow \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|}\,. \end{aligned} \tag{B.7}$$

We may apply the update in (B.7) iteratively to extract each component, taking care to ensure that the resulting demixing matrix $\mathbf{W}$ remains orthonormal after the updates. This essentially implies the constraints

$$\mathbf{w}_q^H\mathbf{w}_{q'} = \begin{cases} 1 & q = q'\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{B.8}$$

Alternatively we may perform the fixed-point update with respect to $\mathbf{W}$ directly as:

$$\begin{aligned} \mathbf{W}^+ &= \nabla_{\mathbf{W}}\mathcal{J}(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}^{(i)}} \\ \mathbf{W}^{(i+1)} &\leftarrow \left(\mathbf{W}^+\mathbf{W}^{+,H}\right)^{-\frac{1}{2}}\mathbf{W}^+\,, \end{aligned} \tag{B.9}$$

where $\nabla_{\mathbf{W}}$ represents the complex matrix gradient operator with respect to $\mathbf{W} = [\mathbf{w}_1\,,\ldots\,,\mathbf{w}_Q]^H$. The update of (6.4) is of this kind.

For a more detailed and rigorous justification of the FASTICA update rule, and especially the derivation for cost functions obtained using non-polynomial approximations to entropy, the interested reader is referred to [56, 57].

Appendix C

Detection of Brake Squeal Presence prior to Localization

The brake-squeal events are narrowband, with most of the spectral energy being concentrated in the squeal frequencies[1], giving rise to a rather well-structured spectrum in the presence of brake-squeal. Thus, modified information theoretic tools (like entropy) may be used to form an *a priori* decision on the presence of a squeal event. The prerequisites for any such approach are that it should be computationally not too demanding and yet reasonably accurate.

The entropy for a discrete source with an alphabet $\mathcal{A}$ and an associated probability distribution function P is defined as [109]:

$$\mathfrak{H}(\mathcal{A}) = -\sum_{\forall a \in \mathcal{A}} P_a \log_2(P_a). \tag{C.1}$$

It characterizes the amount of disorder in the system: the entropy is maximum if all the symbols in the alphabet are equally probable and reduces as the probability distribution becomes more 'peaked'.

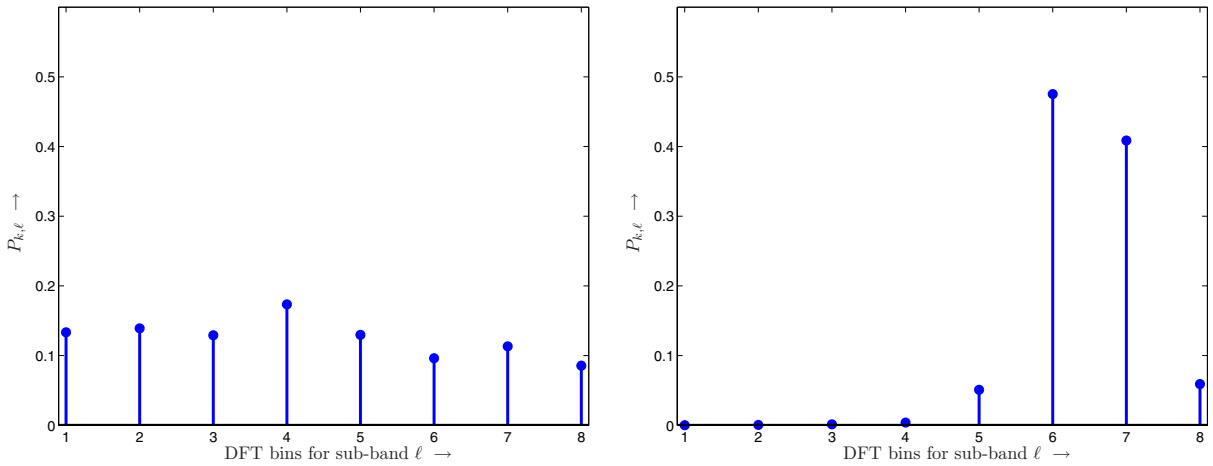
Now, consider small sub-bands of frequencies in the discrete Fourier spectrum of a signal. Let the length of each sub-band be K'. If a sub-band contains a harmonic signal, the power spectrum would be peaked at this frequency. Alternatively, if the sub-band contained only environmental noise, the distribution of power would be more or less equal along all the discrete frequency bins of that band. Similar to [98], let us define a pseudo probability distribution function along the bins k of a sub-band ℓ as:

$$P_{k,\ell} = \frac{|X^{(\ell)}(k)|^2}{\sum_{k=1}^{K'} |X^{(\ell)}(k)|^2}, \tag{C.2}$$

where $X^{(\ell)}(k) = X(k + (\ell - 1)K')$ represents the DFT coefficient of bin k in sub-band ℓ. Such a distribution function is shown in Figure C.1.

It may be seen from Figure C.1 that the distribution defined in (C.2) is a faithful representation of the underlying spectral structure. This 'entropy' can then be calculated using (C.1) and (C.2) and used to predict squeal presence for a fixed detection threshold $\Upsilon_{\text{a priori}}$, as shown below.

[1] The lower frequencies containing motor noise are neglected in our considerations.



(a) Sub-band containing environmental noise only

(b) Sub-band containing squeal

Figure C.1: Probability distribution functions for cases where only noise is present (Figure C.1(a)) and where the sub-band contains an active frequency (Figure C.1(b)). Note the 'peakiness' of the second plot with respect to the almost flat curve of the first.

```
∀ frames b of the STFT, Do
   Divide it into non-overlapping sub-bands ℓ of length K′.
   ∀ℓ Do
      Calculate the entropy, 𝔊(ℓ, b), of that sub-band.
   End
   Find the minimal entropy, min 𝔊(ℓ, b), for the frame b.
   Is min 𝔊(ℓ, b) ≤ ϒ_a priori?
      If Yes :  Squeal detected in frame b.
      If No :  No squeal present in frame b.
   End
End
```

Note that setting the detection threshold $\Upsilon_{\text{a priori}}$ involves a trade-off between sensitivity to events at lower power and the generation of too many false alarms, and depends upon the requirements of the application. If this approach detects a squeal in a signal segment, that segment is analyzed in more detail and the squeal frequencies are extracted using more sophisticated algorithms.

Appendix D

Test Setup and Hardware for the Detection and Localization of Brake Squeal

The data used for the testing and development of the brake-squeal localization algorithms were recorded at the VW research & development testing grounds at Wolfsburg, Germany. The 4-array system was mounted on the base of a VW Touran$^{\text{tm}}$. For the signal pre-amplification and A/D conversion in the first-generation hardware prototype, we used four 8-channel, custom designed pre-amplifiers and two 16-channel Creamware$^{\text{tm}}$ A/D converters. The digitized data from the A/D converter was transmitted via ADAT optical cables to a combination of a 24-channel RME-Digiface$^{\text{tm}}$ and an 8-channel RME Multiface$^{\text{tm}}$ connected to a laptop. The recording software used was Nuendo$^{\text{tm}}$. The arrays were mounted as shown in Figure D.1, with the automobile dimensions included for good measure.

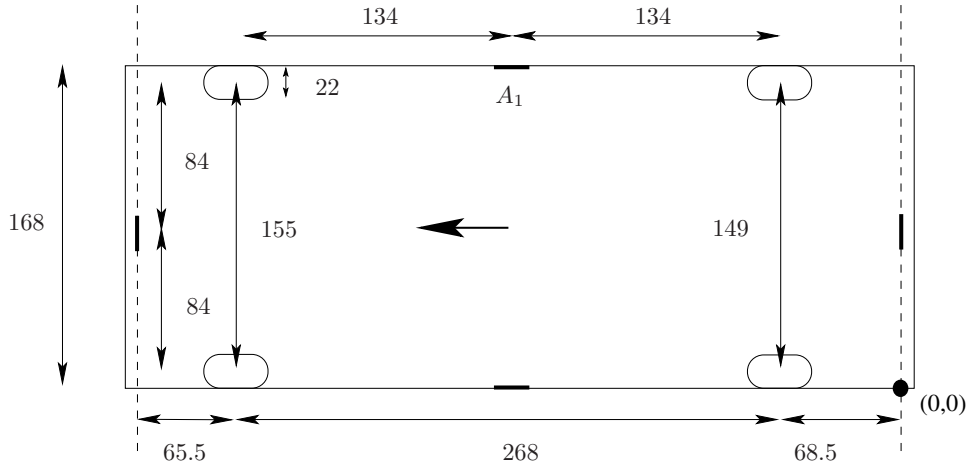
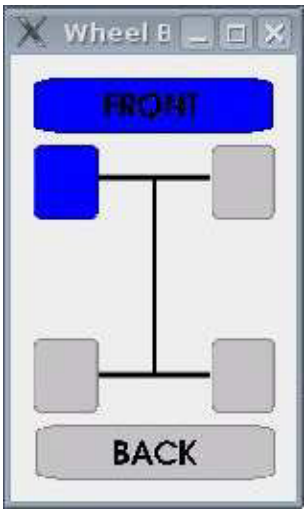
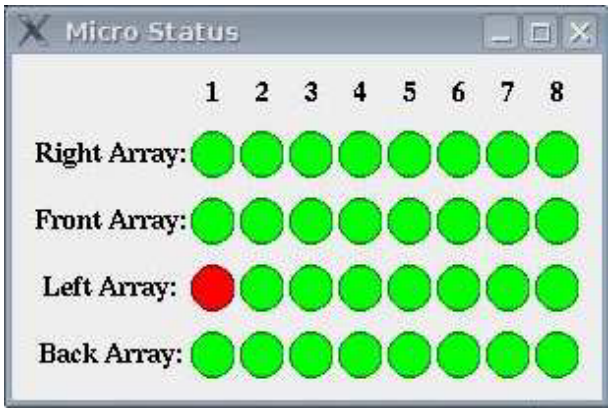


