Spatio-Temporal Analysis of Spontaneous Speech with Microphone Arrays

THÈSE N^o 3689 (2006)

PRÉSENTÉE LE 30 NOVEMBRE 2006

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Laboratoire de l'IDIAP

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Guillaume Lathoud

Ingénieur Diplômé de l'Institut National des Télécommunications, Evry, France

et de nationalité française

acceptée sur proposition du jury:

Prof. J. R. Mosig, président du jury
Prof. H. Bourlard, **Dr J.-M. Odobez**, directeurs de thèse
Dr. C. Faller, rapporteur
Prof. R. Martin, rapporteur
Prof. S. Renals, rapporteur

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Lausanne, EPFL
2006

To Claude Bernard and Alfred Korzybski,

for their inspiring works.

Abstract

Accurate detection, localization and tracking of multiple moving speakers permits a wide spectrum of applications. Techniques are required that are versatile, robust to environmental variations, and not constraining for non-technical end-users. Based on distant recording of spontaneous multi-party conversations, this thesis focuses on the use of microphone arrays to address the question "Who spoke where and when?". The speed, the versatility and the robustness of the proposed techniques are tested on a variety of real indoor recordings, including multiple moving speakers as well as seated speakers in meetings. Optimized implementations are provided in most cases.

We propose to discretize the physical space into a few sectors, and for each time frame, to determine which sectors contain active acoustic sources ("Where? When?"). A topological interpretation of beamforming is proposed, which permits both to evaluate the average acoustic energy in a sector for a negligible cost, and to locate precisely a speaker within an active sector. One additional contribution that goes beyond the field of microphone arrays is a generic, automatic threshold selection method, which does not require any training data. On the speaker detection task, the new approach is dramatically superior to the more classical approach where a threshold is set on training data. We use the new approach into an integrated system for multispeaker detection-localization.

Another generic contribution is a principled, threshold-free, framework for short-term clustering of multispeaker location estimates, which also permits to detect where and when multiple trajectories intersect. On multi-party meeting recordings, using distant microphones only, short-term clustering yields a speaker segmentation performance similar to that of close-talking microphones.

The resulting short speech segments are then grouped into speaker clusters ("Who?"), through an extension of the Bayesian Information Criterion to merge multiple modalities. On meeting recordings, the speaker clustering performance is significantly improved by merging the classical mel-cepstrum information with the short-term speaker location information.

Finally, a close analysis of the speaker clustering results suggests that future research should investigate the effect of human acoustic radiation characteristics on the overall transmission channel, when a speaker is a few meters away from a microphone.

Keywords: Microphone arrays; speaker localization, tracking, segmentation, and clustering; spontaneous multi-party speech processing.

Version abrégée

La détection, la localisation et le suivi dans l'espace de plusieurs locuteurs permet un large spectre d'applications. Les solutions techniques doivent être génériques, robustes aux variations environnementales et non-contraignantes pour les utilisateurs. Cette thèse propose d'utiliser des enregistrements distants de conversations spontanées pour répondre à la question "Qui parle, où et quand ?". La vitesse, la généricité et la robustesse des solutions proposées sont évaluées sur des enregistrements variés, incluant plusieurs locuteurs en déplacement, ou bien plusieurs locuteurs assis dans une réunion. Des implémentations optimisées sont proposées.

Nous proposons de discrétiser l'espace physique en quelques secteurs, et, pour chaque trame temporelle, de déterminer quels secteurs contiennent des sources acoustiques actives ("Quand ? Où ?"). Nous proposons une interprétation topologique du "beamforming", qui permet à la fois d'évaluer l'énergie acoustique moyenne dans un secteur, et de localiser précisément un locuteur dans un secteur actif. Une de nos contributions va au-delà du contexte des antennes de microphones. Il s'agit d'une méthode générale pour la sélection automatique d'un seuil, sans données d'entraînement. Nous utilisons cette approche dans un système intégré de détection-localisation.

Une autre contribution générique est une méthode sans seuil pour le groupage court-terme des positions spatiales de plusieurs locuteurs. Le groupage court-terme permet aussi de détecter où et quand des trajectoires se coupent. Sur des enregistrements de réunions, avec seulement des microphones distants, le groupage court-terme permet une segmentation ayant une performance similaire à celle obtenue avec des microphones placés près de la bouche de chaque locuteur.

Les segments résultants sont ensuite eux-mêmes groupés, pour former idéalement un groupe par personne ("Qui ?"), en étendant le Critère d'Information Bayésienne à des modalités multiples. Sur des enregistrements de réunions, la performance du groupage est améliorée de façon significative en fusionnant l'information mel-cepstrale classique avec l'information court-terme donnée par la position spatiale de chaque locuteur. Une analyse détaillée des résultats du groupage suggère, comme direction pour des recherches futures, d'étudier l'effet de la radiation acoustique humaine sur le canal global de transmission, lorsque le locuteur est à plusieurs mètres d'un microphone.

Mots-clés : Antennes de microphones ; localisation, suivi, segmentation, et groupage de locuteurs ; traitement de la parole spontanée de plusieurs locuteurs.

Contents

Acknowledgment **xxiii**

1 Introduction **1**

1.1 Objective and Motivation 1

1.2 Structure of the Thesis and Contributions 4

1.2.1 AV16.3 Corpus 5

1.2.2 Joint Detection-Localization of Multiple Audio Sources 5

1.2.3 Short-Term Clustering of Instantaneous Location Estimates 7

1.2.4 Speaker Clustering with Distant Microphones 7

1.2.5 Applications to Other Domains 8

1.2.6 Other Contributions 9

2 Notation and Definitions **11**

2.1 Mathematics 11

2.2 Discrete Time Processing of Quasi-Stationary Signals 14

2.3 Multichannel Signals 17

2.4 Probabilities and Random Variables 17

2.5 Glossary of Notations 20

2.6 Abbreviations 24

3 Background **27**

3.1 Speaker Localization 28

3.1.1 Acoustic Waves 28

3.1.2 Microphone Arrays for Localization 30
3.1.3 Detection for Localization 37
3.1.4 Tracking 39
3.2 Multi-Party Speech Segmentation 40
3.2.1 The Task 40
3.2.2 Location for Segmentation 43
3.2.3 A Preliminary Experiment 44

4 The AV16.3 Corpus **51**
4.1 Physical Setup and Camera Calibration 53
4.1.1 Hardware 54
4.1.2 Step One: Camera Placement 54
4.1.3 Step Two: Camera Calibration 56
4.2 Online Corpus 58
4.2.1 Motivations 58
4.2.2 Sequence Names 59
4.2.3 Annotated Contents 60
4.3 Annotation 61
4.3.1 Spatial Annotation Interfaces 61
4.3.2 3-D Mouth Annotation 62
4.3.3 Available Annotation 63
4.3.4 Example 1: Audio Source Localization Evaluation 63
4.3.5 Example 2: Multi-Object Video Tracking 64
4.4 Additional Loudspeaker Sequences 64
4.5 Conclusion 66
4.6 Acknowledgment 66

5 Multisource Joint Detection-Localization **67**
5.1 A Phase Domain Metric Interpretation of SRP 69
5.1.1 Motivation 69
5.1.2 The Proposed Phase Domain Metric (PDM) 70

5.1.3 Equivalence with SRP-PHAT . . . 71
5.2 Sector-Based Activeness . . . 72
5.2.1 Averaging the PDM over a Sector . . . 73
5.2.2 Comparing Sectors: the Sparsity Assumption . . . 75
5.2.3 Sector-Based Activeness: SAM-SPARSE-MEAN . . . 76
5.2.4 Experiments . . . 77
5.3 Threshold Selection for Sector-Based Detection-Localization . . . 79
5.3.1 “Training”: Threshold Selection with Training Data . . . 82
5.3.2 Threshold Selection without Training Data . . . 82
5.3.3 Experiments . . . 86
5.3.4 Openings . . . 88
5.4 Point-Based Localization . . . 91
5.4.1 Proposed Multisource Detection-Localization . . . 92
5.4.2 The Cost Function and its Gradient in Euclidean coordinates . . . 94
5.4.3 Optimization of the Computational Complexity . . . 96
5.4.4 Multiple Microphone Arrays . . . 97
5.4.5 “FULL”, “FAST” and “FASTTDE” Implementations . . . 98
5.4.6 Evaluation Method . . . 100
5.4.7 Experimental Protocol . . . 103
5.4.8 Results and Discussion . . . 104
5.5 Speech/Non-Speech (SNS) Classification . . . 108
5.5.1 Sector-Based MFCCs . . . 108
5.5.2 Low-Cost Speech/Non-Speech Classifier (SNSLOW) . . . 109
5.5.3 Full Covariance GMM Speech/Non-Speech Classifier (SNSGMM) . . . 109
5.6 Conclusion . . . 111

6 Short-Term Spatio-Temporal Clustering **113**
6.1 Short-Term Spatio-Temporal Clustering . . . 115
6.1.1 Assumptions on Local Dynamics . . . 116
6.1.2 Short-Term Clustering (STC) . . . 118

6.1.3 Threshold-Free Maximum Likelihood Clustering 119
6.2 Optimization Algorithms 120
6.2.1 Online: Sliding Window (SW) 120
6.2.2 Offline: Simulated Annealing Optimization (SA) 122
6.3 Application: Threshold-Free Detection of Trajectory Crossings 124
6.3.1 Threshold-Free Confident Clustering 125
6.3.2 Multi-Source Tracking Examples 127
6.4 Application to Detection-Localization of Multiple Speakers 128
6.4.1 Instantaneous Multisource Detection-Localization 128
6.4.2 Speech/Non-Speech (SNS) Classification 129
6.4.3 Experimental Protocol 130
6.4.4 Results and Discussion 131
6.5 Meeting Segmentation Application 132
6.5.1 Test Data 133
6.5.2 Proposed Systems 134
6.5.3 Baseline System using Lapels 134
6.5.4 Performance Measures 135
6.5.5 Results and Discussion 136
6.6 Conclusion 139

7 Speaker Clustering with Distant Microphones 141
7.1 Speaker Clustering with Audio Modalities 143
7.1.1 The Bayesian Information Criterion (BIC) for Speaker Clustering 144
7.1.2 Combining Two Modalities: Location Cues and Acoustic Cues 146
7.1.3 Experimental Results 149
7.1.4 Future Directions 152
7.2 Automatic Audio-Visual Calibration 154
7.2.1 Calibration between Discrete Spaces 154
7.2.2 Calibration without Discretization 157

7.3 Conclusion 161

8 Applications to Other Domains **163**

8.1 Sector-Based Detection for Hands-Free Speech Enhancement in Cars 164

8.1.1 Physical Setups, Recordings and Sector Definition 166

8.1.2 Input SIR Estimation 168

8.1.3 Speech Enhancement 172

8.2 Noise-Robust ASR: Unsupervised Spectral Subtraction 178

8.2.1 Proposed 2-Component Mixture Model 179

8.2.2 Application to Unsupervised Spectral Subtraction (USS) 181

8.2.3 Noise-Robust ASR Experiments 182

8.3 Conclusion 184

9 Conclusion **185**

9.1 Data Acquisition 186

9.2 Multispeaker Detection-Localization 186

9.3 Short-Term Clustering 188

9.4 Speaker Clustering with Distant Microphones 189

9.5 Self-Criticism and Future Directions 190

A Performance Metrics for Detection **193**

B Multidimensional Phase Domain Metrics **195**

B.1 Definition of a PDM 195

B.2 Property 196

B.3 Equivalence Between SRP-PHAT and Δ 196

C Sector-Based Activeness Models and their EM Derivations **199**

C.1 1-dimensional Model 200

C.1.1 Description 201

C.1.2 EM Derivation 203

C.2 Multidimensional Model 208

C.2.1 Description 208
C.2.2 EM Derivation 212

D Comparison of Detection Features for Localization 217

E Some Analytical Formulas for Single Gaussians 221

F Proof of the Rayleigh-Distributed Magnitude Spectrum 223

Curriculum Vitae 237

List of Figures

1.1 (a) Uniform Circular Array (UCA) of microphones. (b) Objective of the thesis: to determine where and when each speaker is talking (dotted lines and brackets). Each estimated azimuth angle is depicted by an arrow in (a) and a dot in (b). 2

1.2 Proposed approach (a,b,c,d,e) and main body of the thesis. 4

2.1 Example of microphone arrays: dots depict microphones, lines depict pairs. (a) Uniform Linear Array with $N_m = 4$ microphones and $N_q = 6$ pairs, (b) Uniform Circular Array with $N_m = 8$ microphones and $N_q = 28$ pairs. 16

3.1 A spherical acoustic wave emitted by a point source at location $\boldsymbol{\ell}^{\mathrm{ps}}$, and recorded by a microphone at location $\boldsymbol{\ell}_m$. 28

3.2 Time Delay Of Arrival (TDOA) τ between two microphones at $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$, of a spherical acoustic wave emitted at $\boldsymbol{\ell}^{\mathrm{ps}}$. 32

3.3 SRP-PHAT point-based search on `seq01` (single human speaker): (a) shows the histogram of azimuth errors; (b) shows a zoom of (a); (c) shows the histogram of log energy values. 37

3.4 Example of segmentation result (top) for a 4-people meeting (bottom). The top figure depicts a speech/silence segmentation for each speaker, with sometimes multiple speakers active at the same time (overlaps). 41

3.5 Single speaker HMM topology. 46

3.6 Dual-speaker HMM topology. 46

3.7 Experimental setup. 47

4.1 Physical setup: three cameras C1, C2 and C3 and two 8-microphone circular arrays MA1 and MA2. The gray, L-shaped area is in the field of view of all three cameras. . . 53

4.2 Snapshots from the cameras at their final positions. "+" designate points in the calibration training set $\mathbb{X}_{\text{train}}$, "x" designate points in the calibration test set $\mathbb{X}_{\text{test}}$. 55

4.3 Snapshots of the two windows of the Head Annotation Interface. 62

4.4 Snapshots from visual tracking on 200 frames of "`seq45-3p-1111`" (initial timecode: 00:00:41.17). Tracking results are shown every 25 frames. 65

4.5 Top view of the recording setup for `loud01-3p-0001`, `loud02-3p-0001` and `loud03-3p-0001`: 3 loudspeakers A,B,C. Loudspeaker A lies at 90^o azimuth relative to the array in `loud01-3p-0001` (radius 0.8 m) and `loud02-3p-0001` (radius 1.8 m), and at 0^o azimuth in `loud03-3p-0001` (radius 1.45 m). Loudspeakers B and C lie respectively at $+25.6^o$ and -25.6^o in all three sequences `loud01-3p-0001`, `loud02-3p-0001` and `loud03-3p-0001` (radius 0.8 m). 66

5.1 Proposed multisource detection-localization. The eight dots in the center represent the microphone array. The three dots in the sectors represent point location estimates. 68

5.2 Illustration of the triangular inequality for the PDM in dimension 1: each point on the unit circle corresponds to an angle value modulo 2π. From the Euclidean metric: $\left|e^{ju_3} - e^{ju_1}\right| \leq \left|e^{ju_3} - e^{ju_2}\right| + \left|e^{ju_2} - e^{ju_1}\right|$. 71

5.3 Two examples of microphone arrays and sector definitions. Each dot corresponds to a $\mathbf{v}_{\check{s},n}$ location. In (a) the sectors are defined in 3-D, following (5.9), but for the sake of clarity, we have only represented the horizontal plane. In (b) the sectors are defined in 2-D. 73

5.4 Example of sector-based activeness pattern (part of `seq01-1p-0000`). For each sector $\mathbb{S}_{\check{s}}$ and each time frame t, a sector-based activeness value $\zeta_{\check{s},t} \geq 0$ is represented, with larger values in white. 76

5.5 Examples of logarithmic ROC curves for the sector-based activeness $\zeta_{\check{s},t}$. 78

5.6 Sector-based detection-localization (20-degree sectors): multichannel waveforms from a microphone array (dots in (a)) are transformed into "activeness" values (a,b), as in (5.22), which are thresholded to obtain the final decision (c). A false alarm happens when the ground-truth is $B_{\check{s},t} = 0$ and the final decision is $\hat{B}_{\check{s},t} = 1$. 79

5.7 ROC curve. The task is to select a threshold Ψ_ζ such that the obtained FAR (triangle) is as close as possible to the target $\mathrm{FAR_T}$ (dot). Ideally $\mathrm{FAR} = \mathrm{FAR_T}$. 80

5.8 (a) Unsupervised fit of a 2-mixture model $\mathcal{M}$ with parameters $\mathbf{\Lambda}(\mathcal{M}) = \{w_0, w_1, f_0, f_1\}$. The histogram in (a) is a 1-D view of all data $\{\zeta_{\check{s},t}\}$, irrespective of sector $\mathbb{S}_{\check{s}}$ or time t. w_0 and w_1 are the priors of inactivity and activity, respectively. (b) "Model-only" threshold selection, using the model $\mathcal{M}$ to match the target $\mathrm{FAR_T}$. 83

5.9 Graphical model for the independence assumptions used in the multidimensional model. The r.v. $\underline{A}$ is the frame state (inactive or active) and the r.v. $\underline{B}_{\check{s}}$ is the state (inactive or active) of a given sector $\mathbb{S}_{\check{s}}$. The r.v. $\underline{\zeta}_{\check{s}} \geq 0$ is the activeness of sector $\mathbb{S}_{\check{s}}$. On an active frame ($\underline{A}$= 1) at least one sector is active ($\exists \check{s} \quad \underline{B}_{\check{s}} = 1$). 85

5.10 Example of fit of the three main distributions used to define the multidimensional model. 85

5.11 FAR curves: comparison between target $\mathrm{FAR_T}$ & obtained FAR. In seq37, the positive bias is due to body noises (breathing, stomps, shuffling paper) marked as "inactivity" in the ground-truth, since their locations are unknown. 87

5.12 Logarithmic ROC curves on the loudspeaker recordings, for the 1-D approaches ("training", "model only", "model+data"), and for the multidimensional approach ("model+data (N-D)"). 88

5.13 Threshold selection with and without training data, applied to loudspeaker recordings (a,b,c) and human recordings (d,e): comparison between desired target and measured False Rejection Rate. Note that all $\mathrm{FRR_T}$ values are shown (from 0 to 1). (d) and (e) illustrate the ground-truthing issue with human data. 89

5.14 Example of frequency selection (dots): we select only the discrete frequencies with magnitude above the geometric mean (horizontal dashed line), and at or next to a magnitude peak. 96

5.15 Proposed 2-step approach, with two microphone arrays. 98

5.16 Example of fit of the Gaussian + Uniform model $(\mathcal{N}+\mathcal{U})$ on the localization error θ^{ERR}. $(\mathcal{N}+\mathcal{U})$ is used for evaluation of the detection-localization. `seq11` is a recording with a single moving speaker, from the AV16.3 Corpus (Chapter 4). The gray histogram represents the distribution of localization errors. The dark curve represents the Gaussian pdf, modelling "correct" location estimates. The uniform pdf is not represented. . 101

5.17 Histogram of $\hat{n}_{\mathrm{C}}(t)$, the estimated number of correctly localized speakers in the 2-speaker sequence `seq18`. Note that $\hat{n}_{\mathrm{C}}(t) \in \mathbb{R}$ can take non-integer values. 102

5.18 Result (red dots) of the detection-localization ("FAST" implementation, followed by short-term clustering and SNSLOW). The ground-truth (black curves) is derived from the cameras, including on silences. Gaps are due to the mouth of a person being occluded on at least one camera (*gaps are not related to silences*). 106

6.1 The goal (a) and the proposed approach (b)(c). Dots depicts instantaneous location estimates $r_i \stackrel{\text{def}}{=} (\theta_i, T_i)$. Dashed lines depict trajectories of the sources (true in (a), estimated in (c)). Square brackets depict the beginning and the end of each speech utterance. Round, continuous lines depict the short-term clusters $\omega_1, \cdots, \omega_{10}$. 114

6.2 Histogram of azimuth angle variations $\theta_i - \theta_j$ over a 2-frame delay ($|T_i - T_j| = 2$), on real data (recording `seq01` from the AV16.3 Corpus, see Chapter 4). The superimposed curves depict the bi-Gaussian mixture model obtained through EM training. 117

6.3 The two types of clusters. This chapter focuses on short-term clusters (a), obtained with location cues only. Long-term clustering (b) requires additional cues, as investigated in Chapter 7. 118

6.4 This two-cluster partition $\Omega = \{\omega_1, \omega_2\}$ of the set of location estimates $r_{1:5}$ (dots) is equivalently defined by six local decisions $H_0(i,j)$ (dotted lines) and four local decisions $H_1(i,j)$ (dashed lines). In this particular case, all location estimates (dots) are within T_{short} time frames of each other. 120

6.5 Example of low confidence decision $H_0(i,j)$ at a trajectory crossing. Each dot is a location estimate. A continuous line depicts each short-term cluster ω_n. 125

6.6 Comparison ML clustering / confident clustering on multiple source cases, where the number of active sources varies over time. Gray dots: location estimates $r_i = (\theta_i, T_i)$. Black lines: clusters ω_k. The ML clustering algorithm takes arbitrary decisions at trajectory crossings. On the contrary, the confident clustering correctly splits the short-term clusters at each trajectory crossing. 127

6.7 2-step implementation for multisource detection-localization (Chapter 5). The eight dots indicate the locations of the microphones. (a) sector-based detection-localization. (b) gradient descent within each active sector. 129

6.8 Detection-localization of multiple speakers, using microphone arrays (systems SW-1, SW-7 and SA). 129

6.9 Recording `seq45` from the AV16.3 Corpus (Chapter 4), with three moving speakers. The 8-microphone array is marked with an ellipse. The ball markers on the heads were used to construct the ground-truth location of each speaker with respect to the array. 130

6.10 Comparison between audio location estimates (dots) and the ground-truth location(s) obtained with three cameras (black lines, when all three cameras are available: *gaps in the ground-truth do not correspond to silences*), on the AV16.3 Corpus (Chapter 4). (a) Raw azimuth estimates for a single moving speaker, result of the multisource detection-localization, (b) After STC and removal of non-speech clusters. (c) Example with three simultaneous speakers. 131

6.11 Histogram of speech segment durations in the ground-truth (M4 Corpus (McCowan et al., 2005)). 134

6.12 Simulated Annealing (SA): Comparison between different initialization methods (Sections 6.2.2 & 6.5.5). In (a) and (b), each point represents a result for one meeting and one initialization method (SA(1), SA(N_r) or SA(SW-1)). For each meeting, all results are normalized through subtraction with respect to the reference SW-1 (Section 6.5.5). In (c), η_0 is the initial temperature. In the case $\eta_0 = 0$, only the ICM optimization is used, without simulated annealing. 137

7.1 The goal (a) and the proposed approach (b)(c). Dots depicts instantaneous location estimates $r_i \stackrel{\text{def}}{=} (\theta_i, T_i)$. Dashed lines depict trajectories of the sources (true in (a), estimated in (b) and (c)). Square brackets depict the beginning and the end of each speech utterance. Round, continuous lines depict the short-term clusters $\omega_1, \omega_2, \omega_3$. . . 142

7.2 Example of speaker clustering: a long-term partition Ω. In this example the data samples $\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}}$ are splitted into $N_\Omega = 4$ *long-term* speaker clusters $\omega_1, \omega_2, \omega_3$ and ω_4. . . 145

7.3 Unsupervised AV calibration in discrete space. The covariance is calculated between each video motion indicator (block of pixels) and each audio activity indicator (sector of space around the microphone array). Numbers represent indices of video blocks and audio sectors. 156

7.4 Unsupervised AV calibration in discrete space. Top row: snapshot of the `seq11` recording (microphone array indicated by a black ellipse). Middle row: AV covariance between the audio activity of sector $\mathbb{S}_{18}$ and the video motion of each camera. Bottom row: global result of the AV covariance analysis. For each block of pixels, the index (1 to 18) of the audio sector with the highest covariance is represented by a gray level (colorbar on the right side). The black background color ("none") appears whenever all audio sectors have a covariance inferior to e^{-6}. 156

7.5 Unsupervised AV calibration without discretization, camera #1. (a)(b)(c) For each pixel, the means μ_1, μ_2 and μ_3 (in degrees) of the three components of the A-GMM. For each pixel, speech components appear first in (a), then possibly in (b), then possibly in (c), followed by non-speech components in the remaining pictures. For each pixel, the number B of speech components of the A-GMM is shown in Figure 7.6a. 159

7.6 Unsupervised AV calibration without discretization, cameras #1, #2, #3. Each pixel depicts the number B of speech components in the corresponding A-GMM. 159

7.7 Unsupervised AV calibration without discretization, camera #3. Example of binary decision on frame #189. Black: background, white: foreground. (a) Video-only adaptive background subtraction (Stauffer and Grimson, 2000). (b) Same, where the weight adaptation is restricted, based on the A-GMM. 160

8.1 Entire acquisition process, from the emitted signals to the enhanced signal (Section 8.1). The focus is on the adaptive filtering block $h(t)$, so that $\mathrm{SIR}_{\mathrm{imp}}(t)$ is maximized when the interference is active (interference cancellation). The s and i subscripts designate contributions of target and interference, respectively. The whole process is supposed to be linear. $\sigma^2[x(t)]$ is the variance or energy of a speech signal $x(t)$, estimated on a short time frame (20 or 30 ms) around t, on which stationarity and ergodicity are assumed. 164

8.2 Proposed explicit and implicit adaptation control. $\mathbf{x}(t) = [x_1(t) \cdots x_{N_m}(t)]^{\mathrm{T}}$ are the signals captured by the N_m microphones, and $\mathbf{h}(t) = [\mathbf{h}_1(t) \cdots \mathbf{h}_{N_m}(t)]^{\mathrm{T}}$ are their associated filters. Double arrows denote multiple signals. 165

8.3 Physical setups I (2 mics) and II (4 mics). 166

8.4 Sector definition. Each dot corresponds to a $\mathbf{v}_{\check{s},n}$ location, as defined in Section 5.2.1. . 167

8.5 Estimation of the input SIR for setups I (left column) and II (right column). Beginning of recordings `train` (top row), `test` (middle row), `test+noise` (bottom row). 170

