# Cross-relation based blind identification of acoustic SIMO systems and applications

by

Mathieu Hu

A Thesis submitted in fulfillment of requirements for the degree of
Doctor of Philosophy of Imperial College London

Speech and Audio Processing Group
Department of Electrical and Electronic Engineering
Imperial College London

2017

"Neque porro quisquam est, qui dolorem ipsum, quia dolor sit, amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt, ut labore et dolore magnam aliquam quaerat voluptatem."

Cicero, *De Finibus Bonorum et Malorum*

# Abstract

Speech signals captured by microphones placed at a distance from the speaker are corrupted by reverberation, i.e. sound waves reflected off hard surfaces such as walls and objects. The spectral distortion caused by reverberation drastically decreases the performance of automatic speech recognition systems and may degrade the intelligibility and the quality of speech for human listeners. The increased use of devices controlled by distant speech therefore induces the need for dereverberation.

A possible approach to dereverberation is that of system equalization, which consists of the blind estimation of the room impulse responses from noisy reverberant signals followed by an inversion of these impulse responses. This thesis investigates the first part of this two-stage approach. The cross-relation method is adopted and exploited in two different ways.

The first way follows the adaptive filter framework, which was first introduced in the context of blind identification of room impulses responses in the Multi-Channel Least Mean Square. By considering a block update of this stochastic gradient algorithm, a noise robust algorithm is developed. The convergence rate of the resulting algorithm is then increased by using a locally optimal adaptive step-size. The cross-relation, expressed in the frequency domain, is then shown to contain the transfer function relating any of the microphone to a reference microphone. This relative transfer function can be used to reduce the number of variables to be estimated.

However, the performance of the previous methods severely degrades when realistically long room impulse responses are considered. An alternative interpretation of the cross-relation, from an annihilation filter perspective, is therefore explored. The resulting

algorithm is shown to be able to estimate room impulse responses of thousands of taps.

From a more practical perspective, the use of room impulses estimated at a poor accuracy is investigated for the problem of speaker diarization. The spatial information captured in the direct-to-reverberant ratio is shown to be robust to high levels of errors in the estimated room impulse responses. Blindly estimated direct-to-reverberant ratios combined with speech features in a single-channel diarization system are shown to provide additional information, which improves the performance of the diarization system.

# Acknowledgment

I would like to express my gratitude to the following people, without whom this thesis would not exist:

- First and foremost, my supervisor Dr. Patrick A. Naylor, for his patience, his support and the time he devoted to guide and nurture me throughout these years. Without him, the research presented here would not have been possible.
- Prof. Dr. Ir. Simon Doclo, who co-supervised me and guided my research during my 6 month secondment at the University of Oldenburg.
- Dr. Dushyant Sharma, who co-supervised me during my 6 month secondment at Nuance Communications, Inc. in Marlow and showed me the different aspects of a career in industry.
- Mr. Mike Brookes for inspiring and fruitful discussions regarding various research ideas.
- Ms. Anne Hough and Charlotte Grady for lightening the burden of administrative formalities.
- The European Unions Seventh Framework Programme (FP7/2007-2013), grant agreement n° ITN-GA-2012-316969 and the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465 for financial support.
- My friends and colleagues for their moral support and the great times we shared.

- Last but not least, I am forever grateful to my mother for her support and for giving me the freedom to pursue my own path despite the uncertainties lying along the way.

# Contents

**Abstract** **3**

**Acknowledgment** **5**

**Contents** **7**

**List of Figures** **10**

**List of Tables** **13**

**List of Notation** **14**

**Statement of Originality** **18**

**Abbreviations** **20**

**Chapter 1. Introduction** **22**

1.1 Motivation to remove reverberation . . . . . 22

1.2 Possible approaches to dereverberation . . . . . 23

1.3 Motivation for BSI . . . . . 24

**Chapter 2. Problem analysis** **26**

2.1 Overview . . . . . 26

2.2 Introduction and literature review . . . . . 26

2.3 Problem formulation and evaluation metric . . . . . 28

2.3.1 System model . . . . . 28

2.3.2 Room impulse response characteristics . . . . . 29

2.3.3 Evaluation metric: the Normalized Projection Misalignment . . . . . 34

2.4 The LS method for BSI . . . . . 36

2.4.1 The cross-relation residual . . . . . 36

2.4.2 Identifiability conditions . . . . . 37

2.4.3 LS approach to BSI . . . . . 38

2.4.4 Stability of the solution to additive noise . . . . . 40

2.5 Adaptive filter approach to BSI . . . . . 44
2.5.1 The MCLMS algorithm . . . . . 44
2.5.2 The NMCFLMS algorithm . . . . . 46
2.5.3 Misconvergence of the adaptive algorithms . . . . . 50
2.6 Summary and scope of the thesis . . . . . 52

**Chapter 3. SIMO structure based blind system identification 54**
3.1 Overview . . . . . 54
3.2 Introduction and literature review . . . . . 54
3.3 Fixed step-size block-based Rayleigh quotient . . . . . 58
3.3.1 Algorithm derivation . . . . . 58
3.3.2 Simulations . . . . . 61
3.3.3 Results . . . . . 61
3.4 Adaptive step-size extension . . . . . 64
3.4.1 Algorithm derivation . . . . . 64
3.4.2 Simulation . . . . . 65
3.4.3 Result . . . . . 66
3.5 RTF constrained BSI . . . . . 69
3.5.1 Cross-relation error in the frequency domain . . . . . 70
3.5.2 Parametrization of the RIR using the RTF . . . . . 73
3.5.3 Full-band RIR estimation . . . . . 74
3.5.4 Implementation details . . . . . 77
3.5.5 Experimental setup and evaluation . . . . . 78
3.5.6 Experimental results . . . . . 80
3.6 Summary . . . . . 83

**Chapter 4. Regularized blind system identification 85**
4.1 Overview . . . . . 85
4.2 Introduction . . . . . 85
4.3 Spectral characteristics of the residual at the solution . . . . . 87
4.3.1 Annihilation filter . . . . . 87
4.3.2 Spectrum of the residual at the solution . . . . . 88
4.3.3 Additive white Gaussian noise case . . . . . 88
4.3.4 Speech corrupted by stationary additive noise . . . . . 92
4.3.5 Experimental results . . . . . 95
4.3.6 Discussion . . . . . 101
4.4 Regularization . . . . . 102
4.4.1 Early reflections . . . . . 103
4.4.2 Late reflections . . . . . 106

4.4.3 Summary of the Regularized Sequential Spectral Flatness (RSSF) cost function . . . 107
4.4.4 Simulation setup . . . 107
4.4.5 Simulation results . . . 110
4.5 Summary . . . 114

**Chapter 5. Environment aware speaker diarization** **118**
5.1 Overview . . . 118
5.2 Introduction . . . 118
5.3 Spatial feature based diarization . . . 120
5.3.1 Signal model . . . 120
5.3.2 The overall diarization system . . . 120
5.3.3 Spatial feature extraction of the baseline . . . 121
5.3.4 Spatial feature extraction of the proposed method . . . 121
5.3.5 Feature labeling . . . 122
5.3.6 Search for the optimal spatial feature . . . 123
5.3.7 Diarization experiment setup . . . 125
5.3.8 Experimental results . . . 129
5.3.9 Discussion . . . 131
5.4 Single-channel environment-aware speaker diarization . . . 131
5.4.1 The information bottleneck principle . . . 132
5.4.2 Feature specification and combination . . . 133
5.4.3 Pre- and post-processing . . . 135
5.4.4 Diarization systems . . . 135
5.4.5 Experimental setup . . . 135
5.4.6 Evaluation . . . 137
5.4.7 Results . . . 138
5.5 Summary and discussion . . . 142

**Chapter 6. Conclusion** **143**
6.1 Summary . . . 143
6.2 Guidelines for future research on blind identification of SIMO acoustic systems 146

# List of Figures

2.1 Block diagram of a SIMO system. . . . . 28
2.2 EDC of a RIR from the database described in [53]. The fitted straight line relied on a $T_{60}$ estimated using [52]. . . . . 31
2.3 Room impulse responses from the database described in [53]. . . . . 31
2.4 Square magnitude of a RIR from the database [53]. . . . . 32
2.5 Zeros of the considered impulse responses. . . . . 34
2.6 Geometric illustration of the NPM in a 2D space. . . . . 35
2.7 Variations of Cond($\boldsymbol{h}$) and the NPM of the estimated stacked RIRs with $L$. 43
2.8 Ground truth and estimated RIRs using the LS method for $L = 800$ and an SNR of 60 dB. . . . . 44
2.9 Condition number of the sought solution for each position and realization of the source. The vertical dashed lines indicate a change in the source location. 45
2.10 Misconvergence of NMCFLMS. . . . . 51

3.1 Box plot of the mean NPMs after the first 100 s of signal and $M$=2. . . . . 62
3.2 Box plots of the mean NPMs after the first 100 s of signal and $M$=5. . . . . 63
3.3 Mean NPM for each algorithm for SNR=15 dB, $L$=500 and $M$=5 together with the 1$^{\text{st}}$ and 9$^{\text{th}}$ decile (shown in dotted lines). . . . . 63
3.4 Average NPM against time, no additive noise and $L$=128. . . . . 67
3.5 Average NPM against time, SNR=30 dB and $L$=128. . . . . 67
3.6 Average NPM against time, no additive noise and $L$=256. . . . . 68
3.7 Average NPM against time, SNR=30 dB and $L$=256. . . . . 68
3.8 Lower bound of NPM when the convolution theorem holds. . . . . 81
3.9 Effect of windowing on the lower bound of achievable NPM. . . . . 82
3.10 NPM of the estimated RIRs with a 100 ms white noise input signal and various levels of additive noise. . . . . 83

3.11 NPM of the estimated RIRs with a 600 ms white noise input signal and various levels of additive noise and $L$=2400. . . . . 83

4.1 Residual at the RIRs estimated via the RNMCFLMS after 3 s of reverberant noisy speech for $L = 300$ and an SNR of 30 dB. . . . . 93
4.2 Spectrogram of the residual at the solution. . . . . 94
4.3 Spectrogram of the input signal and residual at the RIRs estimated by SSF for $L = 300$ samples and an SNR of 30 dB. . . . . 97
4.4 NPM against time for the estimated RIRs obtained via the RNMCFLMS. . 98
4.5 Optimality and NPM of the estimated RIRs obtained from 1 s long noisy reverberant speech signals for $L = 200$. . . . . 99
4.6 Optimality and NPM of the estimated RIRs obtained from 1 s long noisy reverberant speech signals for $L = 500$. . . . . 100
4.7 Optimality and NPM of the estimated RIRs obtained from 1 s long noisy reverberant speech signals for $L = 2400$. . . . . 100
4.8 NPM of the estimated RIRs obtained from 3 s long noisy reverberant speech signals. . . . . 101
4.9 Ground truth and estimated RIRs from the first second long reverberant noisy speech described in Section 4.3.4 with $L = 2400$. The estimated RIRs were obtained by minimizing $J_{\text{SSF}}$. . . . . 102
4.10 Image method generated RIR truncated to $L = 512$ for $f_s = 8$ kHz. . . . . 103
4.11 Image method generated RIR truncated to $L = 512$ taps and for $f_s = 8$ kHz to the power 2. . . . . 103
4.12 Performance of the RNMCFLMS on 1 s long reverberant noisy signals for different step-size for an SNR of 25 dB (▬) and an SNR of 15 dB (▭), $L = 200$. . . . . 110
4.13 Performance of the RNMCFLMS on 1 s long reverberant noisy signals for different step-size for an SNR of 25 dB (▬) and an SNR of 15 dB (▭), $L = 1600$. . . . . 111
4.14 Variation of the median NPM of the RIRs estimated via RSSF with different trade-off parameters for $L = 500$. . . . . 112

4.15 RIRs estimated via the regularized spectral flatness cost function from the first second of the reverberant noisy speech described in Section 4.3.4 with $L = 2400$. . . . . 113
4.16 Spectrogram of the residual for $L = 2400$ and an SNR of 30 dB. . . . . 113
4.17 Performance of RSSF compared against the RNMCFLMS for $L = 500$ and various SNRs. . . . . 115
4.18 Performance of RSSF compared against the RNMCFLMS for $L = 2400$ and various SNRs. . . . . 115
4.19 Speech waveform and achieved NPMs for various lengths and an SNR of 50 dB. . . . . 116
4.20 Histogram of the NPM for SNR=30 dB and $L$=2400. . . . . 116

5.1 Block diagram of the diarization system. . . . . 121
5.2 Scatter plot of the $C_5$ and $C_{50}$ features for each speaker for an NPM of -5 dB. 125
5.3 Example speech from simulated meeting. . . . . 126
5.4 Variations of the DER and Bhattacharyya score with the NPM. . . . . 129
5.5 Scatter plot of the $C_5$ features for NPM=-5 dB and $M = 2$. The feature points corresponding to "2 speakers" refer to features extracted from frames containing 2 speakers, although these speakers were not active at the same time. . . . . 130
5.6 Box diagram of the DER obtained from 100 different speaker locations. The estimated stacked RIR had an NPM of -10 dB. . . . . 131
5.7 Block diagram of the proposed speaker diarization system. . . . . 132
5.8 DER of the development set as a function of DRR weight ($P_{\mathrm{DRR}} = 1 - P_{\mathrm{MFCC}}$). 138
5.9 Relative improvement in DER and speaker time error by inclusion of DRR features. . . . . 140
5.10 Estimated DRR along with the ground truth speaker identity. . . . . 141

# List of Tables

3.1 Rayleigh quotient based algorithms for BSI. . . . . 60

4.1 Variations of the median NPM, in dB, with the duration of the duration of the analysis frame for the proposed method. . . . . 101

4.2 Variations of the median, first and third quartiles of the NPM achieved by the RNMCFLMS with respect to the position of the microphone array for $L = 200$ and an SNR of 25 dB. . . . . 110

5.1 Bhattacharyya distance between the spatial feature clusters for various levels of NPM. . . . . 124

5.2 Performance of the diarization system in terms of speaker change detection. 130

5.3 $T_{60}$ in s and DRR in dB for the *near* and *far* positions in each of the three rooms. . . . . 136

5.4 Mean DER of the baseline and the proposed method for the development set. 138

5.5 Standard deviation in speaker error time of the baseline and proposed systems for the evaluation set together with the RMSD of the estimated DRR. 139

5.6 Mean speaker error time broken down by gender for the evaluation set. . . . 141

# List of Notation

### General notations

| | |
|---|---|
| $a$ | Scalar |
| $\boldsymbol{A}$ | Matrix with real entries |
| $\boldsymbol{a}$ | Vector with real entries |
| $\underline{\boldsymbol{A}}$ | Matrix with complex entries *a priori* |
| $\underline{\boldsymbol{a}}$ | Vector with complex entries *a priori* |
| $\mathfrak{A}$ | Subspace spanned by the columns of $\boldsymbol{A}$ |
| $a[n]$ | Time series indexed by $n$ |

### Operators

| | |
|---|---|
| $\boldsymbol{\nabla}\{\}$ | Gradient of the argument |
| $\frac{\partial\{\cdot\}}{\partial\{\cdot\}}$ | Partial derivative |
| $\mathfrak{A}^{\perp}$ | Orthogonal complement of $\mathfrak{A}$ |
| $\mathcal{P}^{\perp}_{\mathfrak{A}}(\boldsymbol{a})$ | Orthogonal projection of $\boldsymbol{a}$ onto $\mathfrak{A}$ |
| $\{\cdot\}^*$ | Complex conjugate |
| $*$ | Convolution sum |
| $\{\cdot\}^H$ | Hermitian transpose of the argument |
| $\{\cdot\}^{-1}$ | Inverse of the square matrix given in argument |
| $\|\|\cdot\|\|_2$ | Euclidean norm |
| $\|\cdot\|$ | Modulus |
| $\odot$ | Hadamard product |
| $\otimes$ | Kronecker product |
| $\mathrm{rank}\{\cdot\}$ | Rank of the matrix given in argument |
| $\mathrm{span}\{\cdot\}$ | Linear subspace spanned by the set of vectors given in argument |
| $\{\cdot\}^T$ | Transpose the argument |
| $\mathbb{E}\{\cdot\}$ | Expectation operator |
| $\boldsymbol{\mathcal{D}}_{\{\cdot\}}$ | Diagonal matrix, the diagonal entries of which are given by the argument |
| $\boldsymbol{\mathcal{H}}_{A\times B}\{\cdot\}$ | Hankel matrix of size $A \times B$ created from the entries of the vector given in the argument |
| $\boldsymbol{\mathcal{T}}_{A\times B}\{\cdot\}$ | Toeplitz matrix of size $A \times B$ created from the entries of the vector given in the argument |

### Symbols and variables

| | |
|---|---|
| $\mathbb{C}$ | Set of complex numbers |
| $\mathbb{R}$ | Set of real numbers |
| $\mathbf{0}_{a_1 \times a_2}$ | Null matrix of size $a_1 \times a_2$ |
| $\underline{\boldsymbol{\mathcal{A}}}_K$ | Mapping between the spectrum of a real signal and its representation consisting of real numbers |
| $\mathscr{A}_y$ | Set of input variables |
| $\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p$ | Mapping between the space of conjugate symmetric vector of size $2p$ built over real numbers and $\mathbb{R}^{2p}$ |
| $B$ | Number of samples in a frame |
| $b$ | Block index |
| $\boldsymbol{b}_k$ | Canonical basis vector of a real vector space with a 1 at the $k^{\text{th}}$ entry |
| $C_{50}$ | Clarity index |
| $\boldsymbol{\mathcal{E}}_A$ | Extension matrix which adds $A$ zeros at the end of a vector |
| $e^{\{\cdot\}}$ | Exponential function |
| $\underline{\boldsymbol{\mathcal{F}}}_A$ | Discrete Fourier transform matrix of size $A \times A$ |
| $f_s$ | Sampling Frequency |
| $\underline{\hat{\boldsymbol{H}}}$ | Frequency domain matrix imposing the RTF |
| $h[n]$ | Time series corresponding to the RIR |
| $\boldsymbol{h}$ | Room impulse responses stacked in a column vector |
| $h_m(l)$ | Amplitude of the RIR at the $m^{\text{th}}$ microphone and time sample $l$ |
| $\boldsymbol{h}_m$ | Room impulse response between the source and the $m^{\text{th}}$ microphone |
| $\underline{\hat{\boldsymbol{h}}}$ | $\underline{\hat{\boldsymbol{h}}}_m$ stacked in a column vectors for all $m$ |
| $\hat{\boldsymbol{h}}$ | Estimate of $\boldsymbol{h}$ |
| $\hat{\boldsymbol{h}}_m(n)$ | Estimate of $\boldsymbol{h}$ at the $n^{\text{th}}$ iteration of an adaptive algorithm |
| $\underline{\hat{\boldsymbol{h}}}_m$ | DFT of $\hat{\boldsymbol{h}}_m$ |
| $\underline{\hat{\boldsymbol{h}}}_m(b)$ | DFT of $\hat{\boldsymbol{h}}_m(b)$ at the $b^{\text{th}}$ iteration of an adaptive algorithm |
| $\hat{h}_m(l)$ | Estimate of $h_m(l)$ |
| $h_m[n]$ | Time series corresponding to the estimated RIR at the $m^{\text{th}}$ signal |
| $\hat{\boldsymbol{h}}_m$ | Estimate of $\boldsymbol{h}_m$ |
| $\hat{\underline{\boldsymbol{h}}}(k)$ | Estimated RTF at the $k^{\text{th}}$ frequency bin |
| $\boldsymbol{\mathcal{I}}_A$ | Identity matrix of size $A \times A$ |
| $i, j, m$ | Microphone index |
| $J_{\{\cdot\}}$ | Cost function for the algorithm, the name of which is given in subscript |
| $\jmath$ | Square root of -1 |
| $K$ | Number of frequency bins |
| $k$ | Frequency bin index |
| $L$ | Impulse response length |
| $l$ | Room impulse response tap index |
| $\log_a$ | Base $a$ logarithm |
| log | Natural logarithm function |

| | |
|---|---|
| $M$ | Number of microphones |
| $N_r$ | Number of samples in the cross-relation residual |
| $N_s$ | Duration of the source signal in samples |
| $n$ | Time instant in sample |
| $n_{\text{off}}$ | Frame increment |
| $\mathscr{Q}_x$ | Set of relevance variables |
| $q_x$ | Relevance variable |
| $\boldsymbol{R}$ | Square sample cross-relation matrix computed from all the available sample |
| $\boldsymbol{R}(n)$ | Instantaneous approximation of $\boldsymbol{R}$ |
| $\bar{\boldsymbol{R}}(n)$ | Block average of $\boldsymbol{R}(n)$ |
| $\boldsymbol{R}_{\underline{\hat{\boldsymbol{H}}}}$ | RTF constrained square cross-relation matrix |
| $\boldsymbol{R}_{\boldsymbol{y}_i\boldsymbol{y}_j}$ | Sample covariance matrix between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ computed from all the available data |
| $\boldsymbol{R}_{\boldsymbol{y}_i\boldsymbol{y}_j}(n)$ | Instantaneous approximation of $\boldsymbol{R}_{\boldsymbol{y}_i\boldsymbol{y}_j}$ |
| $\boldsymbol{R}_{\boldsymbol{\nu}}$ | Rectangular sample cross-relation matrix applied to the noise signals across all the microphones and for all time instants |
| $\boldsymbol{R}_{\boldsymbol{\nu}}(n)$ | Instantaneous approximation of $\boldsymbol{R}_{\boldsymbol{\nu}}$ |
| $\boldsymbol{R}_{\boldsymbol{x}}$ | Rectangular sample cross-relation matrix applied to the reverberant noiseless signals across all the microphones and for all time instants |
| $\boldsymbol{R}_{\boldsymbol{x}}(n)$ | Instantaneous approximation of $\boldsymbol{R}_{\boldsymbol{x}}$ |
| $\boldsymbol{R}_{\boldsymbol{y}}$ | Rectangular sample cross-relation matrix applied to the reverberant noisy signals across all microphones and for all time instants |
| $\bar{\boldsymbol{R}}_{\boldsymbol{y}}(b)$ | Rectangular cross-relation matrix for frame $b$ |
| $\boldsymbol{R}_{\boldsymbol{y}}(n)$ | Instantaneous approximation of $\boldsymbol{R}_{\boldsymbol{y}}$ |
| $\boldsymbol{r}$ | Cross-relation residual taken across all pairs of microphones and time instants |
| $\boldsymbol{r}(n)$ | Cross-relation residual taken across all pairs of microphones for a time instant |
| $r_{ij}(n)$ | Cross-relation residual between microphones $i$ and $j$ at time instant $n$ |
| $r_{ij}[n]$ | Time series corresponding to the cross-relation residual between microphone $i$ and $j$ |
| $\boldsymbol{r}_{i,j}(n)$ | $r_{i,j}(n)$ and its $L-1$ preceding samples stored in a vector |
| $\bar{\boldsymbol{r}}_{i,j}(b)$, $\bar{\boldsymbol{r}}(b)$ | $r_{i,j}(n)$ and its $B-1$ preceding samples stored in a vector |
| $\bar{\boldsymbol{r}}_i(b)$ | $i^{\text{th}}$ subframe of $\bar{\boldsymbol{r}}(b)$ in frame $b$ |
| $\boldsymbol{\mathcal{S}}_e^A$ | Selection matrix which extracts the $A$ last entries of a vector |
| $\boldsymbol{\mathcal{S}}_s^A$ | Selection matrix which extracts the $A$ first entries of a vector |
| $s(n)$ | Anechoic signal sample at time instant $n$ |
| $s[n]$ | Time series corresponding to the anechoic signal |
| $\boldsymbol{s}(n)$ | $s(n)$ and its $L-1$ preceding samples stored in a vector |
| $T_{60}$ | Reverberation time |
| $\boldsymbol{W}$ | Rectangular matrix representing a window operation |
| $x_m(n)$ | Reverberant signal sample at the $m^{\text{th}}$ microphone and time instant $n$ |

| | |
|---|---|
| $x_m[n]$ | Time series corresponding to the reverberant noiseless signal at the $m^{\text{th}}$ microphone |
| $\boldsymbol{x}_m^L(n)$, $\boldsymbol{x}_m(n)$ | $x_m(n)$ and its $L-1$ preceding samples stored in a vector. Simplified to $\boldsymbol{x}_m(n)$ when the length is inferred from context |
| $\underline{\boldsymbol{Y}}(k)$ | Frequency domain cross-relation matrix at frequency bin $k$ |
| $y(n)$ | Reverberant noisy signal at time instant $n$ |
| $y_m[n]$ | Time series corresponding to the observed signal at the $m^{\text{th}}$ |
| $\boldsymbol{y}_m^L(n)$, $\boldsymbol{y}_m(n)$ | $y_m(n)$ and its $L-1$ preceding samples stored in a vector. Simplified to $\boldsymbol{y}_m(n)$ when the length is inferred from context |
| $\underline{\boldsymbol{y}}_m(n)$ | DFT of $\boldsymbol{y}_m(n)$ |
| $\underline{\hat{\alpha}}(k)$ | Estimated frequency weight to apply to an estimated RTF to compute the time domain RIR |
| $\underline{\hat{\boldsymbol{\alpha}}}$ | Estimated weight vector to be applied to an estimated RTF |
| $\hat{\boldsymbol{\alpha}}$ | Representation of $\underline{\hat{\boldsymbol{\alpha}}}$ by application of $\underline{\boldsymbol{\mathcal{A}}}_K$ |
| $\gamma_{\{\cdot\}}$ | Trade-off parameter for the regularization term the name of which is given in the argument |
| $\lambda_i\{\cdot\}$ | $i^{\text{th}}$ smallest eigenvalue of the matrix given in the argument |
| $\mu, \mu_{\{\cdot\}}$ | Step-size. The name of the corresponding algorithm is given in subscript when there is an ambiguity |
| $\nu_m(n)$ | Additive noise signal sample at time instant $n$ and microphone $m$ |
| $\nu_m[n]$ | Time series corresponding to the additive noise signal at the $m^{\text{th}}$ microphone |
| $\boldsymbol{\nu}_m^L(n)$, $\boldsymbol{\nu}_m(n)$ | $\nu_m(n)$ and its $L-1$ previous values stored in a vector. Simplified to $\boldsymbol{\nu}_m$ when the length is inferred from context |
| $\phi_{\text{ER}}(\cdot)$ | Regularization term penalizing non spiky early reflections |
| $\phi_{\text{LR}}(\cdot)$ | Regularization term penalizing non exponentially decaying tails |
| $\rho$ | Smoothing factor |
| $\sigma$ | Segment |
| $\boldsymbol{\theta}_p\{\cdot\}$ | Non-linear function elevating each entry of the argument to the power $p$ |
| $\boldsymbol{\theta}_p^2(\cdot)$ | Monomial of power $p$ applied to the early reflections of $\hat{\boldsymbol{h}}_m$ |
| $\mathcal{S}$ | Measure of sparsity, decreasing as the sparsity of the argument increases |
| $\boldsymbol{\xi}_\lambda$ | Eigenvector corresponding to the eigenvalue $\lambda$ |
| $\zeta$ | Decay rate of the exponential envelop of a Polack model room impulse response |

# Statement of Originality

As far as the author is aware, the following aspects of the thesis are believed to be original contributions:

1. The study of the smallest eigenvalue of the noiseless square cross-relation matrix as a way to quantify the robustness of the LS method to additive noise.

2. The formulation of the MCLMS and NMCFLMS as adaptive implementations of the LS method.

3. The use of a stochastic gradient descent relying on block estimate of the square cross-relation matrix for a noise robust BSI algorithm.

4. The use of locally optimal adaptive step-size relying on the nature of the problem in the context of adaptive BSI.

5. The exploitation of the cross-relation error in the frequency domain to estimate the RTF and the derivation of a the lower bound on the NPM that can be achieved using noisy RTFs.

6. The exploitation of the cross-relation in a annihilation filter framework for estimation of realistically long RIRs.

7. The use of DRR combined with speech features for an environment-aware single-channel speaker diarization system.

## Copyright declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to

copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

## List of publications

The following publications were produced during the time of this work:

1. M. Hu, D. Sharma, S. Doclo, *et al.*, "Speaker change detection and speaker diarization using spatial information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015

2. M. Hu, S. Doclo, D. Sharma, *et al.*, "Noise robust blind system identification algorithms based on a rayleigh quotient cost function," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sep. 2015

3. M. Hu, P. Peso Parada, D. Sharma, *et al.*, "Single-channel speaker diarization based on spatial features," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015

4. M. Hu, D. Sharma, S. Doclo, *et al.*, "Blind adaptive SIMO acoustic system identification using a locally optimal step-size," in *Proc. Audio Eng. Soc. (AES) Conf.: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, Leuven, Belgium, Feb. 2016

# Abbreviations

| | |
|---|---|
| **ASR** | Automatic Speech Recognition |
| **BNRQ** | Block-based Normalized Rayleigh quotient |
| **BSI** | Blind System Identification |
| **BURQ** | Block-based Unconstrained Rayleigh Quotient |
| **DER** | Diarization Error Rate |
| **DFT** | Discrete Fourier Transform |
| **DRR** | Direct-to-Reverberant Ratio |
| **EDC** | Energy Decay Curve |
| **FFT** | Fast Fourier Transform |
| **FIR** | Finite Impulse Response |
| **GCC-PHAT** | Generalized Cross-Correlation with Phase Transformation method of estimating TDOA |
| **GMM** | Gaussian Mixture Model |
| **HMM** | Hidden Markov Model |
| **IB** | Information Bottleneck |
| **LMS** | Least Mean Square |
| **LOG-BURQ** | Locally Optimal Gradient BURQ |
| **LS** | Least Squares |
| **MCFLMS** | Multi-Channel Frequency domain Least-Mean Square |
| **MCLMS** | Multi-Channel Least Mean Square |
| **MFCC** | Mel-frequency Cepstral Coefficients |
| **MINT** | Multiple-input/output INverse Theorem |
| **NIRA** | Non-Intrusive Room Acoustics |

| | |
|---|---|
| **NMCFLMS** | Normalized Multichannel Frequency Domain Least Mean Square |
| **NPM** | Normalized Projection Misalignment |
| **PLP** | Perceptual Linear Predictive |
| **PSD** | Power Spectral Density |
| **RIR** | Room Impulse Response |
| **RMSD** | Root-Mean-Square Deviation |
| **RNMCFLMS** | Robust NMCFLMS |
| **RSSF** | Regularized Sequential Spectral Flatness |
| **RTF** | Relative Transfer Function |
| **SF** | Spectral Flatness |
| **SIMO** | Single-Input-Multiple-Output |
| **SNR** | Signal-to-Noise Ratio |
| **SSF** | Sequential Spectral Flatness |
| **SURQ** | Sample-based Unconstrained Rayleigh Quotient |
| **TDOA** | Time Difference of Arrival |
| **UDA** | Unsupervised Discriminant Analysis |
| **VAD** | Voice Activity Detector |
| **VSS-UMCLMS** | Variable Step-Size Unconstrained MCLMS |

# Chapter 1

# Introduction

## 1.1 Motivation to remove reverberation

Capturing a sound source within an enclosed environment with a microphone placed at a distance typically results in recordings that are corrupted by reverberation, i.e. acoustic reflections against surfaces within the enclosure. While reverberation is desirable for music reproduction [5], [6], its presence may be detrimental to applications using distant speech.

The performance of speech processing algorithms may indeed decrease in reverberant environment.

The problem of noise reduction, for example, usually requires an estimation of the noise Power Spectral Density (PSD) [7]. However, classical estimators of the noise PSD such as the minimum statistic estimator [8] or the minimum mean squared error estimator [9] do not model reverberation. The mismatch between the model and a recorded distant speech may result in erroneous PSD estimations, which then reduces the performance of the speech enhancement methods [10].

The accuracy of Automatic Speech Recognition (ASR) systems is another example of systems sensitive to reverberation. Automatic speech recognition systems indeed usually rely on features extracted from anechoic or near anechoic speech utterances to transcribe new utterances, see [11] and references therein. However, in the case of distant speech, reverberation introduces a smearing of the spectrogram which results in a mismatch be-

tween the features extracted from the reverberant utterance and those from the training utterances [12]. That mismatch then causes a drastic decrease of the performance [13].

In addition to limiting the performance of speech processing algorithms, reverberation also affects the intelligibility of speech for human listeners. The listening tests carried out by Humes and Roberts [14] showed that both normal hearing and hearing impaired listeners were sensitive to high amount of reverberation.

The need for deverberation methods, i.e. methods removing reverberation from recordings, therefore arose with the increasing use of hearing aids, hand-free devices and distant speech recognition systems to communicate with personal digital assistants or robots [15]. To approach the problem of dereverberation, the reverberant noiseless signal is usually modeled as a convolution between an anechoic speech signal and a time-invariant Finite Impulse Response (FIR) filter representing the Room Impulse Response (RIR). The signal observed at a microphone is the sum of the reverberant signal and some additive noise uncorrelated with the source signal.

## 1.2 Possible approaches to dereverberation

Lebart, Boucher, and Denbigh [16], for example, extend the spectral subtraction method, originally developed for additive noise reduction [17], to dereverberation. This approach aims to remove the late reverberation from the recorded signal provided that the full-band reverberation time is estimated. Habets [18] included the frequency dependency of the reverberation time into the statistical model of [16] for an improved performance. However, because of the nature of the processing, spectral subtraction methods are known to possibly introduce musical noise, which may be more disturbing to the listener than the original distortion [19].

When more than one microphone is available, i.e. in a multi-channel setup, beamforming methods may be used to exploit spatial diversity and extract the signal coming from a certain direction [20]. Although originally formulated to handle uncorrelated additive noise, these methods have been adapted to take reverberation into account, e.g. [21].

However, these methods require a good estimate of the direction of the source, which may be difficult to obtain in a reverberant environment [22].

The problem of dereverberation can also be viewed as a channel equalization problem, where the RIR is to be inverted. In fact, assuming a noiseless multi-channel setup and known RIRs, Miyoshi and Kaneda proposed the Multiple-input/output INverse Theorem (MINT) algorithm in [23], which perfectly inverts the acoustic channels and hence achieve perfect dereverberation. Zhang, Habets, and Naylor [24] showed that the MINT algorithm was sensitive to estimation errors in the RIRs so that, in the presence of estimation errors, the equalized channels had stronger energy in the reverberation tail than the original impulse response. Kodrasi, Goetze, and Doclo [25] and Lim, Zhang, Habets, *et al.* [26] investigated the design of equalization algorithms robust to estimation errors. Kodrasi and Doclo then proposed an equalization algorithm robust to estimation errors in the RIRs, which does not amplify the additive noise present in the recorded signal in [27].

## 1.3 Motivation for BSI

However, the aforementioned algorithms all rely on the assumption that the RIRs are estimated to some degree of accuracy. This thesis therefore investigates the blind identification of RIRs, i.e. the estimation of the RIRs from the noisy reverberant signal without knowledge of the dry or noise signals. Although here motivated by the application of Blind System Identification (BSI) to channel equalization, it should be pointed out that RIRs find many more applications. The reflection patterns can be used for room inference [28], while the cross-correlation between the RIRs contains the Time Difference of Arrival (TDOA) [29]. Moreover, reverberation dependent quantities such as the clarity index [30] can be used in ASR systems trained on multi-condition data to significantly improve their performance [31].

In Chapter 2, the problem is formulated and analyzed. In particular, the fundamental differences between communication channels, which were considered the early development of BSI, and RIRs are outlined to emphasize the new challenges that need to be overcome

for the blind identification of RIRs.

Chapter 3 investigates the noise robustness issue without exploiting assumptions specific to the problem of blind identification of RIRs from reverberant noisy speech signals so that the developed algorithms could be applicable to different input signals or impulse responses of varying nature.

While the cross-relation is still at the heart of the work presented in Chapter 4, a fundamentally different characterization of the sought solution is investigated. The use of regularization terms for the estimation of realistically long RIRs is also investigated.

Chapter 5 investigates the use of spatial features robust to estimation noise in the RIRs for environment-aware speaker diarization.

The main results of the work presented in this thesis are summarized and discussed in Chapter 6. Possible directions for further development in blind identification of Single-Input-Multiple-Output (SIMO) acoustic systems are then outlined.

# Chapter 2

# Problem analysis

## 2.1 Overview

This chapter provides the background material together with a detailed analysis of the problem, which will lay the foundation of the work presented in the other chapters of this thesis.

The Least Squares (LS) method [32] together with the identifiability conditions are described and the robustness of the LS method against additive noise is analyzed. That analysis gives an insight into the intrinsic limitations of the LS method, which underlies state-of-the-art BSI algorithms. These algorithms, namely the Normalized Multichannel Frequency Domain Least Mean Square (NMCFLMS) and Multi-Channel Least Mean Square (MCLMS), are indeed shown to be strongly related to the LS method.

The problem of blind identification using the cross-relation is then put into the context of blind identification of RIRs. In particular, the challenges that arises in that specific setup are outlined.

## 2.2 Introduction and literature review

The problem of BSI is not only limited to dereverberation and indeed has potential applications in various fields. Blind system identification can be applied to image deconvolution

and exploration seismology [33]. In communications, for example, fast and accurate blind estimation of the communication channels would remove the need for pilot signals, which would then increase the throughput [34]. In fact, the literature on BSI of communication channels is rich, e.g. [32], [35], [36]. As it will be made clear in the following section, state-of-the BSI algorithms estimating RIRs are in fact strongly related to the method of [32], referred as the LS method.

In its early developments, the LS approach to BSI was referred to as a "deterministic" approach [37]. That method indeed does not require a statistical model of the dry signal, which is of particular interest for the blind identification of RIRs from reverberant noisy speech signals.

Other BSI methods based on second order statistics such as [38] or the linear prediction based methods such as [33], [39] require the estimation of either the covariance matrix or the autocorrelation sequence of the input signal over a duration at least as long as that of the impulse responses. When the blind identification of RIRs is considered, the input signal may be speech. Speech, however, is known to be quasi-stationary and ergodic only over segments of at most 80 ms [40]. Therefore, when RIRs of several hundreds of milliseconds are considered, accurate estimation of the autocorrelation function or the covariance matrix may not be possible [41, Chapter 9].

This chapter gives an in-depth description and analysis of state-of-the-art BSI algorithms for SIMO acoustic systems. To this end, the characteristics of the considered system together with an evaluation metric, relevant for the problem of channel equalization are given in Section 2.3. The cross-relation and the LS method together with its limitations are then described in Section 2.4. The MCLMS [42] and NMCFLMS [43] are then derived in a way to emphasize their links to the LS method in Section 2.5. Finally, Section 2.6 sets the scope of this thesis while summarizing the problem of BSI of SIMO acoustic systems in a list of challenges that need be overcome for practical implementation.

## 2.3 Problem formulation and evaluation metric

### 2.3.1 System model

A single active sound source captured by $M$ microphones can be modeled by the SIMO system illustrated in Fig. 2.1.

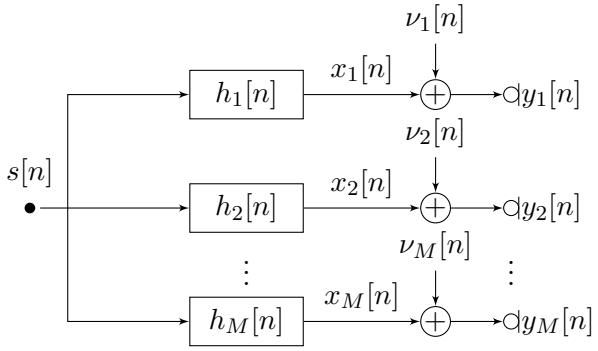



Figure 2.1: Block diagram of a SIMO system.

The time series $x_m[n]$ composed of reverberant signal samples, indexed by the time index $n$, at the $m^{\text{th}} \in \{1, 2, \ldots, M\}$ microphone is given by the convolution of the dry source signal $s[n]$ with the time-invariant RIR $h_m[n]$, which models the acoustic path from the source to the $m^{\text{th}}$ microphone. The time series $y_m[n]$ of observed samples is then given as $x_m[n]$ corrupted by some additive noise $\nu[n]$ uncorrelated with the source signal $s[n]$. In a matrix formulation, we have [44]

$$\boldsymbol{y}_m^L(n) = \boldsymbol{x}_m^L(n) + \boldsymbol{\nu}_m^L(n), \tag{2.1}$$

where $\boldsymbol{y}_m^L(n)$ is a vector of $L$ reverberant noisy samples given by $\boldsymbol{y}_m^L(n) = \begin{bmatrix} y_m(n) & y_m(n-1) & \ldots & y_m(n-L+1) \end{bmatrix}^T$, $\boldsymbol{x}_m^L(n)$ is a vector of $L$ reverberant samples given by $\boldsymbol{x}_m^L(n) = \begin{bmatrix} x_m(n) & x_m(n-1) & \ldots & x_m(n-L+1) \end{bmatrix}^T$ and $\boldsymbol{\nu}_m^L(n)$ is a vector of $L$ noise samples given by $\boldsymbol{\nu}_m^L(n) = \begin{bmatrix} \nu_m(n) & \nu_m(n-1) & \cdots & \nu_m(n-L+1) \end{bmatrix}^T$. The superscript $^L$ will be omitted whenever there is no ambiguity.

The reverberant signal is itself related to the input signal via

$$\boldsymbol{x}_m^L(n) = \boldsymbol{\mathcal{T}}_{L\times 2L-1}\{\boldsymbol{h}_m\}\boldsymbol{s}(n), \tag{2.2}$$

where $\boldsymbol{s}(n) = \begin{bmatrix} s(n) & s(n-1) & \dots & s(n-2L+2) \end{bmatrix}^T$, $\boldsymbol{\mathcal{T}}_{L\times 2L-1}\{\boldsymbol{h}_m\}$ is a Toeplitz matrix of size $L \times (2L-1)$ representing the convolution operation

$$\boldsymbol{\mathcal{T}}_{L\times 2L-1}\{\boldsymbol{h}_m\} = \begin{bmatrix} h_m(1) & h_m(2) & \cdots & h_m(L) & 0 & \cdots & 0 \\ 0 & h_m(1) & \cdots & h_m(L-1) & h_m(L) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & h_m(1) & h_m(2) & \cdots & h_m(L) \end{bmatrix} \tag{2.3}$$

and $\boldsymbol{h}_m = \begin{bmatrix} h(1) & h(2) & \cdots & h(L) \end{bmatrix}^T$.

The aim of BSI is to estimate $\boldsymbol{h}_m$, $m \in \{1, 2, \dots, M\}$, without knowledge of $\boldsymbol{x}_m^L(n)$, $\boldsymbol{\nu}_m^L(n)$ or $\boldsymbol{s}(n)$ but using $\boldsymbol{y}_m^L(n)$ only.

### 2.3.2 Room impulse response characteristics

In the context of dereverberation, the impulse responses to be estimated are of peculiar nature and significantly differ from what has been considered in the literature of blind identification of communication channels. Let us describe here some characteristics of RIRs.

#### Reverberation time

Perhaps, the most striking difference between acoustic and communication channels is the length of the considered channel. The length of a RIR is often determined by the reverberation time $T_{60}$ expressed in samples, e.g. [45], [46]. The reverberation time represents the duration of the decay of a sound signal produced in a room [47, Chapter 3]. The reverberation time does not depend on the source microphone configuration and depends only on the geometry of the room and the absorption coefficients of reflecting surfaces [30].

For typical office-sized rooms, the reverberation time varies between 0.1 s and 1 s [48]. In larger spaces, e.g. churches, the reverberation time can be of several seconds [49, Chapter 3]. Assuming a sampling frequency of 8 kHz, this means that the lengths of RIRs vary between 800 taps to several thousand taps. The number of unknowns to be estimated by the BSI algorithm is therefore of several thousands and strongly differs from communication systems, where each channel usually has a single digit number of taps, see e.g. [32], [50].

When the RIR is available, the reverberation time can be estimated from the Energy Decay Curve (EDC) [51] using methods such as [52]. The EDC, denoted by $\Xi\{\boldsymbol{h}_m\}$, is obtained from the RIR by backward integrating the energy of the RIR and expressing it in dB [52]:

$$\Xi\{\boldsymbol{h}_m\}(n) = 10\log_{10}\left(\frac{\sum_{l=n}^{L} h_m^2(l)}{\sum_{l=1}^{L} h_m^2(l)}\right). \tag{2.4}$$

Figure 2.2 shows the EDC of the first channel of the set of RIRs contained in the file named `Chromebook_503_2_RIR.wav` from the database recorded by Eaton, Gaubitch, Moore, *et al.* and described in [53]. After the initial drop of energy, the observed linear decay can be used to estimate the reverberation time [30]. If the variations of the reverberation time with respect to frequency are desired, a reverberation per frequency band can be obtained by applying the previously described method to subband RIRs.