Figure D.1: Array assembly details for brake squeal localization. All dimensions are in cm. For orientation purposes, array 1 is labelled. The arrays are configured as detailed in Chapter 4. The bold-face arrow indicates the direction of forward motion. The origin of the co-ordinate system, used for computing the search regions in the hierarchical approach, is also depicted.

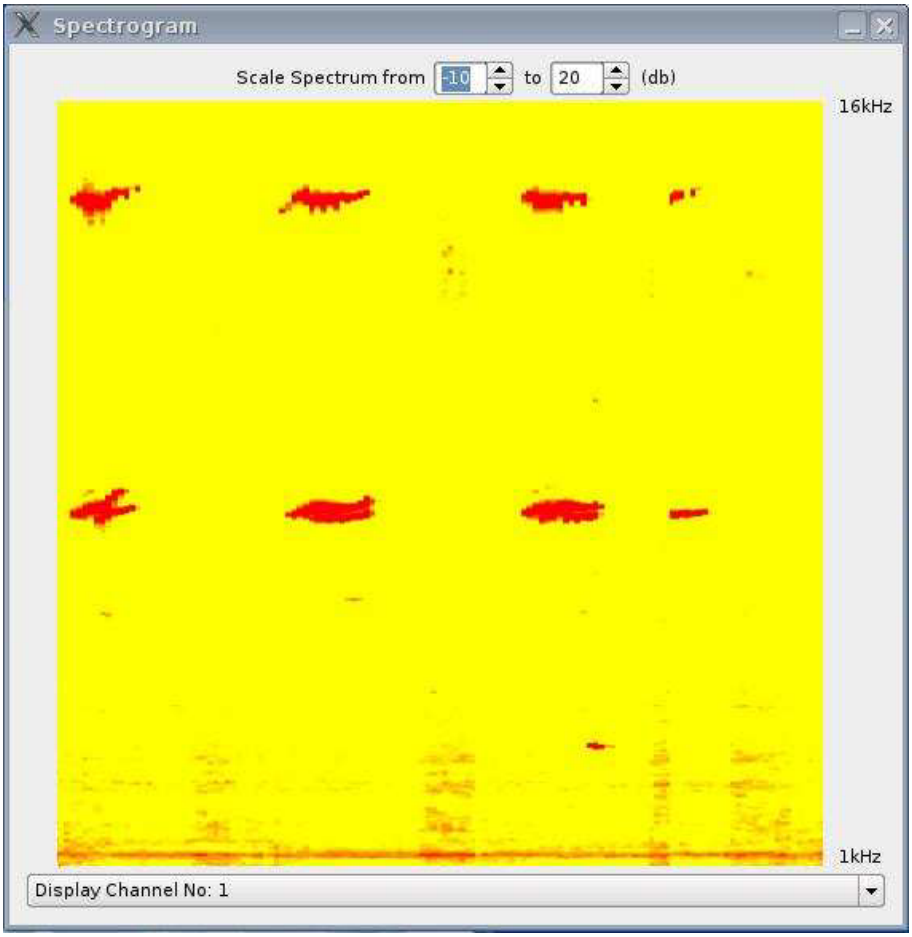
A snapshot of the online system is illustrated in Figure D.2.



(a) Localization result



(b) Microphone status display



(c) Running spectral analysis

Figure D.2: Snapshot of the online system developed for brake squeal localization. The localization results demonstrated in the current figure are for the hierarchical approach. Note that during this measurement microphone 1 of array 3 (left array) was indeed defective and was correctly identified as such by the detection approaches of Chapter 7

Appendix E

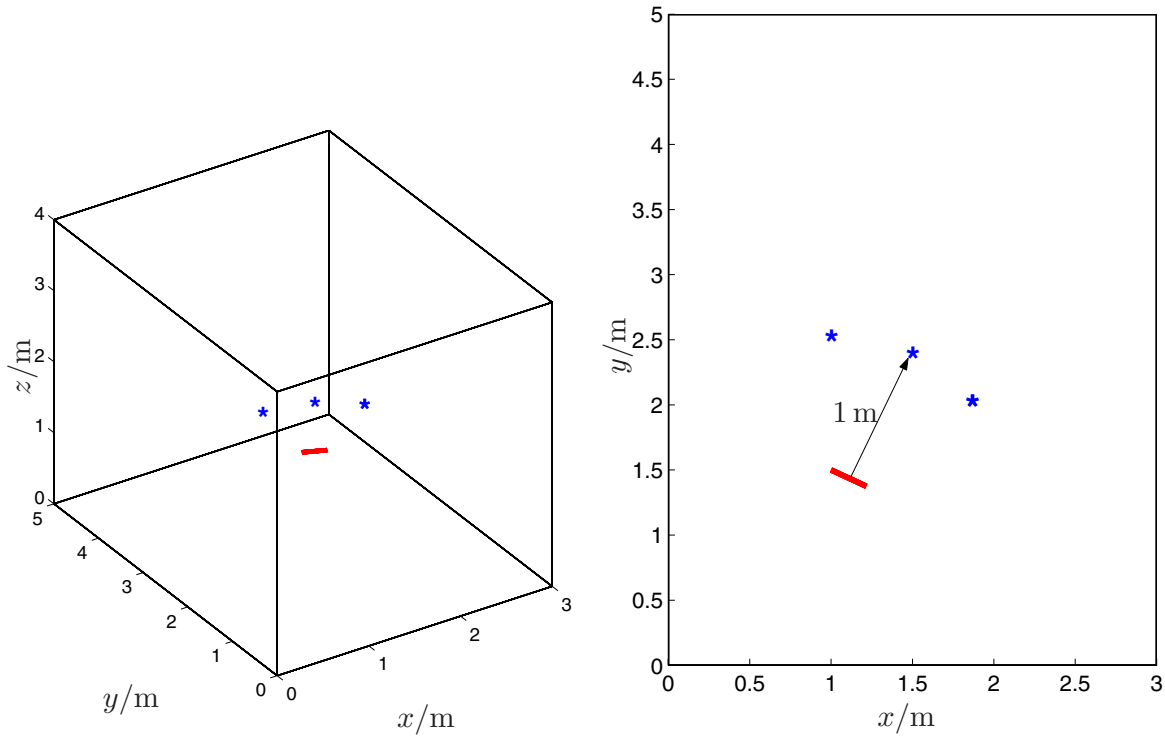
Test Environments for the Evaluation of Localization and Separation Algorithms

The evaluation of the proposed MoG-based speaker localization and enhancement algorithms were carried out on the basis of experiments carried out in two different reverberant environments, under three different *types* of background noise and with three different signal to noise ratios (SNR). The sampling frequency was set to 8 kHz. The source signals comprised 10 talkers from the TIMIT database (5 female and 5 male, each uttering a single sentence) downsampled to the sampling frequency used.