8.6 Linear models for the acoustic channels and the adaptive filtering. 172

8.7 Improvement over input SIR (100 ms moving average, first 3 seconds shown). Column (a) shows results on clean data (`test`), whereas column (b) shows results on noisy data (`test+noise`: 100km/h background road noise). 177

8.8 Model of the problem: recognize speech from the observed signal $x(t) = s(t) + n(t)$, where $s(t)$ is the clean speech signal and $n(t)$ is the additive acoustic noise signal. . . 178

8.9 Observations on real meeting room data (`seq01` in the AV16.3 Corpus, Chapter 4) of a pre-emphasized waveform $y(t) \stackrel{\mathrm{def}}{=} x(t) - 0.97 \cdot x(t-1)$. (a),(c): histograms, (b),(d): phase plots. 179

8.10 Example of fit of the 2-mixture model on noisy data taken from the OGI Numbers 95 database (Factory 0dB condition). All plots show magnitude data in the frequency domain. On spectrogram plots (a) and (d), the largest magnitudes are white, the smallest magnitudes are black. $f = \frac{k-1}{N_{\mathrm{F}}} \cdot \frac{f_s}{2}$ and $f_s = 8\,kHz$. 182

C.1 Graphical model for the 1-dimensional model. The r.v. $\underline{B} \in \{0, 1\}$ is the sector state (inactive or active). The r.v. $\underline{\zeta} \geq 0$ is the sector activeness. 201

C.2 Fit of the 2-component mixture model described in Section C.1: (a) automatic initialization, (b) final model $p(\underline{\zeta}=\zeta)$ after convergence of EM. 207

C.3 Graphical model for the independence assumption (C.46) used in the multidimensional model. The r.v. $\underline{A}$ is the frame state (inactive or active) and the r.v. $\underline{B}_{\check{s}}$ is the state (inactive or active) of a given sector $\mathbb{S}_{\check{s}}$. The r.v. $\underline{\zeta}_{\check{s}} \geq 0$ is the activeness of sector $\mathbb{S}_{\check{s}}$. On an active frame ($\underline{A}=1$) at least one sector is active ($\exists \check{s} \quad \underline{B}_{\check{s}}=1$). 210

C.4 Fit of the multidimensional model described in Section C.2: (a) automatic initialization, (b) final pdfs $\mathcal{G}_{00}, \mathcal{G}_{01}, \mathcal{R}_{11}$ and the mixture, after convergence of EM. 216

D.1 GCC-PHAT localization: Variation of the RMS azimuth error (in degrees) when the detection threshold is varied. "% of frames" is the proportion of active frames that are above a given value of the detection threshold. In (b), only frames with a localization error below 10 degrees are considered. 220

D.2 SRP-PHAT localization: Variation of the RMS azimuth error (in degrees) when the detection threshold is varied. "% of frames" is the proportion of active frames that are above a given value of the detection threshold. In (b), only frames with a localization error below 10 degrees are considered. 220

List of Tables

3.1 Average SNR across meetings and speakers of the M4 Corpus (McCowan et al., 2005), in dB domain. The lapels are worn by each speaker, below the throat. Each speaker is between 0.8 and 3 meters from the microphone array. The average SNR is defined in Section 2.4. 42

3.2 Results for the single speaker system (test set 1). FA stands for Frame Accuracy, PRC, RCL and F for Precision, Recall and F-measure, respectively (the higher, the better for all metrics). 48

3.3 Results for the extended system (test sets 1 and 2). The FA calculated only on actual overlap segments is shown in parentheses. 49

4.1 List of the annotated sequences. Tags mean: [A]udio, [V]ideo, predominant [ov]erlapped speech, at least one visual [occ]lusion, [S]tatic speakers, [D]ynamic speakers, [U]nconstrained motion. 60

4.2 Available annotation, as of December 7th, 2006. "C" means continuous annotation: on all frames of each 25 Hz video. "S" means sparse annotation: on some of the video frames (annotation rate in brackets). "Undersegmented" means that some short silences are included in the segments marked as "speech". 63

5.1 Average FRR for FAR in $[0, 0.1]$ (the lower, the better). Bold face indicates the best result in each column. The FRR values are larger in the case of humans (`seq01` & `seq37`), because of the many short silences between words and syllables that are marked as "speech" in the ground-truth segmentation (see the discussion in Section 5.3.4). On the other hand, the ground-truth segmentation is exact in the case of loudspeakers (`loud01`, `loud02` and `loud03`). 78

5.2 RMS error, as defined by (5.34), for $\mathrm{FAR_T} \in [0.001, 0.05]$. This is the RMS of $(\mathrm{FAR}/\mathrm{FAR_T} - 1)$: the lower, the better. The best result for each recording is indicated in boldface. 87

5.3 RMS error over the interval $\mathrm{FRR_T} = [0.001, 0.05]$. This is the RMS of $(\mathrm{FRR}/\mathrm{FRR_T} - 1)$: the lower, the better. The best result for each recording is indicated in boldface. The rightmost column shows the maximum over all 3 recordings. 90

5.4 Localization precision, in degrees, along with the percentage of correct location estimates. 107

5.5 Distribution of the number of correct simultaneous location estimates (percentage of frames). For recordings with multiple simultaneous speakers, the multispeaker cases are in bold face. 107

5.6 Effective computational complexity: computation duration divided by recording duration *(real time = 1)*. We used a Matlab/C implementation on a Pentium 4, with 3.2 GHz CPU speed and 1 GB of RAM. "SCG" is the time spent doing SCG descent only. "TDE" is the time spent doing TDE-based localization only. "Input" is the time spent reading and buffering wave files (variations due to Matlab). The cost of FFT and GCC-PHAT is very small (around 0.003 real time duration). 107

6.1 The online Sliding Window (SW) maximum likelihood algorithm. The likelihood of a partition is estimated with (6.6). Location estimates are ordered by increasing times ($\forall n \quad T_n \leq T_{n+1}$). 121

6.2 SW algorithm: Number of possible partitions, for each possible number of elements (Step 2 in Table 6.1). 121

6.3 SA algorithm: The MRF optimization (in practice $\lambda = 0.97$). $SA(E, \eta)$ is described in Table 6.4. 123

6.4 SA algorithm: One simulated annealing step $SA(E, \eta)$. 123

6.5 Comparison between the two types of SNS decision, on the AV16.3 Corpus (Chapter 4), including real recordings with multiple moving speakers, simultaneously speaking. Bias and standard deviation (std) are expressed in degrees. 131

6.6 Segmentation results on the M4 Corpus. SW-1 and SW-7 use distant microphones only. Values are percentages, results on overlaps only are indicated in brackets. PRC, RCL, F: the higher the better. DER: the lower, the better. 138

6.7 Comparison with a previous speaker clustering work: segmentation results on 6 meetings, with a silence minimum duration of 2 seconds. Values are percentages: the lower, the better. 139

6.8 F-measure on the M4 Corpus with SW-1, for two types of speech/non-speech decisions. The segmentation post-processing is detailed in Section 6.5.4. 139

7.1 Speaker clustering results on 18 meetings of the M4 Corpus (McCowan et al., 2005) (the lower, the better). Brackets indicate results on overlapped speech only. "ds" stands for delay-sum beamforming. "GMM/HMM" is the speaker clustering algorithm described in (Ajmera and Wooters, 2003). 151

8.1 Results on `test` and `test+noise`. Methods and parameters were selected on `train`. The RMS error of the input SIR estimation was calculated in log domain (dB). *Percentages (the lower, the better)* indicate the ratio between the RMS error and the dynamic range of the true input SIR (max - min). *Values in brackets (the higher, the better)* indicate the correlation between the true and the estimated input SIR. 169

8.2 Average segmental SIR improvement in dB. In Setup I, the reference is the output x_1 of microphone $\boldsymbol{\ell}_1$. In Setup II, the reference is the output of the delay-sum W_0. (W_0 brings a SIR improvement over x_1 of 0.1, 1.6, 2.2 dB respectively in the "co-driver", "both" and "driver" cases.) 176

8.3 Word Error Rate results on Aurora 2 (the lower, the better), per SNR level, averaged on the three noisy test sets A, B and C. Training is done on clean signals. 183

A.1 The four types of results. TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. 193

Acknowledgment

I would like to thank Prof. Hervé Bourlard for giving me a chance to do the PhD, and for his great support of my research, as well as of its publication. Dr Iain McCowan helped me a lot to define the initial direction, and gave me very useful advice on writing. Dr Jean-Marc Odobez provided critical help with the audio-visual corpus, as well as many insightful comments throughout the PhD.

I am particularly grateful to my family and friends in France, who have given me moral support throughout those years. They made the PhD possible, by giving life to the word "home".

My IDIAP colleagues made the PhD journey particularly interesting, Mathew, Todd, Hemant and Bahvna, Jitendra Ajmera, Jithendra Vepa, Vivek, Ikbal, Hayian and Datong, Mohamed, Daniel, Samy, Alessandro, Fabio, Bertrand, David Barber, Guillermo, Christos, Viktoria and many others... I deeply thank Alexei Pozdnoukhov for helping me with vertical research in the Swiss mountains.

Sylvie Millius, Nadine Rousseau, Pierre Dalpont and Ed Gregg have all my gratitude for helping me so diligently and efficiently with all my (sometimes surprising) administrative requests.

I could not have spent so much time doing research, without the continuous, caring support of the system team: Frank Formaz, Norbert Crettol and Tristan Carron. Olivier Masson and Darren Moore gave me very critical hardware and software support for the microphone arrays. Maël Guillemot, Bastien Crettol and Vincent Spano provided very useful help to put my work on the web.

Last but not least, very very special thanks go to five persons who have patiently read, commented (and endured?) this thesis while I was writing it: my supervisors Prof. Hervé Bourlard and Dr Jean-Marc Odobez, as well as Dr Mathew Magimai.-Doss, Bertrand Mesot and Julien Bourgeois.

Guillaume Lathoud

September 2006

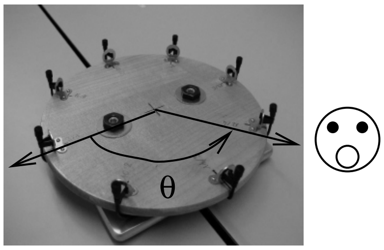
Chapter 1

Introduction

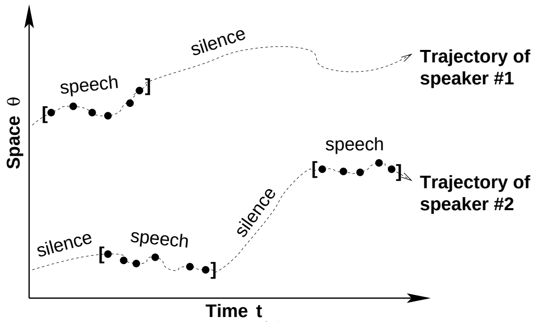
The present chapter first presents the objective and motivation of this thesis. The contributions are then detailed, following the structure of the thesis. For the sake of readability, notation, definitions and a detailed literature review are in separate chapters (2 and 3).

1.1 Objective and Motivation

The research presented in this thesis takes place in the context of "instrumented meeting rooms" and the automatic processing and analysis of multimodal, multi-party meeting recordings. This thesis investigates the analysis of spontaneous multi-party speech in a "non-invasive" manner. The goal is to estimate where and when the various speakers are talking. "Non-invasive" means distant microphones, for example a Uniform Cirular Array (UCA), as illustrated in Figure 1.1a. Comparison between the signals received at the various microphones of the array permits to evaluate the instantaneous locations of multiple acoustic sources (Krim and Viberg, 1996; Brandstein and Ward, 2001; Chen et al., 2006). For example, with a UCA, the instantaneous location of a given acoustic source, at a given instant t_i, is estimated in terms of azimuth angle θ_i, i.e. the source direction in the horizontal plane (round face in Figure 1.1a and dots in Figure 1.1b). Non-invasive methods can be opposed to very efficient but "invasive" methods that use close-talking microphones such as lapels (Wrigley et al., 2005), where one microphone is worn by each speaker, usually near the throat. Lapels permit to know precisely when each speaker is talking, since their signals are much



(a)



(b)

Figure 1.1. (a) Uniform Circular Array (UCA) of microphones. (b) Objective of the thesis: to determine where and when each speaker is talking (dotted lines and brackets). Each estimated azimuth angle is depicted by an arrow in (a) and a dot in (b).

cleaner and have higher energy than those received by distant microphones, due to the difference of distance. For example, in a series of meeting recordings a 10 dB difference of Signal-to-Noise-Ratio (SNR) is reported (see Section 2.4 and Table 3.1). However, the range of applications permitted by lapels is limited, because (1) they require each user to wear a lapel, and (2) they provide practically no information about the location of each speaker.

On the contrary, distant[1] microphones are "non-invasive", which we define as passive (non-emitting) *and* not attached to the human body. Thus, distant microphones put much less constraints on the users. Moreover, arrays of distant microphones permit to estimate speakers' locations based on geometrical considerations (Krim and Viberg, 1996; Brandstein and Ward, 2001; Chen et al., 2006). These two properties allow for a wide range of applications to spontaneous speech processing, including surveillance (Cerwin, 2004), intelligent homes, offices and meeting rooms (AAAI, 2006), hearing aids (Spriet, 2004), hands-free speech processing in cars (Lathoud et al., 2006a), as well as autonomous robots (Sony Corp., 2006). For example, a user browsing a meeting may be interested to jump directly to the presentation of a person, that is when that person stood up and moved to the screen. This would require to determine where and when each speaker is talking (Figure 1.1b). An important issue in spontaneous multi-party speech is "overlapped speech", that is when two or more participants talk at the same time (Shriberg et al., 2001). This, along with the presence of background noise sources, calls for a multisource detection-localization[2] system, that performs *joint* detection and localization – as opposed to first detect, then locate.

[1] In the experiments reported in this thesis, we considered distances from about 0.5 m to about 2.5 m.
[2] See (Korzybski, 1994) on the use of the dash.

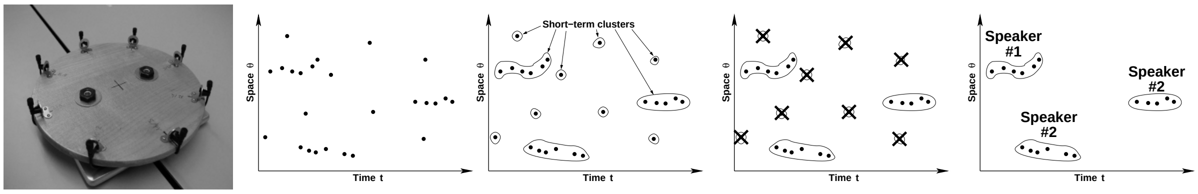
Therefore, the purpose of this thesis is to build and evaluate an integrated system for the detection and localization of multiple speakers with distant microphones. In other words, the aim is to determine who spoke where and when. The integrated system is designed to handle both static scenarios such as seated speakers in a meeting (McCowan et al., 2005), and dynamical scenarios such as multiple moving speakers (Lathoud et al., 2005c). The structure of the thesis is a progression from short-term, low-level analysis (where? when?) to longer-term, higher-level annotation (who?). At each stage, research issues are investigated, and techniques are proposed and tested on a variety of real meeting room recordings, including cases with multiple moving speakers as well as seated speakers in meetings.

The directions taken throughout this thesis correspond to three underlying aims:

- to put the least possible constraints on often non-technical end-users,
- to adapt to varying conditions in a robust manner (one or multiple speakers, clean conditions or background noise, etc.),
- to propose techniques that can be applied to a wider context than microphone arrays and/or meetings.

Our work thus focussed on methods that are non-invasive, and use little or no training data. We have also chosen *not* to address:

- Improvement of the precision of audio localization. Instead, we investigated whether *jointly* detecting and locating speakers could be beneficial, as opposed to first detect, then locate.
- Smooth trajectories in space (e.g., the results of particle filtering, Kalman filtering etc.) over long periods (several seconds or more). We argue that *spontaneous* multi-party speech is too sporadic for long-term tracking. For example a speaker may move while being silent (speaker #2 in Figure 1.1b). Tracking an explicit number of *speech* sources leads to difficult data association issues (Vermaak et al., 2003), often requiring complex birth/death rules. We thus investigated whether *short-term* analysis (on periods shorter than 250 ms) could be helpful for higher-level tasks such as speech/non-speech segmentation and speaker clustering with distant microphones.
- High performance systems that use lots of training data, but may not adapt to new conditions and/or may be difficult to use for non-technical users.



(a) Acquisition of synchronized signals	(b) Detection-local. of multiple audio sources	(c) Short-term clustering	(d) Speech/non-sp. decision	(e) Speaker clustering
Chapter 4 (AV16.3 Corpus)	Chapter 5 (where? when?)	Chapter 6 (speech? non-speech?)		Chapter 7 (who?)

Figure 1.2. Proposed approach (a,b,c,d,e) and main body of the thesis.

Instead, this thesis explicitly addresses the following three tasks:

1. Instantaneous detection-localization: from a single time frame (about 30 ms) of speech recorded with multiple microphones, the active audio sources are *jointly* detected and located. "Audio sources" include human speech and "non-speech" noise.

2. Speech segmentation with distant microphones: determine speech and silence time intervals, called "speech segments" and "silence segments". This implicitly requires to discriminate between speech sources and non-speech sources (machines, body noises). Also, overlaps between speakers (Shriberg et al., 2001) need to be properly detected.

3. Speaker clustering with distant microphones: for each speech segment, estimate who spoke. No enrollment data is available, so speaker "names" are tags #1, #2, etc. (see Figure 1.1b).

The next section briefly summarizes the structure of the thesis, then details each of the main contributions.

1.2 Structure of the Thesis and Contributions

Chapter 2 defines the notation and abbreviations. As mentioned above, this thesis does *not* investigate improvement on instantaneous audio source localization, but rather associated issues, such as joint detection-localization and its applications. Chapter 3 thus summarizes basics of microphone array-based instantaneous audio source localization, along with related work on detection. Existing works on meeting segmentation and speaker clustering are also summarized. The main body of the thesis is organized in a bottom-to-top approach, starting with the creation of the AV16.3 evaluation corpus (Chapter 4), then progressing from instantaneous, low-level analysis (where? when?) to long-term analysis (who?). This progression spans three chapters (5, 6 and 7), as depicted by Figure 1.2, and describes most of the research presented in this thesis. Finally Chapter 8 presents applications of the developed techniques to other domains: hands-free speech enhancement in cars, and noise-robust Automatic Speech Recognition. Chapter 9 concludes the thesis. Some lengthy definitions and laborious derivations lie in the Appendices, to keep the main body as light as possible. The main contributions of the thesis are detailed below, following the structure of the thesis.

1.2.1 AV16.3 Corpus

Chapter 4 presents an audio-visual corpus recorded with two uniform circular microphone arrays similar to the one depicted in Figure 1.2a, and three cameras. A variety of scenarios is included, with multiple moving speakers and overlapped speech. Most recordings were made with real human speakers rather than loudspeakers. This choice is justified by studies on the specificity of human speech radiation (Schwetz et al., 2004). The cameras are calibrated (Bouguet, 2004) and used to define the ground-truth 3-D mouth locations with an error less than 1.2 cm. To the best of our knowledge, this corpus was the first publicly available, annotated audio-visual corpus for speaker localization and tracking.

1.2.2 Joint Detection-Localization of Multiple Audio Sources

Chapter 5 focuses on the instantaneous, static analysis of a time frame of speech (typically about 20 to 30 ms), for the detection and localization of multiple audio sources (Figure 1.2b). The evaluation is conducted on the AV16.3 Corpus. For localization, we have chosen to use Steered Response Power methods, which consist in finding the 3-D location(s) in space that maximize a beamformed power (Krim and Viberg, 1996; DiBiase, 2000). Signal subspace methods such as MUSIC (Schmidt, 1986) are not used here, as they are known to be sensitive to modelling assumptions, which can lead to issues with speech in reverberant environments (DiBiase et al., 2001). In all of the following, "audio source" includes both human and machine sources (laptop, projector, etc.), while "speech source" includes only human speakers. The main contributions are detailed below:

- The Phase Domain Metric (PDM) is proposed. The idea is to interpret beamforming as a comparison between observed phase values, and theoretical phase values associated with a particular speaker location. The PDM is used as a principled framework for *both* detection and localization of multiple audio sources.

  First, the PDM is used in a sector-based *joint* detection-localization approach that drastically reduces the localization search space for a negligible cost, while being able to detect and locate multiple simultaneous speakers. The space around an array is discretized into sectors, and the relative phases between the microphones in the array are compared using the PDM, to determine whether there is audio activity in each sector. This sector-based approach is successfully applied in the meeting room domain, and in cars (Chapter 8). Optimized code is provided, combining Matlab and C.

  Second, within the active sectors, audio sources are precisely located through minimization of the PDM. Optimized code is provided, combining Matlab and C.

- A second contribution is unsupervised probabilistic modelling of the acoustic power in a sector. It consists in modelling background noise and large magnitudes of audio activity *jointly*. No training data is required, therefore the method is adaptive and robust to environmental variations. It is introduced and applied to sector-based detection-localization in Chapter 5. The same idea is also successfully applied to noise-robust ASR (Chapter 8).

- Based on such probabilistic modelling, a third contribution is the automatic selection of a detection threshold, which permits to use the probabilistic models in mismatched conditions. Microphone array experiments validate the approach on mismatched recording conditions. Theoretical investigations show that it can also be applied to multiclass classification tasks.

As a result of Chapter 5, for each time frame separately, zero, one or more audio source location estimates are produced (dots in Figure 1.2b).

1.2.3 Short-Term Clustering of Instantaneous Location Estimates

Chapter 6 proposes to analyze the instantaneous location estimates produced in Chapter 5, to detect speech utterances as segments in time *and* trajectories in space (brackets and dotted lines in Figure 1.2c). The highly changing, sporadic nature of spontaneous speech leads to analyze in the short-term only. Contributions include:

- A principled, unsupervised framework for short-term clustering of location estimates is proposed (Figure 1.2c). Speech/Non-Speech (SNS) decisions are then taken for each cluster (Figure 1.2d), and only "speech" clusters are kept. This approach is shown to be advantageous over individual frame-level SNS decisions, both in terms of final performance and robustness to post-processing.
- Application to meeting segmentation: Short-term clustering is used to segment real recordings of multi-party speech with distant microphones only, in terms of "speech" and "silence" for each speaker. The evaluation is conducted on the M4 Corpus (McCowan et al., 2005). The segmentation performance is comparable to that of close-talking microphones, with a dramatic improvement on overlapped speech.
- Application to multi-target tracking: experiments on synthetic data show that short-term clustering can be used to detect trajectory crossings *in a threshold-free manner*, which may be useful as a prior step to the recombination of pieces of trajectories, such as (Jorge et al., 2004).

To summarize, using short-term clustering, "non-speech" noise sources are rejected, and the beginning and end times of each "speech" utterance are detected: the beginning and the end of each short-term cluster in Figure 1.2d. At this point, we still do not know who spoke a given utterance, which is addressed below.

1.2.4 Speaker Clustering with Distant Microphones

Chapter 7 investigates the determination of the speaker identity of each speech utterance with distant microphones. "Distant" means at least 30-40 cm between mouth and microphone (up to 2.2 meters in our data). As presented in Section 1.1, no enrollment data is available. Thus, we investigate *agglomerative clustering*, where speech utterances from the same speaker are progressively grouped together into a single "long-term" cluster. Contributions include:

- A modification of the Bayesian Information Criterion (BIC) (Chen and Gopalakrishnan, 1998) to merge multiple modalities (location cues and cepstral cues such as MFCCs), resulting in an effective speaker clustering scheme, that uses distant microphones only. A speaker clustering performance is obtained that is superior to that of a state-of-the-art approach. A close analysis of the individual errors showed that further research should investigate the distance-dependent variabilities of the acoustic features (MFCCs).
- Initial investigations on unsupervised audio-visual calibration, as an alternative to audio-only distant speech processing. The goal is to increase the robustness of speaker clustering without asking any technical operation from the users. We investigate the discovery of geometrical links between non-colocated microphone array and camera, as well as the determination of multiple depths area in the image plane of a camera.

To summarize, Chapter 7 addresses the highest-level annotation considered in this thesis, that is to determine *who* spoke. A successful modification of the BIC criterion for speaker clustering is proposed. An analysis of the errors suggests future directions of work, using audio only, or combining audio and video.

1.2.5 Applications to Other Domains

Chapter 8 presents the respective applications of two concepts introduced above, to different tasks, with different hardware and outside the meeting room environment:

- The PDM is used for the detection of two-speaker speech with a linear microphone array, in a car. The application is the adaptation control of filters for the separation of driver and co-driver speech. This was a joint work conducted with Mr. Julien Bourgeois, while he was working at Daimler-Chrysler in Ulm, Germany. This collaboration was part of the HOARSE Research Training Network[3].
- The *joint* probabilistic modelling of speech and background noise, introduced above for detection-localization, was also used to determine the noise level in single-channel spectral subtraction. It was applied to noise-robust speech recognition on telephone channels. This was a joint

[3] `http://www.hoarsenet.org/`

work conducted with my IDIAP colleagues Dr. Mathew Magimai-Doss, Prof. Hervé Bourlard, Bertrand Mesot and Dr. Jithendra Vepa.

1.2.6 Other Contributions

The lapel baseline for speech segmentation (Section 6.5.3) was applied for fast pre-annotation of meetings. This was a joint work with Mr. Maël Guillemot and Ms. Joanne Moore from IDIAP, and Ms. Agnes Lisowska from the University of Geneva. Within the framework of the European AMI project, human annotators had to mark in the AMI meeting recordings the beginning and end times of each sentence, as well as the words spoken in between. According to meeting annotators, starting from an automatic time segmentation makes the work much easier than starting from scratch.

The annotation tools developed for the AV16.3 Corpus (Chapter 4) were also used in the AMI WP4 effort led by Dr Daniel Gatica-Perez from IDIAP.

The work done in the course of this thesis contributed to the Swiss project IM2, as well as the European projects HOARSE, M4 and AMI.

Most of the data and code developed in the course of this work are freely available at:
`http://mmm.idiap.ch/Lathoud/`

Chapter 2

Notation and Definitions

This chapter defines mathematical notation and abbreviations used throughout the thesis. As much as possible, we tried to keep the notations consistent across the chapters. There may be some overlap between the notations of the different chapters, but the context should make clear in which sense a notation is used. Section 2.5 contains a glossary of notations.