Blind $T_{60}$ estimators has been the subject of active research and methods robust to additive noise have been developed, see [54]–[56].

### Stochastical model of the late reflections

Artificial reverberation, which has been an active area of research for more than 50 years [6], provides us with models, which may be relevant for BSI.

Room impulse responses can be decomposed into an early and an reflection part [57]. The early reflections represent individual reflected sound waves while the late reflections, which are the superposition of an increasing number of reflections, can be statistically modeled [57], [58]. The time instant from which the statistical model holds is referred to

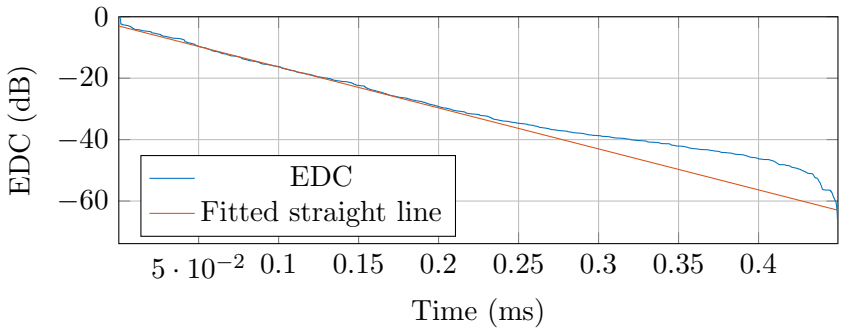
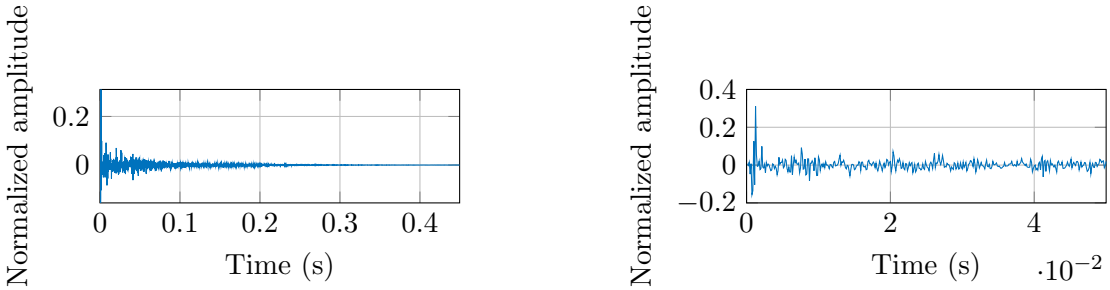



Figure 2.2: EDC of a RIR from the database described in [53]. The fitted straight line relied on a $T_{60}$ estimated using [52].




(a) Whole impulse RIR.

(b) Early reflections.

Figure 2.3: Room impulse responses from the database described in [53].

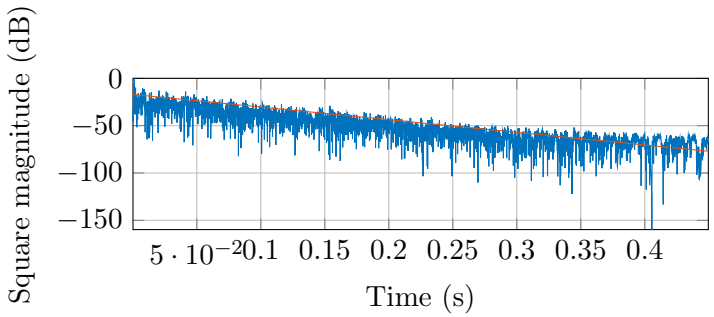



Figure 2.4: Square magnitude of a RIR from the database [53].

as the *mixing time* [59].

The Polack model [60] relates the reverberation time to the decay rate of the reverberation tail:

$$h(l) = \psi(l)e^{-\zeta l}, \tag{2.5}$$

where $\psi(l)$ are independent identically distributed samples drawn from a zero-mean Gaussian distribution and

$$\zeta = \frac{3\log(10)}{T_{60}}. \tag{2.6}$$

That exponential decay envelop is observed in Fig. 2.4, which shows the power of the RIR displayed in Fig. 2.3a in the log domain. In Fig. 2.2 and Fig. 2.4, the straight red line is given by

$$f(l) = \frac{-20}{\log(10)}\zeta l + y_{\text{off}},$$

where the offset $y_{\text{off}}$ has been arbitrarily chosen to make the straight line and curve coincide at the beginning and the value of $T_{60}$ was estimated from the RIR using the method described in [52]. The slope of that straight line is obtained by taking $10\log_{10}(\cdot)$ of the squared envelop of (2.5).

### Early reflections

Unlike the late reflections, which are statistically modeled, the early reflections strongly depend on the geometry of the room and the source/microphone position [6].

Schroeder used the sparse structure of the early reflections for his artificial reverbera-

tion algorithm presented in [61]. Kowalczyk, Habets, Kellermann, *et al.* and later Crocco and Del Bue also used a sparse model of the early reflections to estimate TDOAs from blindly estimated under-modeled RIRs in [62] and [63] respectively.

However, as illustrated in Fig. 2.3b, the early reflections are not strictly sparse in the sense that the reflections are separated by smaller valued taps and not by exact zeros.

### Spatial diversity

The performance of the adaptive BSI algorithms and that of the LS method, in the presence of noise, are believed to be related with the distance between the zeros of the impulse responses [34], [64]. The accuracy of the cross-relation based BSI methods has, in fact, been shown to strongly depend on the zeros of the impulse responses in early work so that impulse responses with near common zeros have been characterized as "ill-conditioned" [65].

In [36], the degradation in performance was experimentally observed by blindly estimating 3 tap long impulse responses, the zeros of which were imposed. An adaptive algorithm robust to near common zeros was proposed in [66]. The proposed algorithm first estimates the non-common part using the MCLMS then applies a Least Mean Square (LMS) filter to a partially dereverberated signal to estimate the common filter. The effect of noise, however, is not considered. While [36] and [66] provide results for a few channels, Schmid and Enzner give extensive simulation results demonstrating how the performance of the adaptive algorithms degrades in the presence of additive noise and near common zeros in [67]. Let us now analyze the zeros of a RIR.

Assuming a Polack model of the RIRs, the theorem proved in [68] states that the zeros of the RIRs cluster around the unit circle with a uniformly distributed phase on the interval $[0, 2\pi)$. In other words, although exact common zeros may not occur, near common zeros are unavoidable. Furthermore, the distribution of the zeros of each channel challenges the successful application of the algorithms proposed in [66] and [69] as these algorithms indeed assumed that the RIRs could be decomposed into two components: one containing the near common zeros and another component with significantly different

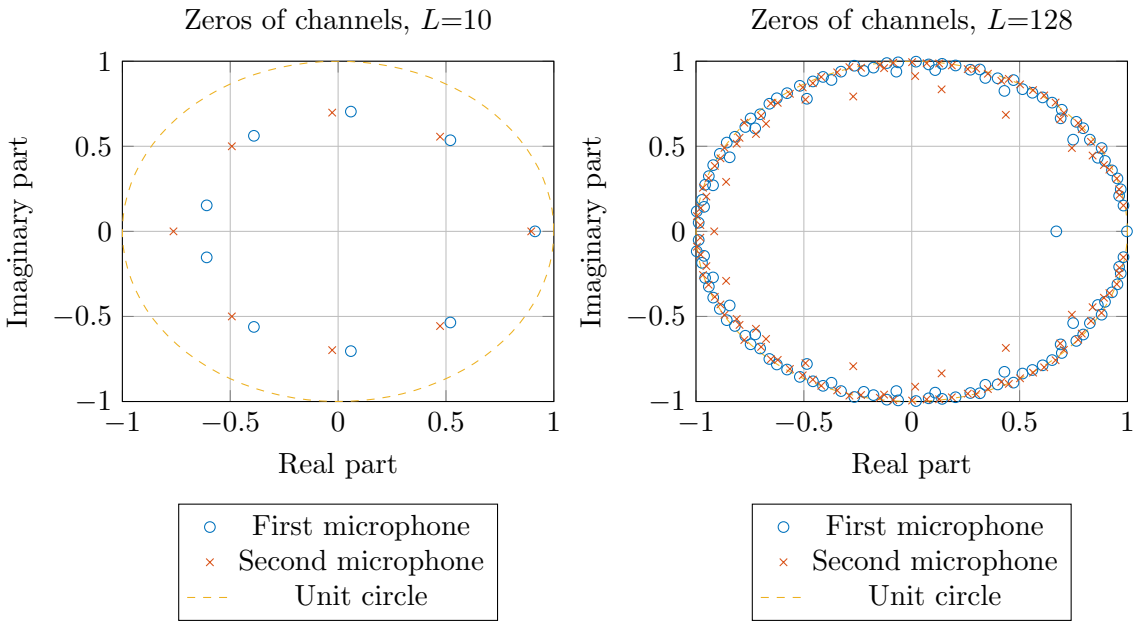



Figure 2.5: Zeros of the considered impulse responses.

zeros.

Figure 2.5 shows that, as $L$ increases from 10 to 128, the zeros of the impulses responses cluster around the unit circle. That increase in length introduced the presence of near common zeros. However, in this case, there is no common zero as, in the absence of noise, the LS method [32] was able to identify the RIRs with a Normalized Projection Misalignment (NPM) of -241 dB and -192 dB for $L = 10$ and $L = 128$ respectively. It should be noted that a pairwise comparison between the zeros of each pair of channels is not a reliable method to determine if the identifiability conditions are satisfied as the computation of the zeros of a RIR is known to be very sensitive to small changes in the coefficients of the RIR [70].

### 2.3.3 Evaluation metric: the Normalized Projection Misalignment

Before presenting the main BSI algorithms, let us describe the metric used in this thesis to evaluate the performance of BSI algorithms.

The primary purpose of BSI is to provide estimates of the impulse responses that can be used to deconvolve the observed signal [71]. For this purpose, any non-zero vector

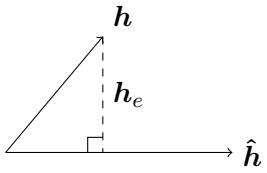



Figure 2.6: Geometric illustration of the NPM in a 2D space.

co-linear with the impulse responses is sufficient [23].

The Normalized Projection Misalignment (NPM) is a scale independent measure, usually expressed in dB, that computes the squared norm of the normalized orthogonal error between two vectors [30]:

$$\mathrm{NPM}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = 10 \log_{10} \left( \frac{\left\| \boldsymbol{h} - \frac{\hat{\boldsymbol{h}}^T \boldsymbol{h}}{\hat{\boldsymbol{h}}^T \hat{\boldsymbol{h}}} \hat{\boldsymbol{h}} \right\|^2}{\boldsymbol{h}^T \boldsymbol{h}} \right), \tag{2.7}$$

where $\boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_1^T & \boldsymbol{h}_2^T & \cdots & \boldsymbol{h}_M^T \end{bmatrix}^T$ is the stacked RIRs and $\hat{\boldsymbol{h}} = \begin{bmatrix} \hat{\boldsymbol{h}}_1^T & \hat{\boldsymbol{h}}_2^T & \cdots & \hat{\boldsymbol{h}}_M^T \end{bmatrix}^T$ the stacked estimated RIRs with $\hat{\boldsymbol{h}}_m^T$ being an estimate of $\boldsymbol{h}_m$.

The definition of the NPM, given in (2.7), essentially states that $\mathrm{NPM}(\boldsymbol{h}, \hat{\boldsymbol{h}})$ is the ratio between the norm of the orthogonal projection of $\boldsymbol{h}$ onto the orthogonal complement of $\hat{\mathfrak{h}}$, the subspace spanned by $\hat{\boldsymbol{h}}$, and the Euclidean norm of $\boldsymbol{h}$. Therefore, the NPM takes value between $-\infty$ and 0 dB. When $\boldsymbol{h}$ is co-linear with $\hat{\boldsymbol{h}}$, $\mathrm{NPM}(\boldsymbol{h}, \hat{\boldsymbol{h}})$ is equal to $-\infty$, whereas any vector $\hat{\boldsymbol{h}}$ orthogonal to $\boldsymbol{h}$ will lead to $\mathrm{NPM}(\boldsymbol{h}, \hat{\boldsymbol{h}})$ of 0 dB.

Figure 2.6 gives an illustration of $\mathrm{NPM}(\boldsymbol{h}, \hat{\boldsymbol{h}})$ for the 2 dimensional case. Using the notations of Fig. 2.6, $\mathrm{NPM}(\boldsymbol{h}, \hat{\boldsymbol{h}})$ is equal to the Euclidean norm of $\boldsymbol{h}_e$ divided by that of $\boldsymbol{h}$, i.e. the distance from $\boldsymbol{h}$ to the subspace spanned by $\hat{\boldsymbol{h}}$.

The numerical value of the NPM has been used to design RIR-dependent algorithms robust to estimation errors. For dereverberation, the MINT algorithm requires $-\infty$ dB of NPM [24], while the relaxed multichannel least squares with envelop constraint [26] has been shown to perform well for an NPM of -33 dB. In Chapter 5, an application requiring NPMs of -5 dB will be presented.

## 2.4 The LS method for BSI

This section describes the cross-relation and its exploitation in the LS method. The LS method takes a deterministic approach to BSI in the sense that the (unknown) dry input signal does not need to be statistically modeled. This is particularly attractive for the problem of BSI of RIRs as, in practice, the dry input signal can be speech or music, which is not stationary [72]. The in-depth analysis of the LS will also allow us to better understand the MCLMS [42] and the NMCFLMS [43], which were developed for the blind estimation of RIRs, as these will be shown to be strongly connected to the LS method. Similarly to the derivations of [32], the derivations are given assuming that no additive noise is present. The effect of additive noise will be analyzed and illustrated with examples shortly after the description of the LS method.

### 2.4.1 The cross-relation residual

Since the convolution sum is commutative, for any pair of microphones $i, j \in \{1, 2, \ldots, M\}$, we have

$$y_i[n] * h_j[n] = s[n] * h_i[n] * h_j[n] = s[n] * h_j[n] * h_i[n] = y_j[n] * h_i[n]. \tag{2.8}$$

For a given pair $(i, j)$ of microphones, estimates $\hat{\boldsymbol{h}}_i$ and $\hat{\boldsymbol{h}}_j$ of $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ respectively, let us define the pairwise cross-relation residual $r_{i,j}(n)$ as

$$r_{i,j}(n) = \hat{\boldsymbol{h}}_i^T \boldsymbol{y}_j(n) - \hat{\boldsymbol{h}}_j^T \boldsymbol{y}_i(n), \tag{2.9}$$

where $\hat{\boldsymbol{h}}_m = \begin{bmatrix} \hat{h}_m(1) & \hat{h}_m(2) & \ldots & \hat{h}_m(L) \end{bmatrix}^T$.

In the noiseless case, it is clear from (2.8) that the cross-relation residual is equal to 0 for all $n$ when $\hat{\boldsymbol{h}}$ is equal to $\boldsymbol{h}$.

### 2.4.2 Identifiability conditions

The aim of BSI is to estimate the RIRs $\{\boldsymbol{h}_m,\, m \in \{1, 2, \ldots, M\}\}$ from the observed signals $\{y_m[n],\, m \in \{1, 2, \ldots, M\}\}$ without knowledge of the input signal $s[n]$. As seen in Section 2.4.1, BSI may be achieved by looking for a non-zero vector $\hat{\boldsymbol{h}}$ such that $r_{i,j}(n)$ is equal to 0 for all $n$. However, the question of uniqueness of the solution arises. There may indeed exist several non-aligned vectors also satisfying $r_{i,j}(n) = 0$ for all $n$.

Xu, Liu, Tong, *et al.* proved in [32] that the solution space to $\sum_{n\in\mathbb{N}} r_{i,j}^2(n) = 0$ such that $\hat{\boldsymbol{h}}$ is non-zero is a one-dimensional subspace spanned by $\boldsymbol{h}$ in the noiseless case provided that the identifiability conditions hold.

These identifiability conditions are given as [32]:

- The RIRs do not have common zeros, i.e. zeros common to all $M$ channels, in the $z$-domain.
- The autocorrelation matrix of size $L \times L$ of the source signal is full-rank.

In other words, the first assumption ensures that the system is sufficiently diverse while the second one ensures that the input signal is sufficiently rich in the frequency domain [34]. Additionally, the length $L$ of the RIRs generating the reverberant signals needs to be known. Under these assumptions, the stacked RIRs $\boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_1^T & \boldsymbol{h}_2^T & \cdots & \boldsymbol{h}_M^T \end{bmatrix}^T$ can be estimated up to a non-zero scaling factor [32].

For an intuitive understanding of these assumptions, let us consider the cross-relation in the $z$-domain for a two channel system. Denoting by $R_{12}(z)$, $H_1(z)$, $H_2(z)$, $S(z)$, $\hat{H}_1(z)$ and $\hat{H}_2(z)$ the $z$-transform of the cross-relation residual $r_{12}$, $h_1[n]$, $h_2[n]$, $s[n]$, $\hat{h}_1[n]$ and $\hat{h}_2[n]$ respectively, the cross-relation residual in the $z$-domain is given as

$$R_{12}(z) = S(z)\left(H_1(z)\hat{H}_2(z) - H_2(z)\hat{H}_1(z)\right). \tag{2.10}$$

From (2.10), we see that if $S(z)$ is equal to 0 for some frequencies, any value of $\hat{H}_i(z)$ at those frequencies will annihilate the cross-relation residual. Hence the source signal needs to have a full-rank autocorrelation matrix. Likewise, if there is a common zero between

the two channel, that zero can be factored out and combined with $S(z)$. Since the length of the estimated RIRs are imposed, $\hat{H}_1(z)$ and $\hat{H}_2(z)$ can only estimate $H_1(z)$ and $H_2(z)$ up to a common zero.

To ensure identifiability of the system, these identifiability conditions are assumed to hold throughout the thesis.

### 2.4.3 LS approach to BSI

Let us now review the BSI method proposed by Xu, Liu, Tong, *et al.* in [32].

Given channel $m$ and a number of input samples $N_s$ from the dry signal $s[n]$, let us consider the pairwise cross-relation residual $r_{m,m+1}(n)$ at time instants $n \in \{L, L+1, ..., N_s\}$. The vector $\boldsymbol{r}_{m,m+1} = \begin{bmatrix} r_{m,m+1}(N_s) & r_{m,m+1}(N_s-1) & \cdots & r_{m,m+1}(L) \end{bmatrix}^T$ is given by

$$\boldsymbol{r}_{m,m+1} = \begin{bmatrix} \boldsymbol{\mathcal{H}}_{N_r \times L}\{\boldsymbol{y}_{m+1}^{N_s}(N_s)\} & -\boldsymbol{\mathcal{H}}_{N_r \times L}\{\boldsymbol{y}_{m}^{N_s}(N_s)\} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{h}}_m \\ \hat{\boldsymbol{h}}_{m+1} \end{bmatrix}, \tag{2.11}$$

where $N_r$ is the number of samples in the residual and

$$\boldsymbol{\mathcal{H}}_{N_r \times L}\{\boldsymbol{y}_m^{N_1}(N_2)\} = \begin{bmatrix} y_m(N_2) & y_m(N_2-1) & \dots & y_m(N_2-L+1) \\ y_m(N_2-1) & y_m(N_2-2) & \dots & y_m(N_2-L) \\ \vdots & \vdots & \ddots & \vdots \\ y_m(N_1-N_2+L) & y_m(N_1-N_2+L-1) & \dots & y_m(N_1-N_2+1) \end{bmatrix} \tag{2.12}$$

is the Hankel matrix [73] of size $N_r \times L$, the entries of which are obtained from the vector given in the argument and $\boldsymbol{y}_m^{N_1}(N_2)$ is the vector of $N_1$ samples from the reverberant noisy signal at the $m^{th}$ channel given by $\boldsymbol{y}_m^{N_1}(N_2) = \begin{bmatrix} y_m(N_2) & y_m(N_2-1) & \dots & y_m(N_2-N_1+1) \end{bmatrix}^T$. Here, $N_r$ is equal to $N_s - L + 1$.

Collecting the cross-relation residual for every channel with an index higher than $m$, the residuals between these channels and the $m^{th}$ channel,

$\begin{bmatrix} \boldsymbol{r}_{m,m+1}^T & \boldsymbol{r}_{m,m+2}^T & \cdots & \boldsymbol{r}_{m,M}^T \end{bmatrix}^T$, are then expressed as

$$\begin{bmatrix} \boldsymbol{r}_{m,m+1} \\ \boldsymbol{r}_{m,m+2} \\ \vdots \\ \boldsymbol{r}_{m,M} \end{bmatrix} = \boldsymbol{R}_{\boldsymbol{y}_m} \hat{\boldsymbol{h}} \tag{2.13}$$

where $\hat{\boldsymbol{h}} = \begin{bmatrix} \hat{\boldsymbol{h}}_1^T & \hat{\boldsymbol{h}}_2^T & \cdots & \hat{\boldsymbol{h}}_M^T \end{bmatrix}^T$ is the estimated stacked RIRs and

$$\boldsymbol{R}_{\boldsymbol{y}_m} = \begin{bmatrix} & \boldsymbol{\mathcal{H}}_{N_s \times L}\{\boldsymbol{y}_{m+1}^{N-L}(N)\} & \\ & \boldsymbol{\mathcal{H}}_{N_s \times L}\{\boldsymbol{y}_{m+2}^{N-L}(N)\} & \\ \boldsymbol{0}_{(M-m)N_s \times (m-1)L} & \vdots & -\boldsymbol{\mathcal{I}}_{M-m} \otimes \boldsymbol{\mathcal{H}}_{N_s \times L}\{\boldsymbol{y}_m^{N-L}(N)\} \\ & \boldsymbol{\mathcal{H}}_{N_s \times L}\{\boldsymbol{y}_M^{N-L}(N)\} & \end{bmatrix} \tag{2.14}$$

with $\boldsymbol{\mathcal{I}}_{M-m}$ the identity matrix of size $(M-m) \times (M-m)$ and $\otimes$ the Kronecker product.

The cross-relation residual

$$\boldsymbol{r} = \begin{bmatrix} \boldsymbol{r}_{1,1}^T & \boldsymbol{r}_{1,2}^T & \cdots & \boldsymbol{r}_{1,M}^T & \boldsymbol{r}_{2,3}^T & \cdots & \boldsymbol{r}_{M-1,M}^T \end{bmatrix}^T, \tag{2.15}$$

for all possible pairs without repetition, is a $\frac{M(M-1)}{2} N_r$ vector given by

$$\boldsymbol{r} = \boldsymbol{R}_{\boldsymbol{y}} \hat{\boldsymbol{h}}, \tag{2.16}$$

where $\hat{\boldsymbol{h}} = \begin{bmatrix} \hat{\boldsymbol{h}}_1^T & \hat{\boldsymbol{h}}_2^T & \cdots & \hat{\boldsymbol{h}}_M^T \end{bmatrix}^T$ is the vector of stacked estimated RIRs and $\boldsymbol{R}_{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{R}_{\boldsymbol{y}_1}^T & \boldsymbol{R}_{\boldsymbol{y}_2}^T & \cdots & \boldsymbol{R}_{\boldsymbol{y}_M}^T \end{bmatrix}^T$.

The stacked impulse responses $\hat{\boldsymbol{h}}$ are then found by solving the linear equations

$$\boldsymbol{R}_{\boldsymbol{y}} \hat{\boldsymbol{h}} = \boldsymbol{0}_{\frac{M(M-1)}{2} N_r \times 1}. \tag{2.17}$$

Under the identifiability conditions, the solutions of (2.17) belong to a subspace spanned by $\boldsymbol{h}$ [32]. In other words, in the absence of additive noise and under the identifiability

assumptions, the solution of (2.17) estimates $\boldsymbol{h}$ up to a scaling factor.

In the presence of noise, it is proposed in [32] to estimate $\boldsymbol{h}$ by minimizing $||\boldsymbol{r}||_2^2$ under some non-nullity constraint on $\hat{\boldsymbol{h}}$. Assuming that the non-nullity constraint is chosen to be the unit norm constraint, $\hat{\boldsymbol{h}}$ is, in effect, a unit-norm right singular vector corresponding to the smallest right singular value of $\boldsymbol{R_y}$, or equivalently, a unit-norm eigenvector corresponding to the smallest eigenvalue of the square cross-relation matrix

$$\boldsymbol{R} = \boldsymbol{R}_{\boldsymbol{y}}^T \boldsymbol{R_y} \tag{2.18}$$

[74, Chapter 17]. The identifiability conditions ensure that $\boldsymbol{R}$ has a smallest eigenvalue with an algebraic multiplicity of 1. If the algebraic multiplicity were greater than 1, any vector the associated invariant subspace would be a solution to (2.17). These vectors, which are written as a linear combination of $\boldsymbol{h}$ and vectors orthogonal to $\boldsymbol{h}$, result in estimates $\hat{\boldsymbol{h}}$, which may be orthogonal to $\boldsymbol{h}$.

### 2.4.4 Stability of the solution to additive noise

The LS method for BSI is essentially an eigenvector estimation problem [32]. In this section, the robustness of the LS to additive noise is analyzed from the point of view of matrix perturbation analysis.

Let us first notice that, because of (2.14) and (2.1), we also have $\boldsymbol{R}_{\boldsymbol{y}_m} = \boldsymbol{R}_{\boldsymbol{x}_m} + \boldsymbol{R}_{\boldsymbol{\nu}_m}$, from which

$$\boldsymbol{R_y} = \boldsymbol{R_x} + \boldsymbol{R_\nu} \tag{2.19}$$

follows. The matrices $\boldsymbol{R}_{\boldsymbol{x}_m}$ and $\boldsymbol{R}_{\boldsymbol{\nu}_m}$ are defined similarly to $\boldsymbol{R}_{\boldsymbol{y}_m}$ and $\boldsymbol{R_x}$ and $\boldsymbol{R_\nu}$ are defined similarly to $\boldsymbol{R_y}$.

#### Condition number

Under the identifiability conditions and in the absence of additive noise, a finite number of samples is sufficient to blindly estimate the RIRs [32]. However, in the presence of additive spatio-temporally white Gaussian noise, accurate estimation of the population

cross-relation matrix is necessary. Indeed, in this case, assuming that the additive noise is zero-mean and uncorrelated with the source signal, taking the expectation of (2.18) using (2.19) yields

$$\mathbb{E}\{\boldsymbol{R_y}\} = \mathbb{E}\{\boldsymbol{R_x}\} + \sigma_\nu^2 \boldsymbol{I}_{ML}, \tag{2.20}$$

where $\mathbb{E}$ denotes the expectation operator, $\boldsymbol{I}_{ML}$ the identity matrix of size $ML \times ML$ and $\sigma_\nu^2$ the power of the additive noise. Hence, the eigenvectors of the population cross-relation matrix obtained from the noisy signals, $\mathbb{E}\{\boldsymbol{R_y}\}$, are identical to those of $\mathbb{E}\{\boldsymbol{R_x}\}$.

Assuming that the input and additive noise signals are ergodic, the sample covariance matrix is a classical approximation of the population covariance matrix [75], [76]. To analyze the effect of additive noise on the stability of the solution provided by the sample cross-relation matrix, let us analyze the spread of the smallest eigenvalues of the square cross-relation matrix.

Inputing (2.19) in (2.18), we see that the analyzed square-cross relation matrix is the sum of $\boldsymbol{R}_{\boldsymbol{x}}^T \boldsymbol{R_x}$ and another symmetric matrix. Hence, in noisy conditions, the LS method aims to estimate an eigenvector associated with the smallest eigenvalue of $\boldsymbol{R}_{\boldsymbol{x}}^T \boldsymbol{R_x}$ via $\boldsymbol{R}$, which is equal to $\boldsymbol{R}_{\boldsymbol{x}}^T \boldsymbol{R_x}$ perturbed by an additive term.

The condition number of an eigenvector associated with an eigenvalue of a matrix, measures the stability of the eigenvector when it is estimated from a noisy observation of the matrix [77, Chapter 3]. In the case of a symmetric matrix with eigenvalues each with an algebraic multiplicity, the condition number of an eigenvector $\boldsymbol{\xi}_{\lambda_i}$ corresponding to the $i^{\text{th}}$ eigenvalue $\lambda_i$ is given by [77, Chapter 3]:

$$\text{Cond}(\boldsymbol{\xi}_{\lambda_i}) = \frac{1}{\min\left(\lambda_i - \lambda_{i-1}, \lambda_{i+1} - \lambda_i\right)}, \tag{2.21}$$

where the eigenvalues have been sorted in ascending order.

Since BSI via the LS method is achieved by estimating an eigenvector associated with the smallest eigenvalue of $\boldsymbol{R}_{\boldsymbol{x}}^T \boldsymbol{R_x}$, the condition number provides us with an insight into the stability of the estimation process with respect to additive noise. The corresponding

condition number, under the identifiability conditions, is given by

$$\text{Cond}(\boldsymbol{\xi}_{\lambda_1}) = \text{Cond}(\boldsymbol{h}) = \frac{1}{\lambda_2}, \tag{2.22}$$

since $\lambda_1$ is equal to 0 under the identifiability condition and in the absence of additive noise.

Since the square cross-relation matrix is symmetric, it should be noted that, unlike the computation of the eigenvectors, the computation of the eigenvalues is well conditioned, i.e. robust to numerical approximations [77, Chapter 3]. Hence the condition number corresponding to an eigenvector associated with the smallest eigenvalue can be safely used to assess the robustness of the LS method to additive noise.

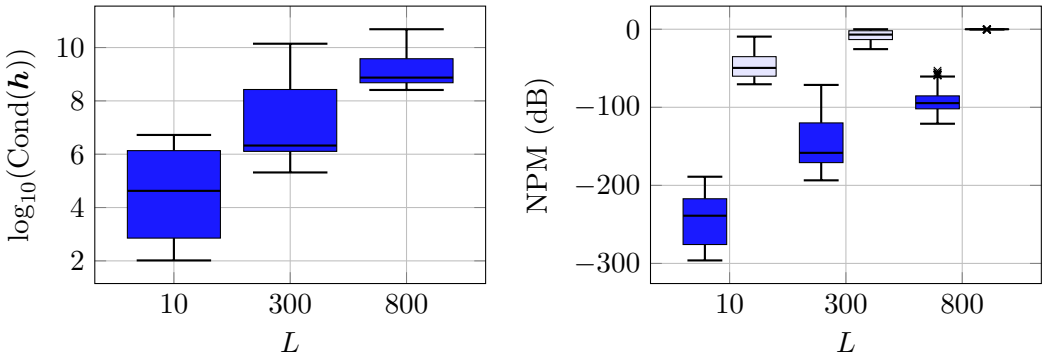
**Experimental results**

Figure 2.7a shows a boxplot of the variations of $\text{Cond}(\boldsymbol{h})$ with respect to the length of the RIRs for a 2 channel SIMO system. For each length, $\text{Cond}(\boldsymbol{h})$ was obtained by eigenvalue decomposition of the noiseless sample cross-relation matrix built from the noiseless reverberant signal. The boxes were obtained by considering 10 different randomly selected locations of the source within the room. To ensure that the full-rank covariance matrix condition of the identifiability assumptions holds, the source was a white Gaussian noise. For each location, 50 realizations of the source are considered. The source was at least 50 cm away from the walls and the centre of the microphone array. The RIRs were generated using the image method for a shoebox shaped room of dimension $4 \times 5 \times 2.5$ m with a reverberation time of 0.2 s. The centre of the two microphone array was placed at (2.5, 3, 2) m. The microphones were separated by 20 cm. The reverberant noiseless signal was of duration 20 times longer than that of the impulse response.

The condition number of $\boldsymbol{h}$ is shown in a $\log_{10}$ scale for clarity. It can be observed in Fig. 2.7a that $\text{Cond}(\boldsymbol{h})$ drastically increases as $L$ increases from 10 to 800, which suggests that the solution provided by the least squares method will become more sensitive to low level of noise as $L$ increases.

Figure 2.7b shows a boxplot of the NPMs of the estimated RIRs in noiseless and noisy cases against the length of the RIRs. The power of the additive spatio-temporally white Gaussian noise was scaled so that the Signal-to-Noise Ratio (SNR) was 60 dB. As expected, the sensitivity of the least squares method to additive noise increases as $L$ increases so that for $L$=800, the stacked RIRs cannot be estimated from the noisy signal. For $L$=300, the NPM of the estimated stacked RIRs varies between -25 dB and 0 dB. In the noisy case, accurate approximation of the population cross-relation matrix by the sample cross-relation matrix is therefore critical in the LS method. The NPMs of the RIRs estimated from the noiseless reverberant signal are always smaller than -50 dB, which shows that the system is identifiable and that the poor accuracy of the RIRs estimated from noisy data is due to the additive noise alone.

Figure 2.8 shows one the estimated RIRs for $L = 800$ and an SNR of 60 dB, which resulted in an NPM of -0.028 dB. It can be observed that the estimated RIRs do not capture the overall shape of $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ and therefore, are not expected to be usable for applications relying on estimated RIRs.




(a) Condition number of the stacked RIRs using the LS method.

(b) NPM of the estimated stacked RIRs for SNR = $\infty$ ( ) and SNR = 60 dB ( ).

Figure 2.7: Variations of Cond($\boldsymbol{h}$) and the NPM of the estimated stacked RIRs with $L$.

Figure 2.9 shows the variation of Cond($\boldsymbol{h}$) for the different source/microphone configurations and the different lengths. Each of the dashed vertical line indicates a change in the source location. It can be observed that Cond($\boldsymbol{h}$) is insensitive to the realization of the source but varies with the source/microphone configuration.

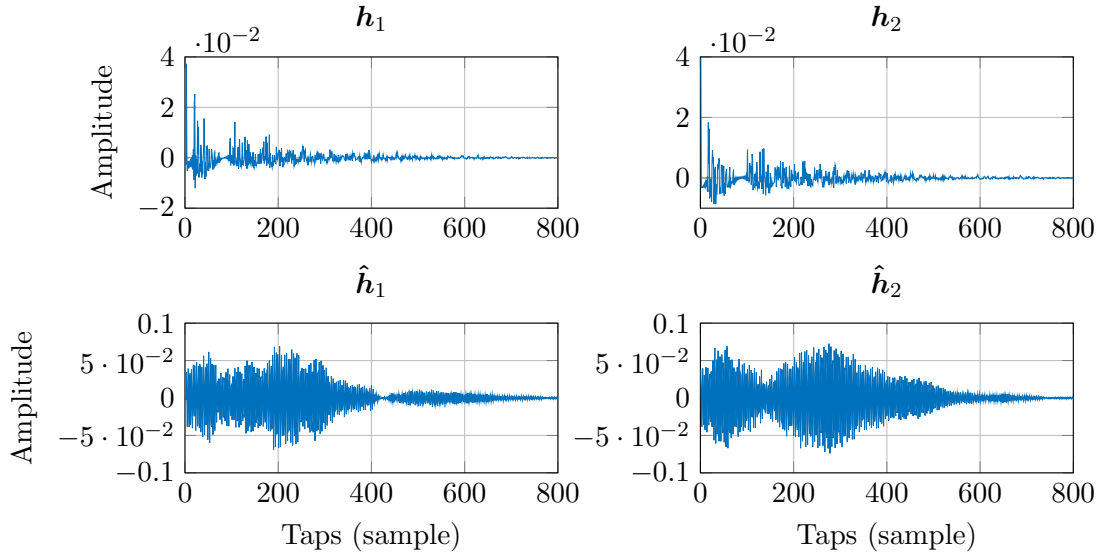



Figure 2.8: Ground truth and estimated RIRs using the LS method for $L = 800$ and an SNR of 60 dB.

## 2.5 Adaptive filter approach to BSI

This section describes two state-of-the-art methods for the blind estimation of RIRs, namely the MCLMS and the NMCFLMS. The derivations slightly differ from those given in [42] and [43] so as to emphasize the relationship between the MCLMS, NMCFLMS and the LS method.

### 2.5.1 The MCLMS algorithm

The need for a low complexity algorithm arose with the estimation of RIRs, which are significantly longer than communication channels [42]. To reduce the computational burden of a singular value decomposition and to enable BSI for real-time applications, Huang et al. proposed the MCLMS [42], which casts the problem of BSI into an adaptive filter framework.

The MCLMS adaptively minimizes the norm of the cross-relation residual $r_{ij}(n)$ across channels by implementing a stochastic gradient descent on the the cost function $J_{\mathrm{MCLMS}}(n)$

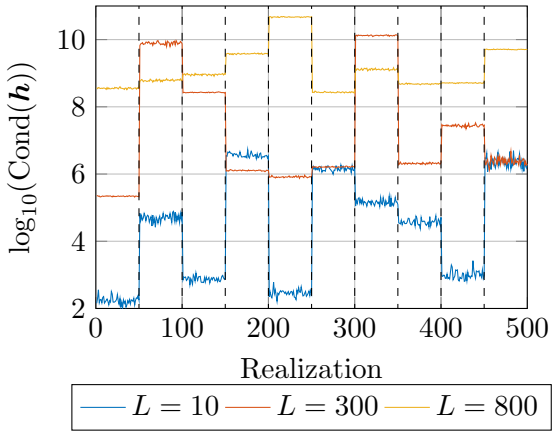



Figure 2.9: Condition number of the sought solution for each position and realization of the source. The vertical dashed lines indicate a change in the source location.

[42] defined as

$$J_{\mathrm{MCLMS}}(n) = \frac{\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} r_{ij}^2(n)}{\hat{\boldsymbol{h}}^T(n)\hat{\boldsymbol{h}}(n)}, \tag{2.23}$$

where $\hat{\boldsymbol{h}}(n)$ is the estimate of $\boldsymbol{h}$ at time instant $n$.

Denoting by $\boldsymbol{r}(n)$ the $\frac{M(M-1)}{2}$ cross-relation residual vector across channels for a single time instant $n$ and by $\boldsymbol{R}_{\boldsymbol{y}}^T(n)$ the rectangular cross-relation matrix for a single time instant, the numerator of (2.23) is rewritten as $||\boldsymbol{r}(n)||_2^2 = \hat{\boldsymbol{h}}^T(n)\boldsymbol{R}_{\boldsymbol{y}}^T(n)\boldsymbol{R}_{\boldsymbol{y}}(n)\hat{\boldsymbol{h}}(n)$. Hence $J_{\mathrm{MCLMS}}(n)$ is, in effect, the Rayleigh quotient [78] for the symmetric matrix $\boldsymbol{R}_{\boldsymbol{y}}^T(n)\boldsymbol{R}_{\boldsymbol{y}}(n)$. The update equation proposed in [42] is then given by a gradient descent step followed by a normalization:

$$\hat{\boldsymbol{h}}(n+1) = \frac{\hat{\boldsymbol{h}}(n) - \mu\boldsymbol{\nabla}J_{\mathrm{MCLMS}}(n)}{||\hat{\boldsymbol{h}}(n) - \mu\boldsymbol{\nabla}J_{\mathrm{MCLMS}}(n)||_2}, \tag{2.24}$$

where

$$\boldsymbol{\nabla}J_{\mathrm{MCLMS}}(n) = \frac{2}{\hat{\boldsymbol{h}}(n)^T\hat{\boldsymbol{h}}(n)}\left\{\boldsymbol{R}_{\boldsymbol{y}}^T(n)\boldsymbol{R}_{\boldsymbol{y}}(n)\hat{\boldsymbol{h}}(n) - J_{\mathrm{MCLMS}}(n)\hat{\boldsymbol{h}}(n)\right\} \tag{2.25}$$

is the gradient of $J_{\mathrm{MCLMS}}(n)$ with respect to $\hat{\boldsymbol{h}}$ obtained by application of the product rule [79] and $\mu$ is a fixed step-size.

In [42], by taking the derivative of the numerator of (2.23) with respect to each channel

separately, the gradient is shown to be equal to

$$\boldsymbol{\nabla} J_{\mathrm{MCLMS}}(n) = \frac{2}{\hat{\boldsymbol{h}}(n)^T \hat{\boldsymbol{h}}(n)} \left\{ \boldsymbol{R}(n)\hat{\boldsymbol{h}}(n) - J_{\mathrm{MCLMS}}(n)\hat{\boldsymbol{h}}(n) \right\} \tag{2.26}$$

with

$$\boldsymbol{R}(n) = \begin{bmatrix} \sum_{m=2}^{M} \boldsymbol{R}_{\boldsymbol{y}_m \boldsymbol{y}_m}(n) & -\boldsymbol{R}_{\boldsymbol{y}_2 \boldsymbol{y}_1}(n) & \dots & -\boldsymbol{R}_{\boldsymbol{y}_M \boldsymbol{y}_1}(n) \\ -\boldsymbol{R}_{\boldsymbol{y}_1 \boldsymbol{y}_2}(n) & \sum_{m\neq 2}^{M} \boldsymbol{R}_{\boldsymbol{y}_m \boldsymbol{y}_m}(n) & \dots & -\boldsymbol{R}_{\boldsymbol{y}_M \boldsymbol{y}_2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{R}_{\boldsymbol{y}_1 \boldsymbol{y}_M}(n) & -\boldsymbol{R}_{\boldsymbol{y}_1 \boldsymbol{y}_M}(n) & \dots & \sum_{m=1}^{M-1} \boldsymbol{R}_{\boldsymbol{y}_m \boldsymbol{y}_m}(n) \end{bmatrix}, \tag{2.27}$$

where

$$\boldsymbol{R}_{\boldsymbol{y}_i \boldsymbol{y}_j}(n) = \boldsymbol{y}_i(n)\boldsymbol{y}_j^T(n). \tag{2.28}$$

Since (2.25) and (2.26) are equal for all $\hat{\boldsymbol{h}} \in \mathbb{R}^{ML}$, we also have the equality

$$\boldsymbol{R}_{\boldsymbol{y}}^T(n)\boldsymbol{R}_{\boldsymbol{y}}(n) = \boldsymbol{R}(n) \tag{2.29}$$

and (2.23) is equivalently rewritten as

$$J_{\mathrm{MCLMS}}(n) = \frac{\hat{\boldsymbol{h}}^T(n)\boldsymbol{R}(n)\hat{\boldsymbol{h}}(n)}{\hat{\boldsymbol{h}}(n)^T\hat{\boldsymbol{h}}(n)}. \tag{2.30}$$

From (2.24), we see that the MCLMS is a LMS-type algorithm. It proceeds in two steps: a gradient descent step followed by a normalization step. The difference with a typical LMS scheme is the normalization step [80, Chapter 11], which is, here, necessary to avoid the null vector $\mathbf{0}_{ML\times 1}$ [42].

### 2.5.2 The NMCFLMS algorithm

#### Algorithm derivation

As shown in (2.23), at each iteration, the MCLMS computes linear convolutions for each pair of channels without repetition. To reduce the computational burden, Huang and Benesty proposed in [43] the NMCFLMS, an extension of the MCLMS which computes

all these convolutions in the frequency domain via the overlap-save method. Here, the derivations of the NMCFLMS are done differently to [43] to emphasize the link between the NMCFLMS and the MCLMS. In particular, it is shown that the NMCFLMS is a block-based implementation of the MCLMS with a slight modification of the cost function.

Denoting by $\boldsymbol{\mathcal{S}}_e^L = \begin{bmatrix} \mathbf{0}_{L\times L} & \boldsymbol{\mathcal{I}}_L \end{bmatrix}$ the selection matrix such that $\boldsymbol{\mathcal{S}}_e^L \boldsymbol{a}$ is a vector, the entries of which are equal to the $L$ last entries of $\boldsymbol{a}$ and using $\boldsymbol{\mathcal{E}}_L = \begin{bmatrix} \boldsymbol{\mathcal{I}}_L & \mathbf{0}_{L\times L} \end{bmatrix}^T$ as the expansion matrix which adds zeros at the end of a vector $\boldsymbol{a}$ of size $L \times 1$ so that $\boldsymbol{\mathcal{E}}_L \boldsymbol{a}$ is as vector of size $2L \times 1$, the residuals $\boldsymbol{r}_{i,j}(bn_{\text{off}})$ for block $b$ and an increment of $n_{\text{off}}$ consisting of $L$ samples taken in reverse time from $bn_{\text{off}}$ are given by

$$\boldsymbol{r}_{i,j}(bn_{\text{off}}) = \boldsymbol{\mathcal{S}}_e^L \underline{\boldsymbol{\mathcal{F}}}_{2L}^{-1} \boldsymbol{\mathcal{D}}_{\underline{\boldsymbol{y}}_j^{2L}(bn_{\text{off}})} \underline{\boldsymbol{\mathcal{F}}}_{2L} \boldsymbol{\mathcal{E}}_L \hat{\boldsymbol{h}}_i(b) - \boldsymbol{\mathcal{S}}_e^L \underline{\boldsymbol{\mathcal{F}}}_{2L}^{-1} \boldsymbol{\mathcal{D}}_{\underline{\boldsymbol{y}}_i^{2L}(bn_{\text{off}})} \underline{\boldsymbol{\mathcal{F}}}_{2L} \boldsymbol{\mathcal{E}}_L \hat{\boldsymbol{h}}_j(b), \tag{2.31}$$

where $\underline{\boldsymbol{\mathcal{F}}}_{2L}$ denotes the non-unitary Discrete Fourier Transform (DFT) matrix of size $2L \times 2L$, i.e. the DFT matrix where each input has unit modulus, $\boldsymbol{\mathcal{D}}_{\{\cdot\}}$ is a diagonal matrix with the entries specified in the argument on its diagonal, $\underline{\boldsymbol{y}}_i^{2L}(bn_{\text{off}})$ is the DFT of $\boldsymbol{y}_i(bn_{\text{off}})$ zero-padded to length $2L$ and $\hat{\boldsymbol{h}}_i(b)$ is an estimate of $\boldsymbol{h}_i$ at block $b$.