The first enviroment is a *simulated* room with a reverberation time of $T_{60} = 300\,\text{ms}$. It has the same dimensions as the room used in Chapter 3 and the room impulse responses were generated with the same software. This room is referred to as Room 1. The room impulse responses were generated for three different source positions with respective directions of arrival of 60°, 90° and 120° measured with respect to the array axis. The radial distance of the source from the mid-point of the array axis was fixed at 1.0 m. The source was set to be at the same height as the array. The setup is illustrated in Figure E.1. With the dimensions of the room set at $3\,\text{m} \times 5\,\text{m} \times 4\,\text{m}$, we have a critical radius of 0.81 m. The critical radius is defined as the distance from the source where the direct and reverberant sound energy densities are just equal. Microphone placement outside the critical radius will contain a higher amount of reverberation. Note that with our positioning, we are *outside* the critical radius.

The source signals were generated by convolving the individual TIMIT signals with the generated room impulse responses. For the multiple-speaker scenario, the source signals were convolved with the respective impulse responses and then additively mixed at a global signal to interference ratio (SIR) of 0 dB.

The second environment is a reverberant seminar room at the IKA, with a $T_{60} = 600\,\text{ms}$ and a critical radius of 0.84 m, which is referred to as Room 2. The approximate layout of the room and the array position are depicted in Figure E.2. To generate the test scenarios in this room, the speaker signals were individually played back through a loudspeaker (Genelec 2029BR) at the same angles of arrival as in the synthetic case, and recorded by the array using a custom designed pre-amplifier frontend and the RME Multiface$^{\text{tm}}$ A/D converter, at a sampling rate of 32kHz. The ambient noise in the room was low during the recordings (no extraenous source of noise, recordings made late in the evening). The recorded signals were downsampled to 8kHz offline. The competing speaker situation was created by additively mixing the single speaker recordings at the corresponding azimuths at 0 dB SIR.



(a) 3-D view of simulated room

(b) 2-D view of simulated room

Figure E.1: Source–array setup in the synthetic room. The distance of the source from the array is $d_s = 1\,\mathrm{m}$.

For each position and each speaker, three different types of background noise at varying SNR was added to the signals. The noise types considered were (a) computer generated white spatially uncorrelated noise at all the sensors, (b) white noise recorded using the microphone array in a *diffuse* environment and (c) cafeteria babble recorded using the microphone array in a diffuse environment. The noises for cases (b) and (c) were recorded in the reverberant chamber at the IKA premises. Note that the noise signals are no longer spatially uncorrelated for cases (b) and (c).

For the separation experiments, a subset of 6 speakers from the original 10 was used. The competing scenario consisted of one source at 60° and the other at 120°. The background noises and the SNRs considered were the same as for the localization experiments. For every combination of background noise type and SNR, each speaker in the subset was simulated at each of these two positions, against every other speaker in the corresponding interference position.

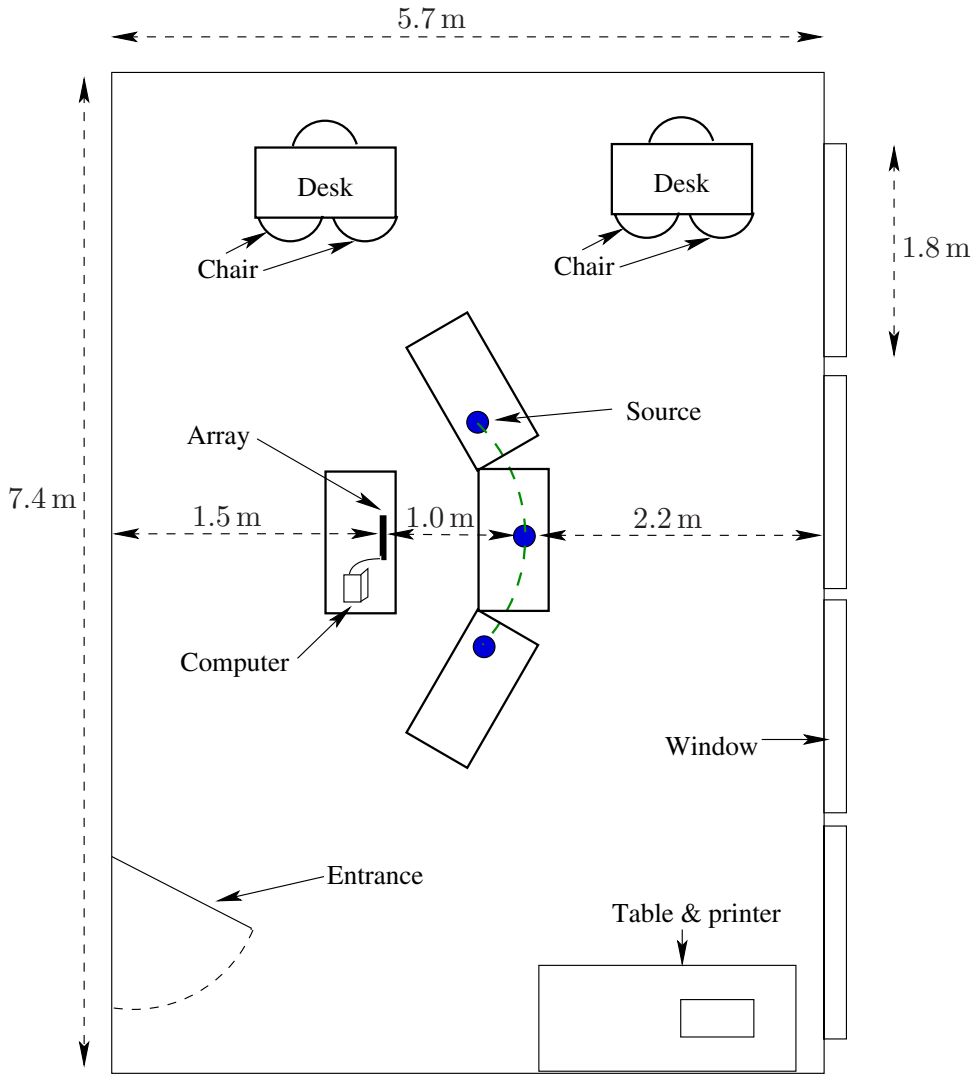


Figure E.2: Source–array setup for the recordings made in room 2, a seminar room at IKA.

Appendix F

Non-Linearity of Time-Frequency Masking

For any complex valued constants α and β, a fundamental property of a linear function $\mathcal{f}$ is:

$$\mathcal{f}(\alpha x + \beta y) = \alpha \mathcal{f}(x) + \beta \mathcal{f}(y)\,. \tag{F.1}$$

Consider a signal mixture of two sources $s_1(n)$ and $s_2(n)$. In its simplest version, the time-frequency *mask* $\mathcal{M}_q(k,b)$ generated for the source q at the T-F point (k,b) depends on the preponderance of the source at that T-F point, i.e.,

$$\mathcal{M}_q(k,b) \triangleq \begin{cases} 1 & |S_q(k,b)| > |S_{q'}(k,b)|\,,\ q \neq q' \\ 0 & \text{otherwise.} \end{cases} \tag{F.2}$$

This data-dependence of the mask generation function makes it susceptible to signal scaling, i.e., for amplitude-scaled versions of the signals it is *not* guaranteed that:

$$\mathcal{M}_{q,\alpha_1,\alpha_2}(k,b) = \mathcal{M}_{q,\beta_1,\beta_2}(k,b)\,, \tag{F.3}$$

where

$$\mathcal{M}_{q,\alpha_q,\alpha_{q'}}(k,b) \triangleq \begin{cases} 1 & \alpha_q|S_q(k,b)| > \alpha_{q'}|S_{q'}(k,b)|\,,\ q \neq q' \\ 0 & \text{otherwise.} \end{cases} \tag{F.4}$$

It is now easy to see that for different amplitude scaling factors the output of the mask function may not satisfy (F.1). The mask estimate is also influenced by the temporal alignments of the signals, i.e., the mask generated at any T-F point for a signal mixture of $s_1(n)$ and $s_2(n)$ will *not* be the same as the mask generated for a signal mixture of, e.g., $s_1(n)$ and $s_2(n-l)$.

Thus, the strong data-dependence on the mask estimation makes the time-frequency mask non-linear by *design*. Some linearization may be obtained, however, by smoothing the estimated masks over time and frequency. Such a smoothing is usually always done in practical applications.

In contrast, the output of beamforming algorithms always satisfy (F.1). This holds to a large extent even for *adaptive* beamforming algorithms which are estimated from the data because the beamforming coefficients are implicitly computed over *averaged* data, reducing the dependence on instantaneous values.