The mathematical tools used in this thesis are also briefly defined. The underlying foundations can be found in (Moon and Stirling, 2000), and more specifically in (Rabiner and Schafer, 1978; Oppenheim et al., 1999) for discrete-time signal processing of speech signals, and (Delmas, 1993; Weisstein, 2006b) for probabilities and random variables.

2.1 Mathematics

A list of mathematical notations is given below. To avoid confusion, we distinguish with an underline:

- Deterministic quantities, such as a time domain signal $x(t)$ and its Fourier transform $X(k)$.
- Random variables, for example $\underline{x}\,(t) \sim \mathcal{N}_{\mu,\sigma}$, a Gaussian random variable.

Such a use of the underline may appear unusual. However, it is justified by the relatively large number of individual notations defined in this thesis (see the glossary in Section 2.5).

Integrals:

$\int_{\mathbb{X}} f(\xi)\, d\xi$ is the integral of $f(\xi)$ over the set $\mathbb{X}$, for example $\mathbb{X} \subset \mathbb{R}$ or $\mathbb{X} \subset \mathbb{R}^3$.

In particular, for a real-valued variable ξ: $\int_{\mathbb{R}} f(\xi)\, d\xi = \int_{-\infty}^{+\infty} f(\xi)\, d\xi$

Mathematical abbreviations:

iff	If and only if.
i.i.d.	Independently and identically-distributed.
pdf	Probability Density Function.
r.v.	Random variable.
r.v.s	Random variables.

Mathematical notations:

$\cdot$	Product operator.
$\overset{\text{def}}{=}$	Definition.
x	x is a variable, taking *one* value in a set $\mathbb{X}$ of possible values (for example $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{C}$).
$=$	The variable is often identified with its value, for example $x = 1$.
$x(t)$ or x_t	The variable x is a function of the variable t: a value $x(t)$ or x_t is associated to each value of t.
$\propto$	Proportional to. $x(t) \propto s(t) \quad \Leftrightarrow \quad \exists \xi > 0 \quad \forall t \in \mathbb{R} \quad x(t) = \xi \cdot s(t)$.
$\mathbb{N}$	The set of natural integers: $\mathbb{N} \overset{\text{def}}{=} \{0, 1, 2, \cdots\}$.
$\mathbb{Z}$	The set of signed integers: $\mathbb{Z} \overset{\text{def}}{=} \{\cdots, -2, -1, 0, 1, 2, \cdots\}$.
$\mathbb{R}$	The set of real numbers.
e	The Euler number ($\log e = 1$).
j	The imaginary unit ($j^2 = -1$).
$\mathbb{C}$	The set of complex numbers $\mathbb{C} \overset{\text{def}}{=} \left\{ \xi_1 + j\xi_2 \;\middle|\; [\xi_1, \xi_2]^{\mathrm{T}} \in \mathbb{R}^2 \right\}$.
$\Re(c), \Im(c)$	The real and imaginary parts of a complex number $c = \Re(c) + j \cdot \Im(c)$.
z^*	Complex conjugate of $z \in \mathbb{C}$: $z^* \overset{\text{def}}{=} \Re(z) - j \cdot \Im(z)$.
$\lvert z \rvert$	Magnitude of $z \in \mathbb{C}$: $\lvert z \rvert \overset{\text{def}}{=} \sqrt{z \cdot z^*} = \sqrt{\Re(z)^2 + \Im(z)^2}$.
$\angle z$	Phase of $z \in \mathbb{C}$, defined modulo 2π: $z = \lvert z \rvert \cdot e^{j \cdot \angle z} = \lvert z \rvert \cdot (\cos \angle z + j \cdot \sin \angle z)$.
$\sum$	Sum.
$\prod$	Product.

$\mathbf{1}_{\text{proposition}}$	Indicator function, equal to 1 iff proposition is true, and 0 otherwise. For example $\mathbf{1}_{x=7.456}$.
$f \otimes g$	Convolution of two functions.
	In the continuous domain: $\forall \tau \in \mathbb{R} \quad (f \otimes g)(\tau) \stackrel{\text{def}}{=} \int_{\mathbb{R}} f(\xi) \cdot [g(\tau - \xi)]^* \, d\xi$.
	In the discrete domain: $\forall \tau \in \mathbb{Z} \quad (f \otimes g)(\tau) \stackrel{\text{def}}{=} \sum_n f(n) \cdot [g(\tau - n)]^*$.
$\delta_{Kr}(x)$	Kronecker function, equal to 1 iff $x = 0$, and 0 otherwise: $\delta_{Kr}(x) = \mathbf{1}_{x=0}$.
$\delta_0(x)$	Dirac distribution, which has the property: $\int_{-\infty}^{x} \delta_0(\xi)\, d\xi = \mathbf{1}_{x \geq 0}$.
$\equiv$	Congruence of angles modulo 2π: $\xi_1 \equiv \xi_2 \quad \Leftrightarrow \quad \exists n \in \mathbb{Z} \quad \xi_1 = \xi_2 + n \cdot 2\pi$
$\hat{x}$	Estimate of x – except in Appendix C, where the hat designates a new parameter value.
$(x_1, x_2, \cdots, x_N)$	An ordered sequence.
$x_{1:N}$	Abbreviation for an ordered sequence: $x_{1:N} \stackrel{\text{def}}{=} (x_1, x_2, \cdots, x_N)$,
	or for a set (unordered): $x_{1:N} \stackrel{\text{def}}{=} \{x_1, x_2, \cdots, x_N\}$.
$\mathbf{x}$	(bold face) Column vector of variables $\mathbf{x} = [x_1, x_2, \cdots, x_N]^{\mathrm{T}}$ of dimension N.
$\lVert\mathbf{x}\rVert$	L^2-norm: $\lVert\mathbf{x}\rVert = \sqrt{\sum_{n=1}^{N} x_n^2}$.
$\mathbf{M}$	(bold face) Matrix of variables $M(a, b)$ where a is the row and b the column.
$\mathbf{I}$	The identity matrix, which has ones on the diagonal and zeroes elsewhere.
$\lvert\mathbf{M}\rvert$	Determinant of a matrix $\mathbf{M}$.
$[\cdot]^{\mathrm{T}}$	Transpose operator, for a matrix or a vector.
$\{x\}$	Set of values: $\{x\}$ contains all possible values of x.
$\left\{x \;\middle|\; x^2 = 5\right\}$	Set of values: $\left\{x \;\middle|\; x^2 = 5\right\}$ contains all values of x such that $x^2 = 5$.
$[\xi_1\ \xi_2]$	Interval set: $[\xi_1\ \xi_2] \stackrel{\text{def}}{=} \{\xi \in \mathbb{R} \mid \xi_1 \leq \xi \leq \xi_2\}$
$]\xi_1\ \xi_2]$	Interval set: $]\xi_1\ \xi_2] \stackrel{\text{def}}{=} \{\xi \in \mathbb{R} \mid \xi_1 < \xi \leq \xi_2\}$
$7 - \left\{[(4 + x)\, 2]\, y^2\right\}$	An expression: in this case $\{\cdot\}$ and $[\cdot]$ have the same role as parentheses.
$\mathbb{A} \setminus \mathbb{B}$	Set subtraction. The set $\mathbb{B}$ is subtracted from the set $\mathbb{A}$: $\mathbb{A} \setminus \mathbb{B} \stackrel{\text{def}}{=} \{\xi \in \mathbb{A} \mid \xi \notin \mathbb{B}\}$.
$\mathbb{X} \times \mathbb{Y}$	Product of two sets $\mathbb{X}$ and $\mathbb{Y}$: $\mathbb{X} \times \mathbb{Y} \stackrel{\text{def}}{=} \{(x, y) \mid x \in \mathbb{X} \text{ and } y \in \mathbb{Y}\}$.
$\mathbb{X}^N$	Set of N-dimensional vectors of values in $\mathbb{X}$:
	$\mathbb{X}^N \stackrel{\text{def}}{=} \left\{[x_1, x_2, \cdots, x_N]^{\mathrm{T}} \;\middle|\; \forall n \in \{1 \cdots N\}\ x_n \in \mathbb{X}\right\}$.

2.2 Discrete Time Processing of Quasi-Stationary Signals

Speech waveforms are usually assumed quasi-stationary over a time frame of a length up to about 30 ms. A speech waveform $x(t) \in \mathbb{R}$ is thus broken into a series of time frames[1], each time frame spans up to about 30 ms [2]. Discrete Fourier Transform (DFT) is then applied to each time frame for spectral analysis. This section details the short-time windowed DFT, as well as the associated notations. A glossary of notations can be found at the end of the chapter. We adopt a matrix notation for the DFT: $\mathbf{X} = \mathbf{F} \cdot \mathbf{x}$.

Continuous Time Domain: $x(t)$, $y(t)$, $s(t)$ and $n(t)$ stand for real-valued waveform signals, as captured by a microphone. $t \in \mathbb{R}$ is a continuous variable, its value is expressed in "sampling periods": the corresponding time value in seconds is $\frac{t}{f_s}$, where f_s is called the "sampling frequency", expressed in Hertz (Hz).

The continuous time domain cross-correlation function is defined as:

$$g_{x,y}(\tau) \stackrel{\text{def}}{=} [x(t) \otimes y(-t)](\tau) = \int_{\mathbb{R}} x(\xi)\, y(\xi - \tau)\, d\xi \tag{2.1}$$

Discrete Time Domain: To apply the Discrete Fourier Transform (DFT), the continuous time signals[3] are sampled at the sampling frequency f_s, and the analysis is restricted to a single discrete time frame centered on t, where a signal $x(t)$ can be assumed to be stationary. A "discrete time frame" means a vector of $2N_{\mathrm{F}}$ samples, denoted $\mathbf{x}^{(t)} \in \mathbb{R}^{2N_{\mathrm{F}}}$:

$$\mathbf{x}^{(t)} \stackrel{\text{def}}{=} \left[x^{(t)}(1), \cdots, x^{(t)}(n), \cdots, x^{(t)}(2N_{\mathrm{F}})\right]^{\mathrm{T}} \tag{2.2}$$

where n is a generic integer index. Unless stated otherwise, the framing process is composed of sampling at frequency f_s, pre-emphasis with a 0.97 coefficient, followed by Hamming windowing:

$$x^{(t)}(n) \stackrel{\text{def}}{=} \left[0.54 - 0.46 \cos\left(\pi \frac{n-1}{N_{\mathrm{F}}}\right)\right] \cdot \left[x(t - N_{\mathrm{F}} + n) - 0.97 \cdot x(t - N_{\mathrm{F}} + n - 1)\right] \tag{2.3}$$

In the following, we abbreviate "discrete time frame centered on t of the continuous time signal $x(t)$" with "time frame $\mathbf{x}^{(t)}$", "time frame t" or "time frame samples".

[1] Unless stated otherwise, we use a 50 % overlap between any two consecutive time frames.
[2] Unless stated otherwise, each time frame spans 32 ms of data.
[3] All signals $x(t), y(t)$ etc. are assumed to be limited to the frequency band $\left[0\ \frac{f_s}{2}\right]$.

For each time frame $\mathbf{x}^{(t)}$, the corresponding vector of DFT coefficients is written $\mathbf{X}^{(t)} \in \mathbb{C}^{2N_{\mathrm{F}}}$:

$$\mathbf{X}^{(t)} \stackrel{\text{def}}{=} \left[X^{(t)}(1), \cdots, X^{(t)}(k), \cdots, X^{(t)}(2N_{\mathrm{F}})\right]^{\mathrm{T}} \tag{2.4}$$

where $k \in \{1, \cdots, 2N_{\mathrm{F}}\}$ is called the discrete frequency index. The discrete frequencies $k \in \{1, \cdots, N_{\mathrm{F}} + 1\}$ correspond to the real frequencies $\frac{k-1}{N_{\mathrm{F}}} \cdot \frac{f_s}{2}$, where f_s is the sampling frequency in Hertz, of the original signal $x(t)$. Time frame samples and DFT coefficients are linearly related, through the DFT and the Inverse Discrete Fourier Transform (IDFT). Using a matrix notation:

$$\mathbf{X}^{(t)} \stackrel{\text{def}}{=} \mathbf{F} \cdot \mathbf{x}^{(t)} \tag{2.5}$$

$$\mathbf{x}^{(t)} = \mathbf{F}^{-1} \cdot \mathbf{X}^{(t)} \tag{2.6}$$

where each term of the matrix $\mathbf{F}$ and its inverse $\mathbf{F}^{-1}$ are respectively:

$$F(a, b) \stackrel{\text{def}}{=} exp\left(-j\pi \frac{(a-1)(b-1)}{N_{\mathrm{F}}}\right) \tag{2.7}$$

$$F^{-1}(a, b) = \frac{1}{2N_{\mathrm{F}}} exp\left(j\pi \frac{(a-1)(b-1)}{N_{\mathrm{F}}}\right) \tag{2.8}$$

where $a \in \{1, \cdots, 2N_{\mathrm{F}}\}$ is the row index and $b \in \{1, \cdots, 2N_{\mathrm{F}}\}$ is the column index.

The *instantaneous* magnitude spectrum estimate $\mathbf{M}_x^{(t)} \in \mathbb{R}^{2N_{\mathrm{F}}}$, and the *instantaneous* energy[4] spectrum estimate $\mathbf{E}_x^{(t)} \in \mathbb{R}^{2N_{\mathrm{F}}}$ are derived from the complex spectrum estimate $\mathbf{X}^{(t)}$:

$$M_x^{(t)}(k) \stackrel{\text{def}}{=} \left|X^{(t)}(k)\right| \tag{2.9}$$

$$E_x^{(t)}(k) \stackrel{\text{def}}{=} \left|X^{(t)}(k)\right|^2 \tag{2.10}$$

Instantaneous and Average Correlation Spectrum: For any two signals $x(t)$ and $y(t)$, the above-described time frame-based DFT analysis produces two series of DFT coefficients $\mathbf{X}^{(t)}$ and $\mathbf{Y}^{(t)}$. The *instantaneous* complex correlation spectrum estimate is defined as $\mathbf{G}_{x,y}^{(t)} \in \mathbb{C}^{2N_{\mathrm{F}}}$, where:

$$G_{x,y}^{(t)}(k) \stackrel{\text{def}}{=} X^{(t)}(k) \cdot \left(Y^{(t)}(k)\right)^* \tag{2.11}$$

[4]Energy is often confused with power, although this is not an issue with finite-length signals. See (Lehmann, 2004, Section 1.5) for a discussion on this topic.

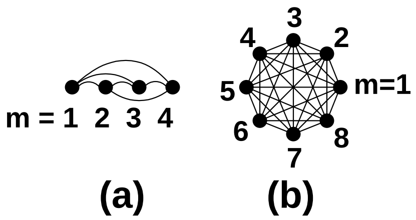


Figure 2.1. Example of microphone arrays: dots depict microphones, lines depict pairs. (a) Uniform Linear Array with $N_m = 4$ microphones and $N_q = 6$ pairs, (b) Uniform Circular Array with $N_m = 8$ microphones and $N_q = 28$ pairs.

The *average* complex correlation spectrum estimate is defined as $\mathbf{\Phi}_{x,y} \in \mathbb{C}^{2N_\mathrm{F}}$, where:

$$\phi_{x,y}(k) \stackrel{\mathrm{def}}{=} \left\langle G^{(t)}_{x,y}(k) \right\rangle_t = \left\langle X^{(t)}(k) \cdot \left(Y^{(t)}(k)\right)^* \right\rangle_t \tag{2.12}$$

where $\langle\cdot\rangle_t$ designates the average operator, applied to several time frames t.

The *average* complex coherence spectrum estimate is defined as $\mathbf{\Gamma}_{x,y} \in \mathbb{C}^{2N_\mathrm{F}}$, where:

$$\Gamma_{x,y}(k) \stackrel{\mathrm{def}}{=} \frac{\phi_{x,y}(k)}{\sqrt{\phi_{x,x}(k) \cdot \phi_{y,y}(k)}} \tag{2.13}$$

Zero-padding: Unless stated otherwise, zero-padding is *not* used in this thesis. However, whenever the time domain GCC-PHAT needs to be evaluated, as defined in (3.9), zero-padding is necessary to avoid circularity issues. In this thesis, zero-padding means concatenating the time frame of pre-emphasized samples with an equal number of zeroes. In such a case, (2.3) is replaced with:

$$x^{(\mathrm{zp},t)}(n) \stackrel{\mathrm{def}}{=} \begin{cases} \left[0.54 - 0.46\cos\left(2\pi\frac{n-1}{N_\mathrm{F}}\right)\right] \cdot \left[x\left(t - N_\mathrm{F} + n\right) - 0.97 \cdot x\left(t - N_\mathrm{F} + n - 1\right)\right] & \text{if } 1 \le n \le N_\mathrm{F} \\ 0 & \text{if } N_\mathrm{F} < n \le 2N_\mathrm{F} \end{cases} \tag{2.14}$$

Whenever zero-padding is used, the value of N_F is doubled, so that the number of samples extracted from $x(t)$ is the same as in (2.3).

2.3 Multichannel Signals

We denote microphone locations, signals and pairs as follows. Locations are denoted $\boldsymbol{\ell} \stackrel{\text{def}}{=} [\mathcal{X}, \mathcal{Y}, \mathcal{Z}]^{\text{T}}$, for example a speaker location $\boldsymbol{\ell}^{\text{ps}}$, or a microphone location $\boldsymbol{\ell}_m$, where $m \in \{1, \cdots, N_m\}$ (Figure 2.1). Two particular types of microphone arrays are often used in this thesis: the Uniform Linear Array (ULA) and the Uniform Circular Array (UCA). The ULA has equispaced microphones placed along a line (Figure 2.1a), and the UCA has equispaced microphones placed along a circle (Figure 2.1b).

The signal received by each microphone m is denoted $x_m(t)$. Following Section 2.2, for each microphone m, each time frame of samples is denoted by the vector $\mathbf{x}_m^{(t)} \in \mathbb{R}^{2N_{\text{F}}}$, and the corresponding vector of DFT coefficients is denoted $\mathbf{X}_m^{(t)} \in \mathbb{C}^{2N_{\text{F}}}$. As explained later in Chapter 3, pairs of microphones can be used for acoustic source localization. With N_m microphones, there are $N_q = N_m \cdot (N_m - 1)/2$ possible pairs of microphones, indexed with $q \in \{1, \cdots, N_q\}$. The two microphones $\boldsymbol{\ell}_{a_q}$ and $\boldsymbol{\ell}_{b_q}$ of the q-th pair are indexed with the integers indices a_q and b_q $(1 \leq a_q < b_q \leq N_m)$. The notation $\tau^{(q)}$ denotes a time delay between the two signals $x_{a_q}(t)$ and $x_{b_q}(t)$, expressed in sampling periods. Note that $\tau^{(q)}$ is not necessarily an integer: $\tau^{(q)} \in \mathbb{R}$.

2.4 Probabilities and Random Variables

This subsection defines the notations of the probabilistic tools that are used in this thesis. We voluntarily simplified the definitions. Please refer to (Delmas, 1993; Weisstein, 2006b) for an exhaustive definition of the underlying concepts, such as probability spaces and measures.

A deterministic variable x, which takes a single value ($x = 1.234$), can be seen as a *realization* of a random variable (r.v.) $\underline{x}$, among a set of possible values. The r.v. $\underline{x}$ is defined by a set $\mathbb{X}$ of *possible* values (for example $\mathbb{X} = \mathbb{R}$), and a Probability Density Function (pdf) on $\mathbb{X}$.

A pdf is a distribution[5] $p_{\underline{x}}(x)$ on realizations $x \in \mathbb{X}$ of the r.v. $\underline{x}$, such that:

$$\begin{cases} \forall x \in \mathbb{X} \quad p_{\underline{x}}(x) \geq 0 \\ \displaystyle\int_{\mathbb{X}} p_{\underline{x}}(x)\, dx \quad = \quad 1 \end{cases} \tag{2.15}$$

[5]Distributions are generalizations of functions, see (Rowland, 2002) for more on this subject. "pdf" are thus sometimes called "probability distributions".

and for any subset $\mathbb{Y} \subset \mathbb{X}$, the probability that a realization x of $\underline{x}$ belongs to the set $\mathbb{Y}$ is given by:

$$P(\underline{x} \in \mathbb{Y}) \stackrel{\text{def}}{=} \int_{\mathbb{Y}} p_{\underline{x}}(x)\, dx \quad \in [0\ 1] \tag{2.16}$$

For the sake of readability, we often use the following abbreviations:

$$\begin{aligned} P(\mathbb{Y}) &= P(\underline{x} \in \mathbb{Y}) &= \int_{\mathbb{Y}} p_{\underline{x}}(x)\, dx \quad &\text{probability of } \mathbb{Y},\ \text{in } [0\ 1] \\ p(x) &= p(\underline{x} = x) &= p_{\underline{x}}(x) \quad &\text{likelihood of } x,\ \text{may be greater than 1} \end{aligned} \tag{2.17}$$

Moreover, any equation written using a set of random variables (r.v.s) means that the equation is valid for *any* realization of the set of r.v.s. For example:

$$p(\underline{a}) = p(\underline{b}) \tag{2.18}$$

is strictly equivalent to:

$$\forall a, b \qquad p(\underline{a} = a) = p(\underline{b} = b) \tag{2.19}$$

A pdf can also be defined with respect to a *prior knowledge* H (model, hypothesis, realization of another r.v., etc.), the likelihood is then a "conditional probability" $p(\underline{x} = x \mid H) = p_{\underline{x}|H}(x)$. Bayes' rule then gives the *posterior* probability of H, given that a realization x was observed:

$$P(H \mid \underline{x} = x) = \frac{p(\underline{x} = x \mid H) \cdot P(H)}{p(\underline{x} = x)} \tag{2.20}$$

where $p(H)$ and $p(\underline{x} = x)$ are referred to as the priors of the "events" H and $\underline{x} = x$, respectively. As above, in all terms of (2.20), we often abbreviate $\underline{x} = x$ with x.
The mean of a function $g(x)$ with respect to a probability distribution $p(x) = p_{\underline{x}}(x)$ is defined as:

$$\langle g(x) \rangle_{p(x)} \stackrel{\text{def}}{=} \int_{\mathbb{R}} g(x) \cdot p(x) \cdot dx \tag{2.21}$$

The expectation of a r.v. $\underline{x}$, also known as the mean of the r.v. $\underline{x}$, is defined as:

$$\mathrm{E}\{\underline{x}\} \stackrel{\text{def}}{=} \langle x \rangle_{p(x)} \tag{2.22}$$

$$= \int_{\mathbb{R}} x \cdot p(x) \cdot dx = \int_{\mathbb{R}} x \cdot p_{\underline{x}}(x) \cdot dx \tag{2.23}$$

Similarly, the conditional expectation of a r.v. $\underline{x}$ with respect to a prior knowledge H, is defined as:

$$\mathrm{E}\{\underline{x} \mid H\} \stackrel{\text{def}}{=} \langle x \rangle_{p(x \mid H)} \tag{2.24}$$

$$= \int_{\mathbb{R}} x \cdot p(x \mid H) \cdot dx = \int_{\mathbb{R}} x \cdot p_{\underline{x}|H}(x) \cdot dx \tag{2.25}$$

The **Signal-to-Noise Ratio (SNR)** is the signal power divided by the noise power. Let us consider a signal r.v. $\underline{s}$ and a noise r.v. $\underline{n}$, both with values in $\mathbb{R}$. The SNR is defined as the ratio of the two second-order moments:

$$\mathrm{SNR}_{\underline{s},\underline{n}} \stackrel{\text{def}}{=} \frac{\mathrm{E}\{\underline{s}^2\}}{\mathrm{E}\{\underline{n}^2\}} \tag{2.26}$$

The SNR is often expressed in decibels (dB), using the formula: $10 \cdot \log_{10}\left(\mathrm{SNR}_{\underline{s},\underline{n}}\right)$.

In the case of speech, we only have observed values of time domain waveforms: speech $s(t)$ and noise $n(t)$. Observe that speech is quasi-stationary over a short time frame (up to about 30 ms), comprising samples $(s(t - N_{\mathrm{F}} + 1), \cdots, s(t + N_{\mathrm{F}}))$. Under second-order stationarity and ergodicity assumptions, the SNR becomes the ratio of the *time domain* second-order moments:

$$\mathrm{SNR}_{\underline{s},\underline{n}}(t) = \frac{\left\langle s(t+a)^2 \right\rangle_{-N_{\mathrm{F}} < a \leq N_{\mathrm{F}}}}{\left\langle n(t+a)^2 \right\rangle_{-N_{\mathrm{F}} < a \leq N_{\mathrm{F}}}} \tag{2.27}$$

The SNR is usually reported in the dB domain: $10 \cdot \log_{10}\left[\mathrm{SNR}_{\underline{s},\underline{n}}(t)\right]$. In this thesis, the *average SNR*, over several time frames t, is also defined in the dB domain: $\left\langle 10 \cdot \log_{10}\left[\mathrm{SNR}_{\underline{s},\underline{n}}(t)\right]\right\rangle_t$.

Note: All definitions involving the DFT also apply to r.v.s. For example, the DFT $\underline{\mathbf{X}} = \mathbf{F} \cdot \underline{\mathbf{x}}$ implies that each DFT coefficient $\underline{X}(k)$ is a r.v., because it is a linear combination of the $2 \cdot N_{\mathrm{F}}$ r.v.s $(\underline{x}(1), \cdots, \underline{x}(2N_{\mathrm{F}}))$.