Since $\boldsymbol{r}_{i,j}(bn_{\text{off}})$ is real, we have $J_{\text{NMCLMS}}(bn_{\text{off}}) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \boldsymbol{r}_{i,j}^H(bn_{\text{off}}) \boldsymbol{r}_{i,j}(bn_{\text{off}})$. From (2.9), we see $\boldsymbol{r}_{i,j}(bn_{\text{off}}) = -\boldsymbol{r}_{j,i}(bn_{\text{off}})$ for all pairs $(i, j)$. Therefore, by completing the sum in $j$, we have

$$2J_{\text{NMCFLMS}}(bn_{\text{off}}) = \sum_{i=1}^{M} \sum_{j=1}^{M} \boldsymbol{r}_{i,j}^H(bn_{\text{off}}) \boldsymbol{r}_{i,j}(bn_{\text{off}}). \tag{2.32}$$

Expanding $\boldsymbol{r}_{i,j}^H(bn_{\text{off}}) \boldsymbol{r}_{i,j}(bn_{\text{off}})$ using (2.31), we have

$$\begin{aligned} \boldsymbol{r}_{i,j}^H(bn_{\text{off}}) \boldsymbol{r}_{i,j}(bn_{\text{off}}) = {} & \underline{\hat{\boldsymbol{h}}}_i^H(b) \underline{\boldsymbol{R}}_{jj}(bn_{\text{off}}) \underline{\hat{\boldsymbol{h}}}_i(b) + \underline{\hat{\boldsymbol{h}}}_j^H(b) \underline{\boldsymbol{R}}_{ii}(bn_{\text{off}}) \underline{\hat{\boldsymbol{h}}}_j(b) \\ & - \underline{\hat{\boldsymbol{h}}}_i^H(b) \underline{\boldsymbol{R}}_{ji}(bn_{\text{off}}) \underline{\hat{\boldsymbol{h}}}_j(b) - \underline{\hat{\boldsymbol{h}}}_j^H(b) \underline{\boldsymbol{R}}_{ij}(bn_{\text{off}}) \underline{\hat{\boldsymbol{h}}}_i(b), \end{aligned} \tag{2.33}$$

where $\underline{\hat{\boldsymbol{h}}}_i(b) = \underline{\boldsymbol{\mathcal{F}}}_L \hat{\boldsymbol{h}}_i(b)$ and

$$\underline{\boldsymbol{R}}_{ij}(bn_{\text{off}}) = (\boldsymbol{\mathcal{S}}_e^L \underline{\boldsymbol{\mathcal{F}}}_{2L}^{-1} \boldsymbol{\mathcal{D}}_{\underline{\boldsymbol{y}}_i^{2L}(bn_{\text{off}})} \underline{\boldsymbol{\mathcal{F}}}_{2L} \boldsymbol{\mathcal{E}}_L \underline{\boldsymbol{\mathcal{F}}}_L^{-1})^H (\boldsymbol{\mathcal{S}}_e^L \underline{\boldsymbol{\mathcal{F}}}_{2L}^{-1} \boldsymbol{\mathcal{D}}_{\underline{\boldsymbol{y}}_j^{2L}(bn_{\text{off}})} \underline{\boldsymbol{\mathcal{F}}}_{2L} \boldsymbol{\mathcal{E}}_L \underline{\boldsymbol{\mathcal{F}}}_L^{-1}). \tag{2.34}$$

Then using (2.33) in (2.32) a formulation of the cost function in the frequency domain similar to (2.23) is found. We indeed have

$$2J_{\text{NMCFLMS}}(bn_{\text{off}}) = 2\sum_{i=1}^{M} \hat{\underline{\boldsymbol{h}}}_i^{\,H}(b) \left( \sum_{j=1}^{M} \underline{\boldsymbol{R}}_{jj}(bn_{\text{off}}) \right) \hat{\underline{\boldsymbol{h}}}_i(b) - 2\sum_{i=1}^{M}\sum_{j=1}^{M} \hat{\underline{\boldsymbol{h}}}_i^{\,H}(b) \underline{\boldsymbol{R}}_{ji}(bn_{\text{off}}) \hat{\underline{\boldsymbol{h}}}_j(b),$$

which can be written as

$$\begin{aligned} J_{\text{NMCFLMS}}(bn_{\text{off}}) &= \hat{\underline{\boldsymbol{h}}}^{\,H}(b) \underline{\boldsymbol{\mathcal{D}}}_{\underline{\boldsymbol{R}}(bn_{\text{off}})} \hat{\underline{\boldsymbol{h}}}(b) - \hat{\underline{\boldsymbol{h}}}^{\,H}(b) \underline{\boldsymbol{C}}_{\underline{\boldsymbol{R}}(bn_{\text{off}})} \hat{\underline{\boldsymbol{h}}}(b) \\ &= \hat{\underline{\boldsymbol{h}}}^{\,H}(b) \underline{\boldsymbol{R}}(bn_{\text{off}}) \hat{\underline{\boldsymbol{h}}}(b), \end{aligned} \tag{2.35}$$

where $\hat{\underline{\boldsymbol{h}}}(b) = (\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{F}}_L)\hat{\boldsymbol{h}}(b)$ is formed of stacked DFTs of the estimated RIRs, $\underline{\boldsymbol{\mathcal{D}}}_{\underline{\boldsymbol{R}}(bn_{\text{off}})}$ is a block diagonal matrix with $\sum_{i=1}^{M} \underline{\boldsymbol{R}}_{ii}(bn_{\text{off}})$ on its diagonal,

$$\underline{\boldsymbol{C}}_{\underline{\boldsymbol{R}}}(bn_{\text{off}}) = \begin{bmatrix} \underline{\boldsymbol{R}}_{11}(bn_{\text{off}}) & \underline{\boldsymbol{R}}_{21}(bn_{\text{off}}) & \cdots & \underline{\boldsymbol{R}}_{M1}(bn_{\text{off}}) \\ \underline{\boldsymbol{R}}_{12}(bn_{\text{off}}) & \underline{\boldsymbol{R}}_{22}(bn_{\text{off}}) & \cdots & \underline{\boldsymbol{R}}_{M2}(bn_{\text{off}}) \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\boldsymbol{R}}_{1M}(bn_{\text{off}}) & \underline{\boldsymbol{R}}_{2M}(bn_{\text{off}}) & \cdots & \underline{\boldsymbol{R}}_{MM}(bn_{\text{off}}) \end{bmatrix}$$

and $\underline{\boldsymbol{R}}(bn_{\text{off}}) = \underline{\boldsymbol{\mathcal{D}}}_{\underline{\boldsymbol{R}}(bn_{\text{off}})} - \underline{\boldsymbol{C}}_{\underline{\boldsymbol{R}}(bn_{\text{off}})}$ is also given by

$$\underline{\boldsymbol{R}}(bn_{\text{off}}) = \begin{bmatrix} \sum_{i,i\neq 1} \underline{\boldsymbol{R}}_{ii}(bn_{\text{off}}) & -\underline{\boldsymbol{R}}_{21}(bn_{\text{off}}) & \cdots & -\underline{\boldsymbol{R}}_{M1}(bn_{\text{off}}) \\ -\underline{\boldsymbol{R}}_{12}(bn_{\text{off}}) & \sum_{i,i\neq 2} \underline{\boldsymbol{R}}_{ii}(bn_{\text{off}}) & \cdots & -\underline{\boldsymbol{R}}_{M2}(bn_{\text{off}}) \\ \vdots & \vdots & \ddots & \vdots \\ -\underline{\boldsymbol{R}}_{1M}(bn_{\text{off}}) & -\underline{\boldsymbol{R}}_{2M}(bn_{\text{off}}) & \cdots & \sum_{i,i\neq M} \underline{\boldsymbol{R}}_{ii}(bn_{\text{off}}) \end{bmatrix}. \tag{2.36}$$

The NMCFLMS aims to minimize $\mathbb{E}\{J_{\text{NMCFLMS}}(bn_{\text{off}})\}$ with respect to $\hat{\underline{\boldsymbol{h}}}(b)$ subject to $||\hat{\boldsymbol{h}}(b)||_2^2 = 1$. The strategy adopted in [43] is the same as in [42] i.e. do a descent step followed by a normalization. However, the cost function, here, only consists of a quadratic form shown in (2.35) whereas the MCLMS used a Rayleigh quotient. The update rule of the resulting algorithm, named Multi-Channel Frequency domain Least-

Mean Square (MCFLMS) [43], is then given by

$$\hat{\underline{\boldsymbol{h}}}(b+1) = \frac{\hat{\underline{\boldsymbol{h}}}(b) - \mu \underline{\boldsymbol{R}}(bn_{\text{off}})\hat{\underline{\boldsymbol{h}}}(b)}{\sqrt{L}||\hat{\underline{\boldsymbol{h}}}(b) - \mu \underline{\boldsymbol{R}}(bn_{\text{off}})\hat{\underline{\boldsymbol{h}}}(b)||_2}, \tag{2.37}$$

where $\sqrt{L}$ is due to the fact that $\underline{\boldsymbol{\mathcal{F}}}_L^H = L\underline{\boldsymbol{\mathcal{F}}}_L^{-1}$ and $\underline{\boldsymbol{R}}(bn_{\text{off}})\hat{\underline{\boldsymbol{h}}}(b)$ is the gradient of the non-analytic function $J_{\text{NMCLMS}}(bn_{\text{off}})$ [79]. Because of the overlap-save method, the residual and the update are computed frame by frame.

To increase the convergence rate, the authors of [43] also proposed to precondition the gradient with an approximation of the inverse of the Hessian matrix of $J_{\text{NMCFLMS}}(n)$. The derivations of the full algorithm are found in [43] and are omitted here as they do not provide additional insight into the connection between NMCFLMS and the previously presented methods.

### Link with MCLMS

Now, let us examine (2.34) in more details and let us draw a link between $\underline{\boldsymbol{R}}(bn_{\text{off}})$ and $\boldsymbol{R}(n)$. The term $\underline{\boldsymbol{\mathcal{F}}}_{2L}^{-1}\boldsymbol{\mathcal{D}}_{\underline{\boldsymbol{y}}_j(bn_{\text{off}})}\underline{\boldsymbol{\mathcal{F}}}_{2L}$, in (2.34), represents the circular convolution matrix the first column of which is given by $\boldsymbol{y}_j(bn_{\text{off}})$ zero-padded to length $2L$ [81, Chapter 5]. Left multiplying by $\boldsymbol{\mathcal{S}}_e^L$ is equivalent to selecting the $L$ last rows while right multiplying by $\boldsymbol{\mathcal{E}}_L$ is equivalent to selecting the first $L$ columns. Hence $\boldsymbol{\mathcal{S}}_e^L\underline{\boldsymbol{\mathcal{F}}}_{2L}^{-1}\boldsymbol{\mathcal{D}}_{\underline{\boldsymbol{y}}_i^{2L}(bn_{\text{off}})}\underline{\boldsymbol{\mathcal{F}}}_{2L}\boldsymbol{\mathcal{E}}_L$ is a $L \times L$ matrix equal to

$$\begin{bmatrix} y_i(bn_{\text{off}}-L+1) & y_i(bn_{\text{off}}-L) & \cdots & y_i(bn_{\text{off}}-2L+2) \\ y_i(bn_{\text{off}}-L+2) & y_i(bn_{\text{off}}-L+1) & \cdots & y_i(bn_{\text{off}}-2L+3) \\ \vdots & \vdots & \ddots & \vdots \\ y_i(bn_{\text{off}}) & y_i(bn_{\text{off}}-1) & \cdots & y_i(bn_{\text{off}}-L+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_i^T(bn_{\text{off}}-L+1) \\ \boldsymbol{y}_i^T(bn_{\text{off}}-L+2) \\ \vdots \\ \boldsymbol{y}_i^T(bn_{\text{off}}) \end{bmatrix}. \tag{2.38}$$

Hence from (2.34), we have

$$\begin{aligned} L\underline{\boldsymbol{\mathcal{F}}}_L^{-1}\underline{\boldsymbol{R}}_{ij}(bn_{\text{off}})\underline{\boldsymbol{\mathcal{F}}}_L &= \sum_{l=1}^{L-1} \boldsymbol{y}_i(bn_{\text{off}}-l)\boldsymbol{y}_i^T(bn_{\text{off}}-l) \\ &= \sum_{l=1}^{L-1} \boldsymbol{R}_{\boldsymbol{y}_i\boldsymbol{y}_j}(bn_{\text{off}}-l) \text{ by definition (2.28)} \end{aligned} \tag{2.39}$$

and from (2.36) we have

$$\underline{\boldsymbol{R}}(bn_{\text{off}}) = \frac{1}{L}(\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L)\left(\sum_{l=1}^{L-1}\boldsymbol{R}(bn_{\text{off}}-l)\right)(\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L^{-1}). \tag{2.40}$$

Then, denoting by $\boldsymbol{\xi}_\lambda$ an eigenvector of $\sum_{l=1}^{L-1}\boldsymbol{R}(bn_{\text{off}}-l)$ corresponding to an eigenvalue $\lambda$, from (2.40), we have

$$\begin{aligned} \underline{\boldsymbol{R}}(bn_{\text{off}})(\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L)\boldsymbol{\xi}_\lambda &= \frac{1}{L}(\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L)\sum_{l=1}^{L-1}\boldsymbol{R}(bn_{\text{off}}-l)\boldsymbol{\xi}_\lambda \\ &= \frac{\lambda}{L}(\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L)\boldsymbol{\xi}_\lambda, \end{aligned}$$

which shows that $(\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L)\boldsymbol{\xi}_\lambda$ is an eigenvector of $\underline{\boldsymbol{R}}(bn_{\text{off}})$. In other words, $(\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L)$ establishes a one-to-one mapping between the eigenvectors of $\underline{\boldsymbol{R}}(bn_{\text{off}})$ and those of $\sum_{l=1}^{L-1}\boldsymbol{R}(bn_{\text{off}}-l)$. Therefore, by definition of $\boldsymbol{\mathcal{F}}_L^{-1}$, by taking the inverse transform of each non-overlapping sub-matrices with $L$ entries of $\underline{\boldsymbol{\xi}}_\lambda = (\boldsymbol{I}_M \otimes \underline{\boldsymbol{\mathcal{F}}}_L)\boldsymbol{\xi}_\lambda$, an eigenvector of $\sum_{l=1}^{L-1}\boldsymbol{R}(bn_{\text{off}}-l)$ is found.

The relationship in (2.40) also shows that the update equation (2.37) of the MCFLMS is a two-step block LMS where the first step implements a stochastic gradient descent on the cost function $\frac{1}{2}\hat{\boldsymbol{h}}^T(b)\left(\sum_{l=1}^{L-1}\boldsymbol{R}(bn_{\text{off}}-l)\right)\hat{\boldsymbol{h}}(b)$ and the second step normalizes $\hat{\boldsymbol{h}}(b)$ to unit norm.

### 2.5.3 Misconvergence of the adaptive algorithms

Although convergence analyses of the MCLMS and the NMCFLMS were respectively provided in [42] and [43], the proof of convergence for the former only investigated the point of convergence under the assumption that the algorithm did converge while the proof

of convergence for the latter only holds for the noiseless case.

When the reverberant signal is corrupted by spatio-temporally white additive Gaussian noise, these adaptive algorithms have been reported to misconverge [82], i.e. they converge to estimated RIRs which are not good estimates of the RIRs in an NPM sense. When the NPM is shown against time, an initial decrease followed by an increase to 0 dB is observed as illustrated in Fig. 2.10.

Figure 2.10 displays the NPMs of RIRs estimated by the NMCFLMS for RIRs generated by the image method [83] for $M$=2 truncated to $L$=128. The input and the additive noise signals were both zero-mean white Gaussian noises. The power of the additive noise was scaled so that the SNR was equal to 30 dB. The source was 2 m away from the centre of the microphone array and the microphones were 20 cm apart. The step-size and the forgetting factor were empirically set to $\mu$=0.9 and $\rho$=0.98 respectively. The misconvergence is intrinsic to the algorithm and cannot be solved by decreasing the fixed step-size as doing so would only decrease the initial convergence rate and delay the time of misconvergence [82].

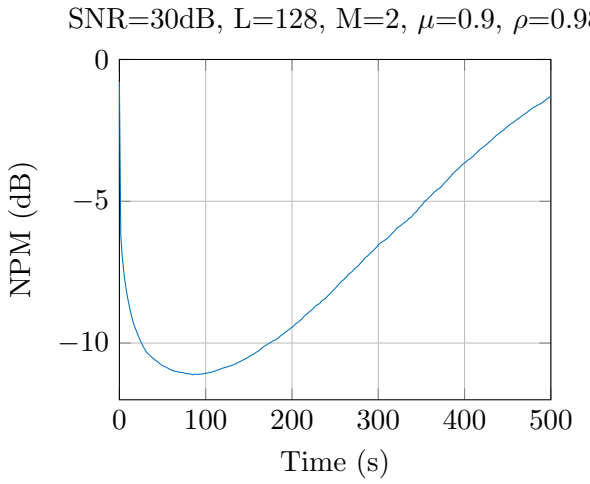



Figure 2.10: Misconvergence of NMCFLMS.

## 2.6 Summary and scope of the thesis

In this chapter, the model adopted in this thesis for the reverberant noisy signals together with the specificities of the impulse responses to be blindly estimated have been presented. The MCLMS and NMCFLMS have been shown to be adaptive implementations of the LS method, the limitations of which have been experimentally demonstrated. These limitations can be summarized as follows:

- Sensitivity to additive noise in the presence of near common zeros. Near common zeros, however, cannot be avoided in the context of blind identification of acoustic systems. As the computation of the roots of high-order polynomial can be numerically unstable [84], the condition number of the sought solution Cond($\boldsymbol{h}$), which is the condition number of an eigenvector associated with the smallest eigenvalue of the noiseless square cross-relation matrix, has been proposed to quantify the stability of the solution of the LS method. Figure 2.9 indeed showed that Cond($\boldsymbol{h}$) strongly depended on the RIRs while being independent of the source signal, which suggests that Cond($\boldsymbol{h}$) is related to the RIRs. However, the explicit relationship relating Cond($\boldsymbol{h}$) to the proximity of the zeros between the channels falls outside the scope of this thesis.

- Knowledge of the length of the impulse responses: The length of the RIRs to be estimated is usually assumed known, e.g. [32], [42], [43]. In the context of RIRs, it is customary to limit the length of the RIRs to $T_{60}f_s$, where $f_s$ is the sampling frequency, e.g. [25], [26]. However that definition of the length of a RIR is perceptually motivated and may not reflect the real length of a RIR. In fact, the length of an RIR may only be a useful concept for artificial reverberation without a physical meaning. Information theory based methods [85], [86] or [87] would require knowledge of a statistical model of the observed signal, which, in our case, means a statistical model of reverberant speech. On the other hand, blind estimators of $T_{60}$, e.g. [54], [55], have been shown to reliably estimate $T_{60}$ [56] and can be used to approximate the length of the RIRs.

In addition to these limitations, the problem of blind identification of acoustic systems also faces the following challenges:

- Fast estimation: Adaptive algorithms are known to converge slowly when long impulse responses are estimated [88]. The ability to quickly track varying RIRs is essential as RIRs change whenever the environment changes [57]. However, in addition to a slower convergence rate than their instantaneous counterpart [89], block adaptive algorithms have an inherent time delay, which may be unacceptable for real-time applications [90].

- Model accuracy: Cross-relation based methods are attractive in that they bypass the need for a statistical model of the input signal [34]. However, the SIMO structure of the problem is the only piece of information that is exploited. Therefore, mismatches between the real model underlying the observed data and the adopted model such as slowly varying RIRs may severely degrade the estimation accuracy.

In this work, the challenge of designing a BSI algorithm robust to additive noise and near-common zeros able to estimate realistically long impulse responses, i.e. with several thousand taps per channel, is investigated. The problem of estimating the RIRs from a limited amount of data is tackled but not of main concern and the problem of length mismatch falls outside the scope of this thesis. The system model shown in (2.1) is assumed to hold and the length of the RIRs is assumed known in the derivation of the different algorithms.

# Chapter 3

# SIMO structure based blind system identification

## 3.1 Overview

This chapter investigates the problem of noise robustness and convergence rate of adaptive algorithms without exploiting information about the nature of the RIRs. This allows to apply the presented algorithms to arbitrary systems provided that the identifiability conditions are satisfied. The different algorithms all rely on the square cross-relation matrix in the sense that they aim at estimating the minimum eigenpair of the population square cross-relation matrix.

## 3.2 Introduction and literature review

Although able to asymptotically estimate the RIRs from reverberant signals corrupted by additive spatio-temporally white Gaussian noise, the analysis of the robustness of the LS method to additive noise in finite time, given in the previous chapter, showed that the estimation accuracy drastically decreases as the length of the considered RIRs increases.

The LS method [32] is essentially deterministic [37] in the sense that it does not need a statistical model of the input signal. In fact, it does not model the additive noise

either. As noted by Xu, Liu, Tong, *et al.* in [32], in the presence of additive noise, the minimization of the Euclidean norm of the residual under a non-zero constraint on the estimated stacked RIRs may not be statistically optimal.

To achieve that statistical optimality, Hua proposed the Two-Step Maximum Likelihood algorithm in [36], where the residual is assumed to be spatio-temporally white Gaussian distributed at the solution. The proposed algorithm maximizes the likelihood of the residual given the noisy reverberant signals with respect to the impulse responses. The optimization problem is then shown to be equivalent to minimizing a quadratic form, the associated matrix of which also depends on the impulse responses. To circumvent that difficulty, Hua proposed a two-step algorithm. In the first step, a coarse estimate of the impulse responses is obtained using [32] or [35] to approximate the matrix associated with the quadratic form and, in the second step, a more accurate estimate of the impulse responses is obtained by eigenvalue decomposition of the matrix obtained in the first step.

For the particular case of 2 microphones, the optimization problem may be solved in one step. For that particular case, the optimization problem can indeed be shown to be an instance of the Structured Total Least Squares problem considered by Markovsky, Huffel, and Pintelon in [91]. Since a coarse estimate of the impulse responses is required in the two-step algorithm, its performance relies on the accuracy of the first step. By employing the one-step algorithm proposed in [91], improvement may be achieved. However, there is no guarantee that the solution to the Structured Total Least Squares problem is a global minimum of the cost function [92].

The misconvergence of the adaptive BSI algorithms has been the subject of recent research and has been investigated along the following aspects: correction of the misconverged estimate, use of an adaptive step-size and constraining the algorithm based on characteristics of RIRs. The use of constraints and regularization terms to improve the robustness of the BSI algorithm will be presented in Chapter 4.

Thomas, Gaubitch, Habets, *et al.* found through simulations that the estimated RIRs, obtained by application of the NMCFLMS to noisy data, shared a common filter even if the ground truth impulse responses did not share common zeros [93]. They also proposed a

supervised algorithm that removes the common component in [93] to increase the accuracy of the estimated impulse responses. However, knowledge of the ground truth impulse responses is needed. In [94], Thomas, Gaubitch, Habets, *et al.* theoretically explained the presence of the common filter by analyzing the behavior of the eigenvectors of the noiseless square cross-relation matrix perturbed by a small additive term.

As seen in Chapter 2, in the 2 microphone case, the cross-relation matrix is actually a covariance matrix formed by the stacked signals of both channels. In this case, the problem of BSI is equivalent to that of minor component analysis [95]. In fact, the update rule of the MCFLMS is very similar to the algorithm proposed by Oja in [96] with the difference that the latter uses a gradient ascend step, written as

$$\begin{aligned} \tilde{\boldsymbol{h}}(n+1) &= \hat{\boldsymbol{h}}(n) + \mu \boldsymbol{R}(n)\hat{\boldsymbol{h}}(n) \\ \hat{\boldsymbol{h}}(n+1) &= \frac{\tilde{\boldsymbol{h}}(n+1)}{||\tilde{\boldsymbol{h}}(n+1)||}, \end{aligned} \tag{3.1}$$

where the notations of [96] have been adapted to the notations used in this thesis, to estimate the principal component. The proof of convergence of the MCFLMS is then given in the analysis of the anti-Hebbian rule proposed in [97], which is a first order Maclaurin expansion of the update rule in $\mu$ with the gradient ascend step replaced by a gradient descent step, i.e. the term $+\mu \boldsymbol{R}(n)\hat{\boldsymbol{h}}(n)$ is replaced by $-\mu \boldsymbol{R}(n)\hat{\boldsymbol{h}}(n)$ in (3.1). Convergence of the anti-Hebbian rule is ensured provided that the sequence of step-sizes is positive decreasing and fulfills the requirement that $\sum_{n=1}^{\infty} \mu(n) = \infty$ but $\sum_{n=1}^{\infty} \mu^2(n)$ is finite, e.g. $\mu(n) = \frac{1}{n}$, and that the smallest eigenvalue is 0 [97]. However, a step-size decreasing to 0 prevents quick adaptation to varying environments. The use of a constant step-size without normalization, on the other hand, has been shown to lead to divergence in [98].

The noise-robust algorithm proposed by Ahmad, Gaubitch, and Naylor [99] can be seen as an adaptive step-size solution for the misconvergence of the NMCFLMS. Assuming that misconvergence is intrinsic to the adaptive algorithms, Ahmad, Gaubitch, and Naylor indeed proposed to estimate the critical time, time instant at which the NPM is minimal, by analyzing the cost function. The algorithm proposed in [99] defines an analysis window

longer than the estimated critical time. At the beginning of each analysis window, the BSI algorithm is reinitialized and only the estimated RIRs at the estimated critical time are kept. In effect, the algorithm proposed by [99] can be seen as an NMCFLMS where the step-size is set to 0 at the moment of detected misconvergence. However, since the algorithm strongly relies on the NMCFLMS, a small step-size is needed to obtain accurate estimates, which reduces the convergence rate. In their simulations, the critical time was indeed larger than 5 s for a 128 tap long impulse responses, 5 microphone system excited by white noise and corrupted by white noise at an SNR of 20 dB.

The problem of fast convergence of the adaptive BSI algorithms has been approached in three different ways: the use of an adaptive step-size, the use of curvature information by employing pre-conditioning matrices and the computation in subbands.

Different step-sizes have been proposed for an increased convergence rate. In [100], Huang, Benesty, and Chen derived an optimal step-size in noiseless conditions for a modified MCLMS algorithm. Gaubitch, Hasan, and Naylor proposed an extension of the work presented in [100] in [64] by deriving a Wiener step-size, which takes noise into account. However, as the derived optimal step-size depends on the ground truth impulse responses, only an approximation of the optimal step-size can be used.

Ahmad, Khong, and Naylor proposed to use step-sizes proportional to each of the estimated taps in the adaptation process in [101]. The tap-dependent step-sizes can be seen as a diagonal preconditioning matrix, the aim of which is to regularize the optimization problem [102, Chapter 9]. In the same family of optimization techniques, the quasi-Newton method developed by Habets and Naylor has shown to drastically improve the convergence rate of the MCLMS algorithm at the expense of an increase of the computational complexity [103]. Habets, Benesty, and Naylor then adopted another approach to cross-relation based BSI in [104] to achieve a similar convergence rate to the quasi-Newton method but at a lower computational complexity. They indeed proposed an alternative cost function which promoted the convergence of the adaptive algorithm. That cost function is then cast into a constrained convex optimization problem where the estimate is constrained to annihilate the cross-relation residual. However, as seen in Chapter 2, in noisy conditions,

at the solution, the cross-relation residual is equal to the additive noise convolved with the solution but not equal to 0.

Long impulse responses are known to reduce the convergence rate of adaptive algorithms [105]. Motivated by the reduction in complexity achieved in subband processing [106], Gannot and Moonen investigated the problem of BSI in subbands and proposed an algorithm shown to estimate the magnitude response of relatively short impulse responses. However, since the proposed method is based on the cross-relation in each subband, a scaling ambiguity between bands arose so that accurate estimation of the full-band impulse response was not straightforward. Assuming that estimates of the subband RIRs with small estimation errors are available, Gaubitch, Lin, and Naylor solved the problem of scaling ambiguities by minimizing the norm of the cross-relation residual in full-band. A method robust to higher levels of estimation errors in the subband estimates of the RIRs is presented in [109]. However, the estimation of RIRs in subbands remains an open question [109].

This chapter investigates the problem of noise-robust BSI algorithms within the framework presented by Xu, Liu, Tong, *et al.* in [32], i.e. the minimization of the norm of the cross-relation residual. In Section 3.3, a noise-robust but slowly converging adaptive algorithm derived from the MCLMS is presented. An adaptive step-size is then derived in Section 3.4 to increase the convergence rate. In Section 3.5, the cross-relation is investigated in the frequency domain and shown to contain the Relative Transfer Function (RTF). An application of the RTF to BSI is then given. Finally, the work presented in this chapter is summarized in Section 3.6.

## 3.3 Fixed step-size block-based Rayleigh quotient

### 3.3.1 Algorithm derivation

As proved in [98], the normalization in the MCLMS is necessary to prevent divergence. In the update equation of an LMS-based procedure, the direction of the descent steers the adaptation towards a point minimizing the cost function. In the case where the

population cross-relation matrix $\mathbb{E}\{\boldsymbol{R}\}$ is approximated by its instantaneous value $\boldsymbol{R}(n)$, as in the MCLMS algorithm, there always exists a high-dimensional subspace over which the positive cost function $J_{\text{MCLMS}}(n)$ is equal to zero. Writing the $ML \times ML$ matrix $\boldsymbol{R}(n)$ as $\boldsymbol{R}(n) = \boldsymbol{D}_{\boldsymbol{R}}(n) - \boldsymbol{C}_{\boldsymbol{R}}(n)$, with

$$\boldsymbol{D}_{\boldsymbol{R}}(n) = \boldsymbol{I}_M \otimes \left( \sum_{m=1}^{M} \boldsymbol{R}_{\boldsymbol{y}_m \boldsymbol{y}_m}(n) \right), \tag{3.2}$$

$$\boldsymbol{C}_{\boldsymbol{R}}(n) = \begin{bmatrix} \boldsymbol{R}_{\boldsymbol{y}_1 \boldsymbol{y}_1}(n) & \boldsymbol{R}_{\boldsymbol{y}_2 \boldsymbol{y}_1}(n) & \cdots & \boldsymbol{R}_{\boldsymbol{y}_M \boldsymbol{y}_1}(n) \\ \boldsymbol{R}_{\boldsymbol{y}_1 \boldsymbol{y}_2}(n) & \boldsymbol{R}_{\boldsymbol{y}_2 \boldsymbol{y}_2}(n) & \cdots & \boldsymbol{R}_{\boldsymbol{y}_M \boldsymbol{y}_2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{R}_{\boldsymbol{y}_1 \boldsymbol{y}_M}(n) & \boldsymbol{R}_{\boldsymbol{y}_2 \boldsymbol{y}_M}(n) & \cdots & \boldsymbol{R}_{\boldsymbol{y}_M \boldsymbol{y}_M}(n) \end{bmatrix}, \tag{3.3}$$

an upper bound of $\text{rank}\{\boldsymbol{R}(n)\}$ is found.

The rank of $\boldsymbol{D}_{\boldsymbol{R}}(n)$ is upper bounded by $M^2$. Indeed, each of the terms in the summation $\sum_{m=1}^{M} \boldsymbol{R}_{\boldsymbol{y}_m \boldsymbol{y}_m}(n)$ is a rank 1 matrix. Hence, by subadditivity of the $\text{rank}\{\cdot\}$ operator [78, Chapter 0], $\sum_{m=1}^{M} \boldsymbol{R}_{\boldsymbol{y}_m \boldsymbol{y}_m}(n)$ has a rank lower or equal to $M$. Then, $\text{rank}\{\boldsymbol{D}_{\boldsymbol{R}}(n)\}$, which is the Kronecker product of $\boldsymbol{I}_M$ of rank $M$ and $\sum_{m=1}^{M} \boldsymbol{R}_{\boldsymbol{y}_m \boldsymbol{y}_m}(n)$ of rank smaller or equal to $M$, is upper bounded by $M^2$ [110, Chapter 13].

Likewise, for a given $n$, any vector of size $ML$, which is the vectorization of a $L \times M$ matrix, the columns of which are in span $\{\{\boldsymbol{y}_m(n)\}\}^{\perp}_{m \in \{1,\cdots,M\}}$, belongs to the null space of $\boldsymbol{C}_{\boldsymbol{R}}(n)$. Hence the dimensionality of the null space of $\boldsymbol{C}_{\boldsymbol{R}}(n)$ is lower bounded by $ML - M$. Then, by the rank plus nullity theorem [74, Chapter 2], $\text{rank}\{\boldsymbol{C}_{\boldsymbol{R}}(n)\}$ is upper bounded by $M$.

By subadditivity of $\text{rank}\{\cdot\}$, $\text{rank}\{\boldsymbol{R}(n)\}$ is upper bounded by $M(M+1)$, which shows that at any time instant $n$, $\boldsymbol{R}(n)$ has a null-space of dimensionality greater or equal to $M(L-M-1)$. In the absence of noise, the stacked RIRs is always in the null space of $\boldsymbol{R}(n)$, as seen in Chapter 2. However, when noise is present, the instantaneous cost function evaluated at $\boldsymbol{h}$ would give the power of the filtered noise divided by $||\boldsymbol{h}||^2$. Therefore, by using a instantaneous cross-relation matrix, the descent direction consistently steers the adaption towards the null space of $\boldsymbol{R}(n)$, which contains $\boldsymbol{h}$ only in the absence of noise.

We hence investigate a class of adaptive algorithms based on the Rayleigh quotient cost function of the MCLMS. The different variants differ in the approximation of $\mathbb{E}\{\boldsymbol{R}(n)\}$ (sample-based vs. block-based) as well as the presence or absence of the normalization step, giving rise to the Sample-based Unconstrained Rayleigh Quotient (SURQ), the Block-based Normalized Rayleigh quotient (BNRQ) and the Block-based Unconstrained Rayleigh Quotient (BURQ) algorithms (see Table 3.1). The corresponding update equations are respectively given by

$$\hat{\boldsymbol{h}}_{\mathrm{SURQ}}(b+1) = \hat{\boldsymbol{h}}_{\mathrm{SURQ}}(b) - \mu\boldsymbol{\nabla} J_{\mathrm{MCLMS}}(b), \tag{3.4}$$

$$\hat{\boldsymbol{h}}_{\mathrm{BNRQ}}(b+1) = \frac{\hat{\boldsymbol{h}}_{\mathrm{BNRQ}}(b) - \mu\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)}{||\hat{\boldsymbol{h}}_{\mathrm{BNRQ}}(b) - \mu\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)||_2}, \tag{3.5}$$

$$\hat{\boldsymbol{h}}_{\mathrm{BURQ}}(b+1) = \hat{\boldsymbol{h}}_{\mathrm{BURQ}}(b) - \mu\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b) \tag{3.6}$$

where

$$J_{\mathrm{BURQ}}(b) = \frac{\hat{\boldsymbol{h}}^T(b)\bar{\boldsymbol{R}}(b)\hat{\boldsymbol{h}}(b)}{\hat{\boldsymbol{h}}^T(b)\hat{\boldsymbol{h}}(b)} \tag{3.7}$$

$$\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b) = \frac{2}{\hat{\boldsymbol{h}}^T(b)\hat{\boldsymbol{h}}(b)}\left\{\bar{\boldsymbol{R}}(b)\hat{\boldsymbol{h}}(b) - J_{\mathrm{BURQ}}(b)\hat{\boldsymbol{h}}(b)\right\}, \tag{3.8}$$

with $\bar{\boldsymbol{R}}(b) = \frac{1}{B}\sum_{n=0}^{B-1}\boldsymbol{R}(bn_{\mathrm{off}}+n)$, $b$, $B$ and $n_{\mathrm{off}}$ respectively representing the iteration index, the block size and a block-increment offset and $\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)$ referring to the derivative of $J_{\mathrm{BURQ}}(b)$ with respect to $\hat{\boldsymbol{h}}$(b). In (3.7) and (3.8), $\hat{\boldsymbol{h}}(b)$ refers to $\hat{\boldsymbol{h}}_{\mathrm{BNRQ}}$(b) and $\hat{\boldsymbol{h}}_{\mathrm{BURQ}}$(b) for the estimates given by (3.5) and (3.6) respectively. In the rest of this chapter the subscripts $\{\cdot\}_{\mathrm{SURQ}}$, $\{\cdot\}_{\mathrm{BNRQ}}$ and $\{\cdot\}_{\mathrm{BURQ}}$ will be dropped when there is no ambiguity.

| Algorithm | Approximation of $\mathbb{E}\{\boldsymbol{R}\}$ | Additional step |
|---|---|---|
| MCLMS (2.24) | Sample-based | Normalization |
| SURQ (3.4) | Sample-based | None |
| BNRQ (3.5) | Block-based | Normalization |
| BURQ (3.6) | Block-based | None |

Table 3.1: Rayleigh quotient based algorithms for BSI.

### 3.3.2 Simulations

The performance of the proposed Rayleigh quotient based algorithms is here investigated and compared with the existing MCLMS, NMCFLMS and Robust NMCFLMS (RNMCFLMS) [111] algorithms. In particular, the robustness to additive spatially white Gaussian noise and the convergence rate are evaluated.

#### Setup

The algorithms were evaluated for a set of illustrative cases involving two RIR lengths ($L$=256 and $L$=500), two different numbers of microphones, ($M$=2 and $M$=5), and two values of SNR, namely 15 dB and 10 dB. For each combination of $L$, $M$ and SNR, the algorithms were evaluated for 50 realizations of additive noise. The RIRs were generated using the image method [83] for a shoebox-shaped room of dimensions $5 \times 6 \times 3$ m and a reverberation time $T_{60}$ of 0.5 s. The lengths $L$ were then fixed by truncating the obtained RIRs to the desired length. The input signal was a white Gaussian noise with a duration of 600 s and uncorrelated with the additive noise.

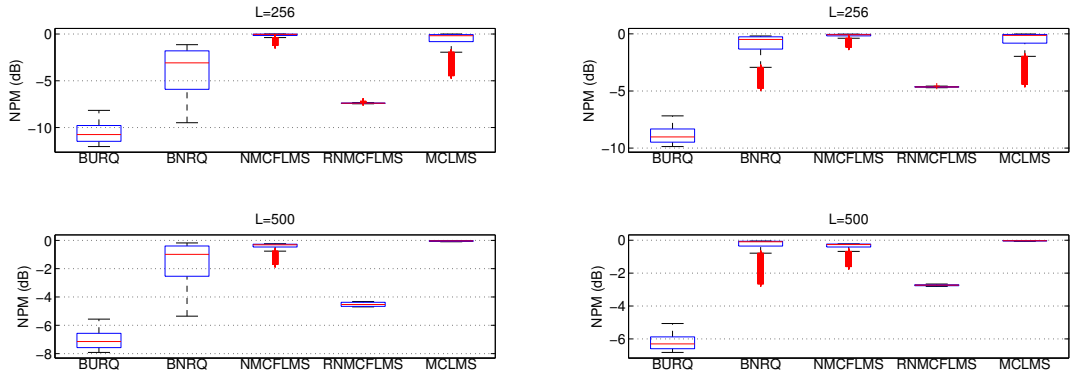
For the BURQ and BNRQ algorithms, a sliding rectangular window of length $B$=2$L$, overlapping by $n_{\text{off}}$=$L$ samples, was used. The different step-sizes were empirically optimized for each SNR such that no misconverge could be observed during the first 20 s of input signal while displaying the highest initial convergence rate. When choosing a step-size, it is usual to look for a step-size which is trade-off between quick initial convergence and low steady-state error [100]. However, since the adaptive BSI algorithms are known to achieve a steady-state of 0 dB of NPM, the step-size is chosen here so that no misconvergence is observed during the first 20 s. The performance of the SURQ algorithm is not shown here as it diverged, as predicted in [98].

### 3.3.3 Results

Figures 3.1 and 3.2 show box plots of the mean NPMs for $M$=2 and $M$=5, where only the samples after 100 s were taken into account (when the algorithms were assumed to

have converged). Unlike the MCLMS, NMCFLMS and BNRQ algorithms, which have misconverged, the RNMCFLMS algorithm has converged in NPM as the interquartile range as well as the spread in NPMs are less than 1 dB. The BURQ algorithm consistently achieves lower NPMs on average.

By comparing the box plots for a fixed SNR and different lengths $L$ in Fig. 3.1 and 3.2, the NPMs increase as $L$ increases. This is due to the fact that near-common zeros are more likely for larger values of $L$ and fixed $M$. By comparing the box plots for fixed SNR and $L$ but different $M$ in Figures 3.1 and Figure 3.2, the NPMs decreases as $M$ increases. This is due to the fact that as $M$ increases, near-common zeros of the RIRs are less likely to appear.

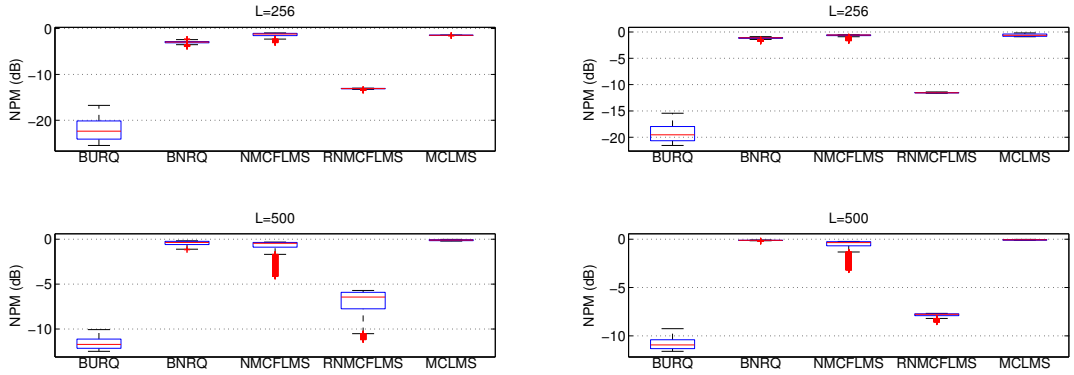



(a) SNR = 15 dB (b) SNR = 10 dB

Figure 3.1: Box plot of the mean NPMs after the first 100 s of signal and $M$=2.

While Fig. 3.1 and 3.2 suggest that the BURQ algorithm consistently outperforms the other algorithms, NPM behavior over time, including the initial convergence rate, cannot be analyzed from these figures. Figure 3.3 provides a more detailed analysis for SNR=15 dB, $L$=500, $M$=5 by displaying the NPMs against time for the considered algorithms. The solid lines represent the mean NPMs, while the dotted lines show the $1^{\text{st}}$ and $9^{\text{th}}$ decile. From this figure it can be observed that, although the MCLMS algorithm has the highest initial convergence rate, it clearly misconverges. In addition, also the NMCFLMS and the BNRQ algorithms misconverge, whereas the RNMCFLMS and the BURQ algorithms have a convergent behavior. Although the RNMCFLMS algorithm

appears to have a faster initial convergence rate than the BURQ algorithm, the BURQ algorithm clearly outperforms the RNMCFLMS algorithm as the NPMs keep decreasing as time increases. For the other combinations of considered SNR, $L$ and $M$, a similar behavior of the NPMs has been observed.




(a) SNR=15 dB (b) SNR=10 dB

Figure 3.2: Box plots of the mean NPMs after the first 100 s of signal and $M$=5.