Appendix G

Performance Evaluation Measures for Source Separation

This section presents the instrumental evaluation criteria used to judge the performance of the adaptive source separation approaches considered in Chapter 6. It should be noted that while a more complete test suite for a specific application should incorporate subjective listening tests, in order to obtain a better perspective on the algorithm performance especially regarding the SINR versus SDR trade-offs involved, the instrumental measures selected are established physical measures that are arguably, albeit imperfectly, related to important aspects of user-perceived signal quality, such as speech intelligibility, signal distortion, and relative loudness of desired and undesired signal components, and are capable of illustrating the extant trends in the algorithm performance. Subsequently, one may use these measures as a guideline for algorithm selection. The 'fine-tuning' of the algorithms with respect to an application, however, would benefit from the subjective listening tests.

G.1 Notation

Since we explicitly consider a two speaker competing situation under varying background noise conditions, we use the following signal processing notation:

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{s}_{\mathrm{t,in}}(n) + \mathbf{s}_{\mathrm{i,in}}(n) + \mathbf{v}_{\mathrm{in}}(n) \\ &\downarrow \\ &\boxed{\begin{array}{c}\text{Separation/}\\ \text{Enhancement}\end{array}} \\ &\downarrow \\ y_{\mathrm{t}}(n) &= s_{\mathrm{t,out}}(n) + s_{\mathrm{i,out}}(n) + v_{\mathrm{out}}(n) \end{aligned} \tag{G.1}$$

Here, $\mathbf{s}_{\mathrm{t,in}}(n)$ is the target signal as received at the microphone *array*[1]. Similarly, $\mathbf{v}_{\mathrm{in}}(n)$ represents the background noise and $\mathbf{s}_{\mathrm{i,in}}(n)$ represents the competing signal, all as received

[1] Propagation effects are implicitly considered.

$$\begin{aligned} \mathbf{s}_{\mathrm{t,in}}(n) &= \big(s_{\mathrm{t,in,1}}(n)\,,\,\ldots\,,\,s_{\mathrm{t,in,M}}(n)\big)^T \\ s_{\mathrm{t,in},m}(n) &= a_{m\mathrm{t}}(n) * s_{\mathrm{t}}(n)\,, \end{aligned}$$

where $a_{m\mathrm{t}}(n)$ is the room impulse response from the target source to microphone m, consistent with the notation derived in Chapter 2.

at the array, hence the vector notation. The signals received at the first channel, denoted as

$$x_1(n) = s_{\mathrm{t,in},1}(n) + s_{\mathrm{i,in},1}(n) + v_{\mathrm{in},1}(n)\,,$$

provide the reference.

The objective measures are based on values computed for short signal segments (analogous to the frames in the short-time Fourier analysis), with duration about 10−20 ms. The smoothed power density spectra are estimated for each segment by a segmental periodogram using the DFT on the windowed segment, giving a uniform linear frequency resolution, followed by summing the power within third-octave bands. The long-term spectra are computed by averaging the segmental power density spectra.

As all the test signals are created by additive mixing, the desired and competing signal components, $\mathbf{s}_{\mathrm{t,in}}(n)$, $\mathbf{s}_{\mathrm{i,in}}(n)$ and $\mathbf{v}_{\mathrm{in}}(n)$, are separately available for evaluation purposes, although the actual algorithm implementation observes only the combined signal $\mathbf{x}(n)$. The performance measures require separate estimates of the corresponding individual ouput components $s_{\mathrm{t,out}}(n)$, $s_{\mathrm{i,out}}(n)$ and $v_{\mathrm{out}}(n)$.

These estimates are obtained by *shadow-filtering*: the adaptive filters are determined based on the combined input signal and their time evolution is stored. During the performance evaluation, these stored filters are applied *individually* to the separately available desired and competing signals, resulting in separate output signal components.

G.2 Long-term SNR

The long-term SNR simply measures the signal to noise ratio over the complete duration of the signals. It is a primitive measure which yields good performance indication if the signals are stationary. For speech, which has only short-term stationarity, such a measure is not capable of predicting the effects of short-term spectral variations.

$$\mathrm{SNR}_{\mathrm{in}} = 10\lg\frac{\sum_n s^2_{\mathrm{t,in}}(n)}{\sum_n v^2_{\mathrm{in}}(n)} \tag{G.2}$$

$$\mathrm{SNR}_{\mathrm{out}} = 10\lg\frac{\sum_n s^2_{\mathrm{t,out}}(n)}{\sum_t v^2_{\mathrm{out}}(n)}\,. \tag{G.3}$$

G.3 Intelligibility-weighted long-term SNR

The intelligibility-weighted long-term SNR is directly related to the standard Speech Intelligibility Index (SII), commonly used to predict speech intelligibility in non-fluctuating noise [5]. For broadband external noise that exceeds the hearing threshold at all frequencies, the SII is based on the frequency-weighted SNR (in dB), calculated using long-term average speech and noise power spectra. Before the frequency-weighted averaging, the SNR is restricted to values between −15 and +15 dB at each frequency, as indicated by the $[]_a^b$ values.

$$\text{IW-SNR}_{\mathrm{in}} = \frac{1}{2\pi}\int_{\Omega_\ell}^{\Omega_u} \mathcal{W}_{\mathrm{SII}}(\Omega)\left[10\lg\frac{\Psi_{S_{\mathrm{t}}S_{\mathrm{t}},\mathrm{in}}(\Omega)}{\Psi_{VV,\mathrm{in}}(\Omega)}\right]_{-15}^{+15}\mathrm{d}\Omega\,. \tag{G.4}$$

$\text{IW-SNR}_{\mathrm{out}}$ is defined in a similar fashion.

Of course, for actual computations the integral is always approximated by a sum. Regardless of the frequency resolution used in the summation, the weighting function is always normalized such that

$$\frac{1}{2\pi}\int_{\Omega_\ell}^{\Omega_u} \mathcal{W}_{\mathrm{SII}}(\Omega)\,\mathrm{d}\Omega = 1\,. \tag{G.5}$$

The standard SII is just a linear transformation of the intelligibility-weighted SNR to an index between 0 and 1, corresponding to IW-SNR values between −15 and +15 dB respectively. The SII standard defines different frequency-weighting factors depending on the linguistic redundancy of different speech test materials. Here, the standard weighting factors for "average speech" are used, giving equal weight for each auditory critical band (CB) between 300 and 4000 Hz.

G.4 Segmental Intelligibility-weighted SNR

The SII standard does not claim to account for the effects of fluctuating noise. It is also possible that the intelligibility-weighted long-term SNR, as defined in the previous section, may obscure some segmental effects of the enhancement algorithms. Therefore, a slightly modified procedure is used to derive a segmental intelligibility-weighted SNR. This measure has not been empirically validated and is used only tentatively. Here the intelligibility-weighted SNR (dB) is first calculated for each short-time segment b and then these values are averaged.

$$\text{IW-SegSNR}_{\mathrm{in}} = \frac{1}{2\pi B}\sum_{b=1}^{B}\int_{\Omega_\ell}^{\Omega_u} \mathcal{W}_{\mathrm{SII}}(\Omega)\left[10\lg\frac{\Psi_{S_\mathrm{t}S_\mathrm{t},\mathrm{in}}(\Omega,b)}{\Psi_{VV,\mathrm{in}}(\Omega,b)}\right]_{-20}^{+35}\mathrm{d}\Omega \tag{G.6}$$

and $\text{IW-SegSNR}_{\mathrm{out}}$ is similarly defined.

G.5 Segmental SNR

The segmental SNR is a temporal average of short-time SNR values (in dB) based on the speech and noise power ratio in each short-time signal segment. The segmental SNR is somewhat related to the relative loudness of desired vs. undesired signal components. However, an improvement in segmental SNR does not necessarily imply a corresponding improvement in speech intelligibility. Before averaging, short-time SNR values are limited to between −20 and +35 dB to avoid bias.

$$\text{SegSNR}_{\mathrm{in}} = \frac{1}{B}\sum_{b=1}^{B}\left[10\lg\frac{\sum_n s_{\mathrm{t,in},b}^2(n)}{\sum_n v_{\mathrm{in},b}^2(n)}\right]_{-20}^{+35} \tag{G.7}$$

$$\text{SegSNR}_{\mathrm{out}} = \frac{1}{B}\sum_{b=1}^{B}\left[10\lg\frac{\sum_n s_{\mathrm{t,out},b}^2(n)}{\sum_t v_{\mathrm{out},b}^2(n)}\right]_{-20}^{+35} \tag{G.8}$$

where $s_{\mathrm{t,out},b}(n)$ defines the corresponding signal values in *segment* b.