Several types of pdf are used in this thesis:

$\underline{\mathbf{x}} \sim \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ — $\underline{\mathbf{x}}$ is a multivariate normal r.v. in $\mathbb{R}^N$ (also known as multivariate Gaussian r.v.), with mean $\boldsymbol{\mu} \in \mathbb{R}^N$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$. The multivariate normal pdf is:

$$\forall \mathbf{x} \in \mathbb{R}^N \quad \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) \stackrel{\text{def}}{=} |2\pi \cdot \boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$\underline{x} \sim \mathcal{N}_{\mu,\sigma}$ — $\underline{x}$ is a normal r.v. in $\mathbb{R}$ (also known as Gaussian r.v.), with mean μ and *standard deviation* σ. The normal pdf is:

$$\forall x \in \mathbb{R} \quad \mathcal{N}_{\mu,\sigma}(x) \stackrel{\text{def}}{=} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\underline{x} \sim \mathcal{G}_{\gamma,\beta}$ — $\underline{x}$ is a Gamma r.v. of parameters $\gamma > 0$ and $\beta > 0$. The Gamma pdf is:

$$\mathcal{G}_{\gamma,\beta}(x) \overset{\text{def}}{=} \begin{cases} \frac{x^{\gamma-1} \cdot e^{-\frac{x}{\beta}}}{\beta^{\gamma} \cdot \Gamma(\gamma)} & \text{if} \quad x \geq 0 \\ 0 & \text{if} \quad x < 0 \end{cases}$$

where Γ is the gamma function: $\Gamma(\gamma) \overset{\text{def}}{=} \int_0^{+\infty} t^{\gamma-1} \cdot e^{-t} \cdot dt$

$\underline{x} \sim \mathcal{R}_{\sigma,V}$ — $\underline{x}$ is a Rice r.v. with parameters $\sigma > 0$ and $V \geq 0$. The Rice pdf $\mathcal{R}_{\sigma,V}(x)$ describes the probability distribution of the envelope of the sum of a sinusoidal wave and a zero mean narrowband Gaussian noise.

$$\mathcal{R}_{\sigma,V}(x) \overset{\text{def}}{=} \begin{cases} \frac{x}{\sigma^2} \cdot e^{-\frac{x^2+|V|^2}{2\sigma^2}} \cdot \mathrm{I}_0\left(\frac{x|V|}{\sigma^2}\right) & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases}$$

where I_0 is the modified Bessel function of the first kind.

In the particular case $V = 0$, the Rice pdf becomes a Rayleigh pdf:

$$\mathcal{R}_{\sigma,0}(x) = \frac{x}{\sigma^2} \cdot e^{-\frac{x^2}{2\sigma^2}}$$

If we assume two zero-mean, uncorrelated Gaussian r.v. $\underline{A} \sim \mathcal{N}_{0,\sigma}$ and $\underline{B} \sim \mathcal{N}_{0,\sigma}$ then $|\underline{A} + j\underline{B}| \sim \mathcal{R}_{\sigma,0}$ (Rice, 1944, 1945).

2.5 Glossary of Notations

There may be some overlap between the notations of the different chapters, but the context should make clear in which sense a notation is used.

General purpose notations.

a, b, i, n, N, Q	General purpose variables, usually indices (signed integers).
$\xi, \boldsymbol{\xi}, \Xi, \boldsymbol{\Xi}$	General purpose variables, usually real-valued (scalars and column vectors).
$f(\cdot), g(\cdot), h(\cdot)$	General purpose functions.
$x(t), y(t), s(t), n(t)$	Continuous time domain signals: real-valued functions of the continuous time $t \in \mathbb{R}$.
$\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \mathbf{s}^{(t)}, \mathbf{n}^{(t)}$	One time frame of discrete time domain samples: vectors in $\mathbb{R}^{2N_{\mathrm{F}}}$.
$\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{S}^{(t)}, \mathbf{N}^{(t)}$	One time frame of DFT coefficients: vectors in $\mathbb{C}^{2N_{\mathrm{F}}}$.
$\mathbf{M}_x^{(t)}, \mathbf{M}_y^{(t)}, \mathbf{M}_s^{(t)}, \mathbf{M}_n^{(t)}$	One time frame of magnitude values: vectors in $\mathbb{R}^{2N_{\mathrm{F}}}$.
$\mathbf{E}_x^{(t)}, \mathbf{E}_y^{(t)}, \mathbf{E}_s^{(t)}, \mathbf{E}_n^{(t)}$	One time frame of energy values: vectors in $\mathbb{R}^{2N_{\mathrm{F}}}$.
$\underline{x}$	Random variable (r.v.), of which x is a realization.

Specific notations: Roman letters.

a_q, b_q	Indices of the microphones $\boldsymbol{\ell}_{a_q}$ and $\boldsymbol{\ell}_{b_q}$ of the q-th pair:
	two integers $a_q \in \mathbb{N}$ and $b_q \in \mathbb{N}$ such that $1 \leq a_q < b_q \leq N_m$.
c	Speed of sound in the air, in m/s.
$d(\mathbf{u}_1, \mathbf{u}_2)$	Phase Domain Metric (PDM) between two vectors $(\mathbf{u}_1, \mathbf{u}_2) \in \left(\mathbb{R}^{N_q}\right) \times \left(\mathbb{R}^{N_q}\right)$.
f_s	Sampling frequency in Hz.
$g_{x,y}(\tau)$	Continuous time domain cross-correlation between $x(t)$ and $y(t)$, where $\tau \in \mathbb{R}$.
$\mathbf{g}_{x,y}^{(\mathrm{PH},t)}$	Instantaneous discrete time domain GCC-PHAT for time frame t: $\mathbf{g}_{x,y}^{(\mathrm{PH},t)} \in \mathbb{R}^{2N_{\mathrm{F}}}$.
$\hat{g}_{x,y}^{(\mathrm{PH},t)}(\tau)$	GCC-PHAT continuous cross-correlation function ($\tau \in \mathbb{R}$), for discrete time frame t.
k	Discrete frequency index: $k \in \{1, \cdots, 2N_{\mathrm{F}}\}$.
$\boldsymbol{\ell}$	Location vector: $\boldsymbol{\ell} \in \mathbb{R}^3$: $\boldsymbol{\ell} = [\mathcal{X}, \mathcal{Y}, \mathcal{Z}]^{\mathrm{T}}$. E.g. microphone loc. $\boldsymbol{\ell}_m$, point source loc. $\boldsymbol{\ell}^{\mathrm{ps}}$.
m, N_m	Microphone index and number of microphones.
$\underline{n}_{\mathrm{C}}$, $\hat{n}_{\mathrm{C}}(t)$	Number of correctly localized speakers: r.v. and its expected value for time frame t.
$p(x)$	Likelihood that $\underline{x} = x$.
$P(\underline{x} \in \mathbb{X})$	Probability that $\underline{x} \in \mathbb{X}$.
q, N_q	Pair index and number of pairs of microphones.
$r_i = (\theta_i, T_i)$	Location estimate, at azimuth θ_i (radians) and frame T_i (integer). $i \in \{1, \cdots, N_r\}$.
$\check{s}, N_{\check{s}}$	Sector index and number of sectors: $\check{s} \in \{1, \cdots, N_{\check{s}}\}$.
$\check{s}_{\min}(k)$	Index of the sector with the minimum PDM, at the discrete frequency k.
$t,\ t_n$	Continuous time values, expressed in sampling periods: $t \in \mathbb{R}$.
$u, \mathbf{u}$	Phase value in radians, and vector of phase values $\mathbf{u} = \left[u_1, \cdots, u_q, \cdots, u_{N_q}\right]^{\mathrm{T}}$.
$u_q^{\mathrm{th}}(k, \boldsymbol{\ell}), \mathbf{u}^{\mathbf{th}}(k, \boldsymbol{\ell})$	Theoretical phase and vector of theoretical phases, for location $\boldsymbol{\ell}$.
$u_q^{(t)}(k), \mathbf{u}^{(t)}(k)$	Observed phase and vector of observed phases, for time frame t.
$\mathbf{v}_{\check{s},n}$	A location $\mathbf{v}_{\check{s},n} \in \mathbb{R}^3$, used to sample the sector $\mathbb{S}_{\check{s}}$. $(\check{s}, n) \in \{1, \cdots, N_{\check{s}}\} \times \{1, \cdots, N_{\mathbf{v}}\}$
w_0, w_1	Weights in a mixture, as in $f(\xi) = w_0 \cdot f_0(\xi) + w_1 \cdot f_1(\xi)$.
$x_m(t), \mathbf{x}_m^{(t)}, \mathbf{X}_m^{(t)}$	For microphone $\boldsymbol{\ell}_m$: received signal, time frame, DFT coefficients.
$\underline{A}$, A_t	Binary frame state of time frame t: r.v. and its realization.
$\underline{B}_{\check{s}}$, $B_{\check{s},t}$	Binary sector state of sector $\mathbb{S}_{\check{s}}$ at time frame t: r.v. and its realization (ground-truth).
$\hat{B}_{\check{s},t}$	Binary decision for sector $\mathbb{S}_{\check{s}}$ at time frame t (result).
$D_{\check{s},k}$	Root Mean Square of the PDM d on sector $\mathbb{S}_{\check{s}}$, at discrete frequency k.

$\mathrm{E}\{\underline{x}\}$, $\mathrm{E}\{\underline{x} \mid H\}$	Expectation of a random variable, and conditional expectation.
$\underline{E}$, E_i	Label field: r.v. and its realization for location estimate r_i.
$\mathbf{F}$, $\mathbf{F}^{-1}$	DFT matrix and its inverse.
$\mathbf{G}_{x,y}^{(t)}$	Instantaneous freq. domain cross-correlation for discrete time frame t: $\mathbf{G}_{x,y}^{(t)} \in \mathbb{C}^{2N_\mathrm{F}}$.
$\mathbf{G}_{x,y}^{(\mathrm{PH},t)}$	Instantaneous freq. domain GCC-PHAT for discrete time frame t: $\mathbf{G}_{x,y}^{(\mathrm{PH},t)} \in \mathbb{C}^{2N_\mathrm{F}}$.
H	Hypothesis.
$\mathrm{I}_0(\cdot)$	Modified Bessel function of the first kind.
$\mathbf{MFCC}$	Vector of Mel-Frequency Cepstral Coefficients (MFCC): $\mathbf{MFCC} \stackrel{\mathrm{def}}{=} [\mathrm{MFCC}_0 \cdots \mathrm{MFCC}_{12}]^T$
K	Number of mixtures in a Gaussian Mixture Model.
N_future	Integer parameter of the short-term clustering. $N_\mathrm{future} \in \mathbb{N} \setminus \{0\}$.
N_t	Number of time frames, each defined by their center time t_n $(1 \leq n \leq N_t)$.
$\mathbb{S}_{\check{s}}$	Sector $\mathbb{S}_{\check{s}}$ of space: $\mathbb{S}_{\check{s}} \subset \mathbb{R}^3$ and $\check{s} \in \{1, \cdots, N_{\check{s}}\}$.
T	Parameter of the adaptive background subtraction.
T_i	Integer time frame index (Chapters 6 & 7): $T_i \in \mathbb{N} \setminus \{0\}$ and $i \in \{1, \cdots, N_r\}$.
T_short	Integer duration parameter of the short-term clustering (number of frames).
$\mathrm{TOF}(\boldsymbol{\ell}, \boldsymbol{\ell}')$	Theoretical Time Of Flight (TOF) of an acoustic wave between locations $\boldsymbol{\ell}$ and $\boldsymbol{\ell}'$.
$\widehat{\mathrm{TOF}}(\boldsymbol{\ell}, \boldsymbol{\ell}')$	Wideband estimate of the TOF of an acoustic wave between locations $\boldsymbol{\ell}$ and $\boldsymbol{\ell}'$.
$\widehat{\mathrm{ntof}}(\boldsymbol{\ell}, \boldsymbol{\ell}', k)$	Narrowband estimate of the TOF of an acoustic wave between locations $\boldsymbol{\ell}$ and $\boldsymbol{\ell}'$.
$U(E)$	Energy of the label field E.
V	Parameter of the Rice distribution.
$Z_{\check{s}}^{q}(k)$	Average theoretical value in $\mathbb{C}$, for sector $\mathbb{S}_{\check{s}}$, pair q and frequency k.
$\mathcal{C}$	Cliques in a Markov Random Field.
$\mathcal{G}$	Neighborhood system in a Markov Random Field.
$\mathcal{G}_{\gamma,\beta}$, $\mathcal{G}_{\gamma,\beta}(\xi)$	Gamma random variable, and Gamma pdf evaluated at ξ.
$\mathcal{M}$	Generic notation for a model.
$\mathcal{N}_{\mu,\sigma}$, $\mathcal{N}_{\mu,\sigma}(\xi)$	Normal/Gaussian random variable, and Gaussian pdf evaluated at ξ.
$(\mathcal{N} + \mathcal{U})$	Gaussian + Uniform model, used to evaluate localization results.
$\mathcal{R}_{\sigma,V}$, $\mathcal{R}_{\sigma,V}(\xi)$	Rice random variable, and Rice pdf evaluated at ξ.
$\mathcal{U}$	Uniform random variable.
$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$	Euclidean coordinates.

Specific notations: Greek letters.

α	Learning parameter of the adaptive background subtraction.
β, γ	Parameters of the Gamma pdf.
$\beta_{ij}^{\text{potts}}$	Potts coefficients.
ϵ	A small value: $0 < \epsilon << 1$.
$\zeta_{\check{s},t}$	Activeness value for sector $\mathbb{S}_{\check{s}}$ and time frame t: $\zeta_{\check{s},t} \geq 0$.
Ψ	Threshold value. For example a threshold $\Psi_\zeta \geq 0$ on activeness $\zeta \geq 0$.
θ	Azimuth value, in radians.
φ	Elevation value, in radians: $\varphi \in \left[-\frac{\pi}{2},\ +\frac{\pi}{2}\right]$.
ρ	Radius value, in meters: $\rho > 0$.
$\mathcal{L}_\rho$	Log radius value: $\mathcal{L}_\rho \overset{\text{def}}{=} \log \rho$.
$\mathbf{\Lambda}(\mathcal{M})$	Parameters of a model $\mathcal{M}$. For example: $\mathbf{\Lambda}(\mathcal{N}_{\mu,\sigma}) = (\mu, \sigma)$.
$\kappa(\mathcal{M})$	Number of free parameters of a model $\mathcal{M}$. For example: $\kappa(\mathcal{N}_{\mu,\sigma}) = 2$.
λ	Generic purpose tuning parameter.
μ	Mean (scalar or vector) of a Gaussian pdf, or step size in the adaptive filtering.
σ, $\mathbf{\Sigma}$	Standard deviation σ and covariance matrix $\mathbf{\Sigma}$.
τ	Time delay in sampling periods.
$\tau_q^{\text{th}}(\boldsymbol{\ell}^{\text{ps}})$	Theoretical time delay for microphone pair q and speaker/point source location $\boldsymbol{\ell}^{\text{ps}}$.
$\hat{\tau}_q^{(t)}$	Estimated time delay for microphone pair q and time frame t.
$\mathbf{\Phi}_{x,y}$	Average correlation spectrum between signals $x(t)$ and $y(t)$. $\mathbf{\Phi}_{x,y} \in \mathbb{C}^{2N_{\text{F}}}$.
$\mathbf{\Gamma}_{x,y}$	Average coherence spectrum between signals $x(t)$ and $y(t)$. $\mathbf{\Gamma}_{x,y} \in \mathbb{C}^{2N_{\text{F}}}$.
$\Gamma(\gamma)$	Gamma function.
η, η_0	Temperature, and initial temperature of the simulated annealing.
ω_n, Ω, N_Ω	Cluster ω_n, partition $\Omega = \{\omega_1, \cdots, \omega_n, \cdots, \omega_{N_\Omega}\}$, number of clusters N_Ω.
O_{ST}	Set of all possible short-term partitions:
	$O_{\text{ST}} \overset{\text{def}}{=} \{\Omega \mid \forall n \in \{1, \cdots, N_\Omega\} \quad \omega_n \text{ is a short} - \text{term cluster}\}$.
Δ	Cost function for gradient descent speaker localization.
$\Delta_{k,q}$	Basic term in the expression of Δ, for frequency k and pair q.
Υ	Subset of the strictly positive discrete frequencies: $\Upsilon \subset \{2, \cdots, N_{\text{F}} + 1\}$.

2.6 Abbreviations

Abbreviation	Full name	Introduced in Appendix or Section
Performance Metrics		
FAR	False Alarm Rate	A
$\widehat{\mathrm{FAR}}$	Estimated FAR value	5.3.2
$\mathrm{FAR}_{\mathrm{T}}$	Target value for FAR	5.3
FRR	False Rejection Rate	A
PRC	Precision	A
RCL	Recall	A
HTER	Half-Total Error Rate	A
F	F-measure	A
Microphone array processing		
BSS	Blind Source Separation	3.1.2
DHBF	Double Hierarchical BeamForming	3.1.2
GCC	Generalized Cross-Correlation	3.1.2
MIMO	Multiple Inputs Multiple Outputs	3.1.2
PDM	Phase Domain Metric	B.1
PHAT	Phase Transform	3.1.2
SIR	Signal-to-Interference Ratio	8.1
SNR	Signal-to-Noise Ratio	2.4
SRP	Steered Response Power	3.1.2
TDE	Time Delay Estimation	3.1.2
TDOA	Time Delay Of Arrival	3.1.2
UCA	Uniform Circular Array	2.3
ULA	Uniform Linear Array	2.3

Short-Term Clustering

MRF	Markov Random Field	6.2.2
SA	Simulated Annealing optimization	6.2.2
SW	Sliding Window optimization	6.2.1

Speaker Clustering

BIC	Bayesian Information Criterion	7.1
GMM	Gaussian Mixture Model	3.2.1
HMM	Hidden Markov Model	3.2.3

Others

%	Percentage notation: $5\% \stackrel{\text{def}}{=} 5/100 = 0.05$	
1-D	1-dimensional	
2-D	2-dimensional	
3-D	3-dimensional	
headset mic.	Microphone worn by a speaker, near the mouth	
i.i.d.	Independently and identically-distributed	
iff	If and only if	
lapel mic.	Microphone worn by a speaker, near the throat	
pdf	Probability Density Function	
r.v.	Random variable	
r.v.s	Random variables	
LHS	Left Hand Side	
LPCC	Linear Prediction Cepstral Coefficients	
MFCC	Mel-Frequency Cepstral Coefficients	
RHS	Right Hand Side	
ROC	Receiver Operating Characteristic	5.2.4
USS	Unsupervised Spectral Subtraction	8.2

Chapter 3

Background

As explained in Chapter 1, the purpose of this thesis is to build an integrated system for non-invasive multispeaker detection-localization and speaker clustering – "non-invasive" meaning that only distant microphones are used. This integration effort led to several research issues, which are addressed by the following chapters of this thesis. In particular, we are trying to build a single system to cope with static data – seated speakers in a meeting – as well as dynamic data – multiple moving speakers in a surveillance scenario. As explained in Section 1.1, this thesis does *not* attempt to improve fundamentals of instantaneous speaker localization or continuous object tracking, but rather explores associated issues, such as detection for speech localization and speaker clustering with distant microphones only.

This chapter sketches the present state-of-the-art in speaker localization, with a particular emphasis on the Steered Response Power methods, also known as beamforming methods (Krim and Viberg, 1996; DiBiase, 2000), because they form the core around which the present work is constructed (Section 3.1). Next, Section 3.2 reviews existing work in meeting segmentation, and summarizes a preliminary experiment where multi-party speech is segmented using location information.

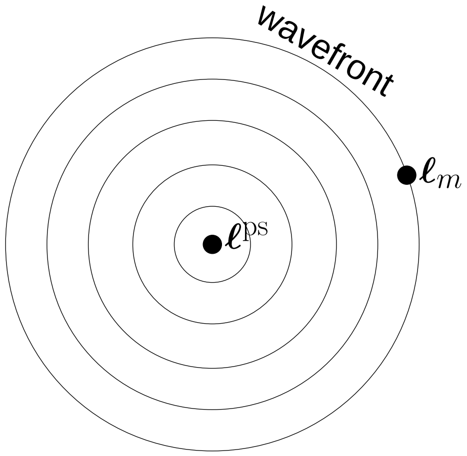


Figure 3.1. A spherical acoustic wave emitted by a point source at location $\boldsymbol{\ell}^{\mathrm{ps}}$, and recorded by a microphone at location $\boldsymbol{\ell}_m$.

3.1 Speaker Localization

3.1.1 Acoustic Waves

The air is usually assumed to be a non-dispersive medium, which means that c, the speed of an acoustic wave, is independent of the wave frequency. This speed is usually assumed to only depend on the air temperature η in Kelvins:

$$c \approx 331.46 * \sqrt{\frac{\eta}{273.15}} \tag{3.1}$$

where c is expressed in m/s (Kinsler et al., 1999). For an ambient room temperature of 18 degrees Celsius (291 Kelvins), we obtain $c = 342\ m/s$. We denote the Euclidean coordinates of a point $\boldsymbol{\ell} \in \mathbb{R}^3$ with the real values $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$: $\boldsymbol{\ell} = [\mathcal{X}, \mathcal{Y}, \mathcal{Z}]^{\mathrm{T}}$. In this thesis, we assume that a speech source can be modelled as a point source, placed at $\boldsymbol{\ell}^{\mathrm{ps}} \stackrel{\mathrm{def}}{=} [\mathcal{X}^{\mathrm{ps}}, \mathcal{Y}^{\mathrm{ps}}, \mathcal{Z}^{\mathrm{ps}}]^{\mathrm{T}}$, emitting a spherical wave that will eventually be received by a microphone m placed at $\boldsymbol{\ell}_m \stackrel{\mathrm{def}}{=} [\mathcal{X}_m, \mathcal{Y}_m, \mathcal{Z}_m]^{\mathrm{T}}$. A spherical wave means that the wavefront is a sphere centered on the point source location $\boldsymbol{\ell}^{\mathrm{ps}}$, with a radius that increases over time with the speed c, as depicted in Figure 3.1. We preferred the spherical wave assumption over the plane wave assumption, in order to accommodate a variety of situations where the speaker can be close to or far from the microphones (near-field or far-field). A plane wave assumption would have precluded the near-field situation.

We can now write the free field signal model, where "free field" means that there is neither obstacle (chair, table) nor reflection (wall) on the wave's path. Because we have assumed a non-dispersive medium, the waveform signal $x_m\,(t)$ recorded by the microphone at $\boldsymbol{\ell}_m$, where t is expressed in sampling periods[1], can be written:

$$x_m\,(t) \quad = \quad A\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right) \cdot x^{\mathrm{ps}}[t - \mathrm{TOF}\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right)] \tag{3.2}$$

where $x^{\mathrm{ps}}\,(t)$ is the waveform signal that would be recorded by a microphone next to the speaker's mouth, $A\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right) > 0$ is an amplitude gain factor, and $\mathrm{TOF}\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right)$ is the time of flight from the acoustic point source to the microphone, expressed in sampling periods. Alternatively, the signal model can be expressed in the frequency domain:

$$X_m^{(t)}\,(k) \quad = \quad A\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right) \cdot X^{(\mathrm{ps},t)}\,(k) \cdot e^{-j \cdot \pi \frac{k-1}{\mathrm{N_F}} \cdot \mathrm{TOF}\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right)} \tag{3.3}$$

Amplitude gain: If we neglect the absorption of acoustic energy by the aerial medium, the spherical wave assumption implies that the total energy carried by the spherical wavefront is constant. Given that the surface of a sphere is proportional to the square of its radius, we obtain the so-called "inverse square law". This law can be equivalently expressed in terms of signal amplitude:

$$A\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right) \quad = \quad \frac{A_1}{\left\|\boldsymbol{\ell}^{\mathrm{ps}} - \boldsymbol{\ell}_m\right\|} \tag{3.4}$$

where $\|\cdot\|$ is the L^2-norm and $A_1 > 0$ is a constant.

Time Of Flight (TOF): The spherical wave assumption implies that the TOF is directly proportional to the distance between the point source location $\boldsymbol{\ell}^{\mathrm{ps}}$ and the microphone location $\boldsymbol{\ell}_m$. We thus define the *theoretical* TOF, expressed in sampling periods:

$$\mathrm{TOF}\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right) \quad = \quad \frac{\left\|\boldsymbol{\ell}^{\mathrm{ps}} - \boldsymbol{\ell}_m\right\|}{c} \cdot f_s \tag{3.5}$$

where f_s is the sampling frequency in Hertz.

In this simplified model, we have assumed a "free field", that is the *absence* of any obstacle and/or any reflection. The free field assumption may not always hold, because real environments contain

[1]As explained in Section 2.2, in this thesis, time variables (t, τ, TOF, etc.) are expressed in sampling periods, and can be real-valued. The corresponding time value in seconds is $\frac{t}{f_s}$, where f_s is the sampling frequency.

surfaces that can be reflecting such as walls and human skin (Conti et al., 2006), or absorbing such as clothes (Conti et al., 2006). However, modelling reverberations is still an open research issue (Gannot et al., 2006), therefore the free field assumption is quite usual in many existing localization studies (Krim and Viberg, 1996; DiBiase et al., 2001; Claudio and Parisi, 2001; Chen et al., 2006). In this thesis, we thus use the free field assumption in most chapters, basing our work on a reverberation-resistant use of the free field assumption (see the next subsection).

3.1.2 Microphone Arrays for Localization

In this thesis, we define "instantaneous localization" as the task to locate the various active acoustic sources in physical space, *from a single, short time frame* on which speech is considered as stationary (typically about 20 to 30 ms). This implicitly means that we do *not* use time averaging across multiple time frames, for example to estimate the instantaneous cross-correlation $\mathbf{G}_{x,y}^{(t)}$. This choice is justified by the very dynamic natures of the speech signal *and* of the human motion, as already shown in (DiBiase, 2000, Section 6.6). We will often abbreviate "instantaneous localization" as "localization". In most of the thesis, we focus on human speech, which is a wideband signal. In spontaneous multi-party speech, overlaps occur often (Shriberg et al., 2001), and real indoor environments also contain noise sources (computer, projector, etc.) as well as reverberant walls. Hence, this subsection focuses on the localization of multiple concurrent wideband sources, from a single time frame. "Concurrent" means "active at the same time".

The speed of sound in the air being relatively low in an indoor environment ($c \approx 342\ m/s$), most practical audio localization/tracking applications rely on the small differences between the waves arriving at multiple microphones in multiple known locations, called microphone arrays (see Figure 1.1a for an example). A three-fold inverse problem then arises: from the multiple recorded signals, and their small differences:

- **Detection:** Infer the number of active acoustic sources at any given time (zero, one or more),
- **Localization:** Infer their instantaneous locations in the physical space,
- **Segmentation/Tracking:** Infer their spatio-temporal trajectories over time (tracking across multiple time frames), including the determination of activity/silence periods (segmentation).