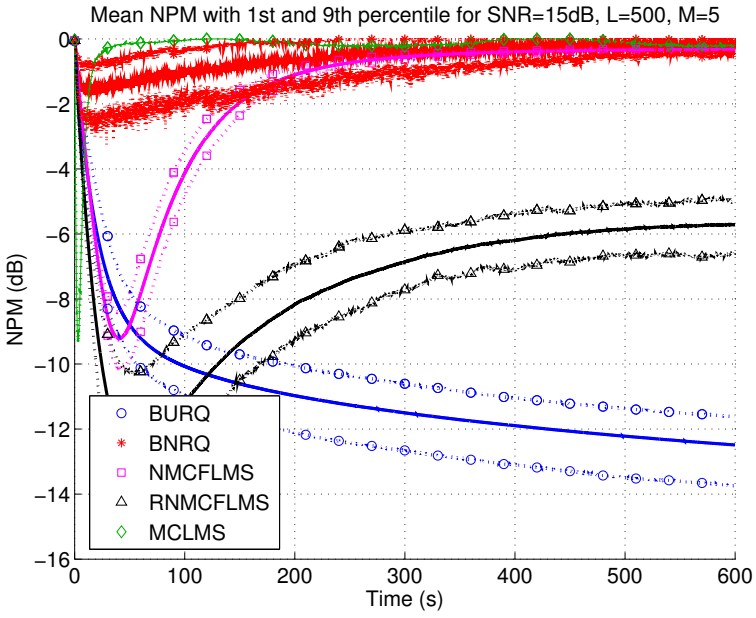



Figure 3.3: Mean NPM for each algorithm for SNR=15 dB, $L$=500 and $M$=5 together with the 1$^{\text{st}}$ and 9$^{\text{th}}$ decile (shown in dotted lines).

Through these simulations, the unit-norm constraint seems to lead to misconver-

gence of the adaptive algorithms. By having a block-based processing, noise robustness is achieved. However, BURQ has a significantly lower initial convergence rate than RNMCFLMS. The use of an empirically chosen step-size is a drawback of the algorithm, which will be addressed in Section 3.4.

## 3.4 Adaptive step-size extension

### 3.4.1 Algorithm derivation

Let us consider the standard update equation (3.6) of BURQ. The updated estimate $\hat{\boldsymbol{h}}(b+1)$ is clearly a linear combination of the previous estimate $\hat{\boldsymbol{h}}(b)$ and the gradient $\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)$ with a coefficient constrained to be 1 on $\hat{\boldsymbol{h}}(b)$. If that constraint is removed, $\hat{\boldsymbol{h}}(b+1)$ can be written as:

$$\hat{\boldsymbol{h}}(b+1) = \begin{bmatrix} \hat{\boldsymbol{h}}(b) & \boldsymbol{\nabla} J_{\mathrm{BURQ}}(b) \end{bmatrix} \boldsymbol{\mu}_{\mathrm{LOG}}(b), \tag{3.9}$$

where $\boldsymbol{\mu}_{\mathrm{LOG}}(b) = \begin{bmatrix} \mu_1(b) \\ \mu_2(b) \end{bmatrix}$ is a $2 \times 1$ vector of weights and $\mu_1(b)$ is equal to 1 in the case of BURQ.

Let us define the Rayleigh quotient residual $\boldsymbol{r}_{\mathrm{RQ}}(b) = \bar{\boldsymbol{R}}(b)\hat{\boldsymbol{h}}(b+1) - \hat{\lambda}_1(b+1)\hat{\boldsymbol{h}}(b+1)$ with $\hat{\lambda}_1(b+1)$ an estimate of the smallest eigenvalue of $\mathbb{E}\{\boldsymbol{R}(n)\}$. Since $\hat{\boldsymbol{h}}(b+1)$ belongs to the subspace spanned by $\hat{\boldsymbol{h}}(b)$ and $\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)$, the new estimate is locally optimal when the residual is orthogonal to that subspace. The weight vector $\boldsymbol{\mu}_{\mathrm{LOG}}$ is therefore such that

$$\begin{bmatrix} \hat{\boldsymbol{h}}^T(b) \\ \boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)^T \end{bmatrix} \bar{\boldsymbol{R}}(b) \begin{bmatrix} \hat{\boldsymbol{h}}(b) & \boldsymbol{\nabla} J_{\mathrm{BURQ}}(b) \end{bmatrix} \boldsymbol{\mu}_{\mathrm{LOG}} =$$
$$\hat{\lambda}_1(b+1) \begin{bmatrix} \hat{\boldsymbol{h}}^T(b) \\ \boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)^T \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{h}}(b) & \boldsymbol{\nabla} J_{\mathrm{BURQ}}(b) \end{bmatrix} \boldsymbol{\mu}_{\mathrm{LOG}}. \tag{3.10}$$

In other words, $\boldsymbol{\mu}_{\mathrm{LOG}}(b)$ is a generalized eigenvector of the $2 \times 2$ matrices

$\begin{bmatrix} \hat{\boldsymbol{h}}^T(b) \\ \boldsymbol{\nabla} J^T_{\mathrm{BURQ}}(b) \end{bmatrix} \bar{\boldsymbol{R}}(b) \begin{bmatrix} \hat{\boldsymbol{h}}(b) & \boldsymbol{\nabla} J_{\mathrm{BURQ}}(b) \end{bmatrix}$ and $\begin{bmatrix} \hat{\boldsymbol{h}}^T(b) \\ \boldsymbol{\nabla} J^T_{\mathrm{BURQ}}(b) \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{h}}(b) & \boldsymbol{\nabla} J_{\mathrm{BURQ}}(b) \end{bmatrix}$ for the generalized eigenvalue $\hat{\lambda}_1(b+1)$.

Provided that $\hat{\boldsymbol{h}}(b)$ and $\boldsymbol{\nabla} J_{\mathrm{BURQ}}(b)$ are made orthonormal before computing $\hat{\boldsymbol{h}}(b+1)$, the matrix on the right hand side of (3.10) is the identity matrix and $\boldsymbol{\mu}_{\mathrm{LOG}}$ is an eigenvector corresponding to the smallest eigenvalue of

$$\bar{\boldsymbol{R}}'(b) = \begin{bmatrix} \hat{\boldsymbol{h}}'^T(b) \\ \boldsymbol{\nabla} J'^T_{\mathrm{BURQ}}(b) \end{bmatrix} \bar{\boldsymbol{R}}(b) \begin{bmatrix} \hat{\boldsymbol{h}}'(b) & \boldsymbol{\nabla} J'_{\mathrm{BURQ}}(b) \end{bmatrix} \tag{3.11}$$

where $\hat{\boldsymbol{h}}'(b) = \frac{\hat{\boldsymbol{h}}(b)}{||\hat{\boldsymbol{h}}(b)||_2}$ and $\boldsymbol{\nabla} J'_{\mathrm{BURQ}}(b)$ is the unit norm vector orthogonal to $\hat{\boldsymbol{h}}'(b)$ obtained via the Gram-Schmidt orthonormalization process [74, Chapter 9].

Since $\boldsymbol{\mu}_{\mathrm{LOG}}$ is defined up to a scaling factor, we chose the one that has a unit-norm so that $\hat{\boldsymbol{h}}(b+1)$ is also unit-norm. The update equation of the proposed Locally Optimal Gradient BURQ (LOG-BURQ) is then given by

$$\hat{\boldsymbol{h}}(b+1) = \mu_1(b)\hat{\boldsymbol{h}}'(b) + \mu_2(b)\boldsymbol{\nabla} J'_{\mathrm{BURQ}}(b) \tag{3.12}$$

where $\mu_1(b)$ and $\mu_2(b)$ are the entries of the unit-norm eigenvector corresponding to the smallest eigenvalue of $\bar{\boldsymbol{R}}'(b)$. Note that by choosing a unit-norm eigenvector for the step-size, the normalization of $\hat{\boldsymbol{h}}(b)$ is reintroduced.

### 3.4.2 Simulation

To illustrate the performance of the proposed algorithm, the LOG-BURQ is evaluated against the NMCFLMS [43] and the Variable Step-Size Unconstrained MCLMS (VSS-UMCLMS) [100] in this section.

#### Setup

The VSS-UMCLMS, NMCFLMS and LOG-BURQ algorithms were evaluated for impulse responses of lengths $L = 128$ and $L = 256$, two different numbers of microphones, ($M$=3 and $M$=5), and two values of SNRs (SNR $\in \{\infty, 30\,\text{dB}\}$). For each combination of $L$, $M$ and SNR, the algorithms were evaluated for 50 realizations of the source and additive noise. The RIRs were generated using the image method [83] for a shoebox-shaped room of dimensions 5 × 6 × 3 m, a reverberation time $T_{60}$=0.5 s and a sampling frequency $f_s$=8 kHz. The length $L$ was then fixed by truncating the obtained RIRs to the desired length. The input signals and the additive noise were white Gaussian noises, uncorrelated with each other. Although longer RIRs are present in practical scenarios, this setup allows us to compare the proposed algorithm against the baseline.

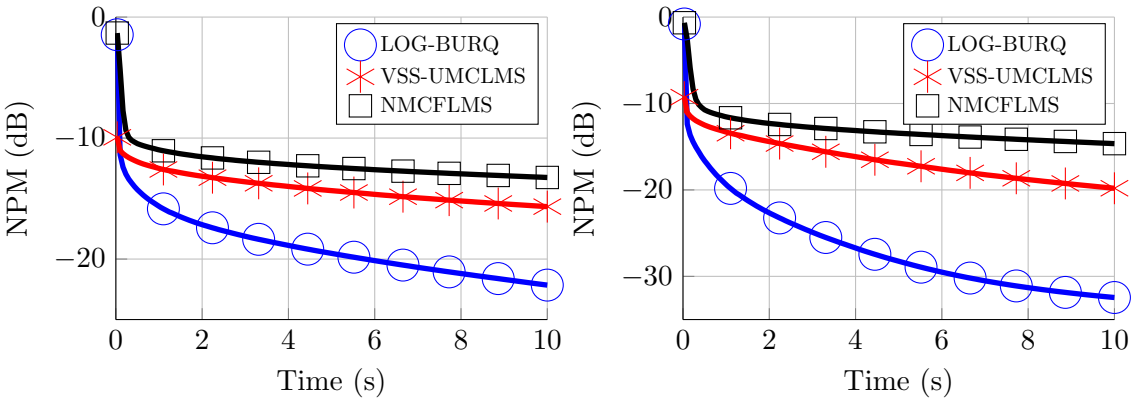
A sliding rectangular window of length $B$=2$L$ overlapping by $n_{\text{off}}$=$L$ samples was used in the LOG-BURQ. The constant step-size of the NMCFLMS was empirically optimized for a toy example and was set to 0.8.

### 3.4.3 Result

Figures 3.4a and 3.4b show that, when no noise is present, the algorithms converge and that an increase of $M$ leads to estimates with lower NPMs as near common zeros between the channels are less likely [112]. In these conditions, the LOG-BURQ achieves significantly lower NPMs on average than the VSS-UMCLMS and the NMCFLMS. For $M$=5, for example, the VSS-UMCLMS and the NMCFLMS respectively achieve average NPMs of -13 dB and -20 dB after 10 s of input data while the proposed algorithm achieves -32 dB.

As shown in Fig. 3.5a and 3.5b, in the presence of additive noise at SNR=30 dB, similar observations can be made although the improvement over the baseline is less significant. However, the initial convergence rate is greatly improved.

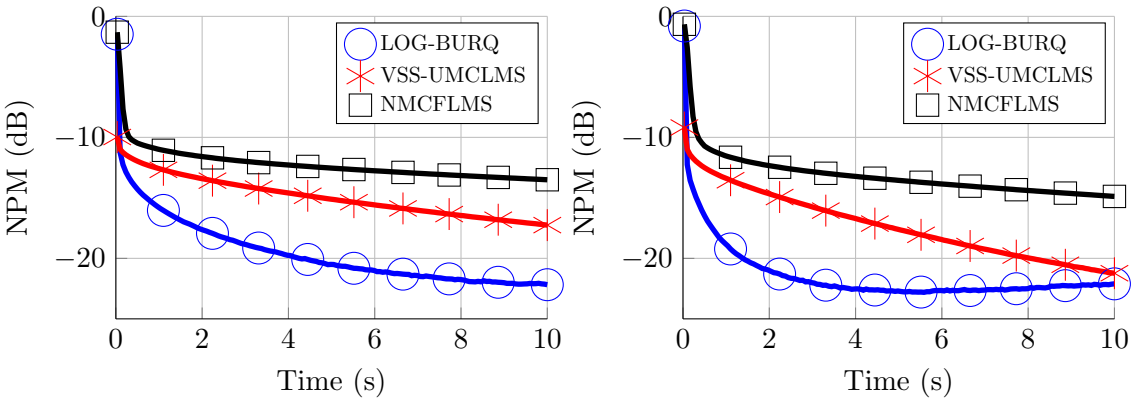
It can be observed that the proposed algorithm outperforms the baseline for $L = 256$ in Fig. 3.6a, 3.6b, 3.7a and 3.7b. Although clearly visible for LOG-BURQ, all the adaptive step-size algorithms have a higher initial convergence rate for an SNR of 30 dB. For $M = 3$




(a) SNR=∞, $M$=3, $L$=128.

(b) SNR=∞, $M$=5, $L$=128.

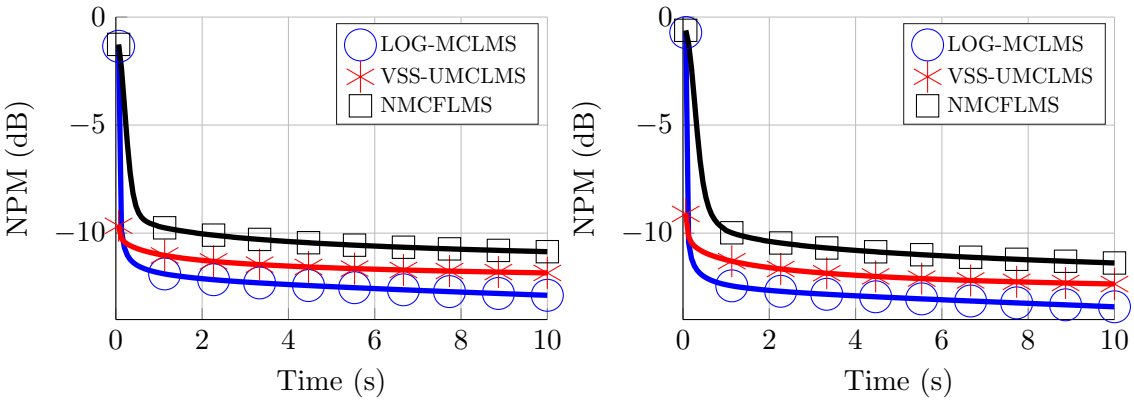
Figure 3.4: Average NPM against time, no additive noise and $L$=128.




(a) SNR=30 dB, M=3, L=128.

(b) SNR=30 dB, M=5, L=128.

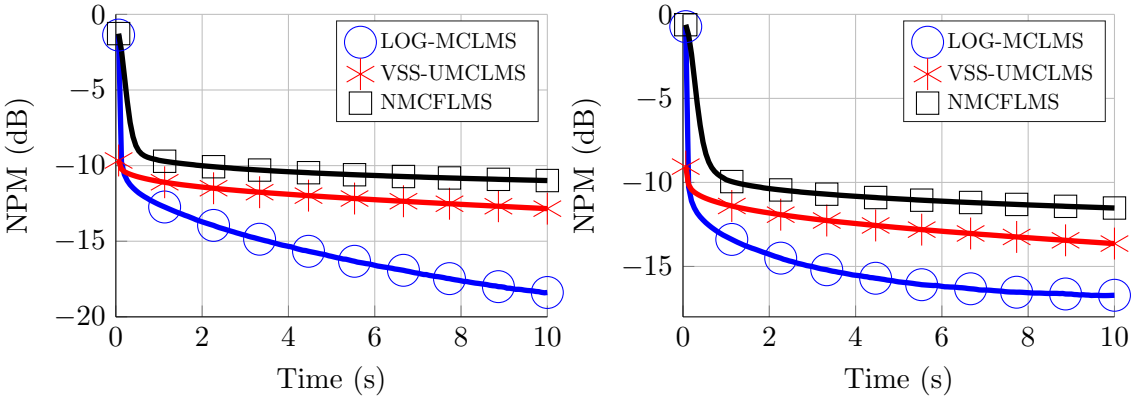
Figure 3.5: Average NPM against time, SNR=30 dB and $L$=128.




(a) SNR=∞, $M$=3, $L$=256.

(b) SNR=∞, $M$=5, $L$=256.

Figure 3.6: Average NPM against time, no additive noise and $L$=256.




(a) SNR=30 dB, M=3, L=256.

(b) SNR=30 dB, M=5, L=256.

Figure 3.7: Average NPM against time, SNR=30 dB and $L$=256.

and $M = 5$, the VSS-UMCLMS indeed achieved NPMs of -13 dB for an SNR of 30 dB, while it achieved an NPM of -12 dB in noiseless conditions.

As is well known with cross-relation error-based algorithms, misconvergence occurs when in the input contains significant noise [113]. As expected, we observed such misconvergence in the proposed as well as the baseline algorithms for SNRs lower than 30 dB. In fact, for $L = 128$, SNR=30 dB and $M = 5$, misconvergence can be seen in Fig. 3.5b. In the presence of noise, the norm of $\hat{\boldsymbol{h}}(n)$ was observed and found to be close to zero after 10 s of input data in the case of VSS-UMCLMS. This can be understood by the fact that the update equation VSS-UMCLMS is exactly that of a gradient descent minimizing a quadratic cost function, the minimum of which is achieved only at the null vector.

## 3.5 RTF constrained BSI

The frequency domain BSI algorithms, namely NMCFLMS [43] and its extensions, only operate in the frequency domain to efficiently compute convolution products between $y_i[n]$ and $\hat{h}_j[n]$ in the calculation of the cross-relation residual. However, the RIRs are still estimated via the time domain cross-relation residual, which means that all of the $ML$ variables of the stacked impulse responses are estimated at the same time. Therefore the number of variables that need to be estimated increases linearly with the number of microphones. In other words, although spatial diversity increases as the number of microphones increases, the convergence rate does not necessarily significantly increase as the dimension of the considered system also increases [114]. We take a different approach here and exploit properties of the reverberant signals in the frequency domain to estimate and use the RTF for BSI.

The RTF is a linear filter, which relates the multi-channel signal at each microphone to that of a reference microphone in the frequency domain [115]. If the RTFs were available, the problem of BSI would then be reduced to estimating the impulse response of the reference channel, which may significantly reduce the number of unknowns. The estimation of the RTF has been investigated in the context of speech enhancement in [115]–[117].

However, in [115], the proposed method requires analysis frames significantly longer than the time-domain RTF while the methods proposed in [116] and [117] were concerned with the estimation of the RTF when the multiplicative transfer approximation does not hold.

In this section, the cross-relation is analyzed in the frequency domain and shown to contain the RTF. The RTF is then exploited to reduce the dimensionality of the problem, which enables the estimation of realistically long RIRs.

Similarly to [42], [43], the derivations are made in the absence of additive noise. However, the performance of the proposed method will be evaluated against noisy data.

### 3.5.1 Cross-relation error in the frequency domain

As stated in Chapter 2, in the noiseless case, the pairwise cross-relation residual $r_{ij}(n)$, evaluated at $h_i[n]$ and $h_j[n]$, is equal to 0 for all $n$, which means that

$$r_{ij}[n] = h_i[n] * y_j[n] - h_j[n] * y_i[n] = 0. \tag{3.13}$$

In the frequency domain, (3.13) becomes

$$\underline{\boldsymbol{r}}_{ij} = \underline{\boldsymbol{h}}_i \odot \underline{\boldsymbol{y}}_j - \underline{\boldsymbol{h}}_j \odot \underline{\boldsymbol{y}}_i, \tag{3.14}$$

where $\underline{\boldsymbol{r}}_{ij}$, $\underline{\boldsymbol{h}}_i$ and $\underline{\boldsymbol{y}}_j$, each of length $K$, are the Fourier transforms of $r_{ij}[n]$, $h_i[n]$ and $y_j[n]$ respectively and $\odot$ denotes the element wise product. The equality in (3.14) holds whenever the Fourier transform of the whole reverberant signal is taken [118]. This imposes the number of frequency bins to be equal to $K = N_s + L - 1$, where $N_s$ is the number of samples in the source signal.

Applying Parseval's theorem [119, Chapter 5] to the sample mean cross-relation residual, we obtain

$$\sum_{n=1}^{N} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} r_{ij}^2(n) = \sum_{k=1}^{K} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} |\underline{r}_{ij}(k)|^2 = 0. \tag{3.15}$$

Although similar to the time-domain expression of the norm of the cross-relation residual, the frequency-domain cross-relation error has a major difference worth highlight-

ing.

For a given time index $n$, $r_{ij}(n)$ is a function of $\hat{\boldsymbol{h}}_i$ and $\hat{\boldsymbol{h}}_j$. In other words, to exploit all the available spatial information, all the channels have to be considered at the same time. The time domain cross-relation residual therefore involves $ML$ variables in total.

For a given frequency bin $k$, on the other hand, the frequency domain cross-relation residual is expressed as

$$\underline{r}_{ij}(k) = \hat{\underline{h}}_i(k)\underline{y}_j(k) - \hat{\underline{h}}_j(k)\underline{y}_i(k). \tag{3.16}$$

Therefore, for a given frequency bin, the frequency domain cross-relation residual only involves $M$ variables in total.

In other words, the cross-relation based BSI approach in the frequency domain consists of independently minimizing a cost function which involves $M$ variables at each frequency bins. Therefore, by exploiting the SIMO structure of the reverberant signal and by expressing the norm of the cross-relation residual in the frequency domain, the complexity of BSI is dramatically reduced as the problem now consists of minimizing $K$ independent problems each involving $M$ variables.

As in the time domain, let us now express $\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}|\underline{r}_{ij}(k)|^2$ in a more compact way than (3.16). This will provide us with an insight into the nature of the considered optimization problem.

For a given frequency bin $k$, using (3.16) we have

$$\begin{aligned}\sum_{i=1}^{M}\sum_{j=1}^{M}|\underline{r}_{ij}(k)|^2 = \sum_{i=1}^{M}\sum_{j=1}^{M}\Big(&|\hat{\underline{h}}_i(k)\underline{y}_j(k)|^2 + |\hat{\underline{h}}_j(k)\underline{y}_i(k)|^2 \\ &- \hat{\underline{h}}_i(k)\hat{\underline{h}}_j^*(k)\underline{y}_j(k)\underline{y}_i^*(k) \\ &- \hat{\underline{h}}_i^*(k)\hat{\underline{h}}_j(k)\underline{y}_j^*(k)\underline{y}_i(k)\Big).\end{aligned} \tag{3.17}$$

Let us introduce the frequency-domain cross-relation matrix $\underline{\boldsymbol{Y}}(k)$ and the frequency-

domain cross-channel estimate $\hat{\underline{\boldsymbol{h}}}(k)$ at frequency bin $k$ respectively defined as

$$\underline{\boldsymbol{Y}} = \begin{bmatrix} \sum_{m,m\neq 1} |\underline{y}_m|^2 & -\underline{y}_2^*\underline{y}_1 & \cdots & -\underline{y}_M^*\underline{y}_1 \\ -\underline{y}_1^*\underline{y}_2 & \sum_{m,m\neq 2} |\underline{y}_m|^2 & \cdots & -\underline{y}_M^*\underline{y}_2 \\ \vdots & \vdots & \ddots & \vdots \\ -\underline{y}_1^*\underline{y}_M & -\underline{y}_2^*\underline{y}_M & \cdots & \sum_{m,m\neq M} |\underline{y}_m|^2 \end{bmatrix}, \tag{3.18}$$

$$\hat{\underline{\boldsymbol{h}}}(k) = \begin{bmatrix} \hat{\underline{h}}_1(k) & \hat{\underline{h}}_2(k) & \dots & \hat{\underline{h}}_M(k) \end{bmatrix}^T, \tag{3.19}$$

where the frequency bin index $k$ has been omitted in (3.18) for conciseness.

The expansion of $2\hat{\underline{\boldsymbol{h}}}^H(k)\underline{\boldsymbol{Y}}(k)\hat{\underline{\boldsymbol{h}}}(k)$ using (3.19) together with (3.18) yields the same expression as the right-hand side of (3.17). However, since $|\underline{r}_{ij}(k)|^2$ is equal to $|\underline{r}_{ji}(k)|^2$ for all pairs $(i,j)$ and $|\underline{r}_{ii}(k)| = 0$, we have

$$\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} |\underline{r}_{ij}(k)|^2 = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} |\underline{r}_{ij}(k)|^2.$$

Hence

$$\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} |\underline{r}_{ij}(k)|^2 = \hat{\underline{\boldsymbol{h}}}^H(k)\underline{\boldsymbol{Y}}(k)\hat{\underline{\boldsymbol{h}}}(k). \tag{3.20}$$

Therefore minimizing the norm of the cross-relation residual in the frequency domain subject to the constraint $||\hat{\underline{\boldsymbol{h}}}(k)|| > 0$ is equivalent to finding an eigenvector corresponding to the smallest eigenvalue of $\underline{\boldsymbol{Y}}(k)$. Therefore, by minimizing the norm of the cross-relation residual in the frequency domain, a relationship between the different channels is found for each frequency bin. However, since the frequency bins are processed independently from each other, a scaling ambiguity between frequency bins arises when the time-domain RIRs are to be estimated. This is similar to the problem investigated in [108], where the authors solved a scaling factor ambiguity between estimates of the subband RIRs. Unlike the approach taken in [108], estimates of the subband RIRs are not assumed known here. However, the estimation of the full-band time-domain RIRs will follow the same idea as in [108] and will be presented in the next section.

In noiseless conditions and for each frequency bin, the minimization of the norm of

cross-relation residual in the frequency domain therefore yields an accurate estimate of the relative relationship between the different channels with respect to an arbitrary reference channel. In other words, an estimate of the RTF has been found.

The RTF allows us to express $M-1$ RIRs as a function of a reference RIR. By using the RTF, the number of unknowns is therefore reduced to the number of frequency bins and made independent of the number of channels.

### 3.5.2 Parametrization of the RIR using the RTF

Because of the scaling inconsistency across frequency bins inherent to the minimization of the norm of the cross-relation residual in the frequency domain, the RIRs cannot be obtained by a straight inversion of the estimated RTFs. However, as in [108], since the time-domain cross-relation binds all the frequency bins together, it can be used to resolve that inconsistency.

Let us first express $\hat{\boldsymbol{h}}$ as a function of the scaling coefficients $\underline{\hat{\boldsymbol{\alpha}}} = \begin{bmatrix} \underline{\hat{\alpha}}(1) & \underline{\hat{\alpha}}(2) & \dots & \underline{\hat{\alpha}}(K) \end{bmatrix}^T$ to be applied to each frequency bin. For that purpose, let us introduce $\underline{\hat{\boldsymbol{h}}}_m = \begin{bmatrix} \underline{\hat{h}}_m(1) & \underline{\hat{h}}_m(2) & \dots & \underline{\hat{h}}_m(K) \end{bmatrix}^T$ obtained by reordering the terms found in $\underline{\hat{\boldsymbol{h}}}(k)$ and the matrix $\underline{\hat{\boldsymbol{H}}}$ defined as

$$\underline{\hat{\boldsymbol{H}}} = \begin{bmatrix} \boldsymbol{\mathcal{D}}_{\underline{\hat{\boldsymbol{h}}}_1} & \boldsymbol{\mathcal{D}}_{\underline{\hat{\boldsymbol{h}}}_2} & \dots & \boldsymbol{\mathcal{D}}_{\underline{\hat{\boldsymbol{h}}}_M} \end{bmatrix}^T. \tag{3.21}$$

We have then the following equality

$$\underline{\hat{\boldsymbol{h}}} = \underline{\hat{\boldsymbol{H}}}\underline{\hat{\boldsymbol{\alpha}}}. \tag{3.22}$$

On the right-hand side of (3.22), the same scaling coefficient $\underline{\hat{\alpha}}(k)$ is indeed applied to $\underline{\hat{h}}_m(k)$ for a given $k$ and for all $m$. The relative gain structure between channels, expressed in the estimated RTF, is therefore preserved while the scaling is independent across frequencies. The product $\underline{\hat{\boldsymbol{H}}}\underline{\hat{\boldsymbol{\alpha}}}$ therefore corresponds to the Fourier transforms of the estimated RIRs, scaled by $\underline{\hat{\alpha}}(k)$ at each frequency bin and stacked in one column

vector.

The time-domain estimated stacked RIRs $\hat{\boldsymbol{h}}$ can be retrieved from $\underline{\hat{\boldsymbol{H}}\hat{\boldsymbol{\alpha}}}$ by applying an inverse Fourier transform to the non-overlapping sub-vectors of length $K$ and by keeping the first $L$ entries of each channel i.e.

$$\hat{\boldsymbol{h}} = \left(\boldsymbol{\mathcal{I}}_M \otimes \left(\boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right)\right) \underline{\hat{\boldsymbol{H}}\hat{\boldsymbol{\alpha}}}, \tag{3.23}$$

where $\boldsymbol{\mathcal{S}}_s^L = \begin{bmatrix} \boldsymbol{\mathcal{I}}_L & \boldsymbol{0}_{L\times(K-L)} \end{bmatrix}$.

Using that parametrization of $\hat{\boldsymbol{h}}$ as a function of $\underline{\hat{\boldsymbol{H}}}$ and $\underline{\hat{\boldsymbol{\alpha}}}$, the optimization problem corresponding to the LS method for BSI is written as

$$\begin{aligned} \min_{\underline{\boldsymbol{\alpha}}} \quad & \left\| \boldsymbol{R}_{\boldsymbol{y}} \left(\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right) \underline{\hat{\boldsymbol{H}}\hat{\boldsymbol{\alpha}}} \right\|_2^2 \\ \text{subject to} \quad & \|\underline{\hat{\boldsymbol{\alpha}}}\|_2 > 0. \end{aligned} \tag{3.24}$$

### 3.5.3 Full-band RIR estimation

The optimization problem (3.24) does not sufficiently constrain the scaling coefficients as the time domain vector representing the estimated stacked RIRs only has real valued entries, which implies that its Fourier transform is conjugate symmetric. Since $\boldsymbol{y}_m$ is real valued, its Fourier transform $\underline{\boldsymbol{y}}_m$ is also conjugate symmetric [81, Chapter 5]. Therefore $\underline{\boldsymbol{Y}}(k)$ is also conjugate symmetric with respect to $k$. Therefore, given the structure of $\underline{\hat{\boldsymbol{H}}}$, the scaling coefficients must be conjugate symmetric.

However, since $\underline{\hat{\boldsymbol{\alpha}}}$ is conjugate symmetric, it can be represented as a linear transformation of a real vector $\hat{\boldsymbol{\alpha}}$. Let us introduce the $2p \times 2p$ matrix $\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p$ defined as

$$\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p = \begin{bmatrix} \boldsymbol{\mathcal{I}}_p & \jmath\boldsymbol{\mathcal{I}}_p \\ \boldsymbol{\mathcal{J}}_p & -\jmath\boldsymbol{\mathcal{J}}_p \end{bmatrix}, \tag{3.25}$$

where $\jmath = \sqrt{-1}$, $p$ is an integer and $\boldsymbol{\mathcal{J}}_p$ is a $p \times p$ matrix with 1 at the entries on the anti-diagonal and 0 elsewhere.

The real and imaginary parts of the first $p$ entries of a vector $\underline{\boldsymbol{a}}$, image of an arbitrary

vector $\boldsymbol{a} \in \mathbb{R}^{2p}$ under the mapping defined by $\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p$, are respectively given by the first and last $p$ entries of $\boldsymbol{a}$. The last $p$ rows of $\underline{\boldsymbol{a}}$ are a complex conjugated version of the first $p$ entries of $\underline{\boldsymbol{a}}$ copied in reverse order. Therefore $\underline{\boldsymbol{a}}$ is conjugate symmetric.

Likewise, for a given conjugate symmetric vector $\underline{\boldsymbol{a}}$, the first and last $p$ entries of $\boldsymbol{a}$, image of $\underline{\boldsymbol{a}}$ under the mapping defined by $\frac{1}{2}\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p^H$, respectively correspond to the real and imaginary parts of the first $p$ entries of $\underline{\boldsymbol{a}}$.

The matrix $\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p$ therefore defines an isomorphism from the vector space $\mathbb{R}^{2p}$ to the real subspace of conjugate symmetric vectors. Using $\boldsymbol{\mathcal{J}}_p^2 = \boldsymbol{\mathcal{I}}_p$, we have $\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p^H \underline{\tilde{\boldsymbol{\mathcal{A}}}}_p = 2\boldsymbol{\mathcal{I}}_p$ and $\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p^{-1} = \frac{1}{2}\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p^H$ follows.

Let us now establish the mapping between $\hat{\boldsymbol{\alpha}}$ and $\underline{\hat{\boldsymbol{\alpha}}}$. Defining the matrix $\underline{\boldsymbol{\mathcal{A}}}_K$ as

$$\underline{\boldsymbol{\mathcal{A}}}_K = \begin{cases} \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & & & & 0 \\ \vdots & & \underline{\tilde{\boldsymbol{\mathcal{A}}}}_{\frac{K-2}{2}} & & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}, & \text{for } K \text{ even}, \\ \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \underline{\tilde{\boldsymbol{\mathcal{A}}}}_{\frac{K-1}{2}} & \\ 0 & & & \end{bmatrix}, & \text{for } K \text{ odd} \end{cases} \tag{3.26}$$

and using our previous observation on $\underline{\tilde{\boldsymbol{\mathcal{A}}}}_p$, we have that for any conjugate symmetric $\underline{\hat{\boldsymbol{a}}}$, there exists a unique vector $\hat{\boldsymbol{a}} \in \mathbb{R}^{2p}$ such that

$$\underline{\hat{\boldsymbol{a}}} = \underline{\tilde{\boldsymbol{\mathcal{A}}}}_p \hat{\boldsymbol{a}}. \tag{3.27}$$

The representation (3.27) of $\underline{\hat{\boldsymbol{\alpha}}}$ reformulates the symmetric conjugate constraint into a classical unconstrained optimization problem where there is no dependency between the different entries of the sought vector.

The symmetric conjugate constraint applied to (3.24) means that the imaginary part of $\boldsymbol{R}_{\boldsymbol{y}} \left(\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right) \underline{\hat{\boldsymbol{H}}\hat{\boldsymbol{\alpha}}}$ is equal to the null vector. Therefore, only the real part of $\boldsymbol{R}_{\boldsymbol{y}} \left(\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right) \underline{\hat{\boldsymbol{H}}\hat{\boldsymbol{\alpha}}}$ needs to be considered in the cost function of (3.24).

Let us denote by $\boldsymbol{R}_{\underline{\hat{\boldsymbol{H}}}}$ the matrix given by

$$\boldsymbol{R}_{\underline{\hat{\boldsymbol{H}}}} = \Re\left\{ \left( \left(\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right) \underline{\hat{\boldsymbol{H}}\boldsymbol{\mathcal{A}}}_K \right)^H \boldsymbol{R}_{\boldsymbol{y}}^H \boldsymbol{R}_{\boldsymbol{y}} \left( \left(\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right) \underline{\hat{\boldsymbol{H}}\boldsymbol{\mathcal{A}}}_K \right) \right\}, \tag{3.28}$$

where $\Re\{\cdot\}$ denotes the real part of the argument.

Minimizing (3.24) subject to a symmetric conjugate and a non-nullity constraint on $\hat{\boldsymbol{\alpha}}$ is therefore equivalent to the following constrained optimization problem

$$\begin{aligned} \min_{\hat{\boldsymbol{\alpha}} \in \mathbb{R}^K} \quad & \hat{\boldsymbol{\alpha}}^T \boldsymbol{R}_{\underline{\hat{\boldsymbol{H}}}} \hat{\boldsymbol{\alpha}} \\ \text{subject to} \quad & \|\hat{\boldsymbol{\alpha}}\| > 0. \end{aligned} \tag{3.29}$$

The optimization problem (3.29) is equivalent to finding an eigenvector corresponding to the smallest eigenvalue of the positive-semidefinite matrix $\boldsymbol{R}_{\underline{\hat{\boldsymbol{H}}}}$.

However, when the number of frequency bins $K$ is larger than the length of the impulse responses $L$, the last $K - L$ terms of each of the estimated RIRs do not contribute to the cost function. Therefore there may exist a high-dimensional subspace of $\mathbb{C}^K$ which minimizes the cost function (3.24). In the noiseless case, however, as the smallest eigenvalue of $\boldsymbol{R}_{\boldsymbol{y}}^T \boldsymbol{R}_{\boldsymbol{y}}$ has algebraic multiplicity one, the first $L$ entries of each estimated RIR obtained via (3.24) are uniquely determined up to a scaling factor. Therefore, in the noiseless case, the solution subspace to (3.29) yields estimate of the RIRs which will differ in the last $K - L$ last entries of each estimated RIR.

The proposed algorithm, which blindly estimates the stacked RIRs from the observed signals $y_m[n]$, is summarised in Algorithm 1.

**Algorithm 1** Blind identification algorithm

**procedure** ALGORITHM($y_m[n], M, L$)

▷ Estimate the RTFs

**for** $m \in \{1, 2, \ldots, M\}$ **do**
Compute $\underline{\boldsymbol{y}}_m$, the DFT of $y_m[n]$
**end for**
**for** $k \in \{1, 2, \ldots, K\}$ **do**
Construct $\underline{\boldsymbol{Y}}(k)$ using (3.18)
Compute $\hat{\underline{\boldsymbol{h}}}(k)$ as the eigenvector corresponding to the smallest eigenvalue of $\underline{\boldsymbol{Y}}(k)$
**end for**

▷ Estimate the RIRs

Compute $\boldsymbol{R}_{\hat{\underline{\boldsymbol{H}}}}$ as defined in (3.28)
Estimate $\hat{\boldsymbol{\alpha}}$ as an eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{R}_{\hat{\underline{\boldsymbol{H}}}}$
Compute $\hat{\underline{\boldsymbol{\alpha}}}$ using (3.27)
Compute $\hat{\underline{\boldsymbol{h}}}$ using (3.22)
**for** $m \in \{1, 2, \ldots, M\}$ **do**
Compute $\hat{\boldsymbol{h}}_m$ as the inverse DFT of $\hat{\underline{\boldsymbol{h}}}_m$
**end for**
**end procedure**

### 3.5.4 Implementation details

Although written in a matrix formulation, the algorithm described in Algorithm 1 can exploit the structure of several of the matrices for an efficient implementation.

The frequency domain cross-relation matrix given in (3.18) is obtained as the diagonal matrix with $\sum_{m=1}^{M} |\underline{y}_m|^2$ on its diagonal from which the outer product of $\begin{bmatrix} \underline{y}_1(k) & \underline{y}_2(k) & \cdots & \underline{y}_M(k) \end{bmatrix}^T$ has been subtracted.

Likewise, the product $\left(\boldsymbol{\mathcal{I}}_M \otimes \left(\boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right)\right) \hat{\underline{\boldsymbol{H}}} \underline{\boldsymbol{\mathcal{A}}}_K$ can be implemented without using matrix products. The term $\boldsymbol{\mathcal{I}}_M \otimes \left(\boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right)$ means that $\boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}$ is applied to $M$ non-overlapping blocks of size $K \times K$ of $\hat{\underline{\boldsymbol{H}}} \underline{\boldsymbol{\mathcal{A}}}_K$. Since each of the $m^{\text{th}}$ block of size $K \times K$ of $\hat{\underline{\boldsymbol{H}}}$ is the diagonal matrix $\boldsymbol{\mathcal{D}}_{\hat{\underline{\boldsymbol{h}}}_m}$, $\underline{\boldsymbol{\mathcal{F}}}_K^{-1} \hat{\underline{\boldsymbol{H}}}$ is the inverse DFT matrix, the $k^{\text{th}}$ column of which has been multiplied by $\hat{\underline{h}}_m(k)$. The application of $\boldsymbol{\mathcal{S}}_s^L$ removes the $(L+1)^{\text{th}}$ to $K^{\text{th}}$ lines of $\underline{\boldsymbol{\mathcal{F}}}_K^{-1} \hat{\underline{\boldsymbol{H}}}$. To analyze the effect of right multiplying a matrix $\boldsymbol{A}$ of size $A \times K$ by

$\underline{\boldsymbol{\mathcal{A}}}_K$, let us assume $K$ is odd and let us decompose $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} & \boldsymbol{a}_{12}^T & \boldsymbol{a}_{13}^T \\ a_{11} & & \\ & \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{a}_{21} & & \\ & \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix},$$

where $\boldsymbol{a}_{21}$ is a $A-1$ entry vector, $\boldsymbol{a}_{12}$ and $\boldsymbol{a}_{13}$ are $\frac{K-1}{2}$ entry vectors and $\boldsymbol{A}_{ij}$, $i,j \in \{1,2\}$ are matrices with $\frac{K-1}{2}$ columns and an arbitrary number of rows.

The product $\boldsymbol{A}\underline{\boldsymbol{\mathcal{A}}}_K$ is then equal to

$$\boldsymbol{A}\underline{\boldsymbol{\mathcal{A}}}_K = \begin{bmatrix} & \boldsymbol{a}_{12}^T + \boldsymbol{a}_{13}^T \boldsymbol{\mathcal{J}}_{\frac{K-1}{2}} & \jmath\left(\boldsymbol{a}_{12}^T - \boldsymbol{a}_{13}^T \boldsymbol{\mathcal{J}}_{\frac{K-1}{2}}\right) \\ a_{11} & \boldsymbol{A}_{11} + \boldsymbol{A}_{12} \boldsymbol{\mathcal{J}}_{\frac{K-1}{2}} & \jmath\left(\boldsymbol{A}_{11} - \boldsymbol{A}_{12} \boldsymbol{\mathcal{J}}_{\frac{K-1}{2}}\right) \\ \boldsymbol{a}_{21} & & \\ & \boldsymbol{A}_{21} + \boldsymbol{A}_{22} \boldsymbol{\mathcal{J}}_{\frac{K-1}{2}} & \jmath\left(\boldsymbol{A}_{21} - \boldsymbol{A}_{22} \boldsymbol{\mathcal{J}}_{\frac{K-1}{2}}\right) \end{bmatrix},$$

which shows that the operation can be expressed as column-wise manipulations. The first column of $\boldsymbol{A}\underline{\boldsymbol{\mathcal{A}}}_K$ is the first column of $\boldsymbol{A}$, the $2^{\text{nd}}$ to $(\frac{K-1}{2}+1)^{\text{th}}$ columns are given by a copy of the second to $(\frac{K-1}{2}+1)^{\text{th}}$ columns of $\boldsymbol{A}$ added to a left/right flip of the $(\frac{K-1}{2}+2)^{\text{th}}$ to $K^{\text{th}}$ columns of $\boldsymbol{A}$. The $(\frac{K-1}{2}+2)^{\text{th}}$ to $K^{\text{th}}$ columns of $\boldsymbol{A}\underline{\boldsymbol{\mathcal{A}}}_K$ are given by $\jmath$ multiplied by a copy of the $2^{\text{nd}}$ to $(\frac{K-1}{2}+1)^{\text{th}}$ columns of $\boldsymbol{A}$ from which a left/right flip of the $(\frac{K-1}{2}+2)^{\text{th}}$ to $K^{\text{th}}$ columns of $\boldsymbol{A}$ is subtracted.

### 3.5.5 Experimental setup and evaluation

#### Setup

Algorithm 1 is evaluated for RIRs generated by the image method, as implemented in [120], for a sampling frequency of $f_s$=8 kHz. The shoebox-shaped room had dimensions of 5 × 6 × 3 m and a reverberation time $T_{60}$ of 0.3 s. The source was 2 m away from the centre of the linear $M$=5 microphone array. The microphones were separated by 0.3 m.

The input signal is chosen to be a white Gaussian noise. The algorithm was evaluated for two different durations of the input signal, namely 100 ms and 600 ms. The reverberant signals, which result from the convolution of the input signal with the RIRs, are then

corrupted by additive white Gaussian noise such that the SNR was equal to $\infty$, 60 dB and 30 dB. The RIRs are truncated to $L$=300 and $L$=$T_{60}f_s$.

When the convolution theorem holds, i.e. when the DFTs of the captured signals are equal to the product of the DFTs of the RIRs with that of the source signal, the analysis in Section 3.5.1 showed that, in the absence of noise, the estimation of the RTF was a well-behaved problem as the number of unknowns can be arbitrarily set. However, the assumption that the convolution theorem holds is restrictive as this means that the number of frequency bins is imposed. The full decay after an offset of the input signal, i.e. the full result of the convolution between the source and the RIRs, has indeed to be observed. Therefore, we also investigate the effect of windowing on the accuracy of the RTF. A measure, appropriate for the use of an estimated RTF in BSI, is consequently proposed.

**Lower bound of NPM given an estimated RTF**

Let us define the lowest achievable NPM on an estimated stacked RIR vector obtained from a given RTF using (3.22). For conciseness, let us denote by $\boldsymbol{G}_{\underline{\hat{\boldsymbol{H}}}}$ the matrix $\left(\boldsymbol{\mathcal{I}}_M \otimes \left(\boldsymbol{\mathcal{S}}_s^L \underline{\boldsymbol{\mathcal{F}}}_K^{-1}\right)\right) \underline{\hat{\boldsymbol{H}}}$.