G.6 Frequency-weighted log-spectral Signal Distortion

The log-spectral signal distortion (SD), as used here, is a measure of the distortion of the desired input signal only, and disregards any distortion of the undesired competing signal. The measure indicates a Euclidean distance between smoothed short-time spectra, in logarithmic (dB) units, of the speech component in the processed and input signals. The distance measure is calculated with a frequency-weighting factor giving equal weight for each auditory critical band, as defined by the equivalent rectangular bandwidth (ERB) of auditory filters [84]. Segments are included in the average only if the input SNR is above -15 dB for this segment.

$$\log_{\mathrm{SD}} = \frac{1}{B}\sum_{b=1}^{B}\sqrt{\frac{1}{2\pi}\int_{\Omega_\ell}^{\Omega_u}\mathcal{W}_{\mathrm{ERB}}(\Omega)\left(10\lg\frac{\Psi_{S_\mathrm{t}S_\mathrm{t},out}(\Omega,b)}{\Psi_{S_\mathrm{t}S_\mathrm{t},in}(\Omega,b)}\right)^2\mathrm{d}\Omega}\,. \tag{G.9}$$

Of course, in actual computations the integral is always approximated by a sum. The weighting factor is normalized in the same way as the intelligibility weighting factor, as defined in (G.5). This measure is dependent on the possibly different room transfer functions to the various microphones involved. However, when the microphone matching is not guaranteed or the source position not accurately known, this measure is a better choice than the deviation of the beamformer from the target direction.

In analogy with (G.4), we may define the log spectral distortion over the average of the complete signal as:

$$\text{long-}\log_{\mathrm{SD}} = \sqrt{\frac{1}{2\pi}\int_{\Omega_\ell}^{\Omega_u}\mathcal{W}_{\mathrm{ERB}}(\Omega)\left(10\lg\frac{\Psi_{S_\mathrm{t}S_\mathrm{t},out}(\Omega)}{\Psi_{S_\mathrm{t}S_\mathrm{t},in}(\Omega)}\right)^2\mathrm{d}\Omega}\,. \tag{G.10}$$

G.7 SIR, SINR and performance measurement

Similar to the SNR measures defined above, we define the signal-to-interference ratio (SIR) and the signal-to-(interference + noise) ratio (SINR). The former measures the effect of the algorithm on the interference suppression and the latter provides an indication of the *overall* enhancement over all interferences (directive and background). The values are computed in the selfsame manner as derived for the SNR metric.

The performance of an algorithm is measured in terms of the *improvement* on the defined metrics. This is computed as the difference between the input and output values of the metric. For example, for the IW-SegSNR we obtain:

$$\Delta\text{IW-SegSNR} = \text{IW-SegSNR}_{\mathrm{out}} - \text{IW-SegSNR}_{\mathrm{in}}\,. \tag{G.11}$$

Similarly, one may define the improvements for the other metrics.

Vetinari sighed. People told him things all the time. Lots of people had been telling him things in the last hour... what it amounted to was not information, but a huge Argus-eyed ball of little niggling factoids, out of which some information could, with care, be teased.

Making money
– Terry Pratchett

Bibliography

[1] Aichner, R., Buchner, H., Wehr, S., and Kellermann, W. Robustness of acoustic multiple-source localization in adverse environments. In *Proceedings of the 7th German Information Technology Conference on Speech Communication (ITG)* (2006).

[2] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automation and Control 19*, 6 (1974), 716–723.

[3] Allen, J. B., and Berkley, D. A. Image method for efficiently simulating small room acoustics. *Journal of the Acoustical Society of America 65*, 4 (1979), 943 – 950.

[4] Anemüller, J., and Kollmeier, B. Amplitude modulation decorrelation for convolutive blind source separation. In *Proceedings of the International Conference on Independent Component Analysis (ICA)* (June 2000), pp. 215–220.

[5] ANSI-S3.5. *American national standard methods for the calculation of the speech intelligibility index*. American National Standards Institute, New York, 1997.

[6] Araki, S., Makino, S., Sawada, H., and Mukai, R. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2005), pp. 81–84.

[7] Araki, S., Sawada, H., Mukai, R., and Makino, S. A novel blind source separation method with observation vector clustering. In *Proceedings of the IWAENC* (Sept. 2005).

[8] Baumann, W., Kolossa, D., and Orglmeister, R. Beamforming-based convolutive blind source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2003), pp. 357–360.

[9] Bechler, D., and Kroschel, K. Confidence scoring of time difference of arrival estimation for speaker localization with microphone arrays. In *Proceedings of the 13th Conference "Elektonische Sprachsignalverarbeitung (ESSV)"* (Dresden, Germany, Sept. 2002).

[10] Bechler, D., and Kroschel, K. Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (2003), pp. 315–318.

[11] Bechler, D., and Kroschel, K. Three different reliability criteria for time delay estimates. In *Proceedings of the European Signal Processing Conference (EUSIPCO)* (2004), pp. 1987 – 1990.

[12] Benesty, J. Adaptive eigenvalue decomposition algorithm for passive source localization. *Journal of the Acoustical Society of America 107*, 1 (Jan. 2000), 384–391.

[13] Bilmes, J. A. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Tech. Rep. TR-97-021, U.C. Berkeley, 1998.

[14] BIRCHFIELD, S. T. A unifying framework for acoustic localization. In *Proceedings of the European Signal Processing Conference (EUSIPCO)* (2004).

[15] BÖHME, J. F. Estimation of spectral parameters of correlated signals in wavefields. *EURASIP Journal on Applied Signal Processing 11*, 4 (1986), 329–337.

[16] BRANDSTEIN, M. S., AND WARD, D. B. Cell-based beamforming for speech acquisition with microphone arrays. *IEEE Transactions on Speech and Audio Processing 8*, 6 (Nov. 2000), 738–743.

[17] BRANDWOOD, D. H. A complex gradient operator and its application in adaptive array theory. *IEE Proceedings 130, Pts.F and H*, 1 (Feb. 1983).

[18] BREITHAUPT, C., GERKMANN, T., AND MARTIN, R. Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. *IEEE Signal Processing Letters 14*, 12 (Dec. 2007).

[19] BRUTTI, A., OMOLOGO, M., AND SVAIZER, P. Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)* (Sept. 2005), pp. 2337–2340.

[20] BUCHNER, H., AICHNER, R., AND KELLERMANN, W. TRINICON: A versatile framework for multichannel blind signal processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2004).

[21] BUCHNER, H., AICHNER, R., AND KELLERMANN, W. A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics. *IEEE Transactions on Speech and Audio Processing 13*, 1 (Jan. 2005), 120–134.

[22] BUCK, M., HAULICK, T., AND PFLEIDERER, H.-J. Self-calibrating microphone arrays for speech signal acquisition: A systematic approach. *EURASIP Journal on Applied Signal Processing 86* (June 2006), 1230–1238.

[23] CERMAK, J., ARAKI, S., SAWADA, H., AND MAKINO, S. Blind speech separation by combining beamformers and a time frequency mask. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (Sept. 2006).

[24] CHAN, Y. T., AND HO, K. C. A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing 42*, 8 (Aug. 1994), 1905–1915.

[25] CHAN, Y. T., RILEY, J., AND PLANT, J. B. Modeling of time delay and its application to estimation of nonstationary delays. *IEEE Transactions on Acoustics, Speech and Signal Processing 29*, 3 (June 1981), 577– 581.

[26] CHEN, J., BENESTY, J., AND HUANG, Y. Robust time delay estimation exploiting redundancy among multiple microphones. *IEEE Transactions on Speech and Audio Processing 11*, 6 (Nov. 2003), 549 – 557.

[27] CHEN, J., BENESTY, J., AND HUANG, Y. Time delay estimation using spatial correlation techniques. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (2003).

[28] CHEN, J., BENESTY, J., AND HUANG, Y. A minimum distortion noise reduction algorithm with multiple microphones. *IEEE Transactions on Audio, Speech and Language Processing 16*, 3 (Mar. 2008), 481–493.

[29] CHEN, J. C., HUDSON, R. E., AND YAO, K. Maximum likelihood source localization and unknown sensor location estimation for wideband signals in the near-field. *IEEE Transactions on Signal Processing 50*, 8 (Aug. 2002), 1843–1854.

[30] COVER, T. M., AND THOMAS, J. A. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[31] Cox, H., Zeskind, R. M., and Owen, M. M. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing 35*, 10 (Oct. 1987), 1365–1375.

[32] Diamantaras, K. I. Blind source separation using principal component neural networks. In *Lecture Notes in Computer Science*, G. Dorffner, H. Bischof, and K. Hornik, Eds., vol. 2130. Springer Verlag, 2001.