In addition to these three tasks, it is also desirable to separate acoustic sources into two groups: speech sources and non-speech sources. As explained in Chapter 1, the main contributions of this thesis involve tasks *associated to* localization, such as detection and tracking. The rest of this subsection reviews existing methods for localization, and clarifies which one is used in this thesis. Detection and tracking are reviewed by Sections 3.1.3 and 3.1.4, respectively.

Given the signal model described by (3.2), (3.4) and (3.5), the above-mentioned "small differences" are tightly linked to the geometrical placement of the microphones. More precisely, the differences are usually measured from several (non-exclusive) viewpoints:

1. Time asynchrony: the value of $\mathrm{TOF}\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_m\right)$ for an acoustic wave travelling between the mouth $\boldsymbol{\ell}^{\mathrm{ps}}$ and the microphone $\boldsymbol{\ell}_m$ is different for different microphones, due to their difference in location $\boldsymbol{\ell}_m$. Audio localization/tracking methods relying on time asynchrony usually require precise knowledge of the microphone array's geometry. However, they do not require any particular knowledge about the room, so they can be used in portable, easy-to-use systems. Omnidirectional microphones are used in most cases. Typical geometries include ULAs and UCAs: a finite number of microphones equally spaced along a line or a circle, respectively (Figures 2.1a and 2.1b). However, a solution particularly designed for meeting rooms is the Huge Microphone Array (HMA) including a large number of microphones on a wall (Silverman et al., 2005). We chose to use UCAs because in the horizontal plane, their characteristics are almost invariant with direction (Fuchs, 2001), therefore imposing no practical constraint on the location of the source. UCAs have most of their discriminative power in azimuth space (Figure 1.1a), but are much less precise in terms of elevation and radius.

2. Level difference: when an object is placed between two microphones – for example the head of a binaural manikin –, the "shadow" cast by the object at higher frequencies – 1500 Hz and above for a 20 cm inter-microphone spacing – determines a noticeable difference of amplitude between the waves received at the two microphones. This is called the Interaural Level Difference (ILD), which is often used in binaural studies (Moore, 1997). However, in microphone arrays there is no object placed between the microphones, so ILDs are out of scope of this thesis. A review of the most important binaural cues, including ILD, can be found in (Baumgarte and Faller, 2003; Faller and Baumgarte, 2003), along with their use in low-bitrate, high-fidelity real-time coding of stereophonic and multichannel signals.

3. Impulse response: the path travelled by the sound from the mouth to the various microphones will vary, depending on the speaker's location and the reflecting/absorbing objects in

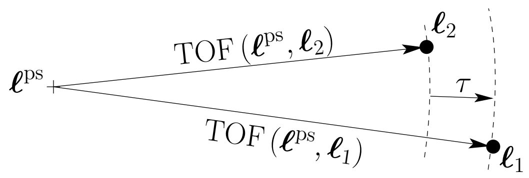


Figure 3.2. Time Delay Of Arrival (TDOA) τ between two microphones at $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$, of a spherical acoustic wave emitted at $\boldsymbol{\ell}^{\text{ps}}$.

the room. Hence, the impulse response from mouth to microphone will vary as well. Assuming the impulse response characteristics of the room to be perfectly known beforehand through a calibration procedure, it is possible to deduce the position of the person. However, the tedious calibration step is often undesirable in a practical application – as in a portable videoconferencing system –, so the impulse responses need to be estimated in an online, automatic fashion. This task is also called blind Multiple Inputs Multiple Outputs (MIMO) channel identification (Chen et al., 2005; Buchner et al., 2004), where geometrical knowledge of the microphone array is not required. This task is tightly linked to the Blind Source Separation approaches (Buchner et al., 2005a). Solving the blind MIMO channel identification problem amounts to retrieve a complete model of the meeting room, which would permit to jointly locate and separate the signals from the various acoustic sources. It is still a difficult, open research issue. A preliminary test of such a method can be found in (Buchner et al., 2005b).

4. Microphone channel: recently, it was proposed to use several directional microphones placed at the same location, but oriented towards different directions (Matsumoto and Hashimoto, 2005). The direction-dependent transfer function of each microphone is assumed to be known, so that the speaker's location can be reconstructed. This type of solution can also be combined with time asynchrony solutions (Rui et al., 2005).

Asynchrony: In the following we focus on the first group of solutions, for which (Krim and Viberg, 1996; Brandstein and Ward, 2001) provide comprehensive introductions. These methods are typically linked, directly or indirectly, to the following observation. Let us consider a point source at location $\boldsymbol{\ell}^{\text{ps}}$, and a pair of microphones at locations $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$, whose continuous time signals are denoted $x_1(t)$ and $x_2(t)$, respectively. Then (3.2) and (3.4) lead to:

$$x_1\left(t + \text{TOF}\left(\boldsymbol{\ell}^{\text{ps}}, \boldsymbol{\ell}_1\right)\right) \quad \propto \quad x_2\left(t + \text{TOF}\left(\boldsymbol{\ell}^{\text{ps}}, \boldsymbol{\ell}_2\right)\right) \tag{3.6}$$

It follows that the continuous time domain cross-correlation function $g_{x_1,x_2}(\tau) = \int_{\mathbb{R}} x_1(\xi)x_2(\xi-\tau)d\xi$ has a maximum at $\tau = \mathrm{TOF}(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_1) - \mathrm{TOF}(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_2)$. See Figure 3.2 for an illustration.

For any pair q of microphones $\left(\boldsymbol{\ell}_{a_q}, \boldsymbol{\ell}_{b_q}\right)$, we thus define the *theoretical* Time Delay Of Arrival (TDOA):

$$\tau_q^{\mathrm{th}}(\boldsymbol{\ell}^{\mathrm{ps}}) \stackrel{\mathrm{def}}{=} \mathrm{TOF}\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_{a_q}\right) - \mathrm{TOF}\left(\boldsymbol{\ell}^{\mathrm{ps}}, \boldsymbol{\ell}_{b_q}\right) \tag{3.7}$$

Note that $\tau_q^{\mathrm{th}}(\boldsymbol{\ell}^{\mathrm{ps}}) \in \mathbb{R}$ is a continuous value expressed in sampling periods (not necessarily an integer). Methods based on the asynchrony between the various microphones can be divided into two types: TDOA methods, which use two steps, and direct methods, which use a single step.

The Time Delay of Arrival (TDOA) methods consist in first estimating the TDOA for each pair of microphones through DFT analysis, second in deriving the location of the source(s) from geometrical considerations, by inverting (3.7), e.g. using the Linear Intersection algorithm (Brandstein, 1995). The main bottleneck is the time delay estimation (TDE) step, which may be affected by reverberations. Thus, TDE is often based on a modified cross-correlation function called GCC-PHAT (Knapp and Carter, 1976), which has reverberation-resistant properties (Gustafsson et al., 2003). For a time frame t, GCC-PHAT-based TDE consists in finding the value of τ that maximizes the GCC-PHAT continuous time domain cross-correlation function $\hat{g}_{x,y}^{(t)}(\tau)$, as detailed below.

In the frequency domain, for any two continuous signals $x_1(t)$ and $x_2(t)$, GCC-PHAT is defined as the DFT-based instantaneous cross-correlation estimate divided by its own magnitude[2]:

$$G_{x_1,x_2}^{(\mathrm{PH},t)}(k) \stackrel{\mathrm{def}}{=} \frac{G_{x_1,x_2}^{(t)}(k)}{\left|G_{x_1,x_2}^{(t)}(k)\right|} = \frac{X_1^{(\mathrm{zp},t)}(k) \cdot \left[X_2^{(\mathrm{zp},t)}(k)\right]^*}{\left|X_1^{(\mathrm{zp},t)}(k) \cdot \left[X_2^{(\mathrm{zp},t)}(k)\right]^*\right|} \tag{3.8}$$

The discrete time domain GCC-PHAT vector $\mathbf{g}_{x_1,x_2}^{(\mathrm{PH},t)} \in \mathbb{R}^{2N_{\mathrm{F}}}$ is obtained through IDFT of $\mathbf{G}_{x_1,x_2}^{(\mathrm{PH},t)} \in \mathbb{C}^{2N_{\mathrm{F}}}$:

$$\mathbf{g}_{x_1,x_2}^{(\mathrm{PH},t)} \stackrel{\mathrm{def}}{=} \mathbf{F}^{-1} \cdot \mathbf{G}_{x_1,x_2}^{(\mathrm{PH},t)} \tag{3.9}$$

[2] Zero-padding is used to avoid circularity issues, as explained in Section 2.2 and defined by (2.14).

The GCC-PHAT continuous time domain cross-correlation function $\hat{g}^{(\mathrm{PH},t)}_{x_1,x_2}(\tau)$ is defined as:

$$\hat{g}^{(\mathrm{PH},t)}_{x_1,x_2}(\tau) \stackrel{\text{def}}{=} \begin{cases} g^{(\mathrm{PH},t)}_{x_1,x_2}(\tau+1) & \text{if } \tau \in \{0,1,\cdots,N_{\mathrm{F}}\} \\ g^{(\mathrm{PH},t)}_{x_1,x_2}(\tau+1+2N_{\mathrm{F}}) & \text{if } \tau \in \{-N_{\mathrm{F}}+1,-N_{\mathrm{F}}+2,\cdots,-1\} \\ \text{through upsampling} & \text{if } \tau \notin \mathbb{Z} \end{cases} \tag{3.10}$$

In this thesis, an upsampling factor[3] of 20 is used, unless stated otherwise. For a microphone pair q of signals $x_{a_q}(t)$ and $x_{b_q}(t)$, GCC-PHAT-based TDE then consists in finding the peak of the GCC-PHAT continuous time domain cross-correlation function:

$$\hat{\tau}^{(t)}_q \stackrel{\text{def}}{=} \arg\max_{\tau} \left[\hat{g}^{(t)}_{x_{a_q},x_{b_q}}(\tau)\right] \tag{3.11}$$

In fact, the GCC-PHAT time-delay estimator was shown to be highly advantageous over other estimators, when reverberations are present, from both theoretical and experimental points of view (Gustafsson et al., 2003). In the case of multiple sources and one microphone pair, an alternative method for the estimation of multiple time-delays is the Adaptive Eigenvalue Decomposition Algorithm (Benesty, 2000), which relies on the eigenanalysis of the covariance matrix of the two signals in one pair of microphones. However, in the case of multiple microphone pairs, multiple sources and multiple sound paths (reverberations), it is not obvious how to pair the various time-delays observed at the various pairs of microphones in order to deduce the exact location of the acoustic sources.

In general, TDE constitutes a bottleneck in the sense that any wrong estimate $\hat{\tau}^{(t)}_q$ leads to increase the error in the subsequent location estimation, that is when inverting (3.7). As already discussed in (DiBiase, 2000, Section 6.6), TDE from a single time frame is not advisable. On the other hand, as explained above, using several consecutive time frames, e.g. through averaging of (3.8), is precluded by the very dynamical natures of speech and of human motion.

The direct methods avoid this bottleneck by directly inferring the source(s) locations from the measured signal. They can be divided into two groups: Coherent Signal Subspace Processing (CSSP) and Steered Response Power (SRP). CSSP methods (Wang and Kaveh, 1985; Stoica and Mose, 1997; Claudio and Parisi, 2001) are extensions of narrowband methods originated in the fields

[3] An upsampling factor N means that N-1 zeros are interleaved between any two consecutive samples. The upsampled signal is always low-passed before any further use. The upsampling + low-pass operation approximates a sine cardinal interpolation (Oppenheim et al., 1999).

of radar and communications (Schmidt, 1986). Examples are the well-known MUSIC (Schmidt, 1986) and ESPRIT (Roy and Kailath, 1989) algorithms, which typically achieve higher resolution than SRP methods. However, these methods were originally designed for narrowband signals and Uniform Linear Arrays (ULAs). Previous work extended CSSP approaches from ULAs to UCAs (Tewfik and Hong, 1992), from narrowband to wideband signals (Su and Morf, 1983), and both (Friedlander and Weiss, 1993; Doron et al., 1993). Only the latter (Friedlander and Weiss, 1993; Doron et al., 1993) are relevant to our problem. Globally, coherent signals such as speech and its reverberations still seem to be a problem with these methods, since reverberations have to be modeled explicitly. Although they allow in theory to estimate jointly the number of sources and their locations, they suffer from sensitivity to reverberant environments. Moreover, they need sufficient amounts of data, which means either long time frames or averaging across several time frames. This is fine with somewhat stationary signals such as vehicle noise (Pham and Fong, 1997), but may be difficult with speech signals (DiBiase, 2000). Also, in the case of (Friedlander and Weiss, 1993), steering matrices have to be defined for each sector of the space. For example, one could define 18 sectors around a microphone array, each sector spanning a 20 degree-azimuth angle. Finding which sector(s) of the space contain active acoustic source(s) is an open issue.

SRP localization methods, also known as beamforming localization methods, are a reasonable alternative to CSSP methods. Indeed, SRP methods can work with a single time frame, and are less sensitive to modelling assumptions (DiBiase et al., 2001). The drawback is a slight decrease of localization precision with respect to CSSP methods. The idea is to estimate the power at any location $\boldsymbol{\ell}^{\text{ps}}$ in space by "steering the array", that is to compensate for the corresponding differences of Time Of Flight $\text{TOF}\left(\boldsymbol{\ell}^{\text{ps}}, \boldsymbol{\ell}_m\right)$ between the microphones $\{\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_m, \cdots, \boldsymbol{\ell}_{N_m}\}$ (Krim and Viberg, 1996). Multiple simultaneous sources will be reflected by multiple power maxima across the search space. However, reverberations will also appear as power maxima ("virtual sources"). A partial solution to this issue is to combine the flexibility of SRP methods with the robustness to reverberations of the PHAT: this is known as SRP-PHAT (DiBiase, 2000). Let us assume N_m microphones, with frequency domain signals $X_m(k)$, for $m \in \{1, \cdots, N_m\}$, and a point source at an unknown location $\boldsymbol{\ell}^{\text{ps}} \in \mathbb{R}^3$, emitting a wideband signal (speech). SRP-PHAT consists in finding the location $\hat{\boldsymbol{\ell}}^{(t)} \in \mathbb{R}^3$ that maximizes the SRP-PHAT power (DiBiase, 2000, Section 6.5):

$$\hat{\boldsymbol{\ell}}^{(t)} = \arg\max_{\boldsymbol{\ell}} \left[\mathrm{P}_{\text{SRP-PHAT}} \left(\boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_M^{(t)} \right) \right] \tag{3.12}$$

where the SRP-PHAT power is obtained by aligning (see (3.3)) and summing the magnitude-normalized signals, in the frequency domain, on the strictly positive discrete frequencies $\{2, \cdots, N_{\mathrm{F}}+1\}$:

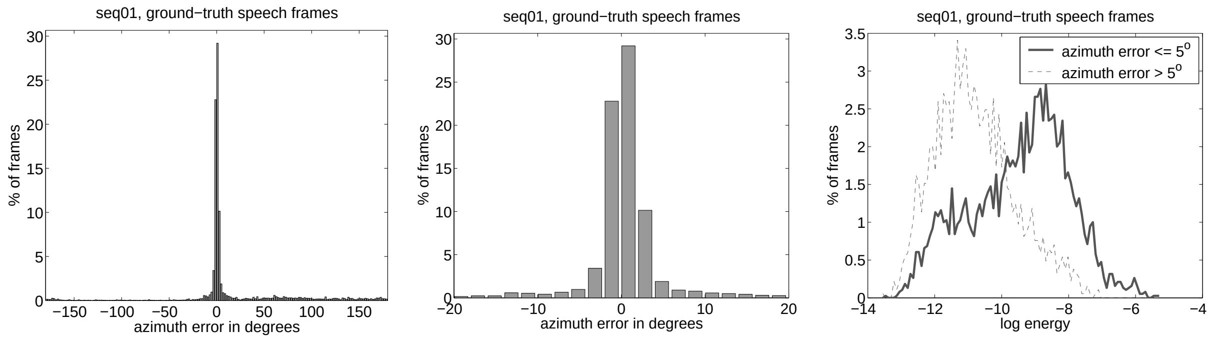
$$\mathrm{P}_{\text{SRP-PHAT}}\left(\boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_M^{(t)}\right) \stackrel{\text{def}}{=} \sum_{k=2}^{N_{\mathrm{F}}+1} \left| \sum_{m=1}^{N_m} \frac{X_m^{(t)}(k)}{\left|X_m^{(t)}(k)\right|} e^{j\pi \frac{k-1}{N_{\mathrm{F}}} \mathrm{TOF}(\boldsymbol{\ell}, \boldsymbol{\ell}_m)} \right|^2 \tag{3.13}$$

An interesting comparison between time averaging – which improves the resolution of GCC-PHAT-based TDOA methods – and spatial averaging – realized in SRP-PHAT is presented in (DiBiase, 2000, Section 6.6). In cases where the speakers may suddenly move a lot, it is preferable *not* to average across consecutive time frames. In this thesis, we thus opted for the SRP-PHAT method.

Search space: In general, the main drawback of most direct methods – both CSSP and SRP – is that the search space can be large (e.g. the whole room in the case of meetings). As suggested in (Friedlander and Weiss, 1993; Fuchs, 2001), one could thus think of discretizing the search space in a few volumes called "sectors". Each sector could be classified as "active" or "inactive", and precise localization could then be applied within active sectors only, as in Double Hierarchical BeamForming (DHBF) (Duraiswami et al., 2001; Zotkin and Duraiswami, 2004). DHBF uses the spectral data from *all* sources to take *each* sector-based decision (active or inactive), which may lead to some noisy decisions. Section 5.2 investigates an alternative to DHBF, that is based on statistical observations of human multi-party speech (Roweis, 2003): for a given discrete frequency, only one speech source is assumed to be dominant in terms of magnitude, and other sources can be neglected.

Three points can summarize this review:

- When localizing acoustic sources from a single time frame, direct methods are more adequate than TDOA methods.
- Within direct methods: although the CSSP methods and, more recently, the MIMO/Blind Source Separation methods have seen very promising developments, they are still not as practically effective as the SRP methods. This thesis is based on the SRP-PHAT method.
- There is a need for a fast and effective search space reduction method, called "sector-based detection-localization" in the following. Within each active sector $\mathbb{S} \subset \mathbb{R}^3$ of space, a classical "point-based" localization method can then be used to produce a point location estimate $\boldsymbol{\ell} \in \mathbb{R}^3$.



(a) (b) = zoom of (a) (c)

Figure 3.3. SRP-PHAT point-based search on `seq01` (single human speaker): (a) shows the histogram of azimuth errors; (b) shows a zoom of (a); (c) shows the histogram of log energy values.

Consequently, Chapter 5 proposes a topological interpretation of SRP-PHAT that serves as a principled foundation for both sector-based detection-localization and point-based localization within the active sectors.

The next two subsections review tasks associated with localization: detection for localization, and tracking (usually viewed as the filtering of instantaneous location estimates).

3.1.3 Detection for Localization

In the above review of localization methods, we have implicitly assumed that *we knew on which time frames acoustic sources should be localized*. Indeed, a received signal contains not only speech but also silences, and it is preferable *not* to estimate any source location on mostly silent time frames, because the resulting location estimate would be noisy or even completely meaningless. To discriminate between speech time frames and silence time frames, traditional Voice Activity Detectors (VADs) typically use single channel features, such as energy and zero-crossing rate. Although well adapted to single channel tasks such as automatic speech recognition, they are suboptimal in terms of localization precision. To highlight this fact, we ran a simple SRP-PHAT point-based single source localization algorithm (detailed in (Gatica-Perez et al., 2003)) on *all* time frames of `seq01` in the AV16.3 Corpus (single human speaker). Figures 3.3a and Figure 3.3b show the distribution of azimuth errors for frames labelled as “speech” in a ground-truth created by a human listener.

These figures can be interpreted as follows:

- On frames containing speech strong enough to be localized, a maximum error of about 5 degrees is achieved, as compared with the true azimuth of the source.
- On frames containing silence or weak speech, the error can be seen as the result of a uniform random process.

A commonly used strategy to select reliable frames for localization is to select frames with high energy only, and to ignore other frames. However, we can see on Figure 3.3c that in terms of energy, there is a large overlap between the two groups "correctly localized" and "incorrectly localized". Thus, energy is not necessarily adapted to the task of detection for acoustic source localization. Alternative methods that rely on the cross-correlation between channels are investigated in Section 5.2.

Let us now assume that we have selected a feature for detection, called "activeness" (energy or other). For each time frame t, the activeness is a scalar value $\zeta_t > 0$, where higher values indicate that localizable acoustic activity is most likely, and lower values indicate that background noise is most likely. In order to build an integrated system, evaluating the activeness ζ_t is not sufficient. Indeed, an additional "hard decision" step is needed, which determines whether there is acoustic activity or not, at each time frame t. This hard decision can be taken by comparing the activeness ζ_t to a threshold Ψ_ζ, set *beforehand*. For example, the end-user would like to have a fixed proportion of falsely detected sound sources, while having a proportion of missed sound sources as small as possible. This means that the False Alarm Rate (FAR, formally defined in Appendix A):

$$\mathrm{FAR} \stackrel{\mathrm{def}}{=} \frac{\text{number of false alarms}}{\text{number of silent time frames}} \tag{3.14}$$

should be constant and equal to a target value: $\mathrm{FAR} = \mathrm{FAR_T}$. We will see in Chapter 5 that setting the detection threshold Ψ_ζ *a priori*, to optimize a desired detection performance metric, and keeping it fixed thereafter, is not necessarily the best solution. Indeed, conditions may vary after the threshold was set: there may be more or less background noise, more or less silence, several speakers or a single one.

Instead, one could think of *adapting* the threshold value Ψ_ζ *to each condition separately*. In a probabilistic modelling context, let us assume *perfect knowledge* of the *types* of pdfs (Gaussian, Gamma, Rice etc.) for ζ_t on active periods, as well as on silence periods. A two-component model $\mathcal{M}$ of the observed ζ_t values can then be built, that has one component for silence and the other for acoustic activity. *On each condition separately,* the parameters $\mathbf{\Lambda}(\mathcal{M})$ are estimated jointly with the threshold value Ψ_ζ, in order to fulfill a user-defined target (e.g. $\mathrm{FAR}(\mathbf{\Lambda}(\mathcal{M}), \Psi_\zeta) = \mathrm{FAR}_\mathrm{T}$). If the types of pdf are *perfectly* known, then $\mathbf{\Lambda}(\mathcal{M})$ and Ψ_ζ can be optimally selected with the Neyman-Pearson and the competitive Neyman-Pearson approaches (Levitan and Merhav, 2002). In practice, the types of pdfs are seldom *perfectly* known, therefore the true $\mathrm{FAR}(\mathbf{\Lambda}(\mathcal{M}), \Psi_\zeta)$ cannot be estimated exactly, and correction procedures are needed, as investigated by Section 5.3.

Once a positive detection decision is taken for a time frame t, the instantaneous localization methods reviewed in Section 3.1.2 can be applied to provide one or more instantaneous location estimates (azimuth angles in the case of a UCA).

3.1.4 Tracking

Tracking can be viewed as the task of filtering instantaneous location estimates provided by the methods mentioned in Section 3.1.2. The Kalman filter (Kalman, 1960; Welch and Bishop, 2004) assumes dynamics to be linear and Gaussian. These assumptions become an issue when dealing with human motion (non linearities such as sharp turns). Moreover, in spontaneous multi-party speech, utterances are often short (typically less than a second), speaker turns are often short as well, and overlaps represent a non-negligible portion of speech (Shriberg et al., 2001).

The Extended Kalman Filter (EKF) was proposed to accommodate non-linear dynamics through a linearization step (Sorenson, 1985), however it is known to be practically difficult to tune its parameters (Julier and Uhlmann, 1997). More recently, the Unscented Kalman Filter (UKF) was proposed to avoid this linearization step and to accommodate non-Gaussian measurement noise sources (Julier et al., 1995; Julier and Uhlmann, 1997; LaViola, 2003). For a recent application of the UKF to acoustic source localization, see (Dvorkind and Gannot, 2005). However, these approaches may encounter difficulties when dealing with spontaneous multi-party speech, which is both highly changing in space (speaker turns) and sporadic over time (short utterances).

As an alternative, Sequential Monte-Carlo (SMC) methods approximate the optimal Bayesian filter by representing each probabilistic distribution through a finite set of particles, as in Particle Filtering (PF) (Gordon et al., 1993). Applications to single acoustic source localization and tracking can be found in (Vermaak and Blake, 2001; Ward and Williamson, 2002; Ward et al., 2003), and a comprehensive review in (Lehmann, 2004). However, the fast-changing speaker turns encountered in spontaneous multi-party speech require either specific multisource models (Larocque et al., 2002) or adapting the single-source model to "switching between speakers" situations (Lehmann, 2005). Estimating the number of active speech sources is still an issue, tightly linked to the data association issue. Although Particle Filters can model multiple objects via multi-modal distributions, deciding which modes are significant and which objects they belong to is an open issue (Vermaak et al., 2003). Moreover, when the number of active objects varies very often along time, as in fast-changing speaker turns, complex birth/death rules are needed.

Chapter 6 investigates an alternative approach called "short-term clustering", where the number of active speech sources does not need to be known. Initial results were presented in (Lathoud et al., 2004), on the multi-party speech segmentation and the multi-object tracking tasks.

3.2 Multi-Party Speech Segmentation

While many existing speaker segmentation/clustering works involve single-channel, broadcast-like environments (Sugiyama et al., 1993; Chen and Gopalakrishnan, 1998; Ajmera and Wooters, 2003; Galliano et al., 2005; Valente, 2006), there are less established standards to deal with spontaneous multi-party speech. This section first defines what the "segmentation" task means within this thesis, and describes the challenges inherent to multi-party spontaneous speech. Next, a review is made of existing works within this scope. Finally, a preliminary meeting segmentation experiment is summarized, which uses location to segment meeting speech.