Let us decompose $\boldsymbol{h}$ into the components $\boldsymbol{h}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}} = \mathcal{P}^{\perp}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}}(\boldsymbol{h})$ and $\boldsymbol{h}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}{}^{\perp}} = \mathcal{P}^{\perp}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}{}^{\perp}}(\boldsymbol{h})$, where $\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}$ is the subspace spanned by the columns of $\boldsymbol{G}_{\underline{\hat{\boldsymbol{H}}}}$. Since these two components belong to two orthogonal spaces, the squared norm of $\boldsymbol{h}$ is equal to $||\boldsymbol{h}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}}||_2^2 + ||\boldsymbol{h}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}_2}{}^{\perp}}||_2^2$.

As explained in Section 2, the numerator of the NPM is the norm of $\boldsymbol{h}$ projected onto the orthogonal complement of $\hat{\mathfrak{h}} = \text{span}(\hat{\boldsymbol{h}})$. However, according to (3.23), $\hat{\boldsymbol{h}}$, estimated from an estimated RTF using (3.22), is a linear combination of the columns of $\boldsymbol{G}_{\underline{\hat{\boldsymbol{H}}}}$. In other words, the optimal estimate obtained by (3.23) must lie in $\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}$.

Therefore, the numerator of the NPM is equal to $||\boldsymbol{h}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}{}^{\perp}}||_2^2 + ||\boldsymbol{h}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}} - \mathcal{P}^{\perp}_{\hat{\mathfrak{h}}}(\boldsymbol{h})||_2^2$. The second term of the previous sum represents the distance between $\boldsymbol{h}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}} = \mathcal{P}^{\perp}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}}(\boldsymbol{h})$ and its projection on to a chosen $\hat{\boldsymbol{h}}$, which also belongs to $\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}$ as $\hat{\boldsymbol{h}}$ belongs to $\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}$. Hence, the choice of the optimal $\hat{\boldsymbol{h}}$, given an estimated RTF and $\boldsymbol{h}$, can only change the second term of the previous sum. That second term can be reduced to 0 if $\hat{\boldsymbol{h}}$ is chosen to be the

orthogonal projection of $\boldsymbol{h}$ onto $\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}$. The first term, on the other hand, represents the error that cannot be recovered from the errors in the RTF by using (3.23). The optimal $\hat{\boldsymbol{h}}$, obtained from an estimated RTF using (3.22), is therefore given by

$$\hat{\boldsymbol{h}}_{\underline{\hat{\boldsymbol{H}}}} = \mathcal{P}^{\perp}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}}(\boldsymbol{h}). \tag{3.30}$$

Any non-zero scaling of $\mathcal{P}^{\perp}_{\mathfrak{G}_{\underline{\hat{\boldsymbol{H}}}}}(\boldsymbol{h})$ is also optimal in the NPM sense as NPM is scale independent.

A lower bound on the achievable NPM using the estimated RTF is then given by computing the NPM between $\hat{\boldsymbol{h}}_{\underline{\hat{\boldsymbol{H}}}}$ and $\boldsymbol{h}$.

### 3.5.6 Experimental results

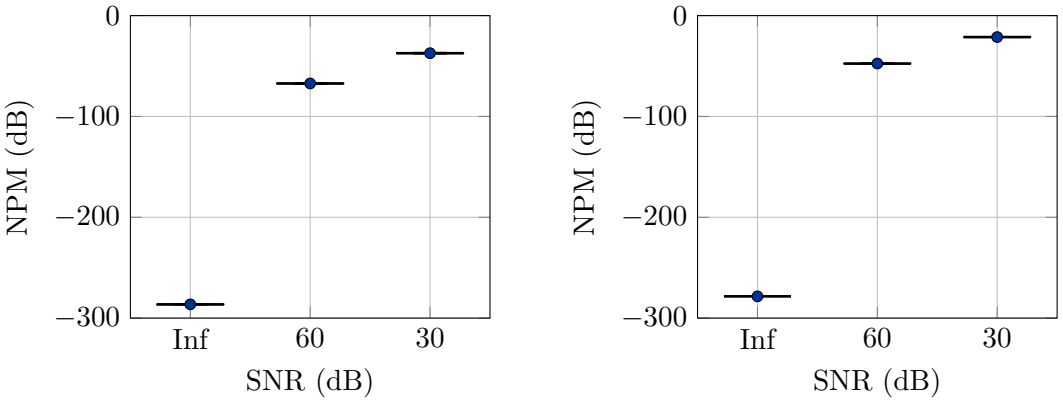
#### Relative transfer function accuracy

The blind estimation algorithm described in Section 3.5.3 strongly relies on an estimated RTF. Although accurate estimation of the RTF is beyond the scope of the presented work, the performance of the estimation method proposed in Section 3.5.1 is analysed in terms of noise robustness and approximation of the convolution theorem.

Figure 3.8 shows the effect of noise on the lower bound of NPM that can be achieved for different values of $L$ when the convolution theorem holds. As the SNR decreases, the lower bound increases. For $L$=300, the best achievable NPM has a median of -37 dB and an interquartile range of 0.6 dB for an SNR of 30 dB. However, as the length of the impulse response increases, the sensitivity to noise also increases. For $L$=2400 and an SNR of 30 dB, the best achievable NPM using the estimated RTF using the method previously described has a median of -21 dB and an interquartile range of 0.1 dB.

However, the full reverberation decay may not be available so that windowing is necessary to obtain an accurate estimate of the spectrum of $y_m$ [121, Chapter 2]. From Fig. 3.8, it seems that additive white noise has mild effects on the lower bound of NPM. Therefore, a Hamming window, which exhibits a main lobe width of approximately $\frac{4}{K}$ in normalized frequency while having a relative sidelobe attenuation of -41 dB [121, Chapter

2], is chosen.

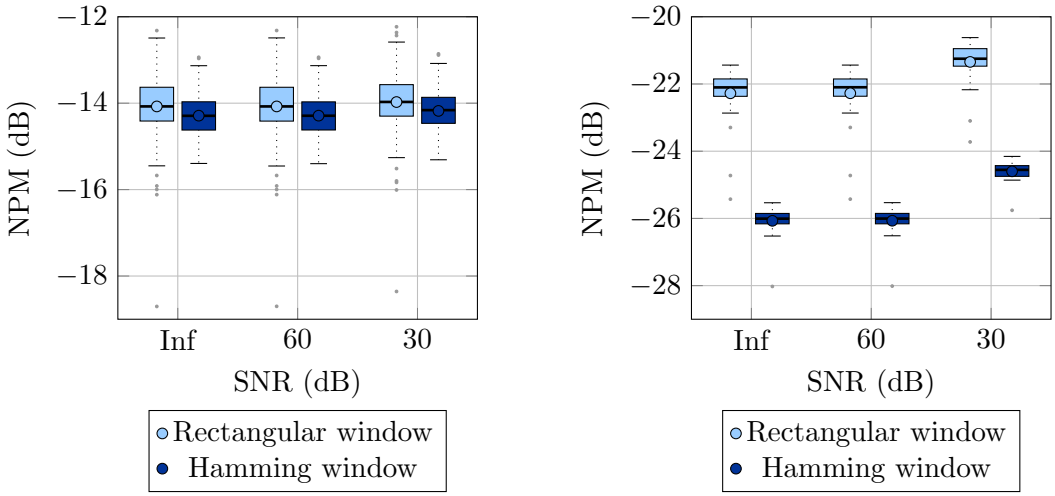



(a) $L$=300, $K$=1099. (b) $L$=2400, $K$=3199.

Figure 3.8: Lower bound of NPM when the convolution theorem holds.

In the noiseless case, for example, the stacked RIRs can be perfectly estimated when the convolution theorem holds. However, when a window is applied, the obtained spectrum is the convolution between the Fourier transform of the reverberant signal and that of the window [121, Chapter 2]. Figure 3.9 shows that the estimation errors introduced by the analysis window result in estimated RTFs that achieve at best an NPM of -15 dB for $L$=300.

Figure 3.9 also shows that the application of a Hamming window instead of a rectangular window may improve the lowest achievable NPM by a few dB and therefore yield a more accurate estimate of the RIRs. When $L$ is equal to 2400, that lower bound is 4 dB lower when a Hamming window is applied. This may be due to the fact that a longer Hamming window has a narrower main lobe width.




(a) $L$=300, $K$=600.

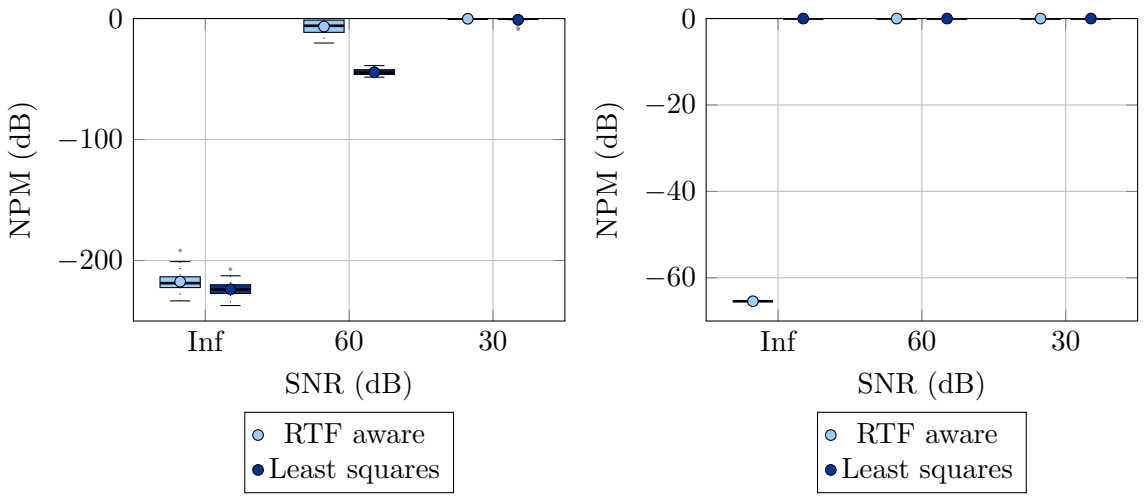
(b) $L$=2400, $K$=4800.

Figure 3.9: Effect of windowing on the lower bound of achievable NPM.

### Full-band room impulse response estimation

Figure 3.10 shows that the RTF aware method is able to estimate the RIRs when the duration of the input signal is short compared to the duration of the RIRs. However, as the LS method, the RTF method is sensitive to noise. It can be observed in Fig. 3.10a that the LS method has higher noise robustness. However that difference in noise robustness disappears as the length of the RIRs increases as shown in Fig. 3.10b.

By comparing Fig. 3.10b and 3.11 for the noiseless case, the least-squares method is able to accurately estimate the RIRs when the duration of the input signal is increased to 600 ms. This suggests that not enough samples were available to compute an accurate cross-relation matrix when the input signal had a duration of 100 ms. However, increasing the duration of the input signal to 600 ms does not improve the noise robustness.




(a) $L$=300. (b) $L$=2400.

Figure 3.10: NPM of the estimated RIRs with a 100 ms white noise input signal and various levels of additive noise.

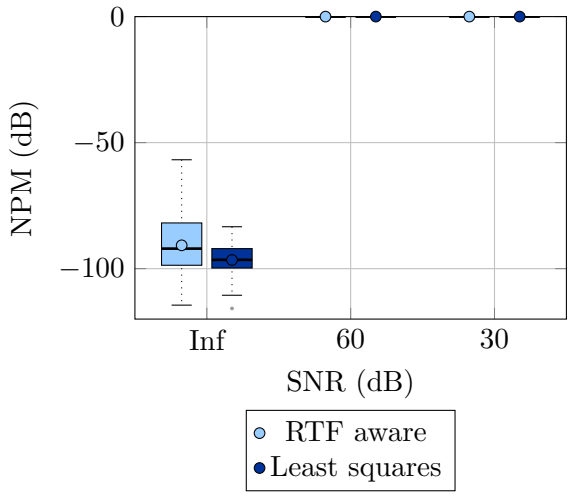



Figure 3.11: NPM of the estimated RIRs with a 600 ms white noise input signal and various levels of additive noise and $L$=2400.

## 3.6 Summary

In this chapter, the noise robustness issue of adaptive BSI algorithms was investigated. Unlike the NMCFLMS and MCLMS, the presented BURQ algorithm showed a converging behavior of the NPM with time. To improve its convergence rate, two approaches were

considered.

The first approach was formulated in the form of a locally optimal step-size. The derivations of the adaptive step-size relied on the structure of the problem, i.e. the estimation of an eigenvector of a symmetric matrix. The resulting algorithm showed a higher convergence rate than what has been proposed in [100]. However, when the level of additive noise increases, misconvergence is observed for both algorithms.

The second approach was motivated by the fact that the convergence rate of adaptive algorithms decreases as the number of unknowns increases. The minimization of the norm of the cross-relation residual, in the frequency domain, was shown to yield an estimate of the RTF. The estimation of the time-domain RIRs using the RTF then involved the estimation of weights for each frequency bin. Simulations showed that the estimation of these weights using the LS method was very sensitive to both additive noise and errors in the estimated RTF. However, the lower bound of achievable NPM using a noisy estimate of the RTF indicates that a trade-off between perfectly estimated RIRs and a lower number of variables to be estimated is possible. By considering a problem with a lower number of variables, the convergence rate of adaptive algorithms can be improved.

The various methods presented here minimized the norm of the cross-relation residual to estimate the RIRs. However, it has been observed that poor estimation, i.e. NPM close to 0 dB, was inevitable whenever a realistic setting was considered, i.e. one with realistically long RIRs and in the presence of additive noise. The next chapter investigates the formulation of an alternative cost function, which better characterizes the sought RIRs.

# Chapter 4

# Regularized blind system identification

## 4.1 Overview

In the previous chapter, the LS method and the adaptive algorithms have been shown to have limited ability for blind identification of long impulse responses such as those encountered in acoustic systems. The increase of the condition number of the eigenvector associated with the smallest eigenvalue of the noiseless cross-relation matrix as the length of the impulse responses increases indicates that the optimization surface considered in BSI becomes flat over a subspace of dimensionality larger than 2. In this chapter, the exploitation of complementary information to the SIMO structure of the problem is investigated. In particular, after describing a novel interpretation of the LS method, the Sequential Spectral Flatness (SSF) algorithm, which exploits the non-stationarity of speech, is proposed. Since RIRs are to be estimated, the cost function of SSF is regularized so as to be able to estimate realistically long RIRs from reverberant noisy speech signals.

## 4.2 Introduction

While the previous chapter investigated BSI algorithms robust to additive noise for arbitrary impulse responses, the problem of blind identification of acoustic SIMO systems finds

application in blindly estimating RIRs from reverberant noisy speech signals. Therefore prior information about RIRs can be used to regularize the cost functions.

The idea of using characteristics of the RIRs to achieve noise robustness is not new. Hasan, Benesty, Naylor, *et al.* [82] constrained the direct path of the estimated RIRs to be exactly equal to that of the ground truth RIRs. This idea was further explored by Ahmad, Khong, Hasan, *et al.* in [122], where the author reformulated the constrained optimization problem of [82] into an unconstrained one. However, both methods require knowledge of the direct paths, both in time instant and amplitude.

Haque and Hasan adopted another approach in the RNMCFLMS proposed in [111]. The authors observed that the RIRs estimated by a misconverged NMCFLMS were band limited in the frequency domain when actual RIRs would be broadband [111]. Therefore, to improve the noise robustness of the NMCFLMS, Haque and Hasan proposed to add a spectral flatness regularization term to the cost function. He, Lu, Chen, *et al.* later changed the quadratic form cost function of the NMCFLMS to an M-estimator and combined that cost function with the regularization term of the RNMCFLMS to achieve robustness against impulsive additive noise such as keyboard noises [123]. However, despite these modifications, the ability of the different algorithms to estimate impulse responses of more than a few hundred taps, e.g. RIRs, may not be reliably possible.

In this chapter, characteristics of RIRs are exploited in an optimization framework to blindly estimate the RIRs. In Section 4.3, an alternative interpretation of the cross-relation in noisy environments is given. That novel interpretation is then exploited together with the non-stationary nature of speech to propose a novel characterization of the sought RIRs. In Section 4.4, two regularization terms are proposed and combined with the objective function found in Section 4.3 so that the estimated RIRs resemble RIRs. Finally the work presented in this chapter is summarized in Section 4.5.

## 4.3 Spectral characteristics of the residual at the solution

### 4.3.1 Annihilation filter

All the BSI algorithms in the class under consideration share the idea of minimizing the norm of the cross-relation residual, an approach which is motivated by an analysis in noiseless conditions. However in noisy conditions, the minimization of the norm of the cross-relation residual may not be statistically optimal, as discussed in [32], [36]. Let us, here, analyze the cross-relation residual at the solution, which can also be seen as the desired cross-relation residual.

Using $\hat{\boldsymbol{h}} = \boldsymbol{h}$ in (2.16) together with (2.19), we find that the residual, at the solution, is equal to $\boldsymbol{R}_{\boldsymbol{\nu}}\boldsymbol{h}$ as $\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{h} = \boldsymbol{0}$, i.e the cross-relation residual that we obtain at the solution is the additive noise filtered by the RIRs. In other words, the ground truth RIRs annihilate the reverberant signals. Therefore, the residual, at the solution, is a colored noise. More precisely, using (2.1) and $\hat{\boldsymbol{h}} = \boldsymbol{h}$ in (2.9) we see that, at the solution, the pairwise residual $r_{ij}(n)$ between channel $i$ and $j$ is given by

$$\begin{aligned} r_{ij}(n) &= \boldsymbol{h}_i^T \boldsymbol{\nu}_j(n) - \boldsymbol{h}_j^T \boldsymbol{\nu}_i(n), \\ r_{ij}(n) &= \begin{bmatrix} \boldsymbol{h}_i \\ \boldsymbol{h}_j \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\nu}_j(n) \\ -\boldsymbol{\nu}_i(n) \end{bmatrix}, \end{aligned} \tag{4.1}$$

which is the stacked noise signal $\boldsymbol{\nu}_{ji} = \begin{bmatrix} \boldsymbol{\nu}_j(n)^T & -\boldsymbol{\nu}_i(n)^T \end{bmatrix}^T$ filtered by the stacked RIRs $\boldsymbol{h}_{ij} = \begin{bmatrix} \boldsymbol{h}_i^T & \boldsymbol{h}_j^T \end{bmatrix}^T$.

In other words, regardless of the level of noise, under the identifiability conditions, there exists a 1-dimensional subspace of $\mathbb{R}^{ML}$, the elements of which annihilate the reverberant signal from the reverberant noisy signal. Hence, in a SIMO acoustic system setup, the sought RIRs remove the speech content from the observed signals by application of the cross-relation residual. This characterization will be exploited in the next section to formulate a new cost function.

### 4.3.2 Spectrum of the residual at the solution

Under the assumption that the entries of $\boldsymbol{\nu}_{ji}$ are drawn from a second order stationary process with PSD $\varphi_{\nu_{ji}}(k)$, the PSD of the residual, at the solution, is therefore given by [121, Chapter 1]:

$$\varphi_{r_{ij}}(k) = |\mathrm{H}_{ij}(k)|^2 \varphi_{\nu_{ji}}(k), \tag{4.2}$$

where $\mathrm{H}_{ij}(k)$ is the frequency response of $\boldsymbol{h}_{ij}$ and $k$ is the frequency bin.

Furthermore, if the stacked RIRs are assumed to have a flat magnitude spectrum, as it has successfully been applied in [111], the power spectrum of the pairwise residual $\bar{\boldsymbol{r}}_{ij}(b)$ at frame $b$ has the same shape as that of the stacked noise.

### 4.3.3 Additive white Gaussian noise case

Let us, here, derive an alternative cost function relying on the previous observation for $M$=2. For conciseness, the subscript $\{\cdot\}_{12}$ is omitted in the equations.

#### Alternative cost function

Under the assumption that the system is corrupted by spatio-temporally white Gaussian noise, $\boldsymbol{\nu}$ is a vector of i.i.d. normally distributed random variables such that the residual, at the solution, has a flat magnitude spectrum. To find an analytical expression of the gradient of the flat magnitude spectrum regularization term, the terms of the cost function are considered in the log domain. Given a finite amount of data, the pairwise stacked RIRs $\hat{\boldsymbol{h}}$ must therefore maximize the log spectral flatness of the $N_r$ data points contained in the windowed residual [124]. The spectral flatness of a signal is defined [124] as the ratio between the geometric and arithmetic means of its power spectrum:

$$J_{\mathrm{SF}}(\hat{\boldsymbol{h}}) = \frac{1}{N_r} \sum_{k=1}^{N_r} \log\left(|\underline{\bar{r}}_b(k)|^2\right) - \log\left(\sum_{k=1}^{N_r} |\underline{\bar{r}}_b(k)|^2\right) + \log(N_r), \tag{4.3}$$

where $\underline{\bar{r}}_b(k)$ is the value of the DFT of a windowed residual $\bar{\boldsymbol{r}}(b)$ at frequency bin $k$ and frame $b$.

By application of the Jensen inequality [125] to the concave function $\log(\cdot)$, $J_{\mathrm{SF}}(\hat{\boldsymbol{h}})$ is always negative. The first term on the right hand side of (4.3) is indeed the logarithm of the geometric mean while the last two terms constitutes the logarithm of the arithmetic mean. Since $J_{\mathrm{SF}}(\hat{\boldsymbol{h}})$ is the ratio of the geometric and arithmetic mean, it is also scale independent.

Let us now derive the gradient $\boldsymbol{\nabla} J_{\mathrm{SF}}(\hat{\boldsymbol{h}})$ of $J_{\mathrm{SF}}(\hat{\boldsymbol{h}})$, which is obtained by successive applications of the chain rule [126, Appendix A].

### Gradient of the geometric mean term

For clarity of the derivations of $\boldsymbol{\nabla} J_{\mathrm{SF}}(\hat{\boldsymbol{h}})$, let us denote by $\boldsymbol{b}_k$ the vector of $\mathbb{R}^{N_r}$ with 1 at the $k^{\mathrm{th}}$ entry and 0 elsewhere. Introducing

$$\underline{\bar{\boldsymbol{r}}}_b = \begin{bmatrix} \underline{\bar{r}}_b(1) & \underline{\bar{r}}_b(2) & \cdots & \underline{\bar{r}}_b(N_r) \end{bmatrix}^T, \tag{4.4}$$

the product $|\boldsymbol{b}_k^T \underline{\bar{\boldsymbol{r}}}_b|^2 = |\underline{\bar{r}}_b(k)|^2$ is then equal to

$$|\underline{\bar{r}}_b(k)|^2 = \left(\boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} \boldsymbol{W} \bar{\boldsymbol{R}}_{\boldsymbol{y}}(b) \hat{\boldsymbol{h}}\right)^H \boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} \boldsymbol{W} \bar{\boldsymbol{R}}_{\boldsymbol{y}}(b) \hat{\boldsymbol{h}}, \tag{4.5}$$

where

$$\boldsymbol{W} = \begin{bmatrix} w(1) & 0 & \dots & 0 \\ 0 & w(2) & \dots & 0 \\ 0 & \vdots & \ddots & 0 \\ 0 & 0 & \dots & w(N_r) \end{bmatrix} \tag{4.6}$$

is a matrix representing a windowing operation and, denoting by $b_{\mathrm{off}} = (b-1)n_{\mathrm{off}}$ for conciseness,

$$\bar{\boldsymbol{R}}_{\boldsymbol{y}}(b) = \begin{bmatrix} \boldsymbol{\mathcal{H}}_{N_r \times L}\{\boldsymbol{y}_2(b_{\mathrm{off}} + N_r + L - 1)\} & -\boldsymbol{\mathcal{H}}_{N_r \times L}\{\boldsymbol{y}_1(b_{\mathrm{off}} + N_r + L - 1)\} \end{bmatrix} \tag{4.7}$$

is the rectangular cross-relation matrix for frame $b$, composed of two $N_r \times L$ Hankel matrices, each built from the samples contained in

$$\boldsymbol{y}_m(b_{\text{off}} + N_r + L - 1) = \begin{bmatrix} y_m(b_{\text{off}} + N_r + L - 1) \\ y_m(b_{\text{off}} + N_r + L - 2) \\ \dots \\ y_m(b_{\text{off}} + N_r) \end{bmatrix},$$

i.e.

$$\begin{aligned} &\boldsymbol{\mathcal{H}}_{N_r \times L}\{\boldsymbol{y}_m(b_{\text{off}} + N_r + L - 1)\} \\ &= \begin{bmatrix} y_m(b_{\text{off}} + N_r + L - 1) & y_m(b_{\text{off}} + N_r + L - 2) & \dots & y_m(b_{\text{off}} + N_r) \\ y_m(b_{\text{off}} + N_r + L - 2) & y_m(b_{\text{off}} + N_r + L - 3) & \dots & y_m(b_{\text{off}} + N_r - 1) \\ \vdots & \vdots & \ddots & \vdots \\ y_m(b_{\text{off}} + L) & y_m(b_{\text{off}} + L - 1) & \dots & y_m(b_{\text{off}} + 1) \end{bmatrix}. \end{aligned}$$

Note that, although square in this section, the matrix $\boldsymbol{W}$ is not necessarily square as we may want to keep only a section of the residual. Rectangular matrices $\boldsymbol{W}$ will, in fact, be extensively used in Section 4.3.4.

Then, omitting the frame index $b$ for conciseness, we have

$$\begin{aligned} \frac{\partial |\bar{\underline{r}}_b(k)|^2}{\partial \hat{\boldsymbol{h}}} &= \left( \bar{\boldsymbol{R}}_{\boldsymbol{y}}^T \boldsymbol{W}^T \boldsymbol{\mathcal{F}}_{N_r}^H \boldsymbol{b}_k \boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} \boldsymbol{W} \bar{\boldsymbol{R}}_{\boldsymbol{y}} + \left( \bar{\boldsymbol{R}}_{\boldsymbol{y}}^T \boldsymbol{W}^T \boldsymbol{\mathcal{F}}_{N_r}^H \boldsymbol{b}_k \boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} \boldsymbol{W} \bar{\boldsymbol{R}}_{\boldsymbol{y}} \right)^T \right) \hat{\boldsymbol{h}} \\ &= \bar{\boldsymbol{R}}_{\boldsymbol{y}}^T \boldsymbol{W}^T \left( \boldsymbol{\mathcal{F}}_{N_r}^H \boldsymbol{b}_k \boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} + \left( \boldsymbol{\mathcal{F}}_{N_r}^H \boldsymbol{b}_k \boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} \right)^T \right) \boldsymbol{W} \bar{\boldsymbol{R}}_{\boldsymbol{y}} \hat{\boldsymbol{h}} \\ &= 2 \bar{\boldsymbol{R}}_{\boldsymbol{y}}^T \boldsymbol{W}^T \Re\{ \boldsymbol{\mathcal{F}}_{N_r}^H \boldsymbol{b}_k \boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} \} \boldsymbol{W} \bar{\boldsymbol{R}}_{\boldsymbol{y}} \hat{\boldsymbol{h}}, \end{aligned}$$

as the DFT matrix is symmetric [127, Chapter 2].

Hence

$$\frac{\partial \log \left( |\bar{\underline{r}}_b(k)|^2 \right)}{\partial \hat{\boldsymbol{h}}} = \frac{2}{|\bar{\underline{r}}_b(k)|^2} \bar{\boldsymbol{R}}_{\boldsymbol{y}}^T \boldsymbol{W}^T \Re\{ \boldsymbol{\mathcal{F}}_{N_r}^H \boldsymbol{b}_k \boldsymbol{b}_k^T \boldsymbol{\mathcal{F}}_{N_r} \} \boldsymbol{W} \bar{\boldsymbol{R}}_{\boldsymbol{y}} \hat{\boldsymbol{h}}$$

and

$$\frac{1}{N_r}\sum_{k=1}^{N_r}\frac{\partial \log\left(|\underline{\bar{r}}_b(k)|^2\right)}{\partial \hat{\boldsymbol{h}}} = \frac{2}{N_r}\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\left(\sum_{k=1}^{N_r}\Re\{\frac{1}{|\underline{r}(k)|^2}\boldsymbol{\mathcal{F}}_{N_r}^H\boldsymbol{b}_k\boldsymbol{b}_k^T\boldsymbol{\mathcal{F}}_{N_r}\}\right)\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}}$$

follows.

However, the product $\boldsymbol{b}_k\boldsymbol{b}_k^T$ is a $N_r \times N_r$ matrix with its only non zero entry at $(k,k)$. Therefore we have

$$\frac{1}{N_r}\sum_{k=1}^{N_r}\frac{\partial \log\left(|\underline{\bar{r}}_b(k)|^2\right)}{\partial \hat{\boldsymbol{h}}} = \frac{2}{N_r}\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\Re\{\boldsymbol{\mathcal{F}}_{N_r}^H\boldsymbol{\mathcal{D}}_{|\breve{\underline{\boldsymbol{r}}}_b|^2}\boldsymbol{\mathcal{F}}_{N_r}\}\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}}, \tag{4.8}$$

where $|\breve{\underline{\boldsymbol{r}}}_b|^2 = \left[\frac{1}{|\underline{\bar{r}}_b(1)|^2} \quad \frac{1}{|\underline{\bar{r}}_b(2)|^2} \quad \cdots \quad \frac{1}{|\underline{\bar{r}}_b(N_r)|^2}\right]$.

**Gradient of the spectral flatness cost function**

Since $\boldsymbol{\mathcal{F}}_{N_r}^* = N_r\boldsymbol{\mathcal{F}}_{N_r}^{-1}$ [127, Chapter 2], the arithmetic mean is given by

$$\log\left(\sum_{k=1}^{N_r}|\underline{\bar{r}}_b(k)|^2\right) = \log\left(N_{N_r}\hat{\boldsymbol{h}}^T\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}}\right),$$

which yields

$$\frac{\partial \log\left(\sum_{k=1}^{N_r}|\underline{\bar{r}}_b(k)|^2\right)}{\partial \hat{\boldsymbol{h}}} = \frac{2}{||\boldsymbol{W}\bar{\boldsymbol{r}}(b)||^2}\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}}, \tag{4.9}$$

as $\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}$ is symmetric.

Hence

$$\begin{aligned}\boldsymbol{\nabla} J_{\mathrm{SF}}(\hat{\boldsymbol{h}}) =& 2\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\Re\{\boldsymbol{\mathcal{F}}_{N_r}^{-1}\boldsymbol{\mathcal{D}}_{|\breve{\underline{\boldsymbol{r}}}_b|^2}\boldsymbol{\mathcal{F}}_{N_r}\}\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}} - \frac{2}{||\boldsymbol{W}\bar{\boldsymbol{r}}(b)||^2}\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}} \\ =& 2\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{W}^T\boldsymbol{\mathcal{F}}_{N_r}^{-1}\left(\boldsymbol{\mathcal{D}}_{|\breve{\underline{\boldsymbol{r}}}_b|^2} - \frac{1}{||\boldsymbol{W}\bar{\boldsymbol{r}}(b)||^2}\boldsymbol{\mathcal{I}}_{N_r}\right)\boldsymbol{\mathcal{F}}_{N_r}\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}}.\end{aligned} \tag{4.10}$$

**Efficient implementation**

The expression of $\boldsymbol{\nabla} J_{\mathrm{SF}}(\hat{\boldsymbol{h}})$ shown in (4.10) indicates that matrix multiplications can be avoided.

The matrix $\boldsymbol{\mathcal{F}}_{N_r}^{-1}\left(\boldsymbol{\mathcal{D}}_{|\breve{\underline{\boldsymbol{r}}}_{\boldsymbol{b}}|^2}-\frac{1}{||\boldsymbol{W}\bar{\boldsymbol{r}}(b)||^2}\boldsymbol{\mathcal{I}}_{N_r}\right)\boldsymbol{\mathcal{F}}_{N_r}$ indeed represents the circular convolution matrix obtained from the vector, the entries of which are given on the diagonal of $\boldsymbol{\mathcal{D}}_{|\breve{\underline{\boldsymbol{r}}}_b|^2}-\frac{1}{||\boldsymbol{W}\bar{\boldsymbol{r}}(b)||^2}\boldsymbol{\mathcal{I}}_{N_r}$ [128, Chapter 3]. Hence the product $\boldsymbol{\mathcal{F}}_{N_r}^{-1}\left(\boldsymbol{\mathcal{D}}_{|\breve{\underline{\boldsymbol{r}}}_b|^2}-\frac{1}{||\boldsymbol{W}\bar{\boldsymbol{r}}(b)||^2}\boldsymbol{\mathcal{I}}_{N_r}\right)\boldsymbol{\mathcal{F}}_{N_r}\boldsymbol{W}\bar{\boldsymbol{R}}_{\boldsymbol{y}}\hat{\boldsymbol{h}}$ is given as the circular convolution between the windowed residual and the vector, the entries of which are given by the inverse DFT of the diagonal entries of $\boldsymbol{\mathcal{D}}_{|\breve{\underline{\boldsymbol{r}}}_b|^2}-\frac{1}{||\boldsymbol{W}\bar{\boldsymbol{r}}(b)||^2}\boldsymbol{\mathcal{I}}_{N_r}$.

If the whole residual is windowed, $\boldsymbol{W}$ is a diagonal matrix so that $\boldsymbol{W}^T\boldsymbol{a}$ and $\boldsymbol{W}\boldsymbol{a}$ are both obtained by an element-wise product.

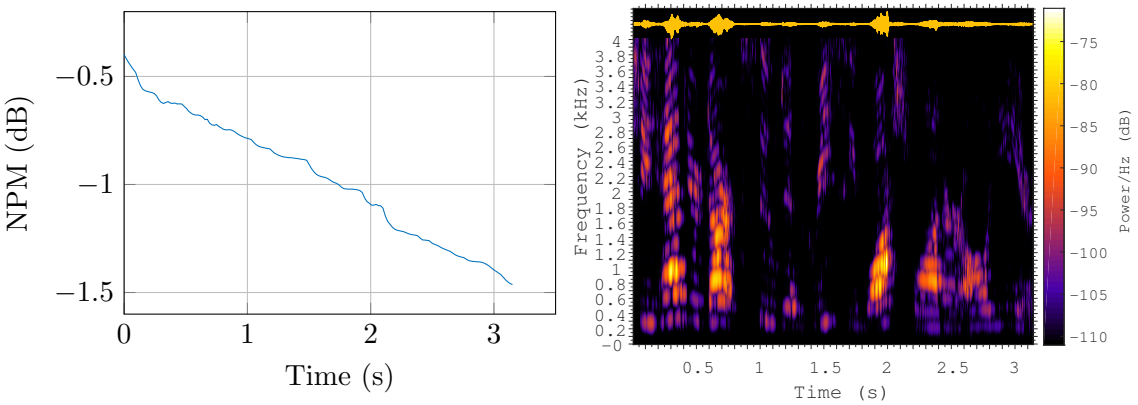
By definition, $\bar{\boldsymbol{R}}_{\boldsymbol{y}}$ is composed of two Hankel matrices put one after the other. Therefore, $\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T$ is composed of two Hankel matrices stacked one after the other. The result of $\bar{\boldsymbol{R}}_{\boldsymbol{y}}^T\boldsymbol{a}$ is then obtained by the computation of two linear convolutions: the first $L$ rows are given by the last $L$ terms of the convolution between the samples $y_2(n)$ considered in $\bar{\boldsymbol{R}}_{\boldsymbol{y}}(b)$ and $\boldsymbol{a}$ while the last $L$ rows correspond to the last $L$ terms of the convolution between the samples $-y_1(n)$ considered in $\bar{\boldsymbol{R}}_{\boldsymbol{y}}(b)$ and $\boldsymbol{a}$.

The gradient of $J_{\mathrm{SF}}(\hat{\boldsymbol{h}})$ is therefore obtained as the result of two element-wise products, two convolutions and one circular convolution.

### 4.3.4 Speech corrupted by stationary additive noise

When the input signal is speech and the additive noise is assumed stationary, the residual resulting from the NMCFLMS or RNMCFLMS may contain speech.

Figure 4.1b shows the spectrogram of the residual when the reverberant noisy signals are filtered by the RIRs estimated by RNMCFLMS. The input signal was the utterance named `SA1.WAV` spoken by speaker `FAKSO` of folder `DR1` of the TIMIT database [129]. The silence preceding the utterance has been removed. The reverberant signal was obtained by convolving the speech signal with two image method generated RIRs of length $L = 300$ for a shoebox-shaped room of dimension $5 \times 6 \times 3$ m. The reverberation time was $T_{60} = 0.3$ s and the centre of the $M = 2$ microphone array was at (2.5, 2, 1.6). The microphones were separated by 20 cm. White Gaussian noise was added so that the overall SNR was 30 dB.




(a) Accuracy of the RNMCFLMS in terms of NPM

(b) Spectrogram of the residual

Figure 4.1: Residual at the RIRs estimated via the RNMCFLMS after 3 s of reverberant noisy speech for $L = 300$ and an SNR of 30 dB.

The step-size used in the RNMCFLMS was $\mu = 0.1$ so that no misconvergence is observed, as shown in Fig. 4.1a. The residual is then obtained by applying the cross-relation with $\hat{\boldsymbol{h}}$ estimated at the end of the 3 s of signal.

As shown in Fig. 4.2, the residual at the solution has a significantly different spectrogram to what is shown in Fig. 4.1b. The spectrogram of the residual at $\boldsymbol{h}$ indeed displays a fairly consistent spectrum across time, whereas the spectrogram of the residual at $\hat{\boldsymbol{h}}$ still contains a significant amount of speech. Therefore, the stationarity of the desired residual can be exploited to further characterize the desired RIRs.

The residual, computed over a frame, can be split into smaller subframes. The spectral flatness cost function characterizing the desired RIRs is then applied to each of these subframes. Since the desired residual is stationary, the spectral flatness cost function should have a small variance within a frame while being small at each subframe. The resulting algorithm will be referred to as the Sequential Spectral Flatness.

Formally, let us denote by $\bar{\boldsymbol{r}}_1(b), \bar{\boldsymbol{r}}_2(b), \ldots, \bar{\boldsymbol{r}}_{B_F}(b)$ the $B_F$ subframes in frame $b$, each containing $N_{\underline{r}}$ samples. The SSF cost function $J_{\mathrm{SSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b))$ is obtained as

$$J_{\mathrm{SSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b)) = \sum_{i=1}^{B_F} J_{\mathrm{SF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}_i(b))^2 \tag{4.11}$$

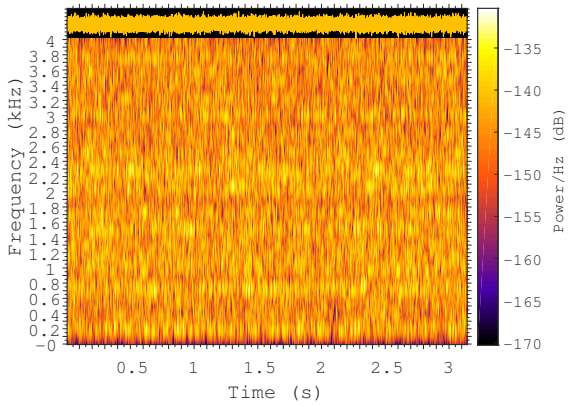



Figure 4.2: Spectrogram of the residual at the solution.

and the gradient is then given by

$$\frac{\partial J_{\mathrm{SSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b))}{\partial \hat{\boldsymbol{h}}} = 2 \sum_{i=1}^{B_F} J_{\mathrm{SF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}_i(b)) \frac{\partial J_{\mathrm{SF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}_i(b))}{\partial \hat{\boldsymbol{h}}}, \tag{4.12}$$

where $J_{\mathrm{SF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}_i(b))$ is the spectral flatness cost function applied to the subframed residual $\bar{\boldsymbol{r}}_i(b)$ at $\hat{\boldsymbol{h}}$, $\frac{\partial J_{\mathrm{SF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}_i(b))}{\partial \hat{\boldsymbol{h}}}$ its derivative with respect to $\hat{\boldsymbol{h}}$ evaluated at $(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}_i(b))$ and $\bar{\boldsymbol{r}}_i(b)$ is related to $\bar{\boldsymbol{r}}(b)$ by

$$\bar{\boldsymbol{r}}_i(b) = \boldsymbol{W}_i \bar{\boldsymbol{r}}(b), \tag{4.13}$$

where

$$\boldsymbol{W}_i = \begin{bmatrix} & w(1) & 0 & \dots & 0 & \\ \mathbf{0}_{N_{\underline{r}} \times (i-1) n_{\mathrm{inc}}} & 0 & w(2) & \dots & 0 & \mathbf{0}_{(N_{\underline{r}} \times (N_r - N_{\underline{r}} - (i-1) n_{\mathrm{inc}})} \\ & \vdots & \vdots & \dots & \vdots & \\ & 0 & 0 & \dots & w(N_{\underline{r}}) & \end{bmatrix}$$

is a $N_{\underline{r}} \times N_r$ rectangular matrix representing a window operation and $n_{\mathrm{inc}}$ is a subframe increment.

Using (4.10), we then have

$$\frac{\partial J_{SSF}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}})}{\partial \hat{\boldsymbol{h}}} = 4 \sum_{i=1}^{B_F} J_{\mathrm{SF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}_i) \bar{\boldsymbol{R}}_{\boldsymbol{y}}^T \boldsymbol{W}_i^T \boldsymbol{\mathcal{F}}_{N_r}^{-1} \left( \boldsymbol{\mathcal{D}}_{|\breve{\bar{\boldsymbol{r}}}|} - \frac{1}{||\bar{\boldsymbol{r}}_i||^2} \boldsymbol{\mathcal{I}}_{N_r} \right) \boldsymbol{\mathcal{F}}_{N_r} \boldsymbol{W}_i \bar{\boldsymbol{R}}_{\boldsymbol{y}} \hat{\boldsymbol{h}}, \tag{4.14}$$

where the frame index $b$ has been omitted for conciseness.

Since $J_{\text{SSF}}$ is the sum of squared spectral flatness scores, it is minimum when each of the terms within the summation is close to 0, which occurs when the magnitude spectrum of each of the subframed residual is flat.

### 4.3.5 Experimental results

The SSF cost function is here evaluated against the LS method [32] and the RNMCFLMS [111] to blindly estimate impulse responses from reverberant noisy speech signals. These two algorithms were chosen as the baseline because the former is known to be robust to additive white Gaussian noise while the latter, similarly to the SSF, is a batch processing. The performance of the algorithms was evaluated for frames of variable durations, RIRs of varying lengths, SNR of different levels and 2 recorded dry speech input signals.

#### Simulation setup

*Speech input.* Two utterances spoken by a male and a female speaker, from the SCRIBE database [130] were used as the anechoic input signal. These respectively corresponded to the free speech files named `MAS_FS` and `FAA_FS` and, in which the speakers were asked to describe a scenery or a picture so as to reproduce the spontaneity of natural speech. Both utterances had a duration of approximately 150 s each and were resampled to $f_s =$ 8 kHz, which corresponds to the sampling rate used in mobile phones. These specific files were chosen to illustrate the performance of the different BSI algorithms. As no prior information on the gender of the speaker is used by the algorithms, the results are not expected to depend on it.

*Room impulse responses.* The RIRs were simulated using the image method [83] for a shoebox shaped room of dimension $5 \times 6 \times 3$ m and reverberation time of 0.3 s. The speaker was placed 2 m away from the centre of the 2 microphone array.

The accuracy, measured in terms of NPM, of the different methods were evaluated for three different lengths: 200, 500 and $T_{60} f_s$ taps.

*Noisy reverberant signal.* For a given speaker, the reverberant signals are obtained by convolving the anechoic speech signal with a set of RIRs. These reverberant signals were then corrupted by additive tempo-spatially white Gaussian noise. To investigate the robustness to additive noise of the different algorithms, the power of the added noise was scaled so that the average SNR was 30 dB and 10 dB.

*Algorithm details.* An overlap of $\lfloor \frac{L}{8} \rfloor$ between the frames was used for the RNMCFLMS as it showed a more accurate estimate of the RIRs in a preliminary experiment. The optimal step-sizes were empirically found to be 0.2 for an SNR 30 dB and 0.1 for an SNR of 10 dB and $L = 500$ taps. These step-sizes are kept for the different values of $L$.

The subframes used in SSF were obtained by applying Hamming windows of duration of 30 ms overlapping by 15 ms to analysis frames of duration $T_{\text{Est}}$. The performance of the algorithms with respect to the duration of the analysis frames is investigated for $T_{\text{Est}} \in \{1 \text{ s}, 3 \text{ s}\}$.