[33] DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. Robust localization in reverberant rooms. In *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, Berlin, 2001, pp. 157–180.

[34] Doclo, S., and Moonen, M. Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP Journal on Applied Signal Processing 11* (2003), 1110–1124.

[35] Doclo, S., Spriet, A., Wouters, J., and Moonen, M. Speech distortion weighted multi-channel Wiener filtering techniques for noise reduction. In *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer Verlag, 2005, pp. 199–228.

[36] Drews, M. Time delay estimation for microphone array speech enhancement systems. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)* (1995), vol. 3, pp. 2013–2016.

[37] Dvorkind, T. G., and Gannot, S. Time difference of arrival estimation of speech source in a noisy and reverberant environment. *EURASIP Journal on Applied Signal Processing 85* (Jan. 2005), 177–204.

[38] Eldar, Y. C., and Oppenheim, A. V. MMSE whitening and subspace whitening. *IEEE Transactions on Information Theory 49*, 7 (July 2003), 1846 – 1851.

[39] Elko, G. Spatial coherence functions for differential microphones in isotropic noise fields. In *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, 2001, pp. 61–85.

[40] Eneman, K., Leijon, A., Doclo, S., Spriet, A., Moonen, M., and Wouters, J. Auditory-profile-based physical evaluation of multi-microphone noise reduction techniques in hearing instruments. In *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. John Wiley & Sons, Ltd., New York, USA, 2008, pp. 431–458.

[41] Etter, D. M., and Stearns, S. D. Adaptive estimation of time delays in sampled data systems. *IEEE Transactions on Acoustics, Speech and Signal Processing 29*, 3 (June 1981), 582–587.

[42] Frost, III, O. L. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE 60*, 8 (Aug. 1972), 926–935.

[43] Gannot, S., Burshtein, D., and Weinstein, E. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing 49*, 8 (Aug. 2001), 1614–1626.

[44] Gannot, S., and Dvorkind, T. G. Microphone array speaker localizers using spatial-temporal information. *EURASIP Journal on Applied Signal Processing*, 1 (2006), 1–17.

[45] Griffiths, L. J., and Jim, C. W. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation AP-30* (1982), 27–34.

[46] Gückel, A. Algorithms for multichannel blind source separation. Master's thesis, Institute of Communication Acoustics, Ruhr-Universität Bochum, April 2006.

[47] GUSTAFSSON, T., RAO, B. D., AND TRIVEDI, M. Source localization in reverberant environments: Modelling and statistical analysis. *IEEE Transactions on Speech and Audio Processing 11*, 6 (Nov. 2003), 791–802.

[48] HABETS, E. A. P. Room impulse response generator. Online resource, `www.sps.ele.tue.nl/members/E.A.P.Habets/rir_generator/default.asp`, 2006.

[49] HAYKIN, S. *Adaptive Filter Theory*, 3 ed. Prentice Hall, 1996.

[50] HERTZ, D. Time delay estimation by combining efficient algorithms and generalized cross-correlation methods. *IEEE Transactions on Acoustics, Speech and Signal Processing 34*, 1 (Feb. 1986), 1–7.

[51] HOSHUYAMA, O., AND SUGIYAMA, A. Robust adaptive beamforming. In *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, Berlin, 2001, pp. 87–109.

[52] HUA, T. P., SUGIYAMA, A., AND FAUCON, G. A new self-calibration technique for adaptive microphone arrays. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (2005).

[53] HUANG, Y., BENESTY, J., AND CHEN, J. Analysis and comparison of multichannel noise reduction methods in a common framework. *IEEE Transactions on Audio, Speech and Language Processing 16*, 5 (July 2008), 957–968.

[54] HUANG, Y., BENESTY, J., ELKO, G. W., AND MERSEREAU, R. M. Real-time passive source localization: A practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing 9*, 8 (Nov. 2001), 943–956.

[55] HUNG, H., AND KAVEH, K. Focusing Matrices for Coherent Signal-Subspace Processing. *IEEE Transactions on Acoustics, Speech and Signal Processing 36*, 8 (Aug. 1988), 1272–1281.

[56] HYVÄRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks 10*, 3 (1999), 626–634.

[57] HYVÄRINEN, A., KARHUNEN, J., AND OJA, E. *Independent Component Analysis*. John Wiley & Sons, Ltd., 2001.

[58] HYVÄRINEN, A., AND OJA, E. A fast fixed-point algorithm for independent component analysis. *Neural Computation 9*, 7 (1997), 1483–1492.

[59] HYVÄRINEN, A., AND OJA, E. Independent component analysis: Algorithms and applications. *Neural Networks 13*, 4–5 (2000), 411–430.

[60] IKRAM, M. Z., AND MORGAN, D. R. A beamforming approach to permutation alignment for multichannel frequency-domain blind source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2002), vol. I, pp. 881–884.

[61] JAFFER, A. G. Maximum likelihood direction finding of stochastic sources: A separable solution. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Apr. 1988), vol. 5, pp. 2893 – 2896.

[62] KAPUR, J. N., AND KESAVAN, H. K. *Entropy Optimization Principles with Applications*. Academic Press, Inc., 1992.

[63] KEMP, T. Personal communication, 2008.

[64] KNAPP, C. H., AND CARTER, G. C. The generalized correlation method for the estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing 24*, 4 (Aug. 1976), 320–327.

[65] KOLOSSA, D., AND ORGLMEISTER, R. *Nonlinear Postprocessing for Blind Speech Separation*, vol. 3195 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, 2004, pp. 832–839.

[66] KRAUS, D. *Approximative Maximum-likelihood-Schätzung und verwandte Verfahren zur Ortung und Signalschätzung mit Sensorgruppen*. PhD thesis, Ruhr-Universität Bochum, 1993.

[67] KRIM, H., AND VIBERG, M. Two decades of array signal processing research. *IEEE Signal Processing Magazine* (July 1996), 67–94.

[68] KROLIK, J. Focussed wideband array processing for spatial spectral estimation. In *Advances in Spectrum Analysis and Array Processing, Vol. II*, S. Haykin, Ed. Prentice-Hall, Upper Saddle River, New Jersey, USA, 1991, ch. 6.

[69] LAAKSO, T. I., VÄLIMÄKI, V., KARJALAINEN, M., AND LAINE, U. K. Splitting the unit delay - tools for fractional delay filter design. *IEEE Signal Processing Magazine 13*, 1 (Jan. 1996).

[70] LATHOUD, G. *Spatio-temporal analysis of spontaneous speech with microphone arrays*. PhD thesis, École Polytechnique Fédérale Lausanne, 2006.

[71] LIU, C., WHEELER, B. C., O'BRIEN, JR., W. D., BILGER, R. C., LANSING, C. R., JONES, D. L., AND FENG, A. S. A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers. *Journal of the Acoustical Society of America 110*, 6 (Dec. 2001), 3218–3231.

[72] LIU, C., WHEELER, B. C., O'BRIEN JR., W. D., LANSING, C. R., BILGER, R. C., AND FENG, A. S. Localization of multiple sound sources with two microphones. *Journal of the Acoustical Society of America 108*, 4 (Oct. 2000), 1888–1905.

[73] MADHU, N. Independent component analysis in multiple input multiple output (MIMO) systems. Master's thesis, Technische Universität München, Munich, Germany, Oct. 2002.

[74] MADHU, N., BREITHAUPT, C., AND MARTIN, R. Temporal smoothing of spectral masks in the cepstral domain for speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2008).

[75] MADHU, N., AND MARTIN, R. Robust speaker localization through adaptive weighted pair TDOA (AWEPAT) estimation. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)* (Sept. 2005).

[76] MADHU, N., AND MARTIN, R. A scalable framework for multiple speaker localization and tracking. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (Seattle, USA, Sept. 2008).

[77] MADHU, N., MARTIN, R., REHN, H.-W., AND FISCHER, A. Brake squeal localization. In *Proceedings of the Annual Meeting of the German Acoustical Society (DAGA)* (Mar. 2006).

[78] MADHU, N., OSWALD, D., AND MARTIN, R. Speaker localization: Novel algorithm and practical aspects. In *Proceedings of the 19th Conference "Elektonische Sprachsignalverarbeitung (ESSV)"* (Frankfurt am Main, Germany, Sept. 2008).

[79] MAIWALD, D. *Breitbandverfahren zur Signalentdeckung und -ortung mit Sensorgruppen in Seismik- und Sonaranwendungen*. PhD thesis, Ruhr-Universität Bochum, 1995.

[80] MARTIN, R. Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2002).