3.2.1 The Task

One application of the techniques developed in this thesis is the segmentation of spontaneous multi-party speech. In this thesis, by "segmentation" we mean, for each speaker, to determine the periods of time when he is silent, and the periods of time when he is speaking. This thesis thus focuses on speech/silence segmentation for each speaker, see Figure 3.4(top) for an example of such segmentation. This thesis does not address higher-level annotation such as sentence boundaries, emotion, interest level, meeting acts etc., as done for example in (Dielmann and Renals, 2004; McCowan

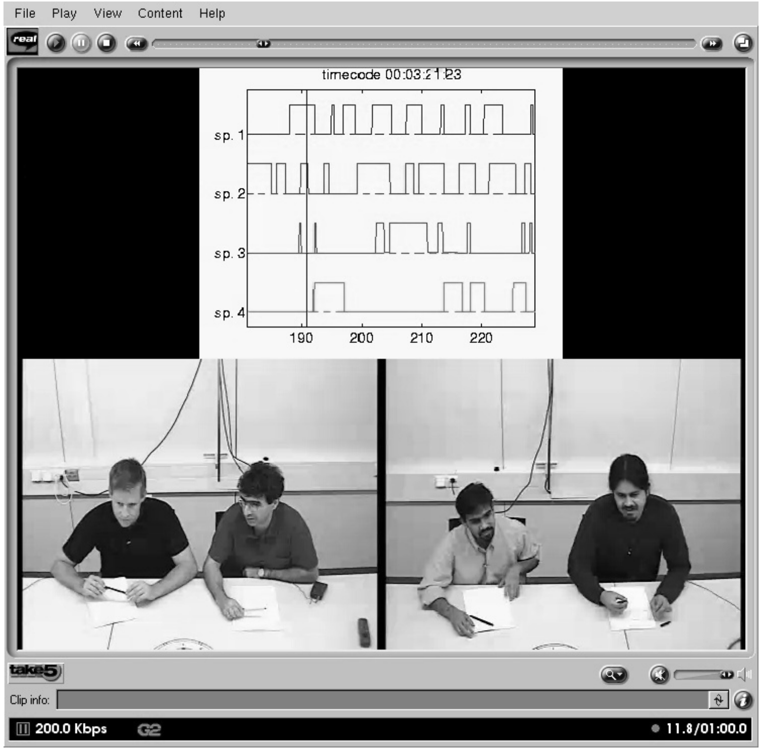


Figure 3.4. Example of segmentation result (top) for a 4-people meeting (bottom). The top figure depicts a speech/silence segmentation for each speaker, with sometimes multiple speakers active at the same time (overlaps).

et al., 2005). However, a precise speaker "segmentation" in terms of speech and silence can be a useful platform for further, higher-level annotation, as in (McCowan et al., 2005).

At first glance, speech/silence segmentation may look like a simple task. Comparing the short-term energy to a threshold comes to mind as an easy, efficient technique. However, speech/silence segmentation is particularly difficult in real multi-party speech, because each speech utterance can be very short ("sporadic" events), which makes "smoothing" difficult, and people often talk over each other ("concurrent" events), as in overlapped speech. In (Shriberg et al., 2001) it was identified that around 10-15% of words, or 50% of contiguous speech segments, in a meeting or telephone conversation contain some degree of overlapping speech. These overlap segments are problematic for speech recognition, producing an absolute increase in Word Error Rate of between 15-30% using close-talking microphones for a large vocabulary task (Shriberg et al., 2001; Morgan et al., 2001; Cetin and Shriberg, 2006). For applications that involve meetings or teleconferences, it is thus important to not only segment the audio into single speaker turns, but also to identify segments of overlapping speech along with their constituent speakers.

	Mean	Std. dev.
Lapel SNR	18.7	2.36
Mic. array SNR	10.7	2.02

Table 3.1. Average SNR across meetings and speakers of the M4 Corpus (McCowan et al., 2005), in dB domain. The lapels are worn by each speaker, below the throat. Each speaker is between 0.8 and 3 meters from the microphone array. The average SNR is defined in Section 2.4.

The sporadic and concurrent nature of spontaneous multi-party speech utterances makes the segmentation task different from other contexts, where each segment has a long duration, and segments are assumed not to overlap, such as Speech/Music segmentation (Williams and Ellis, 1999; Ajmera et al., 2002) and speaker segmentation/clustering in broadcast news (Sugiyama et al., 1993; Chen and Gopalakrishnan, 1998; Ajmera and Wooters, 2003; Galliano et al., 2005; Valente, 2006). The systems used for automatic speaker clustering and segmentation of broadcast news are difficult to apply "as is" in a meeting situation because of the sporadic and concurrent nature of utterances in spontaneous speech:

- Sporadic events: the creation of reliable speaker models, for example with Gaussian Mixture Models (GMMs), is difficult to achieve with utterances shorter than 2 or 3 seconds. Therefore, minimum duration constraints are often included in the modelling. Unfortunately, many spontaneous speech utterances are shorter than 2 seconds.
- Concurrent events: in a lively conversation, not only speaker turns are very short, but they also tend to talk over each other. This is known as "overlapped speech". The above-mentioned minimum duration constraint would lead to incorporate in a segment not only a short utterance but also parts of *other speakers'* utterances, just before and just after. In an agglomerative speaker clustering framework such as (Ajmera, 2004; Galliano et al., 2005), this may well lead to undesirable clusters containing speech from more than one speaker.
- Non-speech concurrent events are also received, including body noises from the participants (body motion, chair motion, paper shuffling, breathing, etc.) and machine noises (fans, projector, etc.). These noises degrade the quality of the signals, therefore spurious clusters appear, and errors are made during the agglomerative speaker clustering.

"Invasive" methods can efficiently detect and segment spontaneous speech from multiple speakers in a meeting (Wrigley et al., 2005), where "invasive" means that a close-talking microphone is *attached* to each speaker (lapel near the throat, or headset near the mouth). Close-talking microphones permit to know precisely when each speaker is talking, because their signals are much cleaner than those received by distant microphones, due to the difference of distance (see the difference in Signal-to-Noise Ratio (SNR) in Table 3.1). However, the range of applications permitted by close-talking microphones is limited, because (1) they require each user to wear a microphone, and (2) they practically provide no information about the location of each speaker. In this thesis, as explained in Chapter 1, the focus is rather on "non-invasive" methods, where only distant microphones are used.

3.2.2 Location for Segmentation

For each pair q of microphones on a table, we can detect the peaks $\hat{\tau}_q^{(t)}$ of the cross-correlation function between the two microphones $\left(\boldsymbol{\ell}_{a_q}, \boldsymbol{\ell}_{b_q}\right)$, using (3.11). Each peak $\hat{\tau}_q^{(t)}$ should be close to the theoretical TDOA $\tau_q^{\mathrm{th}}\left(\boldsymbol{\ell}^{\mathrm{ps}}\right)$ between the two microphones, determined by the location of the speaker $\boldsymbol{\ell}^{\mathrm{ps}}$, as defined in (3.7) and illustrated by Figure 3.2 above. By combining the observed delays from multiple pairs of microphones, it is possible to efficiently detect who is speaking in a meeting, with prior knowledge of the number of speakers N_{ps} and their approximate locations $\left(\boldsymbol{\ell}_1^{\mathrm{ps}}, \cdots, \boldsymbol{\ell}_{N_{\mathrm{ps}}}^{\mathrm{ps}}\right)$ (Lathoud and McCowan, 2003), or without this prior knowledge (Ellis and Liu, 2004). Such location-based approaches for multi-party speech segmentation efficiently deal with the two difficulties mentioned above: (1) short speech utterances are well detected and precisely segmented, (2) including on intervals of overlap between two or more speakers.

However, a central assumption in these two works is that each speaker does not move much around his/her location. Although this is generally a reasonable assumption in meetings, it is desirable to allow speakers to change location over time, to permit a wider range of applications. The speech segmentation task then becomes more difficult, as we cannot rely on a one-to-one mapping between speakers and locations. In general, location is thus not sufficient to determine who spoke when, because a given speaker may move while being silent, or multiple speakers may share locations over time[4]. Addressing this task in the context of moving speakers could be useful in

[4]Even in a meeting, a person is likely to stand up and *move* to a screen for a presentation. People may also swap seats.

meetings themselves, as we would like to know for example when a given speaker stood up and made a presentation, for enriched recording browsing experience. Surveillance applications could also benefit from efficiently tracking a moving speaker. This versatility could not be achieved with a purely static analysis of instantaneous location estimates for speech segmentation, as for example K-means, or the static criteria used in (Lathoud and McCowan, 2003; Ellis and Liu, 2004). In this context, Chapter 6 addresses speech segmentation and short-term tracking, and Chapter 7 addresses the determination of long-term speaker identities. Both rely on instantaneous localization of multiple speakers, which is addressed by Chapter 5.

Segmenting meeting speech based on location is a relatively recent field, as compared to the whole field of speech processing. Thus, the next section summarizes a preliminary experiment.

3.2.3 A Preliminary Experiment

This subsection summarizes a preliminary experiment that uses speaker location for meeting segmentation, where we assume the number of speakers and their approximate location already known. This was a joint work with Dr Iain McCowan, while he was at IDIAP. Full details can be found in (Lathoud and McCowan, 2003). All test data and ground-truth annotations are freely available at: `http://mmm.idiap.ch/Lathoud/2003_locbasedseg/`
and: `http://mmm.idiap.ch/Lathoud/2003_locbasedseg-lpcc/`

We propose a technique that segments a meeting into speaker turns based on their location, essentially implementing a discrete source tracking system. In many multi-party conversations, such as meetings or teleconferences, the location of participants is restricted to a small number of regions, such as seats around a table. In such cases, segmentation according to these discrete regions would be a reliable means of determining speaker turns. We propose a system that uses microphone pair time delay estimates ($\hat{\tau}_q^{(t)}$) as features to represent speaker locations. For each pair q of microphones $(\boldsymbol{\ell}_{a_q}, \boldsymbol{\ell}_{b_q})$ of an array, and each time frame t, a single delay $\hat{\tau}_q^{(t)}$ is estimated from GCC-PHAT, as in (3.11). Each time delay estimate measures the difference in the time of arrival between the signals on a microphone pair. Gaussian distributions are used to model the behaviour of the observed features around a number of speaker locations. These then form the state distributions in a Hidden Markov Model (HMM), which can be used to obtain a maximum likelihood segmentation into speaker turns. The discrimination provided by the location features, coupled with the HMM's ability to model sequences, makes it possible to extend the system to segment the conversation in terms of higher level structure. To demonstrate this, we propose an extension to handle the case of overlapping speech from multiple simultaneous speakers. As mentioned in Section 3.2.1, such speaker overlap has been identified as a significant problem for speech segmentation and recognition of multi-party conversations (Shriberg et al., 2001). The proposed location-based speaker segmentation system is assessed on real recordings from a 4-element microphone array ($N_m = 4, N_q = 6$, see Figure 3.7) in a meeting room. Results are presented comparing the performance of the location features to standard Linear Prediction Cepstral Coefficients (LPCC) features for single speaker segments. In addition, experiments on overlapping speech segments demonstrate the success of the proposed extension to handle dual-speaker overlap.

Assumption: As the basis of our model, we assume the number of speakers $N_{\rm ps}$ known, where each speaker $n \in \{1, \cdots, N_{\rm ps}\}$ is confined to a physical region centered at a known location $\boldsymbol{\ell}_n^{\rm ps} \in \mathbb{R}^3$.

Feature Space: Each estimated time delay $\hat{\tau}_q^{(t)}$ is given by (3.11). A feature space is defined as the vector of GCC-PHAT TDOA estimates (expressed in sampling periods) across N_q pairs of microphones, at time frame t:

$$\hat{\boldsymbol{\tau}}^{(t)} \stackrel{\rm def}{=} \left[\hat{\tau}_1^{(t)}, \cdots, \hat{\tau}_q^{(t)}, \cdots, \hat{\tau}_{N_q}^{(t)}\right]^{\rm T} \tag{3.15}$$

Theoretical Delays: Each theoretical time delay $\tau_q^{\rm th}\left(\boldsymbol{\ell}_n^{\rm ps}\right)$ is given by (3.7). For each speaker location $\boldsymbol{\ell}_n^{\rm ps}$, we define the associated vector of theoretical delays:

$$\boldsymbol{\tau}^{\rm th}\left(\boldsymbol{\ell}_n^{\rm ps}\right) \stackrel{\rm def}{=} \left[\tau_1^{\rm th}\left(\boldsymbol{\ell}_n^{\rm ps}\right), \cdots, \tau_q^{\rm th}\left(\boldsymbol{\ell}_n^{\rm ps}\right), \cdots, \tau_{N_q}^{\rm th}\left(\boldsymbol{\ell}_n^{\rm ps}\right)\right]^{\rm T} \tag{3.16}$$

Proposed Model: For each speaker location $\boldsymbol{\ell}_n^{\rm ps}$, a Gaussian distribution is assumed:

$$\hat{\boldsymbol{\tau}}^{(t)} | \boldsymbol{\ell}_n^{\rm ps} \sim \mathcal{N}_{\boldsymbol{\tau}^{\rm th}\left(\boldsymbol{\ell}_n^{\rm ps}\right), \boldsymbol{\Sigma}} \tag{3.17}$$

where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix, with a variance of 1 sampling period for each pair: $\boldsymbol{\Sigma} = \mathbf{I}$.

HMM Segmentation Framework: To segment the audio signal according to speaker turns, we use a HMM framework similar to that proposed in (Ajmera et al., 2002) for speech/music segmentation. We define a minimum duration left-to-right HMM for each speaker $n \in \{1, \cdots, N_{\rm ps}\}$, where

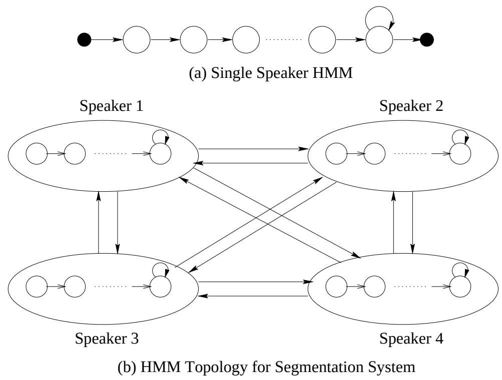


Figure 3.5. Single speaker HMM topology.

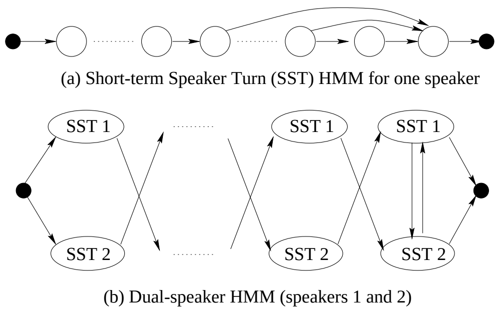


Figure 3.6. Dual-speaker HMM topology.

each state is modelled with the Gaussian pdf $p\left(\hat{\boldsymbol{\tau}}_t \mid \boldsymbol{\ell}_n^{\mathrm{ps}}\right)$ defined by (3.17). This single speaker HMM topology is shown in Figure 3.5a. A grammar is introduced that defines uniform transitions between all speakers, and excludes self-loops. The resulting HMM for the segmentation system is shown in Figure 3.5b for the case of $N_{\mathrm{ps}} = 4$ speakers. Given an observation sequence of feature vectors $\hat{\boldsymbol{\tau}}^{(1:T)} \stackrel{\text{def}}{=} \left(\hat{\boldsymbol{\tau}}^{(1)}, \cdots, \hat{\boldsymbol{\tau}}^{(t)}, \cdots, \hat{\boldsymbol{\tau}}^{(T)}\right)$, the optimal path through the HMM can be found using Viterbi decoding, giving the maximum likelihood segmentation in terms of speaker locations.

Extension to Segments of Speaker Overlap: We propose a dual-speaker HMM topology which can be used to extend the HMM segmentation framework to handle segments of overlapping speech. If there are N_{ps} individual speakers, an overlap segment may be defined as one in which there are n active speakers, where $2 \leq n \leq N_{\mathrm{ps}}$. We restrict this current work to the case of $n = 2$, which we will refer to as dual-speaker overlap.

On segments of dual-speaker overlapped speech, empirical observation of TDOA estimates $\hat{\tau}_q^{(t)}$ shows an alternating sequence of Short-term Speaker Turns (SST's). Each TDOA estimate $\hat{\tau}_q^{(t)}$ successively matches the theoretical TDOA $\tau_q^{\mathrm{th}}\left(\boldsymbol{\ell}_{n_1}^{\mathrm{ps}}\right)$ of a speaker n_1 during a few frames (SST 1 in Figure 3.6b), then the theoretical TDOA $\tau_q^{\mathrm{th}}\left(\boldsymbol{\ell}_{n_2}^{\mathrm{ps}}\right)$ of the other speaker n_2 during a few frames (SST 2 in Figure 3.6b).

These SST's are due to frame-by-frame variations in relative energy levels between the two speakers, as the TDOA estimates are computed from the highest GCC-PHAT peak in each frame. To model this behaviour, we first define a left-to-right HMM that represents a SST, shown in Figure 3.6a. This model imposes a minimum duration to exclude noise, as well as a maximum duration to exclude single-speaker segments. For a given pair of speakers, an alternation of two SST

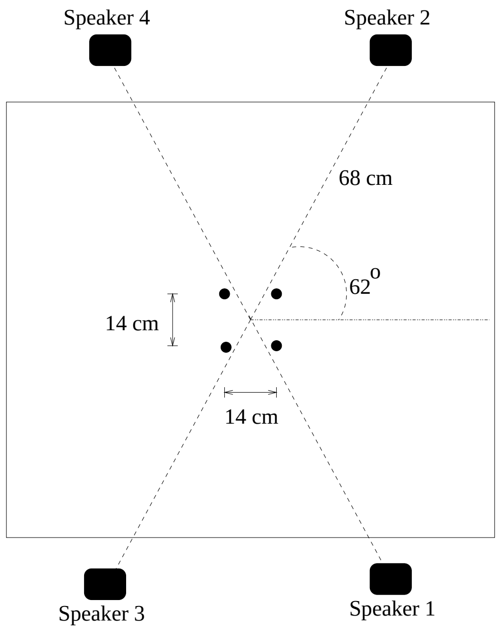


Figure 3.7. Experimental setup.

models forms a dual-speaker HMM, as shown in Figure 3.6b. Similarly to the left-to-right part of the single speaker HMM, a minimum duration constraint is included to eliminate undesired short overlapped speech segments (series of diagonal transition arrows). Similarly to the self-loop in the single speaker HMM, the two vertical transition arrows on the right side of Figure 3.6b model a variable total duration for the dual-speaker HMM.

Subsequently, an audio signal containing a series of single speaker and dual-speaker segments may be segmented using :

- N_{ps} single speaker HMMs, as shown in Figure 3.5a, and
- $N_{\text{ps}}\left(N_{\text{ps}}-1\right)/2$ dual-speaker HMMs, as shown in Figure 3.6b.

These single and dual-speaker classes are combined in an inter-class grammar that forbids self-loops, similar to that shown in Figure 3.5b for the single speaker case.

Experiments were conducted in a meeting room using a 4-element microphone array ($N_m = 4$ microphones, $N_q = 6$ pairs) placed in the center of a table, with speakers seated at 4 different locations around the table, as shown in Figure 3.7. A test database was recorded simultaneously across all microphones at a sampling rate $f_s = 16$ kHz. The total database duration was 20 minutes, consist-

system	FA	PRC	RCL	*F*
TDE features	99.1%	0.98	0.98	0.98
LPCC features	88.3%	0.81	0.73	0.77

Table 3.2. Results for the single speaker system (test set 1). FA stands for Frame Accuracy, PRC, RCL and F for Precision, Recall and F-measure, respectively (the higher, the better for all metrics).

ing of 5 minutes of speech from each speaker/location.

These four 5-minute single speaker/location files were randomly recombined to form two separate test sets. *Test set 1 (non-overlap)* contained only single speaker segments without any overlap segments. Nine files containing 10 speaker turns were constructed in a random manner, with segments varying from 5 to 20 seconds in duration. *Test set 2 (overlap)* was constructed from the same database in a similar manner, however this time a short overlap segment was included at each speaker change. The test set consisted of six files, each containing 10 single speaker segments (of between 5-17 seconds duration), interleaved with 9 segments of dual-speaker overlap (of between 1.5-5 seconds duration). The TDOA estimates were calculated on 32 ms time frames, every 16 ms.

Test Set 1: For the HMM we used a self-loop probability of 0.9, and a minimum duration of 2 seconds. We also compared with a similar system using Linear Prediction Cepstral Coefficients (LPCC)[5], by replacing the single Gaussian used for TDOA with a 8-mixture GMM, trained on separate data, for each speaker.

Test Set 2: For this scenario, the HMM topology from the previous experiments was extended by adding the 6 dual-speaker classes, as described earlier. Each short-term speaker turn (SST) was constrained to a duration of 3 to 10 frames. These SST's were then combined in a minimum duration sequence of 1 second. Once again, transitions in the inter-class grammar were all equally weighted. As this topology was designed directly from observations of the temporal behaviour of the time delay features during overlap segments, direct comparison with the LPCC features was considered inappropriate in this case.

Performance Metrics include Frame Accuracy (FA), the percentage of frames that were correctly labeled. We also evaluated correct and incorrect segment boundaries by calculating Precision (PRC), Recall (RCL) and F-measure (F), as defined in Appendix A. An estimated segment boundary is deemed correct iff within ± 1 second of a true segment boundary.

Results: Table 3.2 presents the results for the location- and LPCC-based single speaker sys-

[5]LPCC features are extracted from one microphone of the array.

test set	FA	PRC	RCL	F
non-overlap (test set 1)	99.1%	0.98	0.98	0.98
overlap (test set 2)	94.1% (85.5%)	0.94	0.86	0.90

Table 3.3. Results for the extended system (test sets 1 and 2). The FA calculated only on actual overlap segments is shown in parentheses.

tems. These results show the improved discrimination provided by the location-based features, as well as the suitability of the proposed HMM framework for segmentation. The improved results of the location-based system are achieved with lower model complexity (one component per GMM, compared to 8 for LPCC's), as well as simpler training, through direct calculation of the state distribution associated with each location.

Table 3.3 presents the results of the location-based dual speaker system. We first observe that the results on test set 1 using the extended system are identical to those obtained using only the 4 single speaker classes, indicating that the addition of the 6 dual-speaker overlap classes does not affect the system's ability to discriminate single speaker segments. Secondly, we see that a high frame accuracy and F-measure are obtained on the overlap test set. This indicates both the suitability of the proposed overlap class topology, as well as the power of the HMM to represent more complex segment structure. We note that part of the decrease in FA for overlap segments may be attributed to the shorter segment duration and difficulty in defining a precise ground truth.

Conclusion: These results suggest that location information can be very helpful to precisely detect speech and discriminate between speakers, including on overlapped speech. However, several limitations arise in this preliminary work:

1. Only dual-speaker overlap is modelled. Following the same approach, $2^{N_{\text{ps}}}$ HMMs would be needed, to model all possible combinations of active/silent speakers.

2. A 2-second minimum duration was used.

3. The number of speakers N_{ps}, and their locations $\left(\boldsymbol{\ell}_1^{\text{ps}}, \cdots, \boldsymbol{\ell}_{N_{\text{ps}}}^{\text{ps}}\right)$, are known in advance.

4. Each speaker is associated with one location, and vice-versa. This assumption is violated as soon as a speaker stands up and moves, or whenever speakers exchange chairs.

(Lathoud et al., 2003) addressed limitation 1. by segmenting each speaker independently, in a lightweight, online manner. Limitations 2. to 4. are addressed by Chapters 5 to 7.

Chapter 4

The AV16.3 Corpus

The objective announced in Section 1.1 includes an integrated system for multisource detection-localization ("Where? When?" questions), that should be able to cope with spontaneous multi-party speech, in both static scenarios (seated speakers) and moving scenarios (moving speakers). Such an effort requires test data, along with precise speaker mouth location annotation. One possibility would be to use loudspeakers playing pre-recorded speech at known locations, or along known trajectories. However, a recent study (Schwetz et al., 2004) has pointed out specificities of the human speech radiation, that suggest wide differences with loudspeaker radiation. Therefore, in the "AV16.3" corpus described by the present chapter, almost all recordings were made with human speakers, recorded in a meeting room context. "AV" stands for audio-visual, and "16.3" stands for 16 microphones and 3 cameras, recorded in a fully synchronized manner. The central idea is to use calibrated cameras to provide a continuous 3-dimensional (3-D) speaker location annotation, that can be used to evaluate the results of audio localization and tracking algorithms. Particular attention is given to multiple moving speakers, and to overlapped speech – when several speakers are simultaneously speaking. Overlap is indeed an important issue in multi-party spontaneous speech (Shriberg et al., 2001), as discussed in Section 3.2.1. Moreover, since visual recordings are available, video and audio-visual tracking algorithms can also be tested. We therefore defined and recorded a series of scenarios so as to cover a variety of research areas, namely audio, video and audio-visual localization and tracking of people in a meeting room. Possible applications range from automatic analysis of meetings to robust speech acquisition and video surveillance, to name a few.

To cover such a broad range of research topics, the "meeting room context" used here includes a high variety of situations, from a single, static speaker, to multiple moving speakers and overlapping speech. This departs from existing, related databases: for example the ICSI database (Janin et al., 2003) contains audio-only recordings of natural meetings, the CUAVE database (Patterson et al., 2002) does contain audio-visual recordings (close-ups) but focuses on multimodal speech recognition. The CIPIC database (Algazi et al., 2001) focuses on Head-Related Transfer Functions. The audio-only ShATR corpus (Crawford et al., 1994) includes close-talking microphones and a binaural manikin. The ShATR corpus should be acknowledged as one of the earliest effort on spontaneous multi-party speech. Although its focus on overlapping speech is relevant to the present objective, it only contains seated speakers, without precise mouth location annotation.

Instead of focusing the entire database on one research topic, we chose to have a single, generic setup, allowing for very different scenarios in the different recordings. The goal is to provide annotation both in terms of "true" 3-D speaker location in the microphone arrays' referent, and "true" 2-D head/face location in the image plane of each camera. Such annotation permits systematic evaluation of localization and tracking algorithms, as opposed to subjective evaluation on a few short examples without annotation. To the best of our knowledge, the AV16.3 corpus was the first publicly available, annotated audio-visual corpus for speaker localization and tracking.