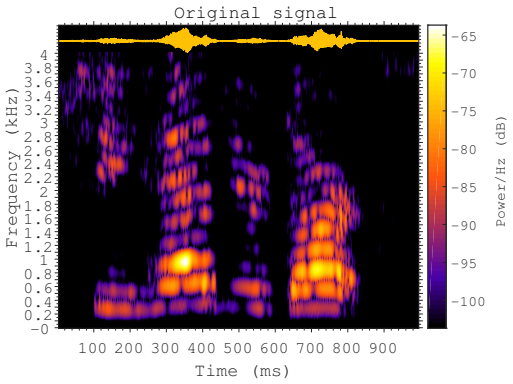
The nonconvex optimization problem was solved by a quasi-Newton method as implemented in the function `fminunc` in MATLAB.

Both the RNMCFLMS and the proposed method were initialized with the vector containing 1 at $mL + 1$ for $m \in \{0, 1, \ldots, M - 1\}$ and 0 elsewhere. Analysis frames containing at least 500 ms of speech are kept in the evaluation. This is done to remove frames containing mostly noise from the evaluation.

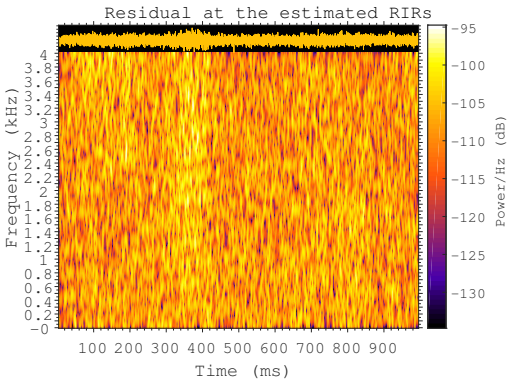
### Results

*Illustrative example.* Before presenting the results for a large number of frames, let us illustrate the performance of the algorithm with a small example.

Figure 4.3b shows the residual obtained from the SSF for a 1 s reverberant noisy speech signals corrupted by additive spatio-temporally white Gaussian noise at an SNR of 30 dB. The RIRs were generated using the image method for the room previously described and truncated to $L = 300$ samples. It can be observed that the estimated RIRs effectively




(a) Spectrogram of the analyzed signal at the first microphone.




(b) Residual at $\hat{\boldsymbol{h}}$.

Figure 4.3: Spectrogram of the input signal and residual at the RIRs estimated by SSF for $L = 300$ samples and an SNR of 30 dB.

remove the speech content from the reverberant noisy signals, the spectrogram of which is given in Fig. 4.3a. The value of $J_{\text{SSF}}$ at $\hat{\boldsymbol{h}}$ was equal to 12 against 32 at the solution, which shows that the solution $\boldsymbol{h}$ was not the global minimum of the cost function used in SSF. However, the NPM was -3.8 dB.

The RNMCFLMS, applied to the same reverberant noisy speech signal with a step-size optimally tuned to 0.17 for this specific example, estimated the RIRs with an NPM of -2.5 dB. Figure 4.4 shows the variations of the NPM against time. The spike observed at approximately 0.38 s is due to the step-size. A step-size, higher by more than 0.01, leads to a misconverging algorithm while a step-size, smaller by more than 0.01, removes the spike but also achieves a higher NPM, i.e. closer to 0 dB, at 1 s. The sensitivity of the RNMCFLMS to the choice of the step-size may be due to the fact that, because of the non-stationarity of speech, the SNR varies locally so that a step-size appropriate for a certain time-frame may be too large for another.

*SCRIBE speech samples, 1 s long frames.* Figures 4.5a, 4.6a and 4.7a shows the performance in terms of NPM of the proposed algorithm compared against the LS method and the RNMCFLMS for $L = 200$, $L = 500$ and $L = 2400$ respectively for different level of SNRs. A total of 340 active speech frames were considered in the evaluation. In Fig. 4.5a, 4.6a and 4.7a, the different boxes, read from left to right show the performance of the SSF,

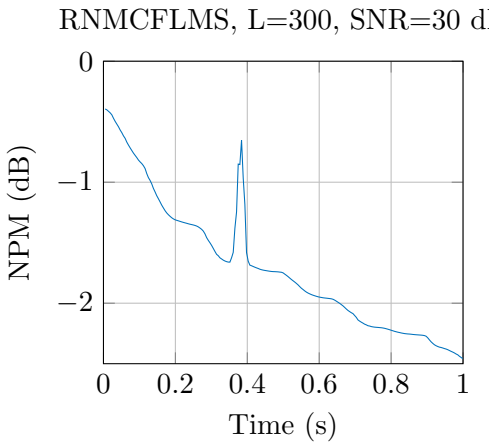



Figure 4.4: NPM against time for the estimated RIRs obtained via the RNMCFLMS.

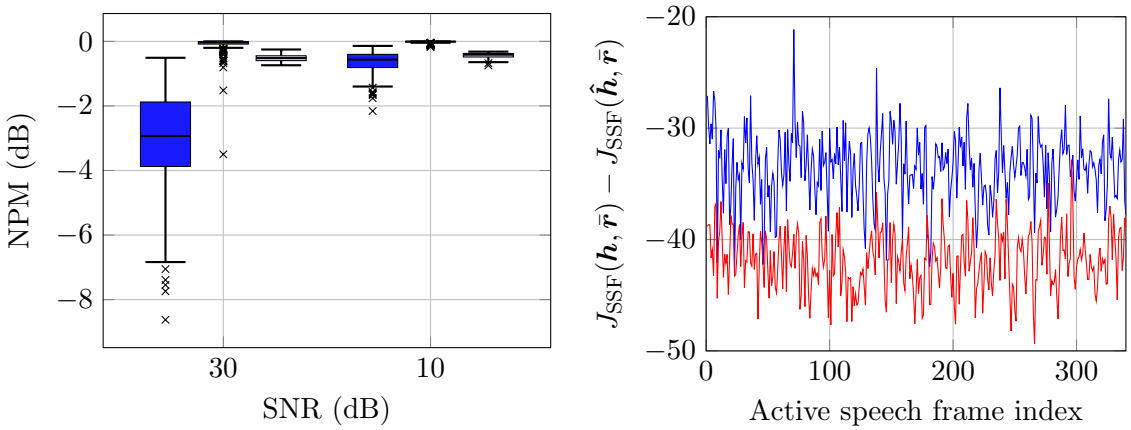
LS and RNMCFLMS respectively. The LS and the RNMCFLMS have little variations of the NPM between frames, which results in flat boxes.

The proposed method achieves a lower median NPM than the RNMCFLMS and the LS method for an SNR equal to 30 dB and $L$ smaller than 500 samples. For example, for $L = 200$ and and SNR of 30 dB, the proposed method indeed achieves a median NPM of -2.9 dB against -0.029 dB and -0.52 dB for the LS and RNMCFLMS respectively. However as the number of taps increases, the performance of all algorithms decreases. For $L = 2400$, the estimated stacked RIRs always display an NPM higher than -0.5 dB, which suggests that none of the algorithms manages to estimate the stacked RIRs.

It can also be observed that the proposed algorithm is sensitive to additive noise. For example, when the SNR decreases from 30 dB to 10 dB, the median NPM increases from -2.9 dB to -0.56 dB. This may be due to the fact that, at low SNRs, since the reverberant noisy signal is buried in noise, the proposed cost function does not sufficiently discriminate the RIRs.

Unlike the SSF and the LS methods, the RNMCFLMS achieves the same range of NPMs regardless of the level of noise, which suggests that the convergence of the RNMCFLMS has not occurred before the end of the 1 s of analyzed signals.

To investigate the optimality of $J_{\text{SSF}}$ for BSI, the difference between $J_{\text{SSF}}(\boldsymbol{h}, \bar{\boldsymbol{r}})$ and $J_{\text{SSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}})$ for each frame is shown in Fig. 4.5b, 4.6b and 4.7b. It can be observed that, for




(a) NPM against SNR for the proposed method (▬), the least squares (▬) and the RNM-CFLMS (▭) algorithms.

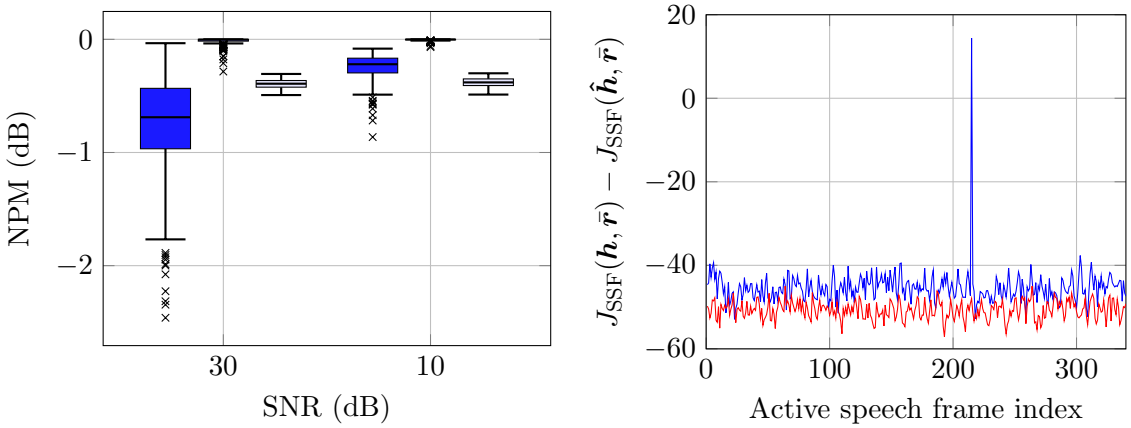
(b) Optimality of the estimated RIRs for SNR=30 dB (—) and SNR=10 dB (—).

Figure 4.5: Optimality and NPM of the estimated RIRs obtained from 1 s long noisy reverberant speech signals for $L = 200$.

most frames, the ground truth stacked RIRs is not a global minimum of the cost function. After convergence of the optimization algorithm, the $J_{\mathrm{SSF}}$ at $\hat{\boldsymbol{h}}$ is indeed almost always smaller than $J_{\mathrm{SSF}}$ at $\boldsymbol{h}$.

*SCRIBE speech samples, 3 s long frames.* Figure 4.8 shows the NPM achieved by the proposed method, the LS method and the RNMCFLMS at the end of a 3 s analysis frame. It can be observed that by increasing the duration of the analyzed signal, the NPM of the stacked RIRs estimated via the SSF decreases. This can be explained by the fact that, as the duration of the analysis frame increases, the solution to the optimization problem has to yield consistent spectral flatness scores across an increased number of subframes.

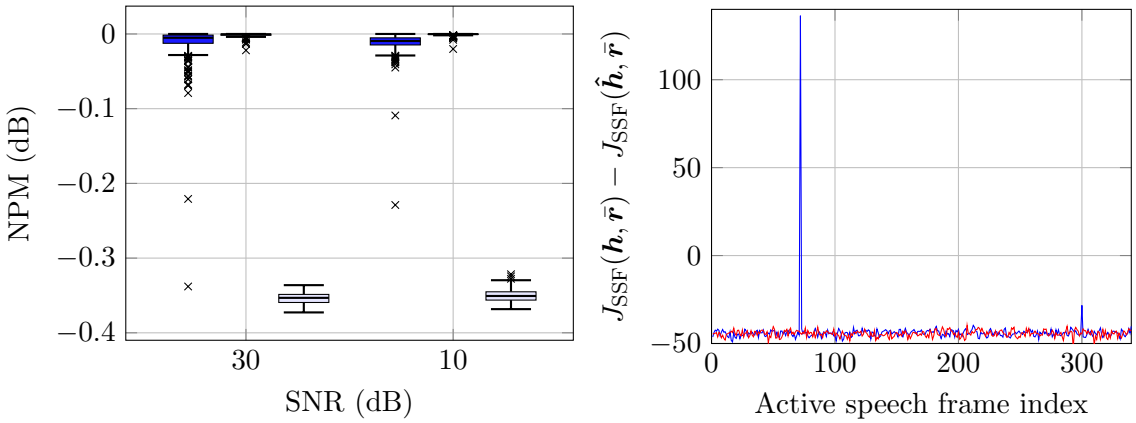
The variations of the median NPM as a function of the duration of the analysis frame for the proposed method is given in Table 4.1. It can be seen that increasing the duration of the analysis frame yields more accurate estimate of the RIRs. Table 4.1 also shows that increasing the estimation time from 1 s to 3 s is not sufficient to estimate realistically long impulse responses. A similar behavior, i.e. increased accuracy of the estimated RIRs with longer frames, has been observed for the RNMCFLMS and the LS method although the respective median NPMs decreased by 0.2 dB and 0.01 dB at best, which means that the




(a) NPM against SNR for the proposed method (■), the least squares (■) and the RNM-CFLMS (■) algorithms.

(b) Optimality of the estimated RIRs for SNR=30 dB (——) and SNR=10 dB (——).

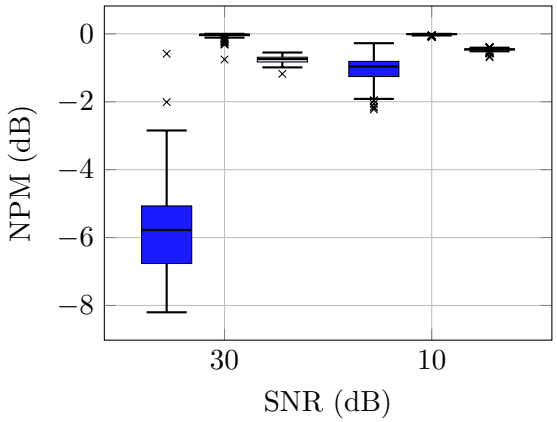
Figure 4.6: Optimality and NPM of the estimated RIRs obtained from 1 s long noisy reverberant speech signals for $L = 500$.




(a) NPM against SNR for the proposed method (■), the least squares (■) and the RNM-CFLMS (■) algorithms.

(b) Optimality of the estimated RIRs for SNR=30 dB (——) and SNR=10 dB (——).

Figure 4.7: Optimality and NPM of the estimated RIRs obtained from 1 s long noisy reverberant speech signals for $L = 2400$.




(a) NPM against SNR for the proposed method (▬), the least squares (▬) and the RNM-CFLMS (▭) algorithms, $L = 200$.

Figure 4.8: NPM of the estimated RIRs obtained from 3 s long noisy reverberant speech signals.

| | SNR=30 dB | | SNR=10 dB | |
|---|---|---|---|---|
| | 1 s | 3 s | 1 s | 3 s |
| $L = 200$ | -2.9 | -5.8 | -0.56 | -0.96 |
| $L = 500$ | -0.68 | -2.5 | -0.22 | -0.48 |
| $L = 2400$ | -0.0057 | -0.075 | -0.0101 | -0.10 |

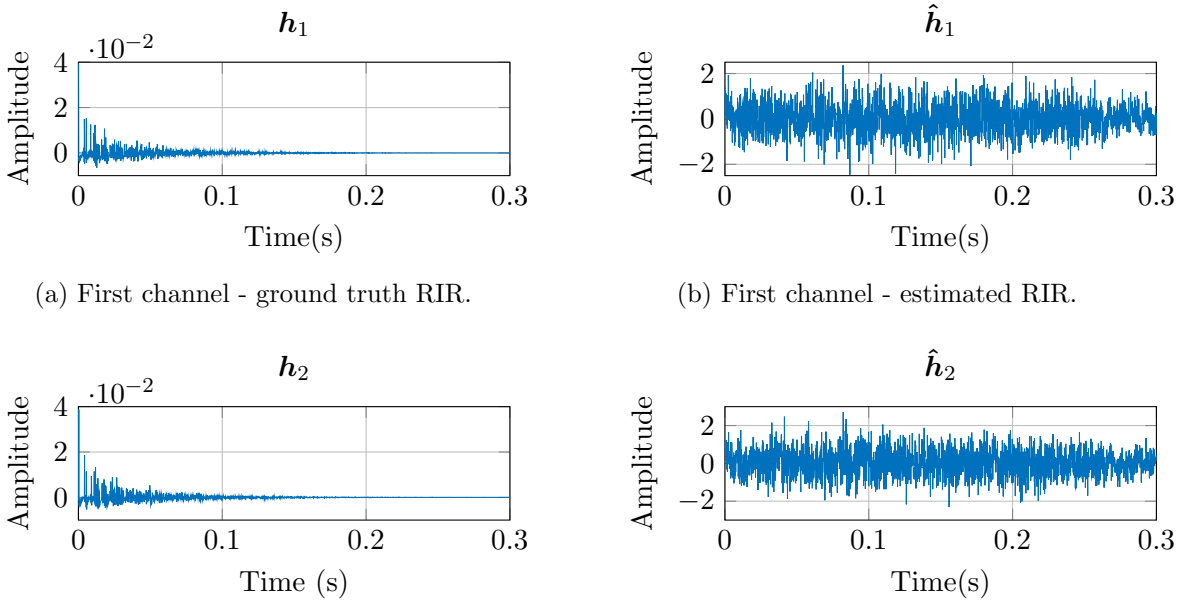
Table 4.1: Variations of the median NPM, in dB, with the duration of the duration of the analysis frame for the proposed method.

estimated stacked RIR vector is orthogonal to the solution.

### 4.3.6 Discussion

While the SSF cost function effectively whitens the residual, its formulation is not optimal in the sense that its global minimum is not achieved at the sought solution. Despite that limitation, minimizing the SSF cost function yields increasingly better estimates of the NPMs as the duration of the analyzed reverberant noisy speech signals increase.

However, the sequential spectral flatness algorithm only relies on the spectral flatness of the stacked RIRs although additional characteristics of RIRs are known. Therefore, these additional characteristics can be used to regularize the cost function such that the estimated RIRs display features characteristic of RIRs.




(a) First channel - ground truth RIR.

(b) First channel - estimated RIR.

(c) Second channel - ground truth RIR.

(d) Second channel - estimated RIR.

Figure 4.9: Ground truth and estimated RIRs from the first second long reverberant noisy speech described in Section 4.3.4 with $L = 2400$. The estimated RIRs were obtained by minimizing $J_{\text{SSF}}$.

## 4.4 Regularization

As derived in Section 4.3.5, the SSF only relies on the spectral flatness characteristic of RIRs to solve the problem of BSI.

Figure 4.9 shows the RIRs estimated via the SSF cost function applied to the 1 s long reverberant noisy speech signals described in the illustrative example of Section 4.3.5. The length of the RIRs were, however, changed to $L = T_{60} f_s = 2400$. By comparing Fig. 4.9a and Fig. 4.9c to Fig. 4.9b and Fig. 4.9d, it can be observed that the estimated RIRs do not actually display the exponential decay in the late reverberation nor impulses localized in time in the early reflections. This section investigates the use of regularization terms to improve the estimation accuracy of the SSF cost function from noisy reverberant speech signals for full length RIRs. The resulting algorithm will be referred to as the RSSF.

### 4.4.1 Early reflections

#### Enforcing sparsity of the early reflections

As mentioned in Chapter 1, a RIR can be decomposed in two parts: the early and the late reflections. The early reflections are characterized by distinct peaks [47, Chapter 4], which correspond to isolated reflections [131], separated by much smaller valued taps.

The presence of small valued taps but not exactly equal to zero may limit the performance of a sparsity promoting regularization term. However, the small valued taps separating the strong reflections can be pushed towards 0 by applying a non-linear function to each estimated RIRs. The application of a measure of sparsity to the transformed early reflections may then be viewed as a measure quantifying the pattern of isolated impulses of the early reflections. That concept is illustrated in Fig. 4.10 and Fig. 4.11. Figure 4.10 shows image method generated RIRs truncated to $L = 512$ while Fig. 4.11 shows these RIRs after applying the non-linear function. For illustrative purposes, the non-linearity is the function that raises each tap of the early reflections to the power 2.

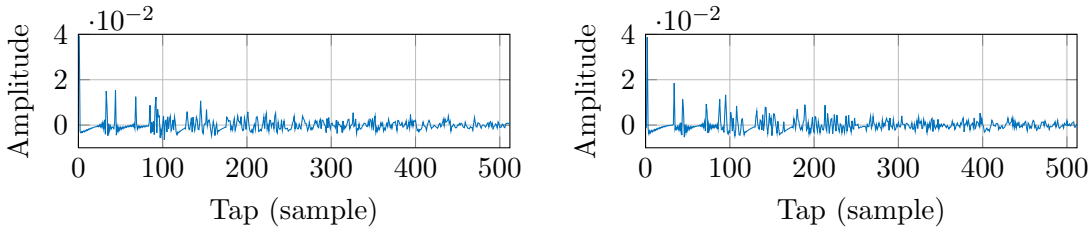



Figure 4.10: Image method generated RIR truncated to $L = 512$ for $f_s = 8$ kHz.

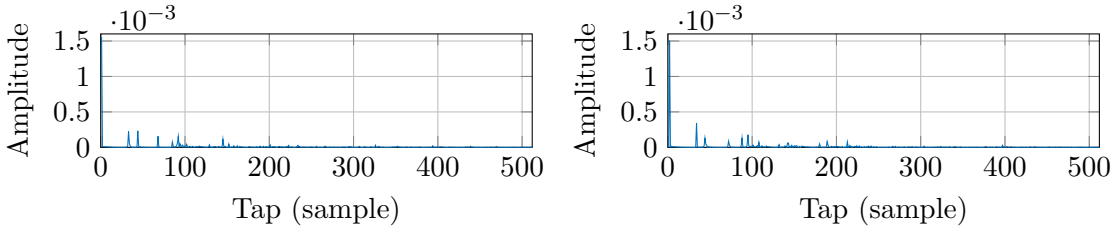



Figure 4.11: Image method generated RIR truncated to $L = 512$ taps and for $f_s = 8$ kHz to the power 2.

### Formulation in an optimization framework

From the previous observations, a regularization term promoting isolated strong reflections in the early part of each RIR can be formulated.

Let us denote by $\boldsymbol{\theta}_p\{\boldsymbol{a}\}$ the monomial that raises the entries of $\boldsymbol{a}$ to a constant positive integer $p$ i.e.

$$\boldsymbol{\theta}_p\{\boldsymbol{a}\}(l) = a^p(l), \tag{4.15}$$

where $a(l)$ is the $l^{\text{th}}$ entry of $\boldsymbol{a}$. Denoting by $n_e$ the mixing time, i.e. the time instant from which the statistical model of the reverberation tail holds [58], and $\boldsymbol{v}_{n_e}$ the $L$ entry long vector with 1 at its first $n_e$ entries and 0 elsewhere, $\boldsymbol{\theta}_p\{\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{b}_m} \otimes \boldsymbol{\mathcal{D}}_{\boldsymbol{v}_{n_e}}\right)\hat{\boldsymbol{h}}\}$ is an $ML$ vector the $l^{\text{th}}$ entry of which is given by

$$\boldsymbol{\theta}_p\{\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{b}_m} \otimes \boldsymbol{\mathcal{D}}_{\boldsymbol{v}_{n_e}}\right)\hat{\boldsymbol{h}}\}(l) = \begin{cases} h^p(l), & \text{if } (m-1)L+1 \leq l \leq (m-1)L+n_e \\ 0 & \text{otherwise} \end{cases}, \tag{4.16}$$

where $\boldsymbol{b}_m$ is a $M \times 1$ vector of zeros, except at the $m^{\text{th}}$ entry, which is equal to 1.

For conciseness, $\boldsymbol{\theta}_p\{\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{b}_m} \otimes \boldsymbol{\mathcal{D}}_{\boldsymbol{v}_{n_e}}\right)\hat{\boldsymbol{h}}\}$ will be referred to as $\boldsymbol{\theta}_p^m(\hat{\boldsymbol{h}})$.

Denoting by $\mathcal{S}(\cdot)$ a cost function, which decreases as the sparsity of the argument increases such as the $\ell_1$-norm [132], the vector $\boldsymbol{\vartheta}_p(\hat{\boldsymbol{h}}) = \begin{bmatrix} \mathcal{S}(\boldsymbol{\theta}_p^1(\hat{\boldsymbol{h}})) & \mathcal{S}(\boldsymbol{\theta}_p^2(\hat{\boldsymbol{h}})) & \cdots & \mathcal{S}(\boldsymbol{\theta}_p^M(\hat{\boldsymbol{h}})) \end{bmatrix}^T$ is formed and the overall sparsity of the early part of each RIR is given by

$$\phi_{\text{ER}}(\hat{\boldsymbol{h}}) = \mathbf{1}_{M\times 1}^T \boldsymbol{\vartheta}_p(\hat{\boldsymbol{h}}). \tag{4.17}$$

Assuming that $\mathcal{S}(\cdot)$ is differentiable at $\boldsymbol{\theta}_p^m(\hat{\boldsymbol{h}}_m)$ for $m \in \{1, 2, \cdots, M\}$, the gradient $\nabla\phi_{\text{ER}}$ of $\phi_{\text{ER}}$ is given by [79]:

$$\nabla\phi_{\text{ER}}(\hat{\boldsymbol{h}}) = \boldsymbol{\nabla}\boldsymbol{\vartheta}_p(\hat{\boldsymbol{h}})\mathbf{1}_{M\times 1}, \tag{4.18}$$

where $\boldsymbol{\nabla}\boldsymbol{\vartheta}_p$ denotes the Jacobian of $\boldsymbol{\vartheta}_p$, defined as

$$\boldsymbol{\nabla}\boldsymbol{\vartheta}_p(\hat{\boldsymbol{h}}) = \begin{bmatrix} \frac{d\mathcal{S}(\boldsymbol{\theta}_p^1(\hat{\boldsymbol{h}}))}{d\hat{\boldsymbol{h}}} & \frac{d\mathcal{S}(\boldsymbol{\theta}_p^2(\hat{\boldsymbol{h}}))}{d\hat{\boldsymbol{h}}} & \dots & \frac{d\mathcal{S}(\boldsymbol{\theta}_p^M(\hat{\boldsymbol{h}}))}{d\hat{\boldsymbol{h}}} \end{bmatrix}. \tag{4.19}$$

Using that approach to regularize the early part of the estimated RIRs, the regularizing term, which aims to quantify the pattern of isolated impulses of the early reflections, is fully specified by the choice of $\mathcal{S}(\cdot)$ and the non-linear function $\boldsymbol{\theta}_p\{\cdot\}$.

Let us derive the gradient for the scale-independent sparsity function $\frac{\ell_2(\cdot)}{\ell_1(\cdot)}$ [133], which will be applied to the non-linearity $\boldsymbol{\theta}_2^m(\cdot)$, i.e. the function that raises the taps before the mixing time to the power 2. Since the value of the function $\frac{\ell_2(\cdot)}{\ell_1(\cdot)}$ increases as the sparsity of the argument increases, $\mathcal{S}(\cdot)$ is chosen as $\mathcal{S}(\cdot) \coloneqq -\frac{\ell_2(\cdot)}{\ell_1(\cdot)}$. Therefore, for any $\boldsymbol{a} \in \mathbb{R}^{ML} \setminus \{\mathbf{0}_{ML\times 1}\}$,

$$\mathcal{S}(\boldsymbol{a}) = -\frac{\sqrt{\sum_{i=1}^{ML} a^2(i)}}{\sum_{i=1}^{ML} |a|(i)},$$

where $\boldsymbol{a} = \begin{bmatrix} a(1) & a(2) & \dots & a(ML) \end{bmatrix}^T$.

With that choice of sparsity function, $\mathcal{S}$ is differentiable on $\mathbb{R}^{ML} \setminus \{\mathbf{0}_{ML\times 1}\}$. Since all the entries in $\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})$ are positive, the denominator of $\mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}}))$ is given by $\mathbf{1}_{ML\times 1}^T \boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})$. The expression of $\mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}}))$ is therefore given by

$$\mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})) = -\frac{\sqrt{\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})^T \boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}}{\mathbf{1}_{L\times 1}^T \boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}. \tag{4.20}$$

The gradient $\frac{d\mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}}))}{d\hat{\boldsymbol{h}}}$ is then found using the chain rule [126, Appendix A], which yields

$$\frac{d\mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}}))}{d\hat{\boldsymbol{h}}} = \frac{d\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}{d\hat{\boldsymbol{h}}} \frac{d\mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}}))}{d\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}, \tag{4.21}$$

where $\frac{d\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}{d\hat{\boldsymbol{h}}}$ is a diagonal matrix, the diagonal entries $\left(\frac{d\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}{d\hat{\boldsymbol{h}}}\right)_{(l,l)}$ of which are given by

$$\left(\frac{d\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}{d\hat{\boldsymbol{h}}}\right)_{(l,l)} = \begin{cases} 2\hat{h}(l), & \text{if } (m-1)L+1 \le l \le (m-1)L+n_e \\ 0, & \text{otherwise} \end{cases} \tag{4.22}$$

and

$$\frac{d\mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}}))}{d\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})} = \frac{1}{\mathbf{1}_{ML\times 1}^T \boldsymbol{\theta}_2^m(\boldsymbol{h})} \left\{ \mathcal{S}(\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}}))\mathbf{1}_{ML\times 1} - \frac{\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}{\sqrt{\boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})^T \boldsymbol{\theta}_2^m(\hat{\boldsymbol{h}})}} \right\}. \tag{4.23}$$

### 4.4.2 Late reflections

The late reflections typically display a decaying exponential envelop, the decay rate of which is inversely proportional to the reverberation time [60]. In this work, a Tikhonov regularization term [126, Chapter 6] is introduced to promote an exponentially decaying tail in the estimated RIRs. This additional term effectively weighs the taps of the estimated RIRs with an increasing exponential so that large taps at the end of the RIRs are more penalized.

Let us denote by $\boldsymbol{e}$ the $L \times 1$ weight vector whose entries are given by

$$\boldsymbol{e} = \begin{cases} e^{\zeta l,} & \text{if } n_e + 1 \leq l \leq L \\ 0, & \text{otherwise} \end{cases} \tag{4.24}$$

where $\zeta = \frac{3\log(10)}{T_{60}}$, is the decay rate defined in Chapter 2. Denoting by $\boldsymbol{\mathcal{D}}_{\boldsymbol{e}}$ the diagonal matrix of size $L \times L$ with the entries of $\boldsymbol{e}$ on its diagonal and zeros elsewhere, the regularization term is given by

$$\phi_{\mathrm{LR}}(\hat{\boldsymbol{h}}) = \frac{||\left(\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{D}}_{\boldsymbol{e}}\right)\hat{\boldsymbol{h}}||_2^2}{\hat{\boldsymbol{h}}^T\hat{\boldsymbol{h}}}, \tag{4.25}$$

the gradient of which is given by

$$\nabla\phi_{\mathrm{LR}}(\hat{\boldsymbol{h}}) = \frac{2}{\hat{\boldsymbol{h}}^T\hat{\boldsymbol{h}}} \left\{ (\boldsymbol{\mathcal{I}}_M \otimes \boldsymbol{\mathcal{D}}_{\boldsymbol{e}}^2)\hat{\boldsymbol{h}} - \phi_{\mathrm{LR}}(\hat{\boldsymbol{h}})\hat{\boldsymbol{h}} \right\}. \tag{4.26}$$

As shown in (4.25), the regularization term weighs the late reflections of each estimated RIRs with a raising exponential so that the last taps of each RIR contribute more to $\phi_{\mathrm{LR}}$. Since the objective function $J_{SSF}$ is scale independent, the denominator in (4.25) is necessary. Otherwise, $\phi_{\mathrm{LR}}$ could be arbitrarily small by simple scaling of $\hat{\boldsymbol{h}}$.

### 4.4.3 Summary of the RSSF cost function

Given a frame of cross-relation residual $\bar{\boldsymbol{r}}(b)$, the RSSF cost function $J_{\mathrm{RSSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b))$ combines the early reflection regularization term $\phi_{\mathrm{ER}}(\hat{\boldsymbol{h}})$ given in (4.17) and the late reflection regularization term $\phi_{\mathrm{LR}}(\hat{\boldsymbol{h}})$ given in (4.25) via two trade-off parameters $\gamma_{\mathrm{ER}}$ and $\gamma_{\mathrm{LR}}$ to regularize $\hat{\boldsymbol{h}}$. The expression of $J_{\mathrm{RSSF}}(\hat{\boldsymbol{h}})$ is given as

$$J_{\mathrm{RSSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b)) = J_{\mathrm{SSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b)) + \gamma_{\mathrm{ER}}\phi_{\mathrm{ER}}(\hat{\boldsymbol{h}}) + \gamma_{\mathrm{LR}}\phi_{\mathrm{LR}}(\hat{\boldsymbol{h}}), \tag{4.27}$$

where $J_{\mathrm{SSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b))$ is the SSF cost function defined in (4.11).

Estimates of the RIRs are then obtained by minimizing $J_{\mathrm{SSF}}(\hat{\boldsymbol{h}}, \bar{\boldsymbol{r}}(b))$, which may use an analytical expression of the gradient of each term in the right-hand side of (4.27). The expression of these gradients were given in (4.12), (4.18) and (4.26).

### 4.4.4 Simulation setup

The regularization terms previously presented are added to the SSF cost function introduced in Section 4.3.

By regularizing the spectral flatness cost function $J_{\mathrm{SSF}}$, the question of choosing appropriate trade-off parameters $\gamma_{\mathrm{ER}}$ and $\gamma_{\mathrm{LR}}$ for $\phi_{\mathrm{ER}}$ and $\phi_{\mathrm{LR}}$ respectively arises [126, Chapter 4]. Here, an appropriate value of the parameter is found by mean of a grid search on a *development* data set. Using the found value of the parameter, the regularized algorithm is evaluated on a *evaluation* data set and its performance is compared to the RNMCFLMS. Similarly, the optimal step-size for the RNMCFLMS is found by means of a grid search.

To assess the robustness of the methods against mismatch in the speech data, room type and level of noise, the development and evaluation datasets contain speech files from different speech copora, impulse responses generated for rooms of different sizes and reverberation times and different SNRs.

According to Fig. 4.11, a power of $p = 2$ applied to the early reflections pushes the

small valued tap to 0 so that the transformed early reflections are sparse. If a higher value of $p$ were used, a more sparse pattern would be achieved but such values would only display a peak at the strongest path and push the other reflections to zero. Therefore, a value of $p = 2$ is chosen.

Figure 4.9 indicates that the pattern of strong reflections separated by much smaller valued taps holds only for, approximately, the first 25 ms of the impulse response. Therefore, a value of $n_e = 0.025 \times 8000 = 200$ is chosen.

### Development set

*Speech data.* Since the SSF algorithm exploits the non-stationarity of the dry signal against the stationarity of the additive noise, white noise cannot be used as the dry signal. Therefore only reverberant speech is considered. The utterance named `SA1` spoken by speaker `FAKS0` of dialect region `DR1` taken from the TIMIT corpus [134] is used as the dry speech signal. The utterance was resampled at $f_s = 8$ kHz, which corresponds to the sampling rate of mobile phone speech, convolved with a set of RIRs and corrupted by additive white noise to achieve an SNR of 25 and 15 dB.

*Room impulse responses.* The image method generated RIRs corresponding to a room of dimension $4 \times 5 \times 2.5$ m with $T_{60} = 0.2$ s and $f_s = 8$ kHz were considered. The centre of the $M = 2$ microphone array was placed at $(1, 3, 1.7)$. The speaker was located 2.5 m away from the centre of the microphone array. The RIRs were then truncated to $L = 500$.

*Tuning of the proposed method.* In order to find appropriate values of $\gamma_{\mathrm{ER}}$ and $\gamma_{\mathrm{LR}}$, a grid search procedure was employed. The performance of the proposed algorithm is evaluated for $\gamma_{\mathrm{ER}}$ and $\gamma_{\mathrm{LR}}$ varying between 0 and 20 with an increment of 1.

*Tuning of the RNMCFLMS.* An appropriate choice of the step-size is crucial for adaptive algorithms [80]. Therefore, a grid search was implemented to search for an appropriate value for the step-size by testing step-sizes varying between 0.1 and 1 with in an increment of 0.1. To ensure that the step-size was appropriate, the training dataset of the RNMCFLMS is larger than the one used for the RSSF.

The speech material was composed of ten different utterances from the TIMIT corpus [134], spoken by two different speakers, namely `FAKS0` and `MDAB0` of dialect region `DR1`, are used as the input signal. In addition to the source/microphone setup previously described, two other positions in the same room were used to estimate an appropriate step-size for the RNMCFLMS. The centre of the microphone array was placed at $(2, 1, 1.6)$ and $(0.5, 1.5, 1.8)$. For each of the three position, the performance of the RNMCFLMS is also evaluated for $L = 200$ and $L = T_{60} f_s = 1600$.

### Evaluation set

*Speech data.* The speech files, namely `FAB_FS` and `MAS_FS` from the free speech section of the SCRIBE corpus [130], used in Section 4.3, are used here. These files were resampled at $f_s = 8$ kHz. The reverberant noisy speech signals were then generated in the same way as in the training set.

*Room impulse responses.* Since the various hyper-parameters of the considered algorithms, i.e. step-size of the RNMCFLMS, regularization parameters for the RSSF, have been fine tuned for a given room, the algorithms need to be evaluated on an unseen room to assess the robustness of the algorithms to new environments. Therefore, the dimensions of the room considered here differ from the room used in the development set. The image method generated RIRs corresponding to a room of dimension $5 \times 6 \times 3$ m with $T_{60} = 0.3$ s and $f_s = 8$ kHz were considered. The centre of the microphone array, with $M = 2$ microphones, was placed at $(1, 1.5, 2.6)$ with 20 cm spacing between the microphones. The speaker was located 3 m away from the centre of the microphone array at $(0.87, 3.95, 1.70)$. The RIRs were then truncated to 500 and $T_{60} f_s = 2400$ taps.

### Evaluation

The performance of the algorithms is evaluated in terms of NPM. As in Section 4.3, out of the 1 s analysis frames, only those containing more than 500 ms of speech are considered in the evaluation of the algorithms.

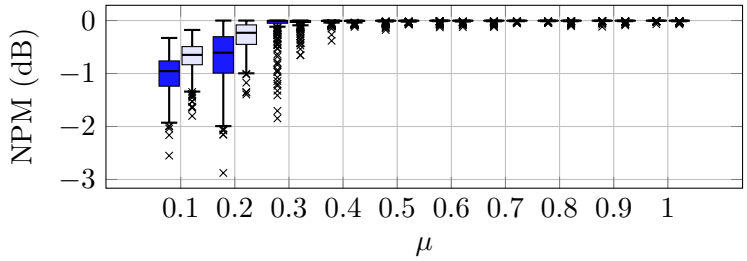



Figure 4.12: Performance of the RNMCFLMS on 1 s long reverberant noisy signals for different step-size for an SNR of 25 dB ( ) and an SNR of 15 dB ( ), $L = 200$.

| | Pos. 1 | Pos. 2 | Pos. 3 |
|---|---|---|---|
| Lower quartile | -1.30 dB | -1.31 dB | -1.23 dB |
| Median | -0.93 dB | -0.96 dB | -0.95 dB |
| Upper quartile | -0.73 dB | -0.76 dB | -0.76 dB |

Table 4.2: Variations of the median, first and third quartiles of the NPM achieved by the RNMCFLMS with respect to the position of the microphone array for $L = 200$ and an SNR of 25 dB.

### 4.4.5 Simulation results

#### Tuning of the RNMCFLMS

Figures 4.12 and 4.13 show the performance of the RNMCFLMS for $L = 200$ and $L = T_{60} f_s = 1600$ respectively on the development set, evaluated over 246 frames of speech. These frames had a duration of 1 s and were overlapping by 50%. Regardless of the level of the additive noise and the length of the impulse responses, it appears that the RNMCFLMS misconverges when $\mu$ is higher than 0.2. For $\mu = 0.2$, the RNMCFLMS converges with $L = 200$. However, as $L$ increases to 1600, the RNMCFLMS misconverges as the median NPM is -0.37 dB for an SNR of 25 dB. The case $L = 500$ displayed NPMs similar to $L = 1600$.

Table 4.2 shows the median, the first and third quartiles of the NPM for the different positions of the microphone array for $L = 200$, $\mu = 0.1$ and an SNR of 25 dB. It can be observed that the NPM does not vary with the position of the microphone array. Similar conclusions can be drawn for $L = 500$ and $L = 1600$. Therefore a step-size of 0.1 will be used on the evaluation set, regardless of the value of $L$ or the SNR.

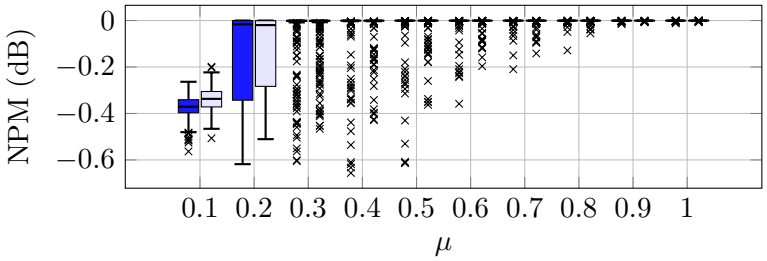



Figure 4.13: Performance of the RNMCFLMS on 1 s long reverberant noisy signals for different step-size for an SNR of 25 dB (▬) and an SNR of 15 dB (▭), $L = 1600$.

### Tuning of the RSSF

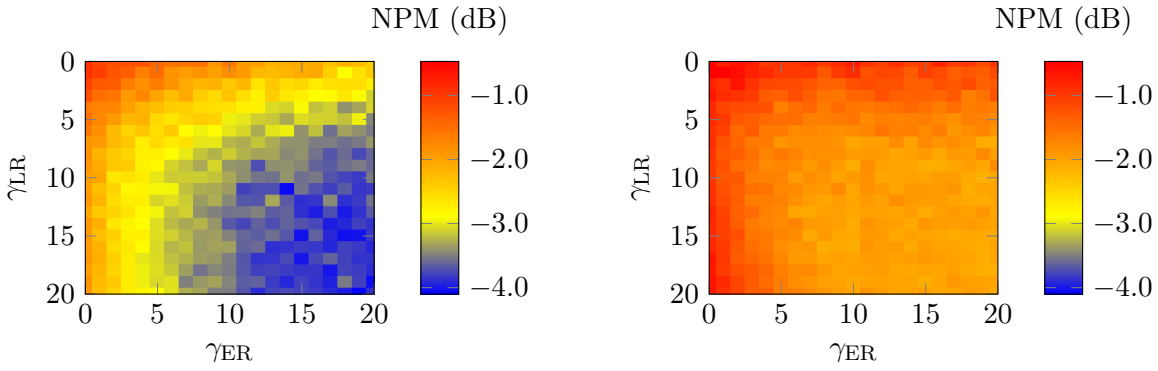
Figure 4.14 shows the variations of the median NPM for different values of $\gamma_{\mathrm{ER}}$ and $\gamma_{\mathrm{LR}}$ for $L = 500$ and SNRs of 25 and 15 dB. It can be observed that the median NPMs decreases as higher weights are given to the regularization terms. In particular, when no regularization is applied at all, the median NPM is close to -0.7 dB while the pair $(\gamma_{\mathrm{ER}}, \gamma_{\mathrm{LR}}) = (10, 9)$ yields a median NPM of -3.4 dB and -2.1 dB for an SNR of 25 dB and 15 dB respectively.

Figure 4.14 also shows that both regularization terms are necessary in the estimation as $\gamma_{\mathrm{ER}} = 0$ yields a minimum median NPM of -2 dB at an SNR of 25 dB while $\gamma_{\mathrm{LR}} = 0$ yields a minimum median NPM of -1.9 dB.

The best, i.e. most negative, values of the median NPM for an SNR of 25 dB and 15 dB are achieved at $(\gamma_{\mathrm{ER}}, \gamma_{\mathrm{LR}}) = (20, 19)$ and $(\gamma_{\mathrm{ER}}, \gamma_{\mathrm{LR}}) = (10, 19)$ respectively. For these optimal values, the median NPMs are respectively equal to -4.1 dB and -2.2 dB. These median NPMs are only a few fractions of dB better than the corresponding values for $(\gamma_{\mathrm{ER}}, \gamma_{\mathrm{LR}}) = (10, 9)$, which shows that precise values for the trade-off parameters are not required in the RSSF. The values of $\gamma_{\mathrm{ER}}$ and $\gamma_{\mathrm{LR}}$ are therefore chosen to be 10 and 9 respectively for the evaluation set.

### Evaluation set

*Illustrative example.* Let us consider the simulation setup described in Section 4.3.4 with $L = 2400$. The RIRs estimated via RSSF, with trade-off parameters of $\gamma_{\mathrm{ER}} = 10$ and $\gamma_{\mathrm{LR}} = 9$ for the early and late regularization terms respectively, is shown in Fig. 4.15.




(a) SNR=25 dB.

(b) SNR=15 dB.

Figure 4.14: Variation of the median NPM of the RIRs estimated via RSSF with different trade-off parameters for $L = 500$.