[81] MATSUO, M., HIOKA, Y., AND HAMADA, N. Estimating DOA of multiple speech signals by improved histogram mapping method. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (Sept. 2005).

[82] MATSUOKA, K., AND NAKASHIMA, S. Minimal distortion principle for blind source separation. In *Proceedings of the International Conference on Independent Component Analysis (ICA)* (Dec. 2001), pp. 722–727.

[83] MCCOWAN, I. A., AND BOURLARD, H. Microphone array post-filter based on noise field coherence. *IEEE Transactions on Speech and Audio Processing 11*, 6 (Nov. 2003), 709–716.

[84] MOORE, B. *An Introduction to the Psychology of Hearing. 5th ed.* Academic Press, London, 2003.

[85] MUKAI, R., SAWADA, H., ARAKI, S., AND MAKINO, S. Real-time blind source separation and DOA estimation using a small 3-D microphone array. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (Sept. 2005).

[86] NGO, K., SPRIET, A., MOONEN, M., WOUTERS, J., AND JENSEN, S. H. Variable speech distortion weighted multichannel wiener filter based on soft output voice activity detection for noise reduction in hearing aids. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (2008).

[87] NORDHOLM, S., CLAESSON, I., AND DAHL, M. Adaptive microphone array employing calibration signals. An analytical evaluation. *IEEE Transactions on Speech and Audio Processing 7*, 3 (May 1999).

[88] OAK, P., AND KELLERMANN, W. A calibration algorithm for robust generalized sidelobe cancelling beamformers. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (2005).

[89] OBRADOVIC, D., MADHU, N., SZABO, A., AND WONG, C. S. Independent component analysis for semi-blind signal separation in MIMO mobile frequency selective communication channels. In *Proceedings of the INNS-IEEE International Joint Conference on Neural Networks (IJCNN)* (2004).

[90] OMOLOGO, M., AND SVAIZER, P. Acoustic event localization using a crosspower-spectrum based technique. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1994), vol. II, pp. 273–276.

[91] OPPENHEIM, A. V., AND SCHAFER, R. W. *Digital Signal Processing.* Prentice Hall, 1975.

[92] O'SHAUGHNESSY, D. *Speech Communications.* IEEE Press, 2000.

[93] PAPOULIS, A. *Probability, Random Variables and Stochastic Processes*, 3 ed. McGraw-Hill Inc., 1991.

[94] PARRA, L., AND ALVINO, C. V. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing* (Sept. 2002), 352–362.

[95] PORTER, J., AND BOLL, S. Optimal estimators for spectral restoration of noisy speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1984).

[96] POTAMITIS, I. Estimation of speech presence probability in the field of microphone array. *IEEE Signal Processing Letters 11*, 12 (Dec. 2004), 956–959.

[97] REED, F., FEINTUCH, P., AND BERSHAD, N. Time delay estimation using the LMS adaptive filter – static behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing 29*, 3 (June 1981), 561–571.

[98] Renevey, P., and Drygajlo, A. Entropy based voice activity detection in very noisy conditions. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)* (Sept. 2001).

[99] Reuven, G., Gannot, S., and Cohen, I. Dual source transfer-function generalized sidelobe canceller. *IEEE Transactions on Audio, Speech and Language Processing 16*, 4 (May 2008), 711–727.

[100] Rickard, S., and Yilmaz, Ö. On the approximate W-Disjoint orthogonality of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2002).

[101] Roman, N., Wang, D., and Brown, G. Speech segregation based on sound localization. *Journal of the Acoustical Society of America 114*, 4 (Oct. 2003), 2236 – 2252.

[102] Saruwatari, H., Kawamura, T., Nishikawa, T., Lee, A., and Shikano, K. Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. *IEEE Transactions on Audio, Speech and Language Processing 14* (Mar. 2006), 666–678.

[103] Sawada, H., Mukai, R., Araki, S., and Makino, S. A polar-coordinate based activation function for frequency domain blind source separation. In *Proceedings of the International Conference on Independent Component Analysis (ICA)* (Dec. 2001), pp. 663–668.

[104] Sawada, H., Mukai, R., Araki, S., and Makino, S. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing 12* (Sept. 2004), 530–538.

[105] Scheuing, J., and Yang, B. Disambiguation of TDOA estimates in multi-path multi-source environments (DATEMM). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2006), vol. 4, pp. 837–840.

[106] Scheuing, J., and Yang, B. Efficient synthesis of approximately consistent graphs for acoustic multi-source location. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2007), pp. 501–504.

[107] Schmidt, R. O. *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*. PhD thesis, Stanford University, 1981.

[108] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics 6*, 2 (1978), 461–464.

[109] Shannon, C., and Weaver, W. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[110] Silverman, H. F., Yu, Y., Sachar, J. M., and Patterson, III, W. R. Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing 13*, 4 (July 2005), 593–606.

[111] Smaragdis, P. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing* (1998), 21–34.

[112] So, H. C., Ching, P. C., and Chan, Y. T. A new algorithm for explicit adaptation of time delay. *IEEE Transactions on Signal Processing 42*, 7 (July 1994), 1816–1820.

[113] Stoica, P., and Nehorai, A. MUSIC, maximum likelihood, and Cramér-Rao bound. *IEEE Transactions on Acoustics, Speech and Signal Processing 37*, 5 (May 1989), 720–741.

[114] STOICA, P., AND NEHORAI, A. On the concentrated stochastic likelihood function in array signal processing. *Circuits, Systems and Signal Processing 14*, 5 (1995), 669–674.

[115] STRANG, G. *Linear Algebra and Its Applications*, 3 ed. Thompson / Brooks Cole, 1988.

[116] TADAION, A. A., DERAKHTIAN, M., GAZOR, S., AND AREF, M. R. A fast multiple-source detection and localization array signal processing algorithm using the spatial filtering and ML approach. *IEEE Transactions on Signal Processing 55*, 5 (May 2007), 1815–1827.

[117] TALANTZIS, F., CONSTANTINIDES, A. G., AND POLYMENAKOS, L. C. Estimation of direction of arrival using information theory. *IEEE Signal Processing Letters 12*, 8 (Aug. 2005), 561–564.

[118] TASHEV, I. Gain self-calibration procedure for microphone arrays. In *Proceedings of the International Conference for Multimedia and Expo (ICME)* (June 2004).

[119] UCLA. 172A. Cognitive psychology of music (introduction). Online resource, `http://www.ethnomusic.ucla.edu/courses/ESM172a/Files172A/Week3.htm`.

[120] VAN TREES, H. L. *Detection, Estimation and Modulation Theory, Part I*. John Wiley and Sons, 1968.

[121] VAN TREES, H. L. *Detection, Estimation and Modulation Theory, Part IV*. John Wiley and Sons, 2002.

[122] VARY, P., AND MARTIN, R. *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons, Ltd., 2006.

[123] WANG, D. Time–frequency masking for speech separation and its potential for hearing aid design. *Trends in amplification* (Oct. 2008), 332–353.

[124] WANG, H., AND KAVEH, M. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources. *IEEE Transactions on Acoustics, Speech and Signal Processing 33*, 4 (Aug. 1985), 823–831.

[125] WARD, D. B., KENNEDY, R. A., AND WILLIAMSON, R. C. Constant directivity beamforming. In *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, 2001, pp. 3–17.

[126] WAX, M. Detection and localization of multiple sources in noise with unknown covariance. *IEEE Transactions on Acoustics, Speech and Signal Processing 40*, 1 (Jan. 1992), 245–249.

[127] WAX, M., AND KAILATH, T. Determining the number of signals by information theoretic criteria. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1984), vol. 9, pp. 232–235.

[128] WAX, M., AND KAILATH, T. Spatio-temporal spectral analysis by eigenstructure methods. *IEEE Transactions on Acoustics, Speech and Signal Processing 32*, 4 (Aug. 1984), 817–827.

[129] WIKIPEDIA. Banach fixed point theorem – Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Banach_fixed_point_theorem`.

[130] WIKIPEDIA. Disc brake – Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Disc_brake#Brake_squealOnlineresource`.

[131] WIKIPEDIA. Fixed point iteration – Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Fixed_point_iteration`.

[132] WIKIPEDIA. Iterated function – Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Iterated_function`.

[133] Xu, G., Liu, H., Tong, L., and Kailath, T. A least-squares approach to blind channel identification. *IEEE Transactions on Speech and Audio Processing 43*, 12 (1995), 2982 – 2993.