While investigating for existing solutions for speaker location annotation, we found various solutions with devices to be worn by each person and a base device that locates each personal device. However, these solutions were either very costly and extremely effective (high precision and sampling rate, no tether between the base and the personal devices), or cheap but with poor precision and/or high constraints (e.g. personal devices tethered to the base). We thus used calibrated cameras, placed at optimized locations, to reconstruct the 3-D location of the speakers. A maximum 3-D error of 1.2 cm was achieved. Moreover, the proposed annotation solution is potentially non-intrusive. Indeed, some recordings of the AV16.3 corpus do not have any marker on the actors.

This chapter is organized as follows: Section 4.1 describes the physical setup and the camera calibration process. Section 4.2 describes the recorded sequences. Section 4.3 presents the annotation interfaces, and details the available annotation. Section 4.4 describes additional loudspeaker recordings. Section 4.5 concludes. The whole corpus, along with annotation tools, files, and Matlab code for 3-D reconstruction, is available at: `http://mmm.idiap.ch/Lathoud/av16.3_v6/`

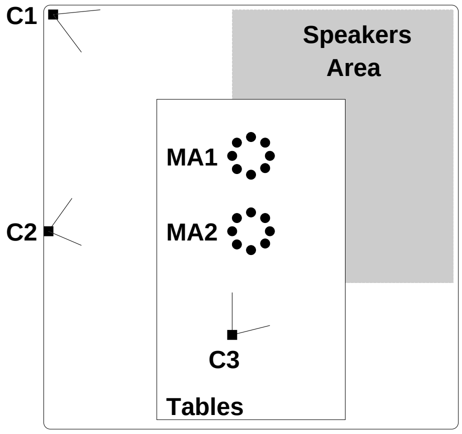


Figure 4.1. Physical setup: three cameras C1, C2 and C3 and two 8-microphone circular arrays MA1 and MA2. The gray, L-shaped area is in the field of view of all three cameras.

4.1 Physical Setup and Camera Calibration

While designing the AV16.3 corpus, two contradicting constraints arose: 1) the area occupied by speakers should be large enough to cover both "meeting situations" and "motion situations", 2) this area should be entirely visible by all three cameras. The second constraint allows for robust reconstruction of 3-D location information, since information from all three cameras can be used. As a compromise between the two constraints, we defined a L-shaped area of possible speakers' locations around the tables in a meeting room, as depicted in Figure 4.1. A general description of the meeting room can be found in (Moore, 2002). The L-shaped area is a 3 m-long and 2 m-wide rectangle, minus a 0.6 m-wide portion taken by the tables. Figure 4.2 contains views taken with the different cameras.

The motivation for precise camera calibration is that if we can track the mouth of a person in each camera's image plane, then we can reconstruct the 3-D trajectory of the mouth using the cameras' calibration parameters. This can be useful as audio annotation, provided that the 3-D mouth location is defined in the referent of the microphone arrays. We thus adopted a 2-step strategy for placing the cameras and calibrating them. First, camera placement (location, orientation, zoom) is optimized, using a looping process that includes sub-optimal calibration of the cameras with 2-D image plane information only. Second, each camera is calibrated in a precise manner, using both 2-D measurements in the image plane, and 3-D measurements in the referent of the microphone arrays. We show that the 3-D reconstruction error is within a very acceptable range.

Section 4.1.1 describes and motivates the choice of hardware. Section 4.1.2 describes the camera placement step. Section 4.1.3 describes the precise camera calibration step. The data itself is described further below, in Section 4.2.

4.1.1 Hardware

We used 3 cameras and two 10 cm-radius, 8-microphone Uniform Circular Arrays (UCA) from an instrumented meeting room (Moore, 2002). The two microphone arrays are placed 0.8 m apart. The motivation behind this choice is threefold:

- Recordings made with two microphone arrays provide test cases for 3-D audio source localization and tracking, as each microphone array can be used to provide an (azimuth, elevation) location estimate of each audio source.
- Recordings made with several cameras generate many interesting, realistic cases of visual occlusion, viewing each person from several viewpoints.
- At least two cameras are necessary to compute the 3-D coordinates of an object from the 2-D coordinates in cameras' image planes. The use of three cameras allows to reconstruct the 3-D coordinates of an object in a robust manner. Indeed, in most cases, visual occlusion occurs in one camera only; the head of the person remains visible from the two other cameras.

Lapels were also worn by the speakers, whenever it was technically feasible.

4.1.2 Step One: Camera Placement

This subsection describes the looping process that optimizes cameras placement (location, orientation, zoom) using 2-D information only. We used a freely available Multi-Camera Self-Calibration (MultiCamSelfCal) software (Svoboda, 2003). "Self-calibration" means that the 3-D locations of the calibration points are unknown. The MultiCamSelfCal uses only the 2-D coordinates in the image plane of each camera. It *jointly* produces a set of calibration parameters[1] for each camera and 3-D location estimates of the calibration points, by minimizing the **"2-D reprojection error"**. For each camera, the **"2-D reprojection error"** is defined as the distance in pixels between the recorded 2-D points and the projection of their 3-D location estimates back onto the camera image

[1] For a description of the camera calibration parameters see (Bouguet, 2004).

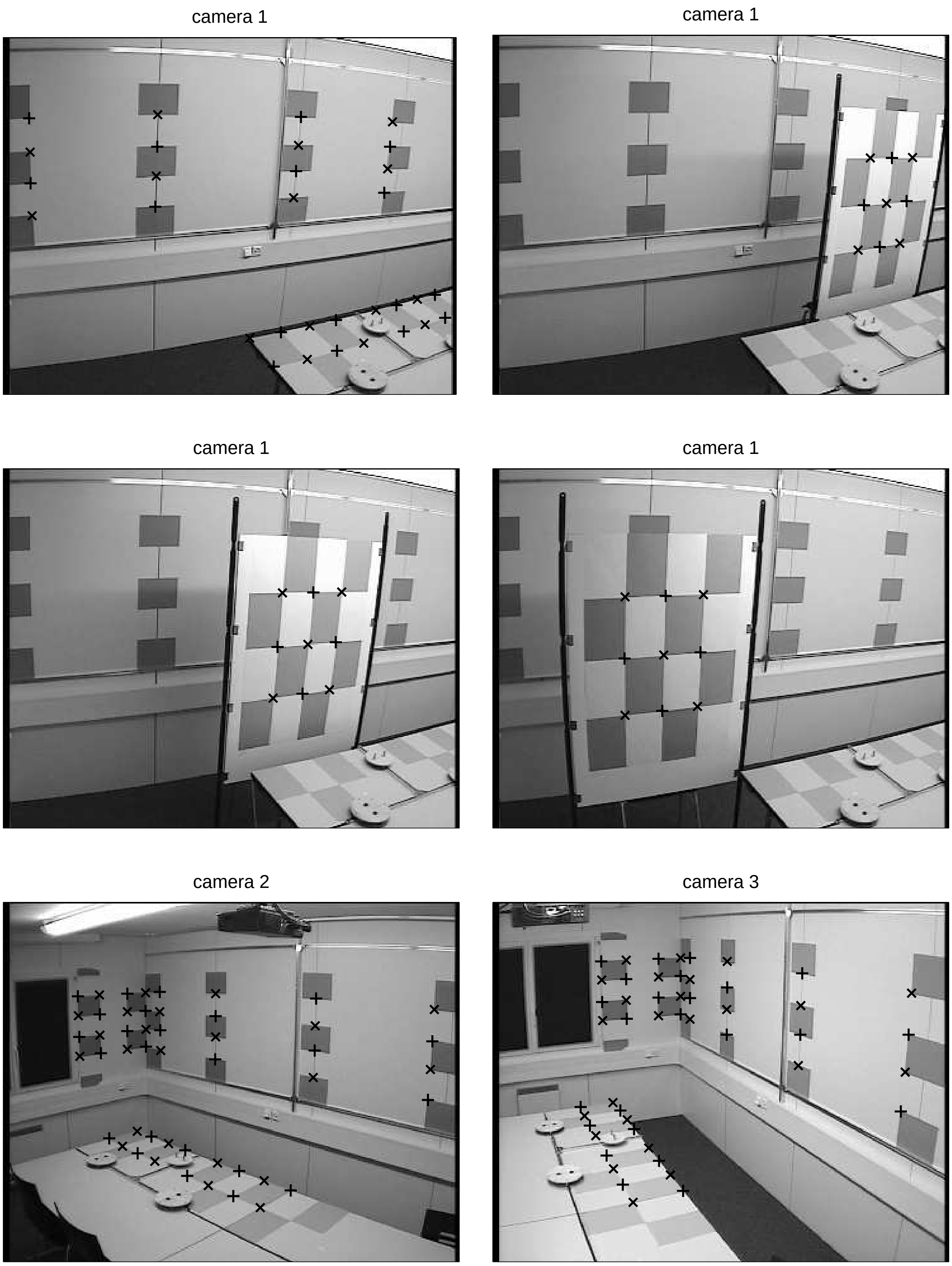


Figure 4.2. Snapshots from the cameras at their final positions. "+" designate points in the calibration training set $\mathbb{X}_{\mathrm{train}}$, "x" designate points in the calibration test set $\mathbb{X}_{\mathrm{test}}$.

plane, using the estimated camera calibration parameters. Although we used the software with the strict minimum number of cameras (three), the obtained 2-D reprojection error was decent: its upper bound was estimated as less than 0.17 pixels. The proposed camera placement procedure consists in an iterative process with three steps: Place, Record and Calibrate:

1. *Place* the three cameras (location, orientation, zoom) based on experience in prior iterations. In practice the various cameras should give views that are as different as possible (different orientation angles), while having as much field of view in common as possible.

2. *Record* synchronously with the 3 cameras a set of calibration points, i.e. 2-D coordinates in the image plane of each camera. As explained in (Svoboda, 2003), waving a modified laser pointer in darkness is sufficient.

3. *Calibrate* the 3 cameras by running MultiCamSelfCal on the calibration points. MultiCamSelfCal finds the calibration parameters that minimize an upper bound of the 2-D reprojection error, given the current camera placement.

4. To further decrease the 2-D reprojection error, loop to 1. Else go to 5. In practice, a 2-D reprojection error below 0.2 pixels is reasonable.

5. Over all iterations, select the camera placement that gave the smallest 2-D reprojection error.

Multi-camera self-calibration is generally known to provide less precision than manual calibration using an object with known 3-D coordinates. However, self-calibration is much faster, precisely because it does not require any 3-D measurements, but only 2-D calibration points that are quickly recorded with a modified laser pointer. One iteration of the Place/Record/Calibrate loop thus takes about 1h30. This process converged to the positioning of the camera depicted in Figure 4.1. For further information, including the multi-camera self-calibration problem statement, the reader is referred to the documentation in (Svoboda, 2003).

4.1.3 Step Two: Camera Calibration

This subsection describes the precise calibration of each camera, assuming the cameras' placement fixed (location, orientation, zoom). This is done by selecting and optimizing the calibration parameters for each camera, with respect to a calibration object. For each point of the calibration object, both true 3-D coordinates *in the microphone arrays' referent* and true 2-D coordinates in each camera's image plane are known. 3-D coordinates were obtained on-site with a measuring tape (measurement error estimated below 0.005 m). Crosses in Figure 4.2 show the 3-D calibration points. These points were splitted into two sets: $\mathbb{X}_{\text{train}}$ (36 points) and $\mathbb{X}_{\text{test}}$ (39 points).

A critical issue was to select the distortion model, that is the type of non-linear distortions assumed to be produced by the optics of each camera. An iterative process was used to select the model, that minimizes the *"3-D reconstruction error"*. The 3-D reconstruction error is defined as the Euclidean distance between the reconstructed 3-D location estimates of points visible from at least two cameras, and their true 3-D location measured with the tape. The complete precise camera calibration procedure can be detailed as follows:

1. *Model selection*: for each camera, select the set of calibration parameters based on experience in prior iterations.

2. *Model training*: for each camera, estimate the selected calibration parameters on $\mathbb{X}_{\text{train}}$ using the software available in (Bouguet, 2004), given the chosen model.

3. *3-D error*: for each point in $\mathbb{X}_{\text{train}}$, compute the Euclidean distance between the true 3-D coordinates and the 3-D coordinates reconstructed from the 2-D coordinates in each camera's image plane, using the trained calibration parameters.

4. *Evaluation*: estimate the "training" maximum 3-D reconstruction error as $\mu + 3\sigma$, where μ and σ stand for mean and standard deviation of the 3-D error, across all points in $\mathbb{X}_{\text{train}}$.

5. To try to decrease the training maximum 3-D reconstruction error, loop to 1. Else go to 6.

6. Over all iterations, select the set of calibration parameters and their estimated values, that gave the smallest maximum 3-D reconstruction error.

The result of this process is a selected set of distortion calibration parameters and their values for each camera. For all cameras the best set of parameters was focal center, focal lengths, r^2 radial and tangential distortion coefficients.

Once the training was over, we evaluated the 3-D error on the unseen test set $\mathbb{X}_{\text{test}}$. The maximum 3-D reconstruction error on this set was 0.012 m. This maximum error was deemed reasonable, as compared to the diameter of an open mouth (about 0.05 m).

4.2 Online Corpus

This section first motivates and describes the variety of recorded sequences, then describes in more details the annotated sequences. “Sequence” means:

- 3 video DIVX AVI files (resolution 288x360), one for each camera, sampled at 25 Hz. Each AVI file also includes one audio signal.
- 16 audio WAV files recorded from the two circular 8-microphone arrays, sampled at 16 kHz.
- When possible, audio WAV files recorded from lapels worn by the speakers, sampled at 16 kHz.

All files were recorded in a synchronous manner: video files carry a time-stamp embedded in the upper rows of each image, and audio files always start at video time stamp `00:00:10.00`. Complete details about the hardware implementation of a unique clock across all sensors can be found in (Moore, 2002). The various sequences of the corpus were recorded over a period of 5 days, including 42 sequences overall, with sequence durations ranging from 14 seconds to 9 minutes (total 1h25). 12 different actors were recorded. Most of the recorded actors did not have any particular expertise in the fields of audio or video localization and tracking. Although only 10 sequences have been annotated, the other 32 sequences are also available. The whole corpus, along with annotation files, camera calibration parameters and additional documentation is accessible[2] at: `http://mmm.idiap.ch/Lathoud/av16.3_v6`

4.2.1 Motivations

A non-limiting list of relevant localization/tracking phenomena includes:

- Overlapped speech.
- Close and far locations, small and large angular separations.
- Object initialization.
- Variable number of objects.
- Partial and total occlusion.
- “Natural” changes of illumination.

[2]Both HTTP or FTP protocols can be used to browse and download the data.

Accordingly, we defined and recorded a set of sequences that contains a high variety of test cases: from short, very constrained, specific cases (e.g. visual occlusion), for each modality (audio or video), to natural spontaneous speech and/or motion in much less constrained context.

Each sequence is useful for at least one of three fields of research: analysis of audio, video or audio-visual data. Up to three people are allowed in each sequence. Human motion can be static (e.g. seated persons), dynamic (e.g. walking persons) or a mix of both across persons (some seated, some walking) and time (e.g. meeting preceded and followed by people standing and moving).

4.2.2 Sequence Names

A systematic coding was defined, such that the name of each sequence (1) is unique, and (2) contains a compact description of its content. For example "`seq40-3p-0111`" has three parts:

- "seq40" is the unique identifier of this sequence. For convenience, we sometimes abbreviate the names: "`seq40`" and "`seq40-3p-0111`" designate the same sequence.
- "3p" means 3 different persons in this sequence – but not necessarily all visible simultaneously.
- "0111" are four binary flags that summarize this sequence. From left to right:

  bit 1: 0 means "very constrained", 1 means "mostly unconstrained" (general behavior: although most recordings follow some sort of scenario, some include very strong constraints such as the speaker facing the microphone arrays at all times).

  bit 2: 0 means "static motion" (mostly seated), 1 means "dynamic motion" (walking).

  bit 3: 0 means "minor occlusion(s)", 1 means "at least one major occlusion": whenever somebody passes in front of somebody else, with respect to one array or camera.

  bit 4: 0 means "little overlapped speech", 1 means "significant overlapped speech". Overlap involves all acoustic sources, including speakers and/or noise sources.

Sequence name	Duration (seconds)	Modalities of interest	Nb. of speakers	Speaker(s) behavior
seq01-1p-0000	217	A	1	S
seq02-1p-0000	189	A	1	S
seq03-1p-0000	242	A	1	S
seq11-1p-0100	30	A, V, AV	1	D
seq15-1p-0100	35	AV	1	S,D(U)
seq18-2p-0101	56	A(ov)	2	S,D
seq24-2p-0111	48	A(ov), V(occ)	2	D
seq37-3p-0001	511	A(ov)	3	S
seq40-3p-0111	50	A(ov), AV	3	S,D
seq45-3p-1111	43	A(ov), V(occ), AV	3	D(U)

Table 4.1. List of the annotated sequences. Tags mean: (A)udio, (V)ideo, predominant (ov)erlapped speech, at least one visual (occ)lusion, (S)tatic speakers, (D)ynamic speakers, (U)nconstrained motion.

4.2.3 Annotated Contents

As mentioned above, the AV16.3 corpus comprises 10 annotated sequences plus 32 unannotated sequences. The 10 annotated sequences were chosen so as to cover a large variety of situations that fulfill interests from various areas of research. "Variety" means different modalities (audio or video) and different speaker behaviors. Table 4.1 gives a synthetic overview.

seq01-1p-0000, seq02-1p-0000, seq03-1p-0000: A single speaker, static while speaking, at each of 16 locations covering the shaded area in Figure 4.1. The speaker is facing the microphone arrays. The purpose of this sequence is to evaluate audio source localization on a single speaker case.

seq11-1p-0100: One speaker, mostly moving while speaking. The only constraint on the speaker's motion is to face the microphone arrays. The motivation is to test audio, video or audio-visual (AV) speaker tracking on difficult motion cases. The speaker is talking most of the time.

seq15-1p-0100: One moving speaker, walking around while alternating speech and long silences. The purpose of this sequence is to 1) show that audio tracking alone cannot recover from unpredictable trajectories during silence, 2) provide an initial test case for AV tracking.

seq18-2p-0101: Two speakers, speaking and facing the microphone arrays all the time, slowly getting as close as possible to each other, then slowly parting. The purpose is to test multi-source localization, tracking and separation algorithms.

`seq24-2p-0111`: Two moving speakers, crossing the field of view twice and occluding each other twice. The two speakers are talking most of the time. The motivation is to test both audio and video occlusions.

`seq37-3p-0001`: Three speakers, static while speaking. Two speakers remain seated all the time and the third one is standing. Overall five locations are covered. Most of the time 2 or 3 speakers are speaking concurrently. (For this particular sequence only snapshot image files are available, no AVI files.) The purpose of this sequence is to evaluate multi-source localization and beamforming algorithms.

`seq40-3p-0111`: Three speakers, two seated and one standing, all speaking continuously, facing the arrays. The standing speaker walks back and forth once, behind the seated speakers. The motivation is both to test multi-source localization, tracking and separation algorithms, and to highlight complementarity between audio and video modalities.

`seq45-3p-1111`: Three moving speakers, entering and leaving the scene, all speaking continuously, occluding each other many times. Speakers' motion is unconstrained. This is a very difficult case of overlapped speech and visual occlusions. The motivation is to highlight the complementarity between audio and video modalities.

4.3 Annotation

Three spatial annotation interfaces were developed, as detailed in Section 4.3.1: for the mouth, for the head, and for an optional ball marker on the head of a speaker. Moreover, some of the sequences were also annotated over time, in the form of a speech/silence segmentation. Section 4.3.3 details the available annotation, and Section 4.3.2 explains the 3-D mouth reconstruction. Sections 4.3.4 and 4.3.5 give two examples of application of the available annotation.

4.3.1 Spatial Annotation Interfaces

BAI: the Ball Annotation Interface, to mark the location of a colored ball on the head of a person, as an ellipse. Occlusions can also be marked, i.e. when the ball is not visible. The BAI includes a simple tracker to interpolate between manual measurements.

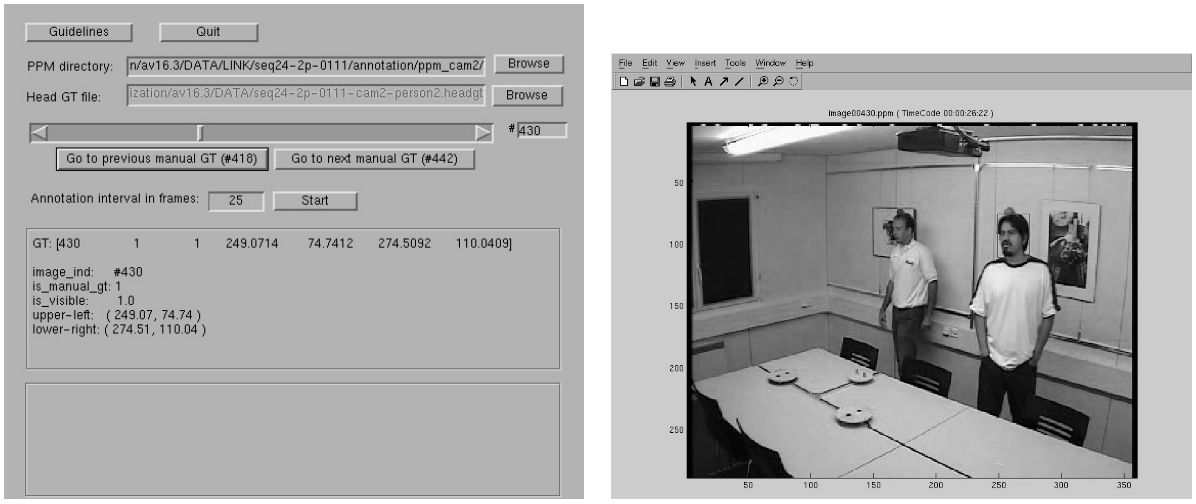


Figure 4.3. Snapshots of the two windows of the Head Annotation Interface.

HAI: the Head Annotation Interface, to mark the location of the head of a person, as a rectangular bounding box. Partial or complete occlusions can also be marked.

MAI: the Mouth Annotation Interface, to mark the location of the mouth of a person as a point. Occlusions can also be marked, i.e. when the mouth is not visible.

All three interfaces share very similar features, including two windows: one for the interface itself, and a second one for the image currently being annotated (Figure 4.3). Each annotation file is a numerical matrix stored in ASCII format[3]. All three interfaces are available and documented online, within the corpus itself. We have already used them to produce continuous 3-D mouth location annotation from sparse manual measurements, as described in Section 4.3.2.

4.3.2 3-D Mouth Annotation

From sparse 2-D mouth annotation on each camera we (1) reconstruct 3-D mouth location using camera calibration parameters estimated as explained in Section 4.1.3, (2) interpolate 3-D mouth location using the reconstructed ball marker 3-D location. The 3-D ball location itself is provided by the 2-D tracker in the BAI interface (see Section 4.3.1), and 3-D reconstruction. The motivation of this choice was twofold: first of all, using simple (e.g. polynomial) interpolation on mouth measurements was not enough in practice, since human motion contains many complex non-linearities

[3] Each format is detailed in `http://mmm.idiap.ch/Lathoud/av16.3_v6/FORMATS`

Sequence	ball		mouth		head	speech/silence
	2-D	3-D	2-D	3-D	2-D	segmentation
seq01-1p-0000	C	C	C	C	S(2 Hz)	precise
seq02-1p-0000			S(1 per location)	S(1 per location)		undersegmented
seq03-1p-0000			S(1 per location)	S(1 per location)		undersegmented
seq11-1p-0100	C	C	C	C	S(2 Hz)	
seq15-1p-0100	C	C	C	C	S(2 Hz)	
seq18-2p-0101	C	C	C	C	S(2 Hz)	
seq24-2p-0111	C	C	C	C	S(2 Hz)	
seq37-3p-0001	C	C	C	C		undersegmented
seq40-3p-0111	C	C	C	C	S(2 Hz)	
seq45-3p-1111	C	C	C	C	S(2 Hz)	

Table 4.2. Available annotation, as of December 7th, 2006. "C" means continuous annotation: on all frames of each 25 Hz video. "S" means sparse annotation: on some of the video frames (annotation rate in brackets). "Undersegmented" means that some short silences are included in the segments marked as "speech".

(sharp head turns and accelerations). Second, visual tracking of the mouth is a hard task in itself. We found that interpolating measurements in the moving referent of an automatically tracked ball marker is effective even at low annotation rates (e.g 2 Hz = 1 video frame out of 12), which is particularly important since the goal is to limit the time spent doing manual measurements. A complete example with all necessary Matlab implementation can be found online[4]. This implementation was used to create all 3-D files available within the corpus.

4.3.3 Available Annotation

Table 4.2 details the available 2-D and 3-D annotations for each of the 10 annotated sequences.

4.3.4 Example 1: Audio Source Localization Evaluation

The online corpus includes a complete example (Matlab files) of single source localization followed by comparison with the annotation, for "seq01-1p-0000". It is based on the Steered Response Power method called SRP-PHAT (DiBiase, 2000), described earlier in Section 3.1.2. All necessary Matlab code to run the example is available online[5]. The comparison shows that the SRP-PHAT localization method provides a precision between -5 and +5 degrees in azimuth. Eight annotated sequences are further evaluated in terms of both detection and localization, in Section 5.4.8.

[4] http://mmm.idiap.ch/Lathoud/av16.3_v6/EXAMPLES/3D-RECONSTRUCTION/README
[5] http://mmm.idiap.ch/Lathoud/av16.3_v6/EXAMPLES/AUDIO/README

4.3.5 Example 2: Multi-Object Video Tracking

The video tracking results of three independent, appearance-based particle filters on 200 frames of the "`seq45-3p-1111`" sequence, using only one of the cameras, are shown in Figure 4.4, and in a video[6]. The sequence depicts three people moving around the room while speaking, and includes multiple instances of object occlusion. Each tracker has been initialized by hand, and uses 500 particles. Object appearance is modeled by a color distribution (Perez et al., 2002) in RGB space. In this particular example we have not done any performance evaluation yet. Suggestions for evaluation can be found in (Smith et al., 2005).