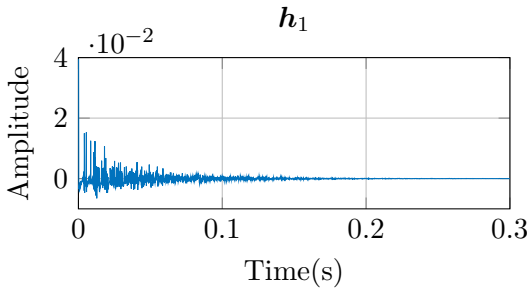
As illustrated in Fig. 4.15, the regularization terms led to estimated RIRs that indeed displayed an exponential decay as well as localized impulses in the early reflections.

By regularizing $J_{\text{SSF}}$ with the two terms previously described, the NPM of the estimated RIRs was equal to -4.5 dB. As a comparison, the NPM of the stacked RIRs estimated via SSF was -0.005 dB while that of the stacked RIRs estimated via the RNMCFLMS was -0.56 dB.

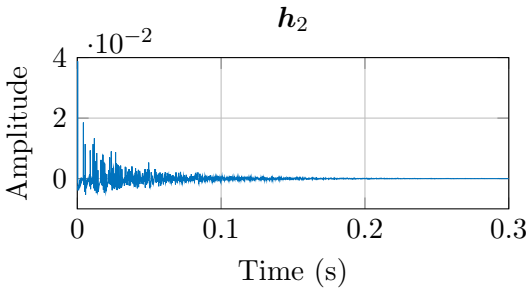
Figure 4.16 shows the residual at the solution and at the estimated stacked RIRs. It can be observed that the estimated stacked RIRs, indeed whiten the residual although the spectrogram of the desired residual, showed in Fig. 4.16a, and that of the residual at $\hat{\boldsymbol{h}}$, showed in 4.16b, are not identical.

*SCRIBE speech samples.* The box plot of the achieved NPMs of the RIRs estimated from frames of 1 s of observed speech signals for various SNRs and $L = 500$ is given in Fig. 4.17a. In Fig. 4.17a, for each SNR, the performance of the RNMCFLMS is displayed right after that of the RSSF. The very narrow boxes indicate that the variance in NPM of the RIRs estimated using the RNMCFLMS is very small.

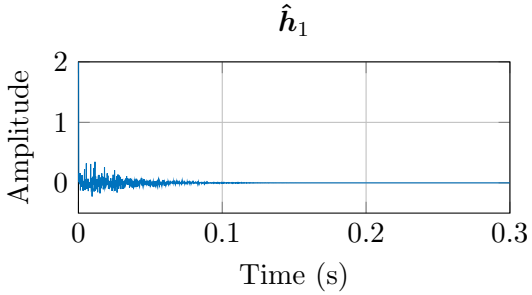
While the RNMCFLMS consistently gives median NPMs around -0.4 dB, the performance of the RSSF decreases as the SNR decreases. For SNRs varying from 50 dB to 30 and 10 dB, the median NPMs are respectively -6.6 dB, -3.0 dB and -0.35 dB. The poor



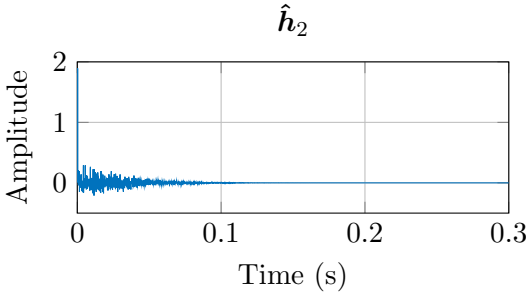

(a) First channel - ground truth RIR.




(b) First channel - ground truth RIR.

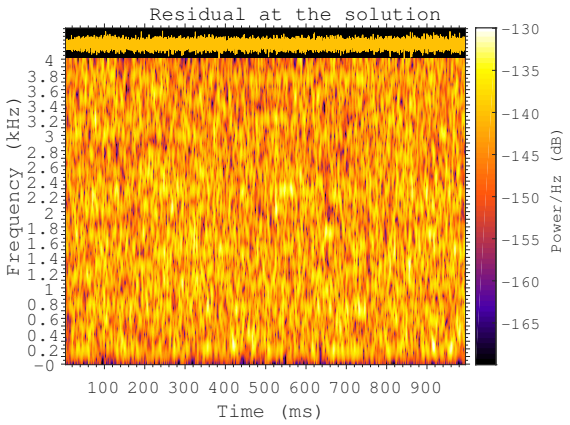



(c) First channel - estimated RIR.

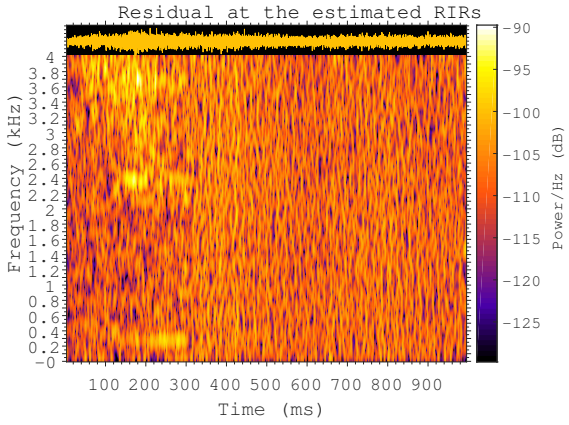



(d) Second channel - estimated RIR.

Figure 4.15: RIRs estimated via the regularized spectral flatness cost function from the first second of the reverberant noisy speech described in Section 4.3.4 with $L = 2400$.




(a) At $\boldsymbol{h}$.




(b) At $\hat{\boldsymbol{h}}$.

Figure 4.16: Spectrogram of the residual for $L = 2400$ and an SNR of 30 dB.

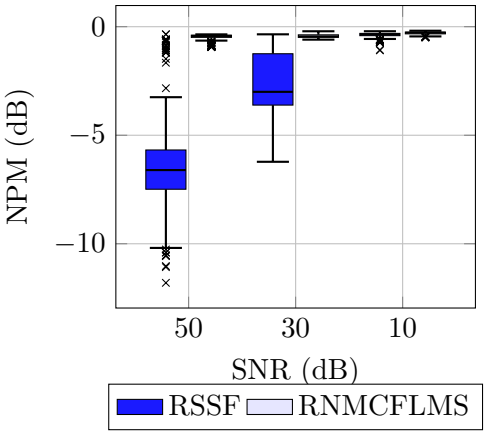
performance of the RNMCFLMS can be explained by the short duration of the frame while the decreasing performance of the RSSF may be explained by the fact that, at lower SNRs, the speech signal is already burried in noise.

To understand the outliers present in Fig. 4.17a for an SNR of 50 dB, it should be pointed out that the additive noise is added with a constant power so that the SNR, which is a ratio of power averaged over the whole utterance, achieves a desired value. However, since the power of speech varies with time, the local SNR computed in a given frame may be different to the average SNR. Figure 4.19a shows that the speech signal present in the frame between 2 and 3 s has a much lower energy than what follows. Figure 4.19c shows that the NPM of the RIR estimated from that frame is equal to -0.78 dB. However, if the 2.5 to 3.5 s frame is considered instead, the analyzed speech signal has a higher overall energy and the achieved NPM is -6.7 dB. A similar observation can be made for $L = 2400$ as shown in Fig. 4.19b and Fig. 4.19d.

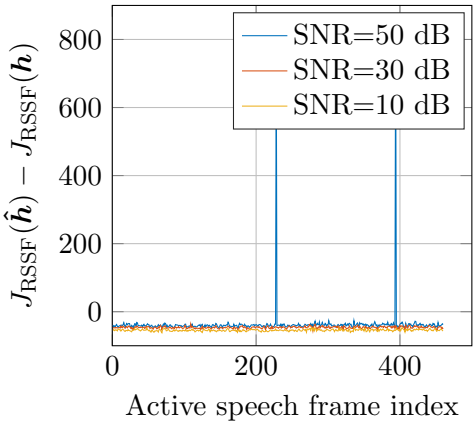
The comparison of Fig. 4.17 to Fig. 4.18 shows that the NPM of the estimated RIRs decreases as the value of $L$ increases. However, the RSSF is still able to estimate the RIRs down to -7.2 dB of NPM with a median of -3.1 dB for an SNR of 50 dB. As the level of noise increases, the performance of the RSSF also decreases. For an SNR of 30 dB, the median NPM is -0.68 dB. Fig. 4.20 shows that, for $L = 2400$ and an SNR of 30 dB, RSSF yields a bimodal distribution of NPM centered around -3 dB and 0 dB. This suggests that there exists local minima in RSSF, which are not necessarily close to $\boldsymbol{h}$ in an NPM sense. However, Fig. 4.17b and 4.18b show that despite minimizing a cost function which is not minimal at $\boldsymbol{h}$, the RIRs estimated by the RSSF can have a low NPM.

## 4.5 Summary

In this chapter, the stacked ground truth RIRs applied to the cross-relation has been shown to yield pair-wise residuals which correspond to the cross-relation applied directly to the additive noise. That characterization of the solution has then been exploited in the spectral flatness cost function. However, since the desired residual must be a filtered

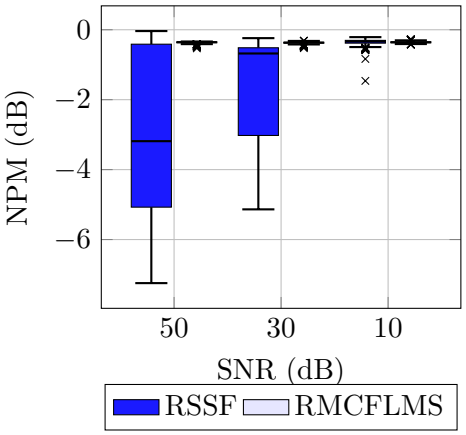



(a) Boxplot of the achieved NPMs.

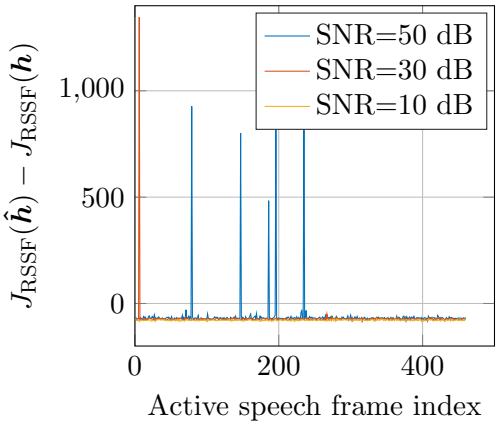



(b) Optimality of the estimated RIRs.

Figure 4.17: Performance of RSSF compared against the RNMCFLMS for $L = 500$ and various SNRs.

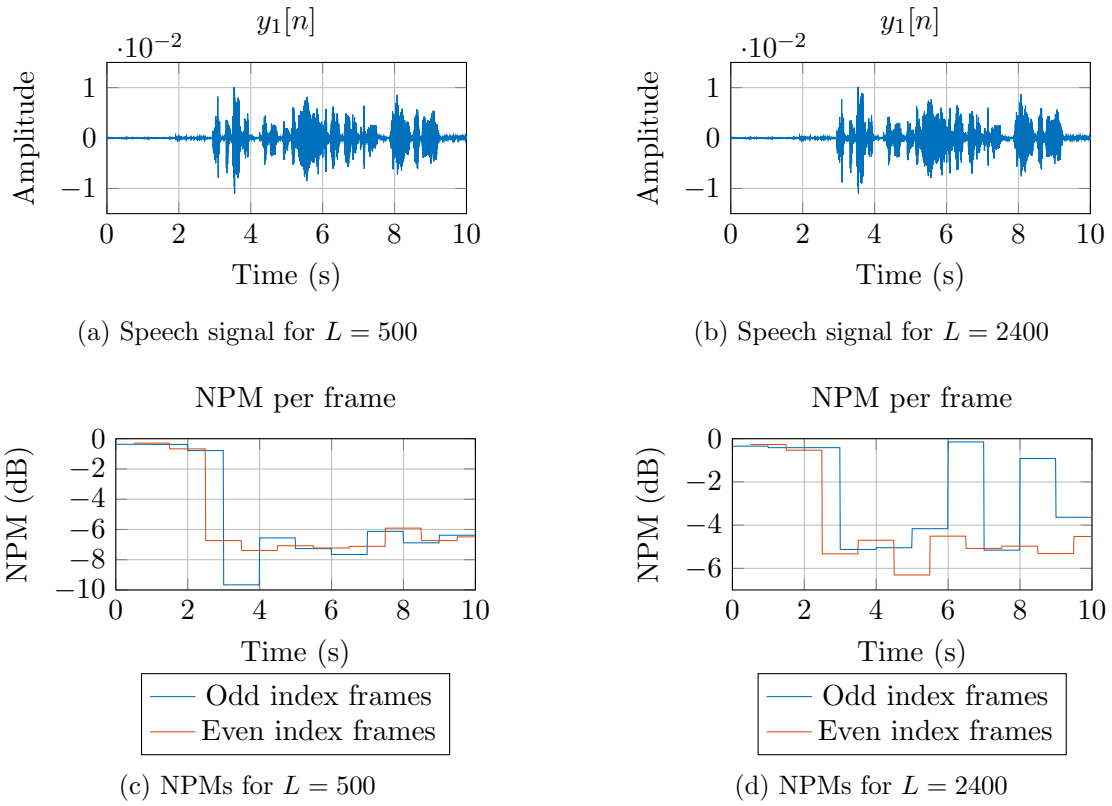



(a) Boxplot of the achieved NPMs




(b) Optimality of the estimated RIRs

Figure 4.18: Performance of RSSF compared against the RNMCFLMS for $L = 2400$ and various SNRs.




(a) Speech signal for $L = 500$

(b) Speech signal for $L = 2400$

(c) NPMs for $L = 500$

(d) NPMs for $L = 2400$

Figure 4.19: Speech waveform and achieved NPMs for various lengths and an SNR of 50 dB.

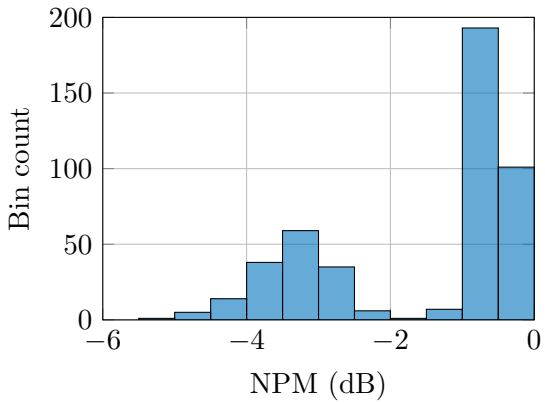



Figure 4.20: Histogram of the NPM for SNR=30 dB and $L$=2400.

noise and since the dry signal is speech, which is non stationary, the spectral flatness cost function is applied in subframes so that the objective function to be minimized aims to yield consistent spectral flatness scores across the subframes. In other words, the resulting SSF algorithm exploits the spectral flatness of the stacked RIRs together with the non stationary nature of speech against the stationarity of the additive noise.

That objective function was then combined with regularization terms in a multi-objective criterion optimization framework. These regularization terms ensured that the estimated RIRs displayed an exponential decay in the reverberation tail as well as isolated strong reflections in the early part of the impulse responses. Simulation showed that this combination enabled the estimation of realistically long RIRs, i.e. RIRs with more than a thousand taps, from 1 second of reverberant noisy speech data, which was out of the reach of previous algorithms. The regularized sequential spectral flatness cost function does, a priori, require additional knowledge than the RNMCFLMS. Indeed, the reverberation time is required to regularize the reverberation tail while the mixing coefficients is needed to combine the two regularization terms. However, the former is needed in any BSI application as the length of the RIRs is directly related to it and the latter does not need to be an accurate estimate of the mixing time as it only determines the time instant from which an exponential decay is observed in the RIRs.

# Chapter 5

# Environment aware speaker diarization[1]

## 5.1 Overview

This chapter is more application-driven than what has been presented earlier in this thesis as it investigates the use of spatial information for the diarization of meeting data.

## 5.2 Introduction

Given an audio recording containing two or more speakers, the objective of diarization is to answer the question 'which speaker spoke when?'. This information can be used to improve the performance of ASR systems by allowing effective speaker adaptation or enable efficient tagging of large transcripts of meetings so that the data can be searched by speaker name, for example.

Originally developed for telephone conversations and broadcast news [135], early diarization systems relied on speech dependent features such as Mel-frequency Cepstral Coefficients (MFCC) or Perceptual Linear Predictive (PLP) coefficients to discriminate between different speakers, e.g. [136]. In meeting scenarios, however, distant microphones

[1]The material of this chapter is the result of an industrial placement at Nuance Communications, Inc which led to the publication of [1] and [3]

may be preferred to headsets or lapel microphones [135]. In this case, the adverse effects of reverberation becomes more prominent so that the performance of MFCC or PLP based systems may be reduced. The combination of reverberation and other adverse phenomena typical of meeting scenarios, e.g. poor SNR or variable speech levels, further reduces the performance of such diarization systems [135].

The extension of single-channel diarization methods to multi-channel environments is not straightforward. The exploitation of the additional information provided by the multi-channel setup is indeed not obvious. In [137], for example, Leeuwen diarized the audio stream based on the data recorded on one of the available microphones. Istrate, Fredouille, Meignier, *et al.*, on the other hand, extracted features relevant for speaker diarization from a weighted average of the multi-channel signals. The use of beamforming as a preprocessor to enhance the speech signal has been explored in [139].

The use of spatial features in diarization systems has also been explored. Wooters and Huijbregts [140], for example, proposed to combine TDOA estimates with MFCC. Since TDOAs estimates are known to sometimes be spurious in the presence of reverberation [141], [142], Evans, Fredouille, and Bonastre [143] post-processed TDOA estimates with an unsupervised discriminant analysis [144] to investigate the performance of a diarization system relying only on TDOA estimates. However, the unsupervised discriminant analysis is possible only when more than 2 microphones are available.

In this chapter, contrary to previous speaker diarization methods, where reverberation would be seen as a hindrance, we propose to turn this hindrance into an advantage. Since the Direct-to-Reverberant Ratio (DRR), which is the ratio between the energy of the direct path and that of the rest of an RIR [30], is known to be strongly correlated with the distance between a microphone and the sound source [145], its use as the sole feature for a two channel diarization system is investigated in Section 5.3.

Furthermore, unlike TDOAs, which only exist when there are more than two microphones, the DRR exists as long as there exists an acoustic channel. Therefore, in many meeting scenarios, such as teleconferencing, where there is only a single microphone, spatial information can be exploited together with speech features in the form of DRR for

diarization. This will be investigated in Section 5.4. A summary of the results is given in Section 5.5.

## 5.3 Spatial feature based diarization

### 5.3.1 Signal model

Typical meeting scenarios involve multiple talkers. Hence, for $N_p$ speakers and $M$ microphones, at any time $n$, the signal $y_m[n]$ recorded at the $m^{th}$ microphone is given by

$$y_m[n] = \sum_{i=1}^{N_P} h_{m,i}[n] * s_i[n] + \nu_m[n], \tag{5.1}$$

where $i$ indexes the identity of a speaker, $h_{m,i}[n]$ is the RIR relating the $i^{th}$ speaker to the $m^{th}$ microphone and $\nu_m[n]$ the additive noise present at the $m^{th}$ microphone. However, most of the time, only one speaker is active at any given moment in time. The presence of simultaneous speakers falls outside the scope of the thesis. Therefore, even though several speakers are present in the audio stream, the system, in practice, has a SIMO structure.

### 5.3.2 The overall diarization system

The diarization system used in this section is that of [143]. Its block diagram is shown in Fig. 5.1. It consists of a Voice Activity Detector (VAD) detecting the non-speech parts, a feature extraction algorithm and a labeling step.

The VAD, based on the P.56 standard [146], [147], takes the summed microphone signals as the input and detects active speech segments. The output consists of estimated time instants indicating the beginning and the end of segments of active speech. A post-processing step is added so that the estimated active speech segments separated by less than 100 ms of estimated pause are merged together so that active speech segments are not separated by natural short pauses occurring between words.

A window of duration $t_e$ sliding with an increment $t_s$ is then applied within each of these active speech segments to obtain frames. Features are extracted from each of these

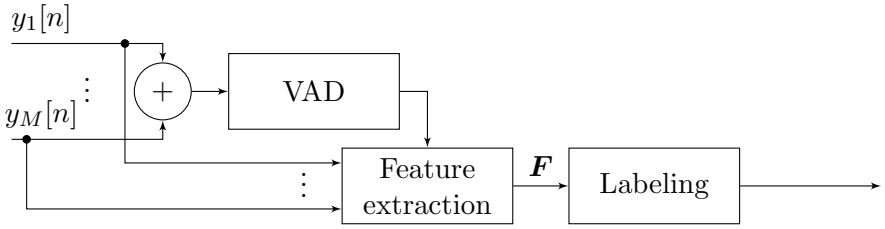



Figure 5.1: Block diagram of the diarization system.

frames. The type of features as well as the method to obtain them from the input signals will be described in Section 5.3.3 and Section 5.3.4.

The features are then labeled using a k-means initialized Hidden Markov Model (HMM), the details of which will be given in Section 5.3.5.

### 5.3.3 Spatial feature extraction of the baseline

The method described in [143] is taken as the baseline as diarization is also achieved based on spatial features only. In that method, the TDOAs between every pair of microphones are estimated using the Generalized Cross-Correlation with Phase Transformation (GCC-PHAT) algorithm [148]. Therefore, for each frame, $\binom{M}{2} = \frac{M(M-1)}{2}$ estimated TDOAs are obtained. In the case where $M$ is greater or equal to 3, dimension reduction techniques aiming at reducing the impact of estimation noise are possible. In [143], a Unsupervised Discriminant Analysis (UDA) [144] is used for that purpose.

### 5.3.4 Spatial feature extraction of the proposed method

The microphone signals $y_m[n]$ can be viewed as a combination of two independent quantities: the dry speech signal $s_p[n]$ and the RIRs $\{h_{m,p}[n]\}, m \in \{1, 2, \ldots, M\}$. While the dry speech contains the characteristics of the speaker, the set of RIRs holds information about the position of that speaker relative to the microphone array. Therefore spatially characterizing localized speakers is possible by blindly estimating the RIRs.

Because of the SIMO structure, BSI is theoretically possible provided that the conditions for the system to be identifiable are fulfilled [32].

However, as shown in Chapter 4, the NPM of estimated RIRs may be close to 0 dB using current state-of-the-art BSI algorithms, which results in estimated RIRs which are not accurate nor consistent enough for a direct use for diarization. We therefore propose a new approach which consists of extracting a feature, referred as $C_t$, which is analogous to the well-known $C_{50}$ [30]. It represents the ratio between the energy in the first $t$ ms of a RIR and that of the remaining taps, i.e.

$$C_t(\hat{\boldsymbol{h}}_m) = \frac{\sum_{j=1}^{n_t-1} \hat{h}_m^2(j)}{\sum_{j=n_t}^{L} \hat{h}_m^2(j)}, \tag{5.2}$$

where $n_t$ is the sample corresponding to $t$ ms, $\hat{\boldsymbol{h}}_m$ the estimated RIR at the $m^{th}$ microphone and $L$ its length in samples.

To assess the robustness of the proposed method to noise in the estimated RIRs, the $C_t$ features were extracted from noisy estimates of the RIRs, which were obtained by adding noise to the ground truth RIRs. The noise samples were drawn from a Gaussian Mixture Model (GMM) fitted onto the errors observed on RIRs estimated by the RNMCFLMS as described in [149]. The power of the noise was then scaled so that a target NPM is achieved.

### 5.3.5 Feature labeling

The extracted features were then labeled using an HMM [150]. The HMM had $P$ states, each representing a speaker. An emission probability, modeled by a Gaussian density with a diagonal covariance matrix, was associated with each state. These emission probabilities represent the density of the features given a state, i.e. the density of some observed features knowing that these features were produced by a specific speaker. The structure of the covariance matrix, i.e. diagonal, was chosen to reduce the computational complexity. This choice implicitly assumes that the input of the features are uncorrelated with each other.

For the problem of feature labeling, the HMM is used to determine the most likely sequence of states. To this end, in addition to the emission probability, we need to define

the transition probabilities, which quantify the probability of changing states, i.e. speaker, and prior probabilities, which represent a prior belief about the activity of the states relative to each other. The transition and prior probabilities were initialized with a uniform distribution.

The following iterative scheme was then used to estimate the most likely sequence of states:

1. Given an assignment of each feature to a state, compute the emission probabilities.
2. Given the prior and transition probabilities as well as the emission probabilities, compute the most likely sequence of states using the Viterbi algorithm [151].
3. Given the new feature-to-state assignment, update the parameters of the Gaussian distributions of each state.
4. Given the estimated path and the new statistics of each state, update the prior and transition probabilities using the Baum-Welch algorithm [152].

To initialize the emission probabilities, the $k$-means algorithm [153, Chapter 9] was used to create $P$ initial clusters of features, each associated with a state the HMM. The initial values of the means and covariance matrices for the emission probabilities corresponding to each state were then obtained based on the clusters.

### 5.3.6 Search for the optimal spatial feature

To determine the value of $t$ which yields a spatial feature robust to errors in the estimated RIR and to speaker locations, a Monte-Carlo simulation is implemented.

#### Setup

Two sets of RIRs, i.e. one set for each source, is generated using the image method [83] for two different source-microphone configurations in a shoebox shaped room, simulating large conference room, of dimension $6 \times 6 \times 2.5$ m with a reverberation time $T_{60}$ of 0.6 s, $M = 2$ microphones and a sampling frequency of $f_s = 8$ kHz. The obtained RIRs,

| | $C_5$ | $C_{10}$ | $C_{15}$ | $C_{50}$ |
|---|---|---|---|---|
| NPM=-10 dB | 12.5 | 13.3 | 9.9 | 3.8 |
| NPM=-5 dB | 2.9 | 3.2 | 2.0 | 0.57 |

Table 5.1: Bhattacharyya distance between the spatial feature clusters for various levels of NPM.

truncated down to $L = T_{60} f_s = 4800$ samples, were then corrupted by an additive noise following the model described in [149] so that the NPM achieved a desired value. The robustness of $C_t$ against errors in the estimated RIRs was evaluated for NPMs of -10 dB and -5 dB.

The value of $C_t$ was computed for $t \in \{5, 10, 15, 50\}$ ms for 100 realizations of the additive noise. The clusters of $C_t$ corresponding to each source, each modeled by a 2-dimensional Gaussian distribution, were then assessed in terms of separability using the Bhattacharyya distance [154], the expression of which is given by [155]

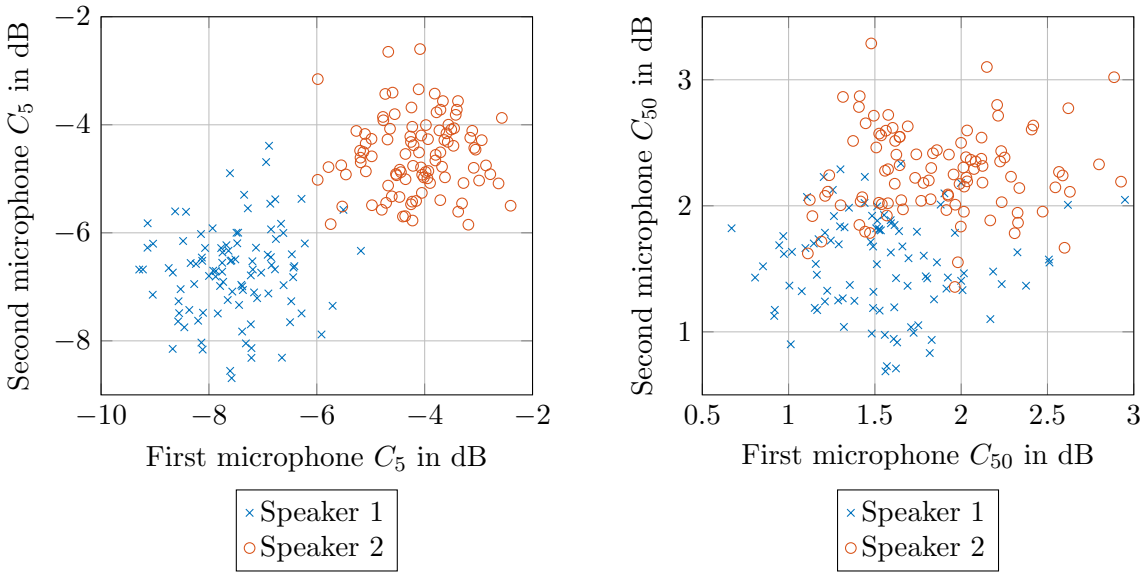
$$d_B = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\log\left(\frac{\left|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}\right), \tag{5.3}$$

where $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively represents the means and covariance matrices of the clusters and $|\boldsymbol{\Sigma}_1|$ denotes the determinant of $\boldsymbol{\Sigma}_1$. The first term of (5.3) increases as the difference between the means increases while the second term, which depends only on the covariance matrices, is equal to zero when the covariance matrices are equal. Hence, the larger the value of $d_B$, the more separable the two clusters will be. This distance therefore allows us to assess which value of $t$ used in $C_t$ leads to the most noise robust features.

### Results

Table 5.1 gives the Bhattacharyya distance between clusters corresponding to each speaker for the different locations and NPMs. It can be observed that $C_5$ and $C_{10}$ yield similar separability while the Bhattacharyya distance for clusters of $C_{15}$ and $C_{50}$ are much smaller.

It can be observed in Fig. 5.2 that for an NPM of -5 dB, the clusters of $C_{50}$ points corresponding to each speaker are merging while the clusters of $C_5$ points show less overlap.




(a) $C_5$. (b) $C_{50}$.

Figure 5.2: Scatter plot of the $C_5$ and $C_{50}$ features for each speaker for an NPM of -5 dB.

### 5.3.7 Diarization experiment setup

#### Speech input generation

The simulated meeting data were obtained using 20 utterances spoken by 2 speakers from the test set of the TIMIT database [134]. Two sets of RIRs were generated using the image method [83], one set corresponding to each speaker. Each utterance was then convolved with the corresponding set of RIRs. The reverberant utterances were then combined to produce an interleaved signal, where the speakers speak in turn. The whole speech data had a duration of approximately 60 s. The simulated data were free of time instants where both speakers are talking at the same time.

White Gaussian noise was added to the dry reverberant meeting signal to achieve a SNR of 30 dB. An excerpt of the simulated speech signal at the first microphone is shown in Fig. 5.3.

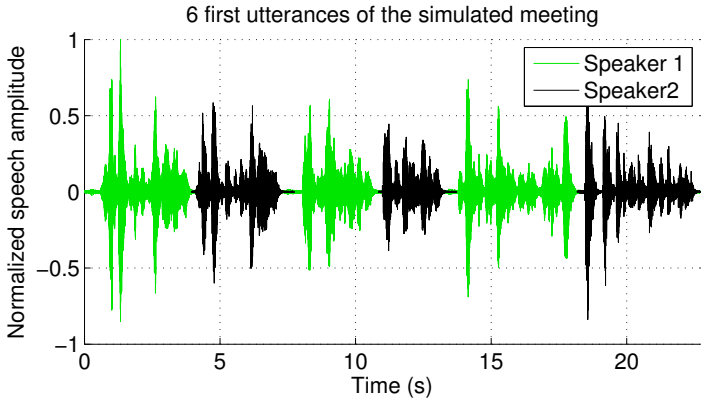



Figure 5.3: Example speech from simulated meeting.

### Experimental setup

Since the optimal value, in the Bhattacharrya sense, of $t$ used in $C_t$ was determined for a specific room, that value of $t$ may not be informative for other rooms. Therefore, in this setup, to illustrate the usefulness of $C_t$ for different rooms, a room of different dimensions and different $T_{60}$ is considered.

The considered room is of dimension $5 \times 6 \times 3$ m. Throughout the experiments, the reverberation time is set to $T_{60} = 0.5$ s, leading to RIRs of length $L = 4000$ for a sampling frequency $f_s = 8000\,\text{kHz}$. The microphones of the microphone array with $M = 2$ were placed at coordinates $(2 \pm 0.2, 3, 1.5)$ expressed in a Cartesian system.

For each of the estimated active speech segments, a sliding analysis window of $t_e = 1$ s is applied with a sliding increment of $t_s = 100$ ms. This leads to frames of duration $t_e = 1$ s overlapping by 900 ms. Although the data were free of overlapping speech, i.e. time instants when both speakers are active, some frames contained both speakers, e.g. the frame from 3 s to 4 s in Fig. 5.3.

A given frame contains either no speaker, only one of the speakers or both. When no speaker is present in the frame, i.e. the VAD failed in detecting the pause, the estimation of the RIRs should not correspond to any of the ground truth RIRs. Therefore, the estimated RIRs are given by one of the two sets of ground truth RIRs, randomly chosen and corrupted by additive noise following the model described in [26] so that the NPM has

a small value ($10^{-6}$ dB). As seen in Chapter 2, such a low value means that the estimate is almost orthogonal to the RIRs and therefore holds no information.

In the case where only one speaker is present in the frame, the estimated RIRs were given by the ground truth RIRs corresponding to that speaker corrupted by an additive noise so that a desired NPM $\epsilon_s$ is achieved.

In the case where both speakers are present in the frame, the estimated RIR at each microphone is given as an average impulse response weighted by the proportion of the active time of each speaker in the considered frame. Noise was also added in the latter case to achieve an NPM of $\epsilon_s$. A different realization of the additive noise is computed for each frame so that RIR estimates obtained from a BSI algorithm periodically reinitialized are simulated. Estimates of $C_5$, which is the ratio of the energy contained in the first 5 ms of the RIR to the energy of the remainder of the RIR, are then obtained from these sets of RIRs, one per microphone.

*Accuracy of BSI.* In the first experiment, the speakers were respectively localized at coordinates $(3.18, 4.88, 1.57)$ and $(2.33, 3.98, 1.53)$. For that particular configuration, the true TDOAs were -0.398 ms and 0.199 ms for the first and second speaker respectively. The values of $C_5$ were respectively $(-5.82, -4.37)$ and $(-1.05, -3.50)$, in dB, for the first and second speakers. In that setup, the robustness of the proposed method to BSI errors was investigated by evaluating the Diarization Error Rate (DER) for $\epsilon_s$ taking 20 linearly spaced values between -10 dB and -1 dB.

*Monte-Carlo simulation.* In the second experiment, the locations of the speakers were randomly drawn under the constraint that they had to be at least 50 cm away from the walls, the microphone array and each other. The accuracy of the estimated RIRs for frames effectively containing speech was set to achieve $\epsilon_s = -10$ dB. The performance of the system was evaluated over 100 different speaker locations.

Since the implemented method to estimate the TDOA features operates in the frequency domain, a Hamming window was applied to reduce windowing artifacts. As that method outputs integers and that the UDA cannot be applied due to the small number of microphones ($M = 2$), it is not always possible to directly fit a Gaussian distribution

model over the estimated TDOA in the HMM. To overcome that issue, a small amount of white Gaussian noise, the variance of which was equal to 0.01, was added.

### Evaluation

The performance of the diarization system was evaluated in terms of DER [156]. The DER corresponds to the fraction of time that is not attributed to the correct speaker. That score can be broken down into the missed speaker time, false alarm time and speaker error time. The first two solely depend on the VAD as they respectively correspond to the fraction of time of estimated silence when speech was present and the fraction of time of estimated speech when there was a silence.

The DER is computed as follows [156]:

$$\mathrm{DER} = \frac{\sum_{\sigma \in S_{\mathrm{seg}}} \tau(\sigma) \left\{ \max\left(N_{\mathrm{ref}}(\sigma),\, N_{\mathrm{sys}}(\sigma)\right) - N_{\mathrm{correct}}(\sigma) \right\}}{\sum_{\sigma \in S_{\mathrm{seg}}} \tau(\sigma) N_{\mathrm{ref}}(\sigma)} \tag{5.4}$$

where $S_{\mathrm{seg}}$ is the set of speech segments taken between 2 consecutive speaker change points after merging the estimated and the reference diarization files, $\sigma$ is a segment and $\tau(\sigma)$ is the duration of $\sigma$. The symbols $N_{\mathrm{ref}}(\sigma)$ and $N_{\mathrm{sys}}(\sigma)$ respectively denote the number of reference and system speakers in $\sigma$. The number of reference speakers speaking in $\sigma$ for whom their mapped system speakers are also speaking in $\sigma$ is denoted by $N_{\mathrm{correct}}(\sigma)$. To take the inaccuracy of the hand labels into account, the default tolerance threshold of 250 ms used in [156] was adopted.

Hit, miss and false alarm rates were used to evaluate the performance of the system for speaker change detection. These were defined as follows:

- The Hit Rate (HR) corresponds to the percentage of estimated speaker changes lying within 250 ms around a true speaker change.
- The Miss Rate (MR) corresponds to the percentage of true changes not estimated within 250 ms.
- The False Alarm Rate (FAR) is the percentage of estimated speaker changes that

do not correspond to a true speaker change.

A speaker change refers to either when the identity of the active speaker effectively changes, when silence follows speech or when the speaker starts talking after a period of silence.

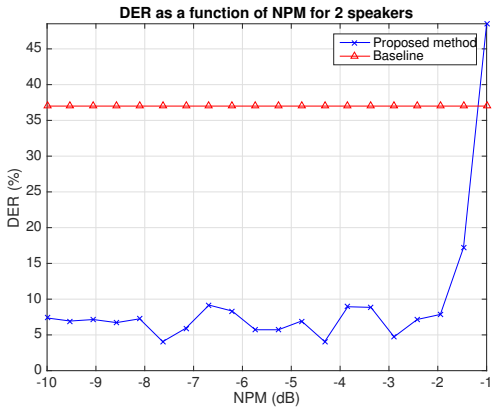
### 5.3.8 Experimental results

#### Fixed location, varying NPM

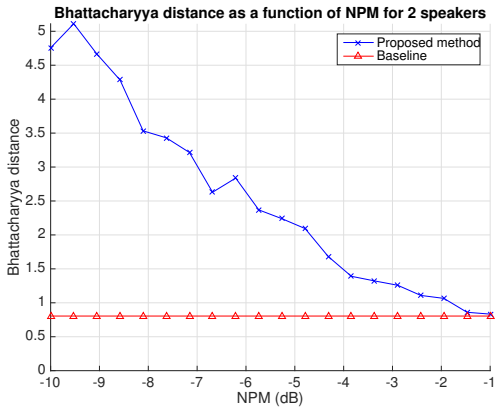
Figure 5.4 shows the variation of the DER of the proposed method for each value of NPM from -10 dB to -1 dB . Although the values of the NPM decreases from -10 dB to -1 dB, the proposed diarization system seems to be strongly affected for values of NPM below -2 dB. However, the Bhattacharyya score decreases as the NPM increases as shown in Fig. 5.4b.

Figure 5.5 is an example of the $C_5$ feature points for an NPM of -5 dB. As the NPM increases, the clusters seem to merge, which results in a higher DER and a lower Bhattacharyya score.

The TDOA based diarization system achieved a DER of 37% with a Bhattacharyya distance of 0.8.




(a) DER as a function of NPM for 2 speakers at given locations




(b) Bhattacharyya score as a function of NPM for 2 speaker at given locations

Figure 5.4: Variations of the DER and Bhattacharyya score with the NPM.

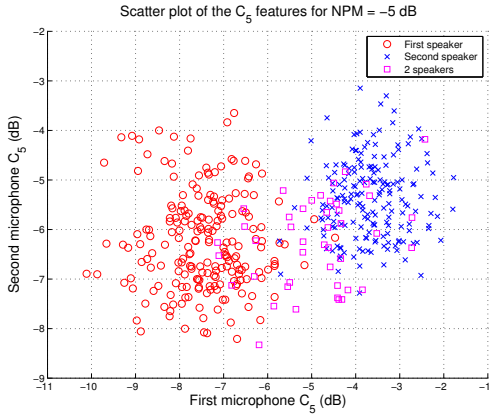



Figure 5.5: Scatter plot of the $C_5$ features for NPM=-5 dB and $M = 2$. The feature points corresponding to "2 speakers" refer to features extracted from frames containing 2 speakers, although these speakers were not active at the same time.

### Fixed NPM, changing locations

Figure 5.6 shows the DER obtained from 100 different speaker locations for an NPM of $-10$ dB. The proposed method leads to less variability of the DER than that of the approach using TDOA features only and has a mean DER of 8.95% against 17.5% for the baseline.

Table 5.2 shows the mean and standard deviation of the diarization system evaluated using the hit, miss and false alarm rate metrics. It can be seen that, on average, the proposed method yields a 3.7% absolute improvement in HR over the baseline and a MR and FAR lower by 7.7% and 14% in absolute value respectively than that of the baseline, while consistently yielding a smaller standard deviation.

| | Method | HR | MR | FAR |
|---|---|---|---|---|
| mean | Proposed | 72.8% | 32.83% | 27.82% |
| | Baseline | 69.06% | 39.15% | 41.39% |
| std. | Proposed | 6.54% | 6.55% | 6.18% |
| | Baseline | 8.25% | 7.64% | 16.13% |

Table 5.2: Performance of the diarization system in terms of speaker change detection.

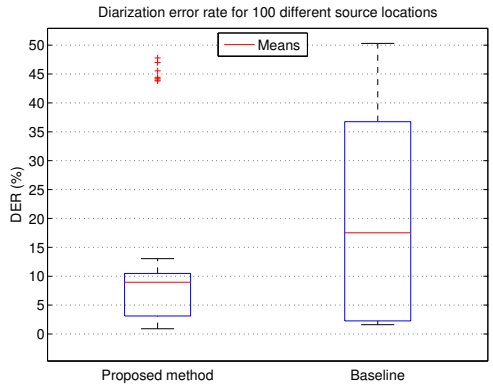



Figure 5.6: Box diagram of the DER obtained from 100 different speaker locations. The estimated stacked RIR had an NPM of -10 dB.

### 5.3.9 Discussion

From these experimental results, $C_5$ alone can be used to discriminate different speakers.

Because of the similarity between $C_5$ and DRR, in cases for which the RIRs, and therefore $C_5$ cannot be estimated with sufficient accuracy, it is possible to blindly estimate the spatial features from the reverberant noisy speech data [56]. In the next section, the combination of blindly estimated DRRs and speech features is explored to diarize single-channel meeting recordings.

## 5.4 Single-channel environment-aware speaker diarization

To simplify the notations, the multi-microphone model adopted in (5.1) is rewritten here as

$$x_i[n] = h_i[n] * s_i[n], \tag{5.5}$$

$$y[n] = \sum_{i=1}^{N_P} x_i[n] + \nu(n) \tag{5.6}$$

where $i$ represents a speaker index. As previously, the RIRs are assumed to not change during the utterances in the period of interest.

The novel diarization system proposed in this section adopts the framework presented

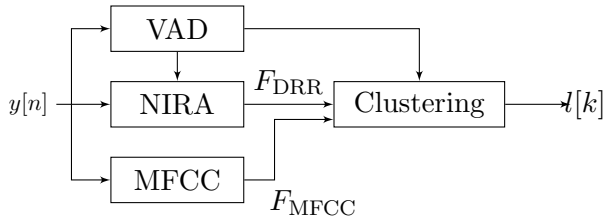



Figure 5.7: Block diagram of the proposed speaker diarization system.

in [157], which follows the block diagram shown in Fig. 5.7. That framework solves the speaker diarization problem by taking an information theoretic approach based on the Information Bottleneck (IB) principle [158].

### 5.4.1 The information bottleneck principle

The signal $y[n]$ is partitioned into non-overlapping frames from which features are extracted. These features are then clustered so that a label, which identifies a speaker, is assigned to each cluster. In this subsection, the adopted clustering algorithm, i.e. the one presented in [157], is described. As this clustering algorithm relies on the IB principle [158], let us describe that principle first.

Let us consider a set $\mathscr{A}_y$ of input variables extracted from a uniform segmentation of the observed signal $y[n]$ and a set $\mathscr{Q}_x$ of relevance variables, correlated with $\mathscr{A}_y$ and relevant to the clustering problem. The information bottleneck aims to find the most compact representation $\mathscr{C}$ of $\mathscr{A}_y$ while preserving as much information about $\mathscr{Q}_x$ as possible [158]. Formally, denoting by $I(\cdot,\cdot)$ the mutual information of two random variables and $\beta$ a trade-off parameter, the cost function

$$J_{\text{IB}} = I(\mathscr{A}_y, \mathscr{C}) - \beta I(\mathscr{C}, \mathscr{Q}_x) \tag{5.7}$$

is minimized with respect to the mapping $p(c|q_x)$, with $c \in \mathscr{C}$ and $q_x \in \mathscr{Q}_x$ [158].