[134] Yilmaz, Ö., Jourjine, A., and Rickard, S. Blind separation of disjoint orthogonal signals: Demixing N sources from two mixtures. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2000).

[135] Yilmaz, Ö., Jourjine, A., and Rickard, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing 52*, 7 (July 2004).

[136] Yoon, Y., Kaplan, L. M., and McClellan, J. H. TOPS: New DOA estimator for wideband signals. *IEEE Transactions on Signal Processing 54*, 6 (June 2006), 791–802.

[137] Zelinski, R. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1988), pp. 2578–2581.

[138] Ziskind, I., and Wax, M. Maximum likelihood localization of multiple sources by alternating projection. *IEEE Transactions on Acoustics, Speech and Signal Processing 36*, 10 (Oct. 1988), 1553–1560.

[139] Zoubir, A. M., and Boashash, B. The bootstrap and its application in signal processing. *IEEE Signal processing magazine 15*, 1 (Jan. 1998), 56–76.

Kurzfassung

Räumlich verteilte Mikrofone, sogennante Mikrofonarrays, sind ein geschätztes Hilfsmittel in vielen akustischen Anwendungen und haben bereits Einzug in unser tägliches Leben gehalten. Beispiele hierfür sind Mensch-Maschine-Kommunikation, Video-Konferenzsysteme und Freisprecheinrichtungen im Automobil. Mikrofonarrays finden auch Verwendung außerhalb der Sprachsignalverarbeitung, so z.B. bei der akustischen Erkennung fehlerhafter Maschinen, in der Fahrzeugakustik und im Bereich autonomer Roboter.

In den meisten Fällen stellt die Lokalisation akustischer Quellen eine notwendige Voraussetzung für weitere Verarbeitungsschritte und Entscheidungsprozesse dar und steht deswegen nach wie vor im Mittelpunkt der Forschung. Die vorliegende Arbeit widmet sich sowohl der Quellenlokalisation an sich als auch deren Erweiterungen für die Quellentrennung.

Die Arbeit beginnt mit der Einführung einer Taxonomie für die Lokalisationsalgorithmen und gibt einen umfassenden Überblick über den gegenwärtigen Stand der Lokalisationsalgorithmen für unterschiedliche Anwendungsfelder. Dann wird die Beziehung zwischen diesen Ansätzen hergeleitet und es wird gezeigt, dass der Kern aller Ansätze eine Mehrkanal-Kreuzkorrelation ist. Die Unterschiede der verschiedenen Algorithmen liegen in der der Kreuzkorrelation vorangehenden Signalverarbeitung oder in den Annahmen bezüglich der Signale oder der akustischen Umgebung. Gerade diese Freiheitsgrade erlauben den Entwurf eines für eine gegebene Anwendung optimierten Lokalisationsalgorithmus. Dieser Aspekt wird anhand zweier sehr unterschiedlicher Anwendungsbeispiele illustriert, die sich im *a priori* Wissen über das Quellenmodell und das Modell der akustischen Umgebung unterscheiden. Die hier betrachteten Anwendungen sind a) Lokalisation von Bremsenquietschen eines fahrenden Kraftfahrzeuges und b) Lokalisation von gleichzeitig sprechenden Sprechern. Obwohl die den beiden Anwendungen zugrunde liegende Statistik die Mehrkanal-Kreuzkorrelation ist, sind die Algorithmen nicht austauschbar. Es wird gefolgert, dass es bei dem Entwurf eines Algorithmus wesentlich darauf ankommt, die verfügbaren anwendungsspezifischen Eigenschaften und Randbedingungen zu berücksichtigen. Die einfache Umsetzung eines für eine vorliegende Klasse von Problemen generischen Algorithmus führt dagegen in aller Regel nicht zu einem optimalen Lokalisationsergebnis.

Im zweiten Teil der Arbeit wird die Lokalisationsinformation für die Separierung eines Nutzsprechersignals in einer gestörten Umgebung mit konkurrierenden Sprechern (Cocktail-Party-Problem) verwendet. Es wird zunächst eine Taxonomie für die Quellentrennung eingeführt und der gegenwärtige Stand der Algorithmen sowie deren Zusammenhang kurz beschrieben. Es wird danach ein auf Independent Component Analysis basierter Ansatz zur Trennung zweier Quellen bei Verwendung zweier Mikrofone betrachtet. Die Verwendung von Lokalisationsinformation führt in diesem Fall zu einer Reduktion des Rechenaufwands. Als nächstes wird der zuvor entwickelte Lokalisationsalgorithmus für die Quellentrennung im allgemeinen Fall von Q Quellen und M Mikrofonen erweitert. Diese Erweiterung erlaubt die Nutzung einer Vielzahl von Trennungsalgorithmen und funktioniert, im Gegensatz zu vielen konventionellen Algorithmen einschließlich ICA, auch bei unterbestimmten ($M < Q$) oder gestörten Systemen. Weiterhin erfordert der neue Trennungsansatz weder eine anhaltende Aktivität der Sprecher noch eine Detektion der Aktivität des Nutzsprechers.

Ein funktionsfähiges Array ist die Voraussetzung für eine gute Lokalisationsschätzung und Quellentrennung. Im letzten Teil der Arbeit werden zwei eng verwandte Algorithmen vorgestellt, mit denen die Funktionstüchtigkeit des Arrays festgestellt werden kann. Das Ziel ist die automatische Detektion fehlerhafter Mikrofone durch *in situ* Tests und der Ausschluss von Signalen beeinträchtigter Mikrofone von der weiteren Verarbeitung. Zusätzlich können die Algorithmen zur online-Kalibrierung des Amplitudengangs der funktionierenden Mikrofone des Arrays verwendet werden.

Zum Abschluss der Arbeit werden neu aufgeworfene Fragestellungen und mögliche zukünftige Forschungsaspekte vorgestellt.

Resume – Nilesh Madhu

Personal	Born on 3rd February, 1979 in Mumbai (Bombay), India
Citizenship	Indian
Schooling	
06.1983 – 04.1994	Primary schooling St. Mary's High School Mumbai, India
06.1994 – 04.1996	Pre-university college SIES College of Arts, Science and Commerce Mumbai, India
University education	
06.1996 – 04.2000	Bachelor of Engineering (Electrical) University of Bombay Mumbai, India
08.2000 – 10.2002	Master of Science (Communications Engineering) Technische Universität München Munich, Germany
Internships	
08.2001 – 10.2001	Siemens AG, CT IC 3 (Corporate Technology, Security) Munich, Germany
Work experience	
11.2002 – 09.2003	Research engineer, Institut für Nachrichtentechnik Technische Universität Braunschweig Braunschweig, Germany
10.2003 – 02.2009	Research engineer, Institut für Kommunikationsakustik Ruhr-Universität Bochum Bochum, Germany
07.2007 – 10.2007	Research intern, Speech Technology Group Microsoft Research Redmond, USA

Microphone array technology is a valued tool in many acoustics based applications, some of which are already visible in our day-to-day life. These include the use of microphone arrays for improving human-machine interfaces, video-conferencing, noise-reduction in hands-free systems in automobiles, etc. The key issues in these applications are the localization and enhancement of the desired source and the suppression of unwanted interference, and these issues form the focus of this book.

We first introduce a localization taxonomy and provide an exhaustive overview of the contemporary state-of-the-art algorithms for source localization, drawn from different application areas deriving, in the process, the relations between these approaches. We then present design considerations for localization algorithms and highlight the need to take into account all the application specific characteristics and constraints.

We next address the use of source localization in the problem of active speaker separation and enhancement. After presenting an overview of the state-of-the-art, we develop a versatile localization based framework for source separation that allows us to realize a variety of robust algorithms for target speaker enhancement.

As a good localization estimate or separation performance is contingent upon a functional microphone array, we examine, in the last part, algorithms for determining the health of the array. The goal is to perform *in situ* self-tests and remove, if required, degraded microphones from further processing. In addition to detecting degradation, the algorithms may also be used to perform online gain calibration of the healthy microphones of an array — thereby reducing the effects of gain-mismatch in practical applications.

Nilesh Madhu has a Bachelors degree in Electrical Engineering from the University of Bombay and a Master of Science degree in Communications Engineering from the Technische Universität München. During his stay at the Institute of Communication Acoustics, Ruhr-Universität Bochum, he has been involved in research in the field of digital signal processing, with particular emphasis on audio signals. His primary interests are in the development and implementation of algorithms for source localization and enhancement, acoustic echo control, beamformer design and (semi-) blind signal separation.

DER ANDERE VERLAG ISBN: 978-3-89959-969-5 +++EUR[D] 35.90