The minimization of (5.7) is achieved via the agglomerative information bottleneck algorithm proposed by Slonim and Tishby in [159]. The agglomerative information bottleneck is a bottom-up iterative algorithm, which starts with the trivial partition

$\mathscr{C} = \cup_{a_y \in \mathscr{A}_y} \{\{a_y\}\}$, i.e. the partition where each element of $\mathscr{C}$ is a singleton containing only one feature point, and merges, at each iterations, the two elements of $\mathscr{C}$, which yield the smallest increase in $J_{\mathrm{IB}}$. The computations required to carry out the agglomerative information bottleneck algorithm only depend on the posterior distribution $p(q_x|a_y)$ [160], where $a_y \in \mathscr{A}_y$.

To determine the number of clusters in $\mathscr{C}$, i.e. the number of speakers, Vijayasenan, Valente, and Bourlard proposes to stop the agglomerative information bottleneck algorithm when the normalized mutual information $\frac{I(\mathscr{C},\mathscr{Q}_x)}{I(\mathscr{A}_y,\mathscr{Q}_x)}$ decreases below a certain threshold, which is empirically chosen based on the training set used in [161].

### 5.4.2 Feature specification and combination

#### Input features

The whole meeting recording is segmented into non-overlapping segments $a_y$ of 3 s. The set $\mathscr{A}_y$ of input variables is defined as the set of these segments. The speech and spatial features are extracted from sliding windows of 20 ms overlapping by 50% applied to each segment in $\mathscr{A}_y$. For clarity, let us denote by $D$ the total number of windows within a segment of $\mathscr{A}_y$.

*Speech features:* The speech features, denoted by $s_j^{\mathrm{MFCC}}$, $j \in \{1, 2, \ldots, D\}$, are 19 MFCCs extracted from the 20 ms sliding windows after application of a Hamming window.

*Spatial features:* The spatial features, denoted by $s_j^{\mathrm{DRR}}$, $j \in \{1, 2, \ldots, D\}$ are estimates of the DRR, which is defined as [53]

$$\mathrm{DRR}_i = 10 \log_{10} \left( \frac{\sum_{n=n_d-n_0}^{n=n_d+n_0} h_i^2(n)}{\sum_{n=0}^{n=n_d-n_0} h_i^2(n) + \sum_{n=n_d+n_0}^{\infty} h_i^2(n)} \right) \mathrm{dB}, \tag{5.8}$$

where $n_0$ is the number of samples in a rectangular window of 8 ms, $n_d$ the time index (in sample) of the direct path in the RIR. However, unlike in Section 5.3, the DRRs are here blindly estimated from the noisy reverberant speech signal via the Non-Intrusive Room

Acoustics (NIRA) estimator developed by Parada, Sharma, Lainez, *et al.* [162].

### Relevance variables

Motivated by the success of GMMs [163] in the task of speaker recognition and verification, see e.g. [164], Vijayasenan, Valente, and Bourlard proposed to use the components of a GMM trained on the MFCC data as the set of relevance variables for the speech features. The number of components is equal to the number of segments in $\mathscr{A}_y$. The probability $p(q_x = 1|s_j^{\text{MFCC}})$ can then be viewed as the contribution of component $q_x$ in explaining $s_j^{\text{MFCC}}$ [153, Chapter 9]. The probability $p(q_x = 1|a_y^{\text{MFCC}})$, needed to carry out the agglomerative information bottleneck, is then obtained as $p(q_x = 1|a_y^{\text{MFCC}}) = \sum_j p(q_x = 1|s_j^{\text{MFCC}})\delta_{\{s_j^{\text{MFCC}} \in a_y^{MFCC}\}}$, where $\delta_{\{s_j^{\text{MFCC}} \in a_y^{\text{MFCC}}\}}$ is equal to 1 when $s_j^{\text{MFCC}}$ has been extracted from the segment $a_y^{\text{MFCC}}$ and 0 otherwise.

Given the GMM trained on the MFCC features, each of the $s_j^{\text{MFCC}}$ is assigned to a component of the GMM. A separate GMM is then trained on the $s_j^{\text{DRR}}$ using that mapping between $s_j^{\text{MFCC}}$ and $q_x$ to obtain $p(q_x|a_y^{\text{DRR}})$. This ensures that the data used to train the GMMs are perfectly aligned.

### Feature combination

Similarly to what has been proposed in [160], given the posterior $p(q_x = 1|a_y^{\text{MFCC}})$ and $p(q_x = 1|a_y^{\text{DRR}})$, the combination of the two types of features takes place in the probability space by computing $p(q_x = 1|a_y)$ as a weighted sum between $p(q_x = 1|a_y^{\text{DRR}})$ and $p(q_x = 1|a_y^{\text{MFCC}})$:

$$p(q_x = 1|a_y) = p(q_x = 1|a_y^{\text{DRR}})P_{\text{DRR}} + p(q_x = 1|a_y^{\text{MFCC}})P_{\text{MFCC}}, \tag{5.9}$$

where the constant weights $P_{\text{DRR}}$ and $P_{\text{MFCC}}$, estimated from a training set, are such that $P_{\text{MFCC}} + P_{\text{DRR}} = 1$.

### 5.4.3 Pre- and post-processing

#### Voice activity detector

A VAD, based on the P.56 method [146], [147], is applied to the reverberant noisy speech signal so that features estimated from non-speech 20 ms frames are excluded in the fitting of the GMMs.

#### Realignment

At the output of the agglomerative IB algorithm, the clusters are aligned with the boundaries of the initial arbitrarily selected 3 s segments. Those boundaries are realigned by computing the sequence of clusters that minimizes the cost function based on the posterior distribution of the relevance variables [160].

### 5.4.4 Diarization systems

#### Baseline

The baseline diarization system only uses the speech features in the clustering step. This system is similar to what has been proposed in [161] with difference that the realignment step described in Section 5.4.3 and in [160] instead of the HMM proposed in [161].

#### Proposed diarization system

The novel system combines blindly estimated DRRs in addition to the MFCC features, as described in Section 5.4.3, to diarize the audio signal.

### 5.4.5 Experimental setup

This section describes the experimental setup used to evaluate the diarization systems. NIRA was trained on the training data of the REVERB Challenge [165], which consists of simulated noisy reverberant signals created by convolving measured RIRs with dry speech from the clean WSJCAM0 training set [166]. The diarization system is evaluated on

| | **Room 1** | **Room 2** | **Room 3** |
|---|---|---|---|
| $T_{60}$ (s) | 0.25 | 0.5 | 0.7 |
| *Near* DRR (dB) | 16 | 11 | 11 |
| *far* DRR (dB) | 6 | 0 | 2 |

Table 5.3: $T_{60}$ in s and DRR in dB for the *near* and *far* positions in each of the three rooms.

simulated meeting data, obtained by convolving speech utterances from the WSJCAM0 corpus with measured RIRs from the evaluation set of the REVERB Challenge [165] database. The speech utterances used to train NIRA and those used to create the simulated meeting data were different. Similarly, the measured RIRs used to generate the simulated meeting data and those used in the training phase of NIRA were captured in different rooms.

**Simulated meeting data**

The simulated meetings were generated by convolving clean speech with measured RIRs taken from the evaluation set of the REVERB Challenge database [165]. Recorded fan noise was added at 20 dB SNR.

*Speech data.* The nearly-anechoic speech data consisted of utterances taken from the WSJCAM0 corpus [166] which were recorded by a close-talking noise-cancelling head-mounted microphone. There were, in total, 14 speakers: 7 male and 7 female. Each utterance was only used once across the simulated meetings.

*RIRs.* The RIRs of the REVERB Challenge were captured in 3 rooms of different size by a circular microphone array. For each room, the RIR that was recorded at a distance of 0.5 m is labeled as *near* and the RIR that was recorded at a distance of 2 m is labeled as *far*. The ground truth DRRs together with the reverberation time $T_{60}$ are given in Table 5.3.

In each simulated meeting, the reverberant signal consisted of 10 to 14 utterances spoken by 2 different speakers speaking in turn. Spatial differences between the 2 speakers are simulated by convolving each speakers' signal with different RIRs, i.e. the utterances

spoken by one speaker are convolved with the *near* RIR while the utterances spoken by the other speaker are convolved with the *far* RIR.

In total, 42 simulated meeting audio streams were generated, 14 streams per room. Among the 42 audio streams, 18 streams contained speakers of different gender, 12 streams contained only male speakers and 12 streams contained only female speakers.

*Noise.* Fan noise signals, recorded in each room, were added to the simulated meeting data at a mean SNR of 20 dB, defined as [165]

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{N}\sum_{i=1}^{N_P}\tilde{x}_i^2(n)}{\sum_{n=1}^{N}\nu^2(n)}\right), \tag{5.10}$$

$$\tilde{x}_i[n] = \tilde{h}_i[n] * s_i[n] \tag{5.11}$$

where $N$ is the total number of samples in the considered recording, $\tilde{h}_i = [h_i(1), h_i(2), \ldots, h_i(n_{50})]$ is the truncated RIR containing the $n_{50}$ taps within the first 50 ms and $\tilde{x}_i[n]$ is the direct sound together with the early reflections up to 50 ms. That specific expression of the SNR in effect states that, in the considered scenario, the "signal" is the dry speech together with the speech components brought by the reflections contained in the first 50 ms of the RIRs. We adopt this definition of the SNR here because the first 50 ms reflections are known to increase the intelligibility of speech [44], which may help in discriminating between different speakers.

*Development and Evaluation sets.* The simulated data were broken down into two non-overlapping subsets: a development and an evaluation set.

The development set was used to determine the optimum weights $P_{\mathrm{MFCC}}$ and $P_{\mathrm{DRR}} = 1 - P_{\mathrm{MFCC}}$ to use in the IB based diarization system to minimize the DER defined below. The optimum weights were then used on the evaluation set.

### 5.4.6 Evaluation

As in Section 5.3.7, the performance of the diarization systems is evaluated in terms of DER. To remove the errors introduced by the VAD, the results are also given in terms of speaker error time.

### 5.4.7 Results

#### Development set

Figure 5.8 shows the variation of DER with $P_{\text{DRR}}$ on the development set. It can be observed that the use of DRR features alone in the IB based diarization system does not achieve low DER. However, by combining them with MFCCs, the performance of the diarization system can be improved to outperform the similar system solely relying on MFCCs.

The lowest DER in the development set is achieved for $P_{\text{DRR}} = 0.18$. This weight reduces the DER from 1.95% to 1.2%, achieving a relative improvement of 38%.

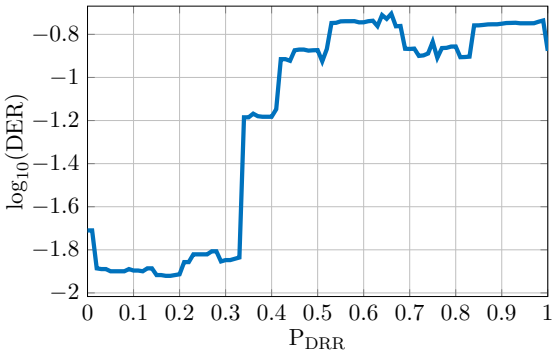



Figure 5.8: DER of the development set as a function of DRR weight ($P_{\text{DRR}} = 1 - P_{\text{MFCC}}$).

As shown in Table 5.4, the inclusion of DRR features in the diarization system reduces the DER on average. Although the baseline and the proposed system perform similarly in Room 1, significant improvements are observed in Room 2 and Room 3, i.e. rooms with higher amount of reverberation. A relative improvement of 56% and 46% are respectively seen in the 2 latter rooms.

| | **Overall** | **Room 1** | **Room 2** | **Room 3** |
|---|---|---|---|---|
| **Baseline** | 1.95% | 0.95% | 2.80% | 2.10% |
| **Proposed** | 1.20% | 0.94% | 1.22% | 1.44% |

Table 5.4: Mean DER of the baseline and the proposed method for the development set.

Analyzing the breakdown of the DER in terms of gender, the proposed system achieves lower DER for simulated meeting data containing mixed genders or 2 female speakers. In these cases, the mean DERs were respectively 1.09% and 1.79% for the proposed system against 2.28% and 3.06% for the baseline.

### Evaluation set

The weights $P_{\text{MFCC}} = 0.82$, $P_{\text{DRR}} = 0.18$ determined from the development set are now applied to evaluate the performance of the diarization system on the evaluation set.

Figure 5.9 shows that the inclusion of DRR features improves by 23% the performance of the diarization system on average. The DER as well as the speaker error time in Room 2 and Room 3 are decreased while the DER in Room 1 is increased by 9% on average. This degradation in performance is due to the limitations of the NIRA DRR estimator, which is reported to have high estimation errors for environments with low reverberation [167]. Analyzing the Root-Mean-Square Deviation (RMSD) of the estimated DRR in each room, the RMSD indeed decreased as $T_{60}$ increased. The RMSD values were respectively 2.7 dB, 2.5 dB and 2.2 dB for Room 1, Room 2 and Room 3.

Similar observations can be made about the speaker error time, which decreases from 2.35% and 1.34% down to 0.85% and 1.03% for Room 2 and Room 3 respectively. The speaker error time increases from 1.07% to 1.24% in Room 1. The variance of the DER is also significantly reduced in Room 2 and Room 3 and slightly increased in Room 1 as shown in Table 5.5.

| | **Room 1** | **Room 2** | **Room 3** | **Overall** |
|---|---|---|---|---|
| **Baseline** | 0.68% | 4.03% | 1.06% | 2.50% |
| **Proposed** | 0.75% | 1.16% | 0.69% | 0.91% |
| **RMSD (dB)** | 2.7 | 2.5 | 2.2 | 2.5 |

Table 5.5: Standard deviation in speaker error time of the baseline and proposed systems for the evaluation set together with the RMSD of the estimated DRR.

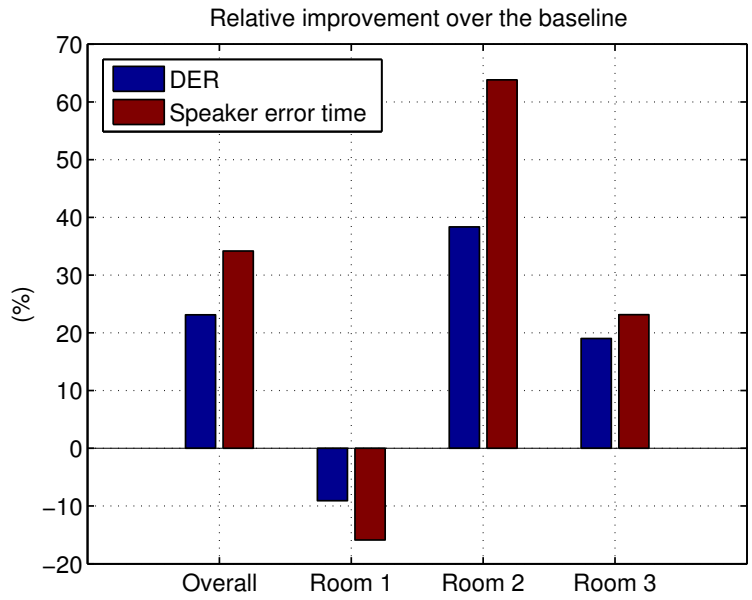



Figure 5.9: Relative improvement in DER and speaker time error by inclusion of DRR features.

Figure 5.10 shows that the estimated DRRs mostly take values around 10 dB and 2 dB, depending on the identity of the active speaker. This figure corresponds to the simulated meeting which had the highest DER and speaker error time, which were respectively 16.73% and 13.7%. The presence of the DRR features respectively decreased the DER and speaker error time to 7.13% and 4.1%, achieving a relative improvement of 57% and 70% respectively.

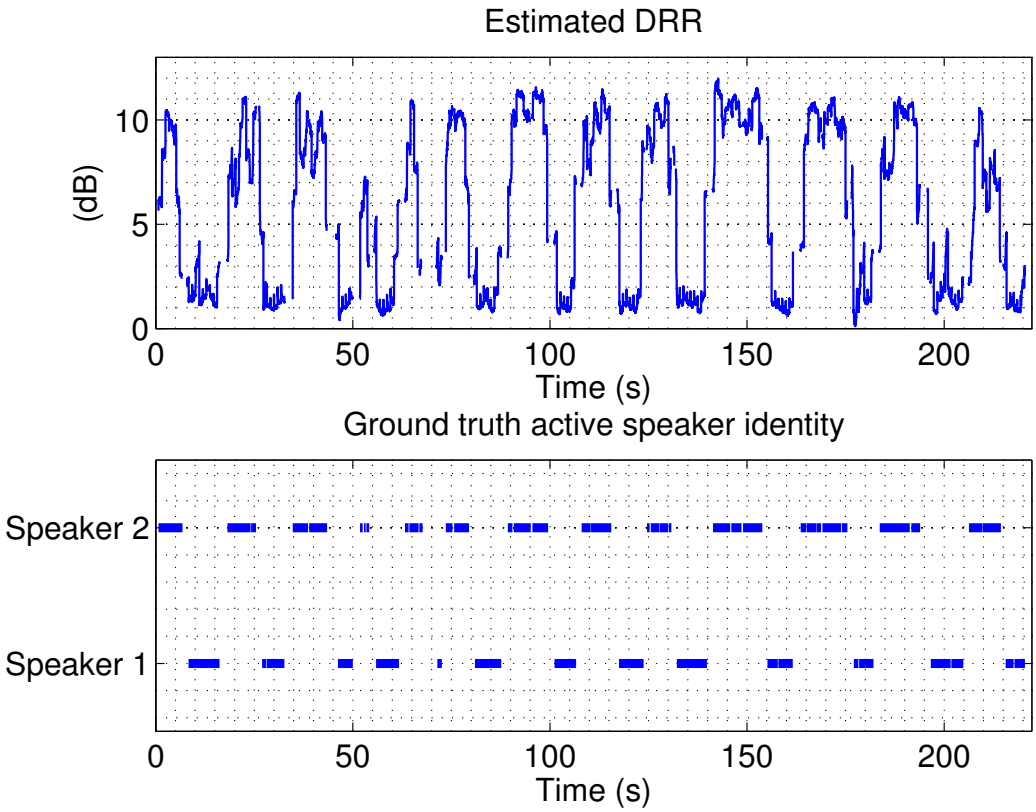



Figure 5.10: Estimated DRR along with the ground truth speaker identity.

Table 5.6 shows that the proposed system decreases the mean speaker error time when the speakers have the same gender while the mean speaker error time slightly increases when the speakers have different gender. This shows that the DRR feature is beneficial when the MFCCs have low discrimination capabilities.

| | **Male - Male** | **Mixed** | **Female - Female** |
|---|---|---|---|
| **Baseline** | 1.07% | 0.76% | 3.20% |
| **Proposed** | 0.74% | 0.77% | 1.68% |

Table 5.6: Mean speaker error time broken down by gender for the evaluation set.

## 5.5 Summary and discussion

In this chapter, the use of the DRR as a feature for speaker diarization has been investigated.

Preliminary experiments with DRRs estimated from noisy RIRs showed that the DRR was suited for the task of speaker diarization. A novel diarization system, which combines blindly estimated DRRs with MFCCs, has then been proposed for single channel speaker diarization. By combining non-intrusive estimates of the DRRs with MFCC features, the proposed diarization system was able to use spatial information for single-channel speaker diarization. Through simulations, the proposed diarization system has been shown to give a relative improvement of 34% in average in terms of speaker error time. The standard deviation of the speaker error time was also reduced in the 2 rooms with higher $T_{60}$.

We might suggest that a good way forward would be selective combination, such that if the $T_{60}$ indicates an environment with low reverberation, spatial features are not used, otherwise if the $T_{60}$ indicates a moderately or highly reverberant environment, we use the spatial features as proposed in this paper. We note that the NIRA algorithm can also be used to estimate the $T_{60}$ in a non-intrusive manner.

# Chapter 6

# Conclusion

This final chapter summarizes and gives possible extension of the work presented in this thesis. Section 6.1 summarizes the problem addressed together with the different adopted approaches and their respective results. Section 6.2 gives guidelines for future research in blind identification of SIMO acoustic systems.

## 6.1 Summary

This thesis has investigated the problem of blind identification of SIMO acoustic systems, that is the estimation of the RIRs relating a single active sound source to several microphones using the recorded signals at each microphone only. The reverberant signal was modeled by a time-invariant FIR filter, also referred to as a RIR, convolved with a dry source. The estimation of the RIRs, in a multi-channel, setup relied on the cross-relation originally found by Xu, Liu, Tong, *et al.* The cross-relation approach allows us to overcome the problem of statistically modeling speech at the expense of the following identifiability assumptions:

- The source signal has a full-rank covariance matrix
- The RIRs do not share common zeros in the $z$-domain
- The length of the RIRs is known.

The analysis of the eigenvalues of the cross-relation matrix showed that the condition number of the eigenvector associated with the smallest eigenvalue of the noiseless cross-relation matrix drastically increases as the length of the RIRs increases. According to perturbation theory, this increase in the condition number leads to an ill-conditioned associated invariant subspace such that a small perturbation on the cross-relation matrix results in eigenvectors misaligned with the stacked RIRs, which correspond to an eigenvector associated with the smallest eigenvalue of the unperturbed cross-relation matrix. This highlighted the limit of a straight application of the LS method to blind identification of acoustic systems.

A reformulation of the MCLMS and NMCFLMS showed that these algorithms, in fact, cast the LS method into an adaptive filter framework. The MCLMS is indeed an adaptive filter relying a Rayleigh quotient cost function with an additional normalization step in the algorithm. The NMCFLMS, on the other hand, can be seen as a computationally efficient implementation of the MCLMS. The curvature of the optimization surface is also exploited in the NMCFLMS by using an approximation of the inverse of the Hessian matrix of the cost function. The cost function is the norm of the cross-relation residual and the processing takes place in the frequency domain to exploit the computational efficiency of the Fast Fourier Transform (FFT).

That shift of framework, i.e. batch to adaptive, enables the MCLMS and NMCFLMS to estimate much longer impulse responses than the LS approach. However, according to results found in the literature of neural networks, the NMCFLMS is guaranteed to converge in noisy conditions only when a non-summable step-size decreasing to 0 is employed, which is detrimental to adaptation to time-varying environments.

A semi-batch formulation of the MCLMS has therefore been investigated in the algorithm named BURQ. Simulations showed that an adaptation rule, based on the Rayleigh quotient and a constant step-size, computed frame by frame and without normalization resulted in a noise robust algorithm which converges to the stacked RIRs. However, the initial convergence rate has also been shown to be much slower than that of the MCLMS, NMCFLMS and RNMCFLMS.

To increase that initial convergence rate, a locally optimal adaptive step-size has been proposed. The proposed adaptive step-size, which exploits the fact that the estimated RIRs must be an eigenvector of the cross-relation matrix, drastically improves the convergence rate of the BURQ algorithm. However, similarly to the other adaptive step-size methods proposed in the literature, the resulting algorithm remains sensitive to noise.

Another approach to increase the convergence rate was to reduce the number of unknowns via the RTF, which relates the taps of a multi-channel system to those of a reference channel. The minimization of the norm of the cross-relation residual in the frequency domain was shown to yields estimates of the RTF and a reconstruction algorithm, that is the estimation of the stacked RIRs given a RTF has been proposed. However, it has also been observed that the estimation of the RTF is sensitive to windowing although a lower bound of achievable NPM of less than -15 dB has been found. Given a noisy estimate of the RTF, the estimation of the full RIRs however remains an open question.

Facing the stability issue of the invariant subspace associated with the smallest eigenvalue of the population cross-relation matrix, which seems to be an intrinsic limitation of LS or LMS approaches, an alternative way of exploiting the cross-relation in noisy conditions has been investigated. The residual obtained by application of the cross-relation to noisy reverberant signals has been shown to be equal to noise signals filtered by pairs of RIRs at the solution, i.e. the solution to the problem of BSI annihilates the dry source signal from the cross-relation residual. This motivated the exploitation of the non-stationary nature of speech against the stationarity of the additive noise signals for BSI.

The combination of the resulting cost function with characteristics of RIRs, namely spectral flatness, isolated impulse pattern in the early reflection and exponential decay of the reverberation tail, yielded an algorithm which is able to blindly estimate RIRs of realistic lengths, which was out of reach of previous methods.

While reverberation introduces a smearing of the speech power spectrum, which may be detrimental to speech feature based diarization systems, the use of spatial information encoded in RIRs has been investigated. In this approach, reverberation is therefore seen as

a source of information about a speaker, while previous methods usually relied on features sensitive to reverberation.

The DRR has been found to be beneficial for that purpose. The combination of estimated DRRs with speech features, namely MFCCs, resulted in a diarization system that has lower DERs than that of a diarization system relying on speech features only. Moreover, the need for BSI has been circumvented by blindly estimating DRRs via the NIRA estimator, which also enabled the use of spatial information in single-channel diarization systems.

## 6.2 Guidelines for future research on blind identification of SIMO acoustic systems

From the work presented here, approaching the problem of blind identification of RIRs from an annihilation filter point of view seems to be more promising than applying methods that have been developed for blind identification of communication channels. The latter methods indeed assume FIR filters of a few coefficients and work best when the system is excited by white Gaussian noise.

By adopting an annihilation filter approach, the solution of the problem of BSI of RIRs can then exploit the statistical differences between speech and the additive noise. Of particular interest is the question of a measure of stationarity. The cost function proposed in this thesis indeed required a small energy of the residual with small variance across time. However, the optimality of that measure has not been investigated.

Models of RIRs have also been shown to be beneficial to BSI. While the exponential decay regularization term proved to effectively yield estimated RIRs with appropriate decay rates, the spectral flatness assumption and the early reflection regularization term appeared to not accurately model RIRs. This opens two research questions:

- How could the estimated RIRs be regularized to show strong isolated spikes in the early reflections?

- How could the frequency dependency of the reverberation time be exploited for more accurate estimates of the RIRs?

While the assumptions that the RIRs do not share common zeros and that the source signal has a full-rank covariance matrix seem reasonable, the knowledge of the length of the RIRs is a fundamental drawback to practical applications. Therefore the extension of existing BSI algorithms to length mismatch scenarios needs to be investigated. In such cases, reference RIRs, against which the estimated RIRs need to be compared, need to be defined. Possible reference RIRs would be the first taps of each RIRs used to create the reverberant signal, which essentially states that BSI algorithms operating on underestimated RIRs should estimate the early reflections of each RIR.

However, when applied to recorded reverberant signal, the RIRs relating the sound source to the microphones is not known. A metric therefore needs to be developed to evaluate the performance of BSI algorithms applied to recorded reverberant speech signals.

The term "misconvergence", which has been used to refer to the behavior of the adaptive algorithms under noisy conditions, implies that these algorithm effectively converge to a point in the search space. While this is often admitted, no study of such convergence can be found in the literature. Knowing the exact behavior of the algorithms, i.e. convergent to an unwanted point in the search space or divergent, would provide us with an in-depth understanding of the issue encountered by the adaptive algorithms.

These guidelines are by no-mean exhaustive but the profusion of open questions shows that the field of BSI still holds unexplored ideas which may enable the use of various speech processing algorithms relying on the existence of accurately estimated RIRs, such as channel equalization or room inference algorithms.

# Bibliography


[1] M. Hu, D. Sharma, S. Doclo, M. Brookes, and P. Naylor, "Speaker change detection and speaker diarization using spatial information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015.

[2] M. Hu, S. Doclo, D. Sharma, M. Brookes, and P. A. Naylor, "Noise robust blind system identification algorithms based on a rayleigh quotient cost function," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sep. 2015.

[3] M. Hu, P. Peso Parada, D. Sharma, S. Doclo, T. van Waterschoot, M. Brookes, and P. A. Naylor, "Single-channel speaker diarization based on spatial features," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015.

[4] M. Hu, D. Sharma, S. Doclo, M. Brookes, and P. A. Naylor, "Blind adaptive SIMO acoustic system identification using a locally optimal step-size," in *Proc. Audio Eng. Soc. (AES) Conf.: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, Leuven, Belgium, Feb. 2016.

[5] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *J. Audio Eng. Soc. (AES)*, vol. 35, pp. 299–306, 1954.

[6] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1421–1448, Jul. 2012.

[7] T. Gerkmann and R. C. Hendriks, "Improved mmse-based noise PSD tracking using temporal cepstrum smoothing," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 105–108.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[9] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[10] N. Faraji and R. C. Hendriks, "Noise power spectral density estimation for public address systems in noisy reverberant environments," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[11] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[12] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[13] B. E. D. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Apr. 1997, pp. 1259–1262.

[14] L. E. Humes and L. Roberts, "Speech-recognition difficulties of the hearing-impaired elderlythe contributions of audibility," *J. of Speech, Language, and Hearing Research*, vol. 33, no. 4, pp. 726–735, 1990.

[15] M. Doostdar, S. Schiffer, and G. Lakemeyer, "A robust speech recognition system for service-robotics applications," in *RoboCup 2008: Robot Soccer World Cup XII*, L. Iocchi, H. Matsubara, A. Weitzenfeld, and C. Zhou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–12.

[16] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[17] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[18] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. Workshop Circuits, Systems and Signal Processing (ProRISC)*, Veldhoven, The Netherlands, Nov. 2004, pp. 250–254.

[19] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[20] W. Kellermann, "Beamforming for speech and audio signals," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, and M. Vorländer, Eds. New York, NY: Springer New York, 2008, ch. 35, pp. 691–702.

[21] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[22] G. Huang, J. Chen, and J. Benesty, "Direction-of-arrival estimation of passive acoustic sources in reverberant environments based on the Householder transformation," *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 3053–3060, Nov. 2015.

[23] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[24] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[25] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[26] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1379–1390, Sep. 2014.

[27] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 680–693, Apr. 2016.

[28] A. H. Moore, M. Brookes, and P. A. Naylor, "Room identification using roomprints," in *Proc. Audio Eng. Soc. (AES) Conf. on Audio Forensics*, Jun. 2014.

[29] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localisation in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, Oct. 2003.

[30] P. A. Naylor., E. A. P. Habets., J. Y. C. Wen, and N. D. Gaubitch, "Models, measurement and evaluation," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds., Springer, 2010, ch. 2, pp. 21–56.

[31] P. P. Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[32] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[33] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, 1997.

[34] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951–1968, 1998.

[35] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Process.*, vol. 43, no. 2, pp. 516–525, Feb. 1995.

[36] Y. Hua, "Fast maximum likelihood for blind identification of multiple FIR channels," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 661–672, Mar. 1996.

[37] H. Liu, G. Xu, and L. Tong, "A deterministic approach to blind identification of multi-channel FIR systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1994, pp. 581–584.

[38] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 340–349, Mar. 1994.

[39] K. Abed Meraim, P. Duhamel, D. Gesbert, P. Loubaton, S. Mayrargue, E. Moulines, and D. Slock, "Prediction error methods for time-domain blind identification of multichannel fir filters," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 1995, pp. 1968–1971.

[40] V. Tyagi, H. Bourlard, and C. Wellekens, "On variable-scale piecewise stationary spectral analysis of speech signals for ASR," *Speech Communication*, vol. 48, no. 9, pp. 1182 –1191, 2006.

[41] G. Grimmett and D. Stirzaker, *Probability and random processes*, ser. Probability and Random Processes. Oxford University Press, UK, 2001.

[42] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Processing*, vol. 82, pp. 1127–1138, Aug. 2002.

[43] ——, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[44] P. A. Naylor and N. D. Gaubitch, "Introduction," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds., Springer, 2010, ch. 1, pp. 1–19.

[45] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Apr. 2012.

[46] F. Lim and P. A. Naylor, "Relaxed multichannel least squares with constrained initial taps for multichannel dereverberation," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[47] H. Kuttruff, *Room Acoustics, Fifth Edition*. CRC Press, Jun. 2009.

[48] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *J. of Electrical and Computer Engineering*, vol. 2010, pp. 1–5, 2010.

[49] B. Blesser and L. Salter, *Spaces speak, are you listening?: Experiencing aural architecture*. MIT Press, 2009.

[50] H. Gazzah, P. A. Regalia, and J. P. Delmas, "Asymptotic eigenvalue distribution of block toeplitz matrices and application to blind simo channel identification," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1243–1251, Mar. 2001.

[51] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.

[52] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc. (AES)*, vol. 11, pp. 867–878, 2002.

[53] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE Challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[54] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[55] H. W. Löllmann, A. Brendel, P. Vary, and W. Kellermann, "Single-channel maximum-likelihood T60 estimation exploiting subband information," in *Proc. ACE Challenge Workshop, a satellite of IEEE-WASPAA*, New Paltz, NY, USA, 2015.

[56] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE Trans. Audio, Speech, Lang. Process.*, Jun. 2016.

[57] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Proc. Audio Eng. Soc. (AES) Convention*, Aug. 1997.

[58] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual Evaluation of Model and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," *J. Audio Eng. Soc. (AES)*, vol. 60, no. 11, pp. 887–898, Dec. 2012.

[59] A. Billon and J.-J. Embrechts, "Numerical evidence of mixing in rooms using the free path temporal distribution," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1381–1389, 2011.

[60] J. D. Polack, "La transmission de l'énergie sonore dans les salles," PhD thesis, Université du Maine, Le Mans, France, 1988.

[61] M. R. Schroeder, "Digital simulation of sound transmission in reverberant spaces," vol. 47, no. 2A, pp. 424–431, 1970.

[62] K. Kowalczyk, E. A. P. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Process. Lett.*, vol. 20, pp. 653–656, 2013.

[63] M. Crocco and A. Del Bue, "Room impulse response estimation by iterative weighted L1-norm," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015, pp. 1895–1899.

[64] N. D. Gaubitch, M. K. Hasan, and P. A. Naylor, "Generalized optimal step-size for blind multichannel LMS system identification," *IEEE Signal Process. Lett.*, vol. 13, no. 10, pp. 624–627, Oct. 2006.

[65] W. Qiu and Y. Hua, "Special issue on subspace methods, part i: Array signal processing and subspace computations performance analysis of the subspace method for blind channel identification," *Signal Processing*, vol. 50, no. 1, pp. 71 –81, 1996.

[66] N. D. Gaubitch, J. Benesty, and P. A. Naylor, "Adaptive common root estimation and the common zeros problem in blind channel identification," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2005.

[67] D. Schmid and G. Enzner, "Cross-relation-based blind simo identifiability in the presence of near-common zeros and noise," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 60–72, Jan. 2012.

[68] C. Hughes and A. Nikeghbali, "The zeros of random polynomials cluster uniformly near the unit circle," *Compositio Mathematica*, vol. 144, no. 3, pp. 734–746, May 2008.

[69] X. Lin, N. D. Gaubitch, and P. A. Naylor, "Two-stage blind identification of SIMO systems with common zeros," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[70] S. Gudvangen and D. Florencio, "Comparison of pole-zero and all-zero modelling of acoustics tranfer functions," *IEE Electronics Lett.*, vol. 28, no. 21, pp. 1976–1978, 1992.

[71] I. Fijalkow, "Multichannel equalization lower bound: A function of channel noise and disparity," in *Proc. IEEE Signal Process. Workshop on Statistical Signal and Array Process. (SSAP)*, Jun. 1996, pp. 344–347.

[72] L. Rabiner and R. Schafer, *Digital processing of speech signals*. New Jersey: Prentice Hall, 1988.

[73] R. Partington J., in *An Introduction to Hankel Operators*, ser. London Mathematical Society Student Texts. Cambridge University Press, Feb. 1989, pp. 1–3.

[74] S. Roman, *Advanced linear algebra*, ser. Graduate texts in mathematics. Springer, 2005.

[75] R. Vershynin, "How close is the sample covariance matrix to the actual covariance matrix?" *Journal of Theoretical Probability*, vol. 25, no. 3, pp. 655–686, 2012.

[76] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[77] Y. Saad, *Numerical methods for large eigenvalue problems*, Second. SIAM, 2011.

[78] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 1985.

[79] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*, Version 20121115, Nov. 2012.

[80] S. Haykin, *Adaptive filter theory*, Fitth, ser. Always learning. Pearson, 2014.

[81] S. Mitra and S. Mitra, *Digital signal processing: A computer-based approach*, ser. Connect, learn, succeed. McGraw-Hill, 2011.

[82] M. K. Hasan, J. Benesty, P. A. Naylor, and D. B. Ward, "Improving robustness of blind adaptive multichannel identification algorithms using constraints," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2005.

[83] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[84] J. H. Wilkinson, "The perfidious polynomial," *Mathematical Association of America*, vol. 24, pp. 1–28, 1984.

[85] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[86] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[87] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[88] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer handbook of speech processing*. Springer, 2008.

[89] J. Lee and S.-C. Chong, "On the convergence properties of multidelay frequency domain adaptive filter," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Phoenix, AZ, USA, Mar. 1999, 1865–1868 vol.4.

[90] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.

[91] I. Markovsky, S. V. Huffel, and R. Pintelon, "Block-toeplitz/hankel structured total least squares," *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 4, pp. 1083–1099, 2005.

[92] A. Beck and Y. C. Eldar, "Structured total maximum likelihood: An alternative to structured total least squares," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 5, pp. 2623–2649, 2010.

[93] M. R. P. Thomas, N. D. Gaubitch, E. A. P. Habets, and P. A. Naylor, "Supervised identification and removal of common filter components in adaptive blind SIMO system identification," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[94] ——, "An insight into common filtering in noisy SIMO blind system identification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2012.

[95] C. Chatterjee, "Adaptive algorithms for first principal eigenvector computation," *Neural Networks*, vol. 18, no. 2, pp. 145 –159, 2005.

[96] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of Mathematical Biology*, vol. 15, no. 3, pp. 267–273, 1982.

[97] L. Xu, E. Oja, and C. Y. Suen, "Modified hebbian learning for curve and surface fitting," *Neural Networks*, vol. 5, no. 3, pp. 441 –457, 1992.

[98] A. Taleb and G. Cirrincione, "Against the convergence of the minor component analysis neurons," *IEEE Transactions on Neural Networks*, vol. 10, no. 1, pp. 207–210, Jan. 1999.

[99] R. Ahmad, N. D. Gaubitch, and P. A. Naylor, "A noise-robust dual filter approach to multichannel blind system identification," in *Proc. European Signal Processing Conf.*, 2007.

[100] J. Huang, J. Benesty, and J. Chen, "Optimal step size of the adaptive multichannel LMS algorithm for blind SIMO identification," *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 173–176, Mar. 2005.

[101] R. Ahmad, A. W. H. Khong, and P. A. Naylor, "Proportionate frequency domain adaptive algorithms for blind channel identification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, Toulouse, France, 2006, pp. 2932–2935.

[102] O. Axelsson, *Iterative solution methods*. Cambridge University Press, Mar. 1994.

[103] E. Habets and P. Naylor, "An online quasi-newton algorithm for blind SIMO identification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, Mar. 2010.

[104] E. A. P. Habets, J. Benesty, and P. A. Naylor, "A cross-relation based affine projection algorithm for blind simo system identification," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2011.

[105] D. R. Morgan and J. C. Thi, "A delayless subband adaptive filter architecture," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1819–1830, Aug. 1995.

[106] A. Spriet, M. Moonen, and J. Wouters, "A multichannel subband GSVD approach for speech enhancement," *European Transactions on Telecommunications*, vol. 13, Special Issue on Acoustic Echo and Noise Control, no. 2, pp. 149–158, Apr. 2002.

[107] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[108] N. D. Gaubitch, X. Lin, and P. A. Naylor, "Scale factor ambiguity correction for subband blind multichannel identification," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Seattle, Washington, USA, Sep. 2008.

[109] B. Castro, N. D. Gaubitch, E. A. P. Habets, S. Gannot, P. A. Naylor, and S. Grant, "Subband scale factor ambiguity correction using multiple filterbanks," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, 2010.

[110] A. J. Laub, *Matrix analysis for scientists and engineers*. Philadelphia, PA, USA: SIAM, 2004.

[111] M. Haque and M. Hasan, "Noise robust multichannel frequency-domain LMS algorithms for blind channel identification," *IEEE Signal Process. Lett.*, vol. 15, pp. 305–308, 2008.

[112] X. S. Lin, A. W. H. Khong, and P. A. Naylor, "A forced spectral diversity algorithm for speech dereverberation in the presence of near-common zeros," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 888–899, Mar. 2012.

[113] M. K. Hasan and P. A. Naylor, "Analyzing effect of noise on LMS-type approaches to blind estimation of SIMO channels: Robustness issue," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[114] J. Benesty and Y. Huang, Eds., *Adaptive signal processing: Application to real-world problems*, ser. Signals and Communication Technology. Berlin, Germany: Springer, 2003.

[115] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[116] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[117] M. Taseska and E. A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[118] Y. Katznelson, *An introduction to harmonic analysis*, 3rd ed. Cambridge University Press, 2004.

[119] C. D. Meyer, *Matrix analysis and applied linear algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2000.

[120] E. A. P. Habets. (2003). Room impulse response generator for MATLAB, [Online]. Available: `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`.

[121] P. Stoica and R. Moses, *Spectral analysis of signals*. New Jersey, USA: Prentice Hall, 2005.

[122] R. Ahmad, A. W. H. Khong, M. K. Hasan, and P. A. Naylor, "An extended normalized multichannel FLMS algorithm for blind channel identification," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[123] H. He, J. Lu, J. Chen, X. Qiu, and J. Benesty, "Robust blind identification of room acoustic channels in symmetric alpha-stable distributed noise environments," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 693–704, 2014.

[124] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[125] J. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, Dec. 1906.

[126] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, Mar. 2004.

[127] W. Briggs and V. Henson, *The DFT: An owner's manual for the discrete fourier transform*. SIAM, 1995. eprint: `http://epubs.siam.org/doi/pdf/10.1137/1.9781611971514`.

[128] P. J. Davis, *Circulant matrices*, ser. Pure and applied mathematics. Wiley, 1979.

[129] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Nat. Inst. of Standards and Tech. (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[130] M. Huckvale. (2004). SCRIBE - spoken corpus recordings in British English, [Online]. Available: `http://www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm`.

[131] J. O. Smith. (2010). Physical audio signal processing. accessed 11.2016, [Online]. Available: `https://ccrma.stanford.edu/~jos/pasp/Early_Reflections.html`.

[132] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," vol. 19, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds., pp. 801–808, 2007.

[133] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[134] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium (LDC), Philadelphia, Corpus LDC93S1, 1993.

[135] X. Anguera Miró, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[136] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. McLaughlin, and M. A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics.," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, 1998.

[137] D. A. van Leeuwen, "The TNO speaker diarization system for NIST RT05s meeting data," in *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, ch. 8, pp. 440–449.

[138] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J. F. Bonastre, "Nist rt'05s evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings," in *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, ch. 8, pp. 428–439.

[139] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, 2007.

[140] C. Wooters and M. Huijbregts, "The icsi rt07s speaker diarization system," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 509–519.

[141] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, Mar. 1996.

[142] A. Stéphene and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, Jun. 1997.

[143] N. W. D. Evans, C. Fredouille, and J. F. Bonastre, "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 4061–4064.

[144] J. Yang, D. Zhang, Z. Jin, and J. yu Yang, "Unsupervised discriminant projection analysis for feature extr," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, 2006, pp. 904–907.

[145] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1793–1805, Sep. 2010.

[146] D. M. Brookes, *VOICEBOX: A speech processing toolbox for MATLAB*, `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997–2016.

[147] *Objective measurement of active speech level*, Recommendation, Mar. 1993.

[148] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[149] F. Lim and P. Naylor, "Statistical modelling of multichannel blind system identification errors," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 119–123.

[150] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[151] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[152] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, Dec. 1966.

[153] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[154] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of Cal. Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.

[155] K. Fukunaga, "Chapter 3 - hypothesis testing," in *Introduction to Statistical Pattern Recognition (Second Edition)*, K. Fukunaga, Ed., Second Edition, Boston: Academic Press, 1990, pp. 51 –123.

[156] NIST. (2006). Spring 2006 (rt-06s) rich transcription meeting recognition evaluation plan, [Online]. Available: `http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/docs/rt06s-meeting-eval-plan-V2.pdf`.

[157] D. Vijayasenan and F. Valente, "DiarTk : An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2012.

[158] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[159] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. Neural Information Processing Conf*, S. A. Solla, T. K. Leen, and K. Müller, Eds., MIT Press, 2000, pp. 617–623.

[160] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic combination of MFCC and TDOA features for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 431–438, Feb. 2011.

[161] ——, "An information theoretic approach to speaker diarization of meeting data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1382–1393, Sep. 2009.

[162] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. v. Waterschoot, and P. A. Naylor, "A single-channel non-intrusive C50 estimator correlated with speech recognition performance," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 719–732, Apr. 2016.

[163] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. 2015.

[164] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[165] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The Reverb Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2013, pp. 1–4.

[166] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcam0: A british english speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1995, pp. 81–84.

[167] P. P. Parada, D. Sharma, and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4718–4722.

[168] P. A. Naylor and N. D. Gaubitch, Eds., *Speech dereverberation*. Springer, 2010.