4.4 Additional Loudspeaker Sequences

Spatial location can be precisely annotated using calibrated cameras, within the 1.2 cm maximum 3-D error mentioned above. On the contrary, the "true" speech/silence segmentation of human speech is difficult to define in an exact manner. Inter-word silences may be extremely short, and even inside a word, there may be parts with very low energy. For some performance metrics, this precludes the *exact* evaluation of a speech detection system, as confirmed by the results in Section 5.3.4. Therefore, we recorded three additional sequences, each containing three loudspeakers playing constant-power synthesized vowels, around a circular microphone array. By construction, the speech/silence segmentation of each loudspeaker is *exactly* known, therefore any detection performance evaluation is *exact*. The sequences `loud01-3p-0001`, `loud02-3p-0001` and `loud03-3p-0001` each contain 20 minutes of synthetic speech, as an alternation of 4 seconds of stationary vowel sound followed by 2 seconds of silence. In each sequence, all 200 possible combinations of 2 and 3 active loudspeakers and 5 different vowels are played sequentially. Vowels are synthesized using a LPC vocoder[7] and constant LPC coefficients, estimated from real speech. Figure 4.5 depicts the locations of the three loudspeakers A, B, C. Only A was moved between sequences, so as to test equal distances (`loud01-3p-0001`), A further than B and C (`loud02-3p-0001`), and low angular separation (`loud03-3p-0001`). The loudspeaker sequences are used in the detection tests reported in Chapter 5.

[6] `http://mmm.idiap.ch/Lathoud/av16.3_v6/EXAMPLES/VIDEO/av-video.mpeg`
[7] Available at `http://www.tcts.fpms.ac.be/cours/1005-08/speech/lpcvocoder.zip`

Figure 4.4. Snapshots from visual tracking on 200 frames of "seq45-3p-1111" (initial timecode: 00:00:41.17). Tracking results are shown every 25 frames.

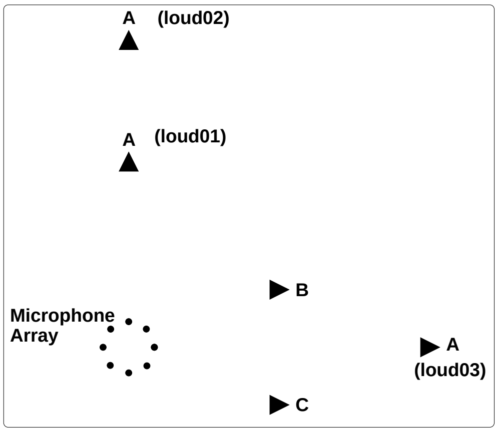


Figure 4.5. Top view of the recording setup for `loud01-3p-0001`, `loud02-3p-0001` and `loud03-3p-0001`: 3 loudspeakers A,B,C. Loudspeaker A lies at 90^o azimuth relative to the array in `loud01-3p-0001` (radius 0.8 m) and `loud02-3p-0001` (radius 1.8 m), and at 0^o azimuth in `loud03-3p-0001` (radius 1.45 m). Loudspeakers B and C lie respectively at $+25.6^o$ and -25.6^o in all three sequences `loud01-3p-0001`, `loud02-3p-0001` and `loud03-3p-0001` (radius 0.8 m).

4.5 Conclusion

This chapter presented the AV16.3 corpus for speaker localization and tracking. AV16.3 focuses mostly on the context of meeting room data, acquired synchronously by 3 cameras, 16 far-distance microphones, and lapels. It targets various areas of research: audio, visual and audio-visual speaker tracking. In order to provide audio annotation, camera calibration is used to generate "true" 3-D speaker mouth location, using freely available software. To the best of our knowledge, this is the first attempt to provide synchronized audio-visual data for extensive testing on a variety of test cases, along with spatial annotation. AV16.3 is intended as a step towards systematic evaluation of localization and tracking algorithms on real recordings. Eight annotated sequences are used in Chapter 5 to evaluate multisource detection-localization, and in Chapter 6 and Appendix D to evaluate the speech/non-speech classification in the localization context.

4.6 Acknowledgment

The AV16.3 Corpus was created in the fall of 2003 and first made publicly available in June 2004, in collaboration with Dr Jean-Marc Odobez and Dr Daniel Gatica-Perez (Lathoud et al., 2005c). The author would like to thank Olivier Masson for the great help with the 3-D measurements, the AMI partners who contributed to some of the annotations (AMI WP4), and all the recorded actors.

Chapter 5

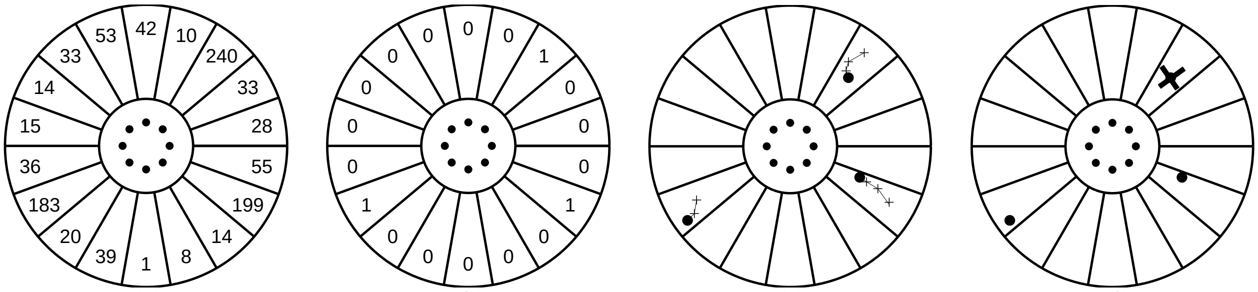
Multisource Joint Detection-Localization

This chapter focuses on *instantaneous* detection-localization of multiple acoustic sources, as announced in Chapter 1 (Figure 1.2b). The goal is to answer the "Where? When?" questions:

- Detect how many acoustic sources are active: zero, one or more.
- Locate in space the active acoustic source(s).

By instantaneous we mean *static* analysis of the signals, that is from a *single* time frame on which speech is assumed quasi-stationary (20 to 30 ms). The *dynamical* analysis (on multiple consecutive time frames) is addressed by Chapter 6.

As explained in Section 3.1.2, we have opted for Steered Response Power (SRP) localization, where a source location estimate is obtained by finding a local maximum of the beamforming power (Krim and Viberg, 1996; DiBiase, 2000). On one hand, Section 3.1.2 showed that a fast search space reduction would be beneficial, where the active sectors of a discretized space are detected. On the other hand, Section 3.1.3 showed that detection for localization may require different approaches than detection in other contexts such as ASR. Following these two points, the present chapter investigates the **sector-based detection-localization** task, where we try to determine – to detect – which sectors – or discrete locations – are active. Classical **point-based localization** is then applied within each active sector. This chapter describes each step of the proposed system:



(a) Sector-based activeness $\zeta_{\check{s},t} \geq 0$ Section 5.2

(b) Sector-based detection-localization $\zeta_{\check{s},t} \gtrless \Psi_\zeta$ Section 5.3

(c) Point-based localization $(\theta, \varphi, \rho) \in \mathbb{R}^3$ Section 5.4

(d) Speech/Non-Speech classification (sector-based MFCCs) Section 5.5

Figure 5.1. Proposed multisource detection-localization. The eight dots in the center represent the microphone array. The three dots in the sectors represent point location estimates.

1. Section 5.1 proposes to view the maximization of the SRP as a minimization of an equivalent metric, called Phase Domain Metric (PDM). The PDM interpretation of SRP serves as a foundation for both sector-based detection-localization and point-based localization.

2. Figure 5.1a: Based on the PDM, Section 5.2 proposes a measure of "acoustic activeness" in a sector of space. The search space $\mathbb{R}^3$ is divided into $N_{\check{s}}$ sectors $\{\mathbb{S}_1, \cdots, \mathbb{S}_{\check{s}}, \cdots, \mathbb{S}_{N_{\check{s}}}\}$. The acoustic activeness is a value $\zeta_{\check{s},t} \geq 0$, which becomes higher when a sector $\mathbb{S}_{\check{s}} \subset \mathbb{R}^3$, at time frame t, contains at least one active acoustic source.

3. Figure 5.1b: Sector-based detection-localization is a binary decision, for example by comparing the activeness $\zeta_{\check{s},t}$ to a threshold Ψ_ζ: $\zeta_{\check{s},t} \gtrless \Psi_\zeta$. Section 5.3 proposes to *jointly* model the small magnitudes (background noise) and the large magnitudes (speech) of the acoustic activeness $\zeta_{\check{s},t}$. No training data is required, so the method is adaptive and robust to environmental variations. Based on this model, a generic method for the automatic selection of the threshold Ψ_ζ is proposed, which further improves the robustness to environmental variations.

4. Figure 5.1c: Based on the PDM, Section 5.4 describes a gradient descent implementation of the point-based localization of the acoustic sources within the active sectors.

5. Figure 5.1d: Based on the PDM, Section 5.5 examines location-dependent measures for the classification of the localized acoustic sources in two groups: speech sources and non-speech sources (SNS classification).

Comparative experiments on real multichannel recordings are included in items 2., 3. and 4. As for the SNS classification (item 5.), it is presented in this chapter because of its link with the PDM. However, comparative SNS experiments are included in Chapter 6, because we compared static and dynamical analysis for SNS classification.

5.1 A Phase Domain Metric Interpretation of SRP

For a given spatial location $\boldsymbol{\ell} \in \mathbb{R}^3$, delay-sum beamforming consists in aligning the time domain signals received at the various microphone locations $\{\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_m, \cdots, \boldsymbol{\ell}_{N_m}\}$, by their theoretical delays $\mathrm{TOF}(\boldsymbol{\ell}, \boldsymbol{\ell}_m)$, and summing them:

$$x_{\mathrm{ds}}(t, \boldsymbol{\ell}) \stackrel{\mathrm{def}}{=} x_1(t) + \sum_{m=2}^{N_m} x(t + \mathrm{TOF}(\boldsymbol{\ell}, \boldsymbol{\ell}_m) - \mathrm{TOF}(\boldsymbol{\ell}, \boldsymbol{\ell}_1)) \tag{5.1}$$

where N_m is the number of microphones and $\mathrm{TOF}(\boldsymbol{\ell}, \boldsymbol{\ell}_m)$ is the Time Of Flight – expressed in sampling periods – of an acoustic wave between locations $\boldsymbol{\ell}$ and $\boldsymbol{\ell}_m$, as defined by (3.5). For a time frame t containing $2 \cdot N_{\mathrm{F}}$ samples, SRP localization consists in finding the location $\boldsymbol{\ell}$ that maximizes the power $\left\langle x_{\mathrm{ds}}(t + a, \boldsymbol{\ell})^2 \right\rangle_{-N_{\mathrm{F}} < a \leq N_{\mathrm{F}}}$, where $\langle \cdot \rangle_{.}$ is the average operator.

Section 5.1.1 shows that *in the frequency domain*, the power maximization is tightly related to a *minimization* of the differences between the observed phases and the theoretical phases. Based on this analysis, Section 5.1.2 formally introduces the proposed PDM, and Section 5.1.3 shows its equivalence with SRP-PHAT (DiBiase, 2000, Sections 6.3 and 6.4). The PDM and its application to sector-based detection-localization were originally introduced in (Lathoud and Magimai.-Doss, 2005; Lathoud et al., 2006a).

5.1.1 Motivation

Let $q \in \mathbb{N}$ be the microphone pair index: $1 \leq q \leq N_q$, where $N_q = N_m \cdot (N_m - 1)/2$ is the number of microphone pairs. Let $a_q \in \mathbb{N}$ and $b_q \in \mathbb{N}$ be the indices of the two microphones in the q-th pair: $1 \leq a_q < b_q \leq N_m$.

For the q-th pair of microphones $\left(\boldsymbol{\ell}_{a_q}, \boldsymbol{\ell}_{b_q}\right)$, let us consider the received signals $x_{a_q}(t)$ and $x_{b_q}(t)$. For a given time frame t, and a given discrete frequency k, the *frequency domain* delay-sum energy can be defined by aligning the two signals, with respect to the theoretical phase $u_q^{\mathrm{th}}(k,\boldsymbol{\ell})$:

$$\begin{aligned} E_q^{(\mathrm{ds},t)}(k,\boldsymbol{\ell}) &\stackrel{\mathrm{def}}{=} \left| X_{a_q}^{(t)}(k) + X_{b_q}^{(t)}(k) \cdot \exp\left[j \cdot u_q^{\mathrm{th}}(k,\boldsymbol{\ell}) \right] \right|^2 \\ &= \left| X_{a_q}^{(t)}(k) \right|^2 \cdot \left| 1 + \left| \frac{X_{b_q}^{(t)}(k)}{X_{a_q}^{(t)}(k)} \right| \cdot \exp\left[j \left(-u_q^{(t)}(k) + u_q^{\mathrm{th}}(k,\boldsymbol{\ell}) \right) \right] \right|^2 \end{aligned} \tag{5.2}$$

where $k \in \{1, \cdots, N_{\mathrm{F}}\}$, and the observed phase $u_q^{(t)}(k)$ is defined as:

$$u_q^{(t)}(k) \stackrel{\mathrm{def}}{=} \angle X_{a_q}^{(t)}(k) - \angle X_{b_q}^{(t)}(k) \tag{5.3}$$

and the theoretical phase $u_q^{\mathrm{th}}(k,\boldsymbol{\ell})$ is defined as:

$$u_q^{\mathrm{th}}(k,\boldsymbol{\ell}) \stackrel{\mathrm{def}}{=} -\pi \cdot \frac{k-1}{N_{\mathrm{F}}} \cdot \tau_q^{\mathrm{th}}(\boldsymbol{\ell}) \tag{5.4}$$

where the theoretical TDOA $\tau_q^{\mathrm{th}}(\boldsymbol{\ell})$ is expressed in sampling periods, as defined by (3.7).

Assuming the received magnitudes to be similar $\left| X_{a_q}^{(t)}(k) \right| \approx \left| X_{b_q}^{(t)}(k) \right|$, (5.2) becomes:

$$\begin{aligned} E_q^{(\mathrm{ds},t)}(k,\boldsymbol{\ell}) &\propto \left| 1 + \exp\left[j \left(-u_q^{(t)}(k) + u_q^{\mathrm{th}}(k,\boldsymbol{\ell}) \right) \right] \right|^2 \\ &\propto \left\{ 1 + \cos\left[-u_q^{(t)}(k) + u_q^{\mathrm{th}}(k,\boldsymbol{\ell}) \right] \right\}^2 + \sin^2\left[-u_q^{(t)}(k) + u_q^{\mathrm{th}}(k,\boldsymbol{\ell}) \right] \\ &\propto 2 + 2\cos\left[-u_q^{(t)}(k) + u_q^{\mathrm{th}}(k,\boldsymbol{\ell}) \right] \\ &\propto 1 - \sin^2\left[\frac{-u_q^{(t)}(k) + u_q^{\mathrm{th}}(k,\boldsymbol{\ell})}{2} \right] \end{aligned} \tag{5.5}$$

Therefore the maximization of the delay-sum energy at the discrete frequency k is equivalent to the minimization of $\sin^2\left[\frac{-u_q^{(t)}(k)+u_q^{\mathrm{th}}(k,\boldsymbol{\ell})}{2}\right]$, which can be seen as a "metric" in phase space. Section 5.1.2 formally defines the metric in the general case of multiple microphone pairs.

5.1.2 The Proposed Phase Domain Metric (PDM)

For speech array applications (DiBiase, 2000, Section 6.6), the highly dynamical natures of both the speech signals *and* the human motion justify the preference for spatial averaging over time averaging. The amount of data in a single time frame is limited, but this limitation is compensated by spatial averaging across multiple microphone pairs.

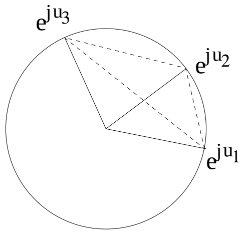


Figure 5.2. Illustration of the triangular inequality for the PDM in dimension 1: each point on the unit circle corresponds to an angle value modulo 2π. From the Euclidean metric: $\left|e^{ju_3} - e^{ju_1}\right| \le \left|e^{ju_3} - e^{ju_2}\right| + \left|e^{ju_2} - e^{ju_1}\right|$.

Thus, for a given time frame t, we define the vector of observed phase values:

$$\mathbf{u}^{(t)}(k) \quad \stackrel{\text{def}}{=} \quad \left[u_1^{(t)}(k), \cdots, u_q^{(t)}(k), \cdots, u_{N_q}^{(t)}(k)\right]^{\mathrm{T}} \tag{5.6}$$

and we define the vector of theoretical phase values:

$$\mathbf{u}^{\mathbf{th}}(k, \boldsymbol{\ell}) \quad \stackrel{\text{def}}{=} \quad \left[u_1^{\text{th}}(k, \boldsymbol{\ell}), \cdots, u_q^{\text{th}}(k, \boldsymbol{\ell}), \cdots, u_{N_q}^{\text{th}}(k, \boldsymbol{\ell})\right]^{\mathrm{T}} \tag{5.7}$$

We then propose the following function to compare two vectors of phase values:

$$\boxed{d(\mathbf{u}, \mathbf{u}') \stackrel{\text{def}}{=} \sqrt{\frac{1}{N_q} \sum_{q=1}^{N_q} \sin^2\left(\frac{u_q - u'_q}{2}\right)}} \tag{5.8}$$

$d(\cdot,\cdot)$ is similar to the Euclidean metric, except for the sine, which accounts for the "modulo 2π" definition of angles. The $1/N_q$ normalization factor ensures that $0 \le d(\cdot,\cdot) \le 1$. $d(\cdot,\cdot)$ is a true Phase Domain Metric (PDM), as defined in Appendix B.1. This is straightforward for N_q=1 by representing any angle u with a point e^{ju} on the unit circle, as in Figure 5.2, and observing that $\left|e^{ju_1} - e^{ju_2}\right| = 2 \cdot \left|\sin\left(\frac{u_1 - u_2}{2}\right)\right| = 2 \cdot d(u_1, u_2)$. Appendix B proves it for higher dimensions $N_q > 1$.

5.1.3 Equivalence with SRP-PHAT

Section 5.1.1 has indicated a rough equivalence between SRP maximization and PDM minimization. Concerning SRP-PHAT, Appendix B proves the following exact result (B.14):

$$\mathrm{P}_{\text{SRP-PHAT}}\left(\boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right) = N_{\mathrm{F}} \cdot (N_m + 2 \cdot N_q) - 4 \cdot N_q \cdot \sum_{k=2}^{N_{\mathrm{F}}+1} d^2\left(\mathbf{u}^{(t)}(k), \mathbf{u}^{\mathbf{th}}(k, \boldsymbol{\ell})\right)$$

Hence, maximizing SRP-PHAT is *strictly* equivalent to minimizing the squared PDM $d^2(\cdot,\cdot)$. This strict equivalence with a metric can be seen as a topological interpretation of SRP-PHAT localization. For sector-based detection-localization, this topological interpretation suggests and justifies *averaging* the metric over a sector of space, as well as *comparing* various metric values for a given frequency k (Section 5.2). For point-based localization, this topological interpretation suggests using gradient descent approaches for precise localization of a source within a sector (Section 5.4).

5.2 Sector-Based Activeness

The review conducted in Section 3.1.2 showed that a fast search space reduction could be beneficial to localization, where the active sectors of a discretized space are detected. The active sources can then be localized within the reduced space formed by the active sectors. Existing works going in that direction include (Kellermann, 1991; Zotkin and Duraiswami, 2004). They essentially rely on the beamformed power calculated *at a point* in the middle of each sector. This can be problematic when a true source location is close to the limit between two sectors. As an alternative, the present section proposes to evaluate the *average* acoustic activeness within a sector (volume of space). Experiments confirm the advantage of the proposed approach.

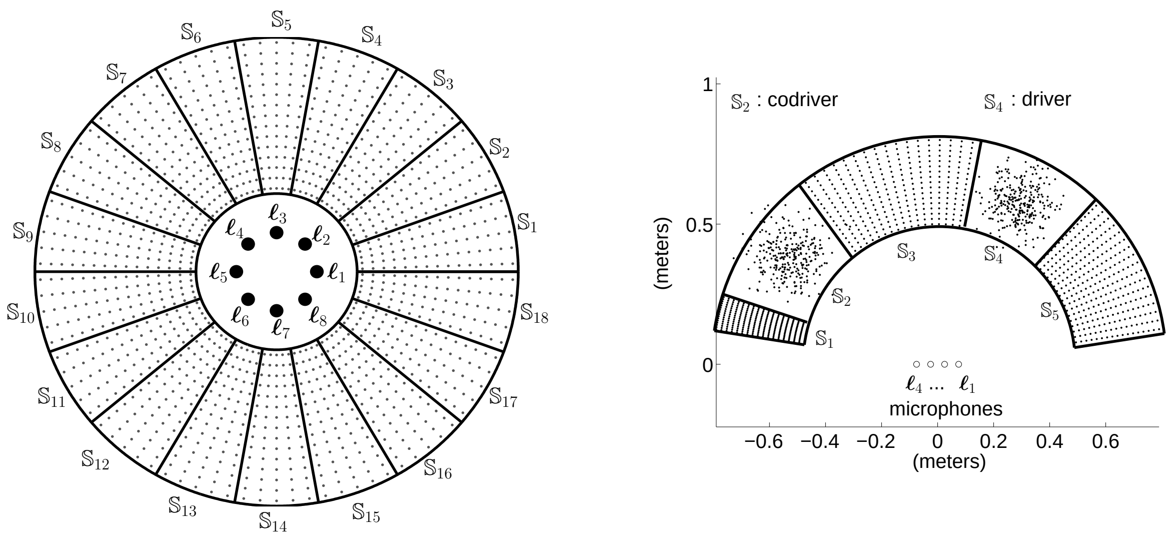
The search space around the microphone array is partitioned into $N_{\check{s}}$ connected volumes called "sectors". For example, the space around a horizontal UCA can be partitioned into $N_{\check{s}}$ "pie slices":

$$\forall \check{s} \in \{1, \cdots, N_{\check{s}}\} \quad \mathbb{S}_{\check{s}} = \left\{ (\theta, \varphi, \rho) \in \mathbb{R}^3 \;\middle|\; 2\pi \tfrac{\check{s}-1}{N_{\check{s}}} \leq \theta < 2\pi \tfrac{\check{s}}{N_{\check{s}}},\;\; 0 \leq \varphi \leq \tfrac{\pi}{2},\;\; \rho \geq \rho_0 \right\} \tag{5.9}$$

where here θ, φ, ρ designate azimuth (in radians), elevation (in radians) and radius (in meters) with respect to the microphone array center; microphones are all in the sphere $\rho < \rho_0$. See Figure 5.3a.

This section proposes a measure of wideband acoustic activity – the activeness $\zeta_{\check{s},t} \geq 0$ in a given sector $\mathbb{S}_{\check{s}}$, at a given time frame t (Figure 5.1a). The proposed activeness measure, called SAM-SPARSE-MEAN[1], is the number of discrete frequencies where a given sector is dominant over other sectors. Based on an sector-based average of the PDM (Section 5.2.1) and a comparison between sectors (Section 5.2.2), the SAM-SPARSE-MEAN activeness measure is defined in Section 5.2.3. Experiments compare SAM-SPARSE-MEAN with other measures in Section 5.2.4.

[1] "SAM" stands for "Sector-based Activeness Measure", "SPARSE" stands for "sparsity assumption", and "MEAN" stands for "average over a sector".



(a) Meeting room setup (Chapters 4 to 7)

- Uniform Circular Array (UCA)
- 8 microphones $\{\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_m, \cdots, \boldsymbol{\ell}_8\}$
- 18 sectors $\{\mathbb{S}_1, \cdots, \mathbb{S}_{\check{s}}, \cdots, \mathbb{S}_{18}\}$
- Uniform prior $P_{\check{s}}(\boldsymbol{\xi})$

(b) In-car setup (Chapter 8)

- Uniform Linear Array (ULA)
- 4 microphones $\{\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_m, \cdots, \boldsymbol{\ell}_4\}$
- 5 sectors $\{\mathbb{S}_1, \cdots, \mathbb{S}_{\check{s}}, \cdots, \mathbb{S}_5\}$
- Uniform or Gaussian prior $P_{\check{s}}(\boldsymbol{\xi})$

Figure 5.3. Two examples of microphone arrays and sector definitions. Each dot corresponds to a $\mathbf{v}_{\check{s},n}$ location. In (a) the sectors are defined in 3-D, following (5.9), but for the sake of clarity, we have only represented the horizontal plane. In (b) the sectors are defined in 2-D.

5.2.1 Averaging the PDM over a Sector

This subsection proposes to compute the Root Mean Square (RMS) of the PDM over each sector of space. This operation is shown to be equivalent to calculate the *average* SRP-PHAT power over a sector. A low-cost practical implementation is given.

RMS over a Sector: For a given sector $\mathbb{S}_{\check{s}}$, a given time frame t and a given discrete frequency k, let us express the root mean square ("MEAN") of the PDM defined in (5.8), between the observed phase vector $\mathbf{u}^{(t)}(k)$ and *all* the theoretical phase vectors $\mathbf{u}^{\mathbf{th}}(k, \boldsymbol{\xi})$ associated with *all* points $\boldsymbol{\xi} \in \mathbb{S}_{\check{s}}$:

$$\boxed{\overline{D}_{\check{s}}^{(t)}(k) \stackrel{\text{def}}{=} \left\{ \int_{\mathbb{S}_{\check{s}}} \left[d\left(\mathbf{u}^{(t)}(k), \mathbf{u}^{\mathbf{th}}(k, \boldsymbol{\xi}) \right) \right]^2 \cdot P_{\check{s}}(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi} \right\}^{\frac{1}{2}}} \tag{5.10}$$

where $P_{\check{s}}(\boldsymbol{\xi})$ is the prior distribution of the active source locations within sector $\mathbb{S}_{\check{s}}$ (for example a uniform or a Gaussian distribution, as in Figure 5.3). $\boldsymbol{\xi}$ can be expressed in any coordinate system (Euclidean or spherical), as long as the expression of $d\boldsymbol{\xi}$ is consistent with this choice.

Physical Interpretation: For a given sector $\mathbb{S}_{\check{s}} \subset \mathbb{R}^3$, at a given time frame t:

- $\left(\overline{D}_{\check{s}}^{(t)}(k)\right)^2$ is roughly equivalent to the average delay-sum energy over $\mathbb{S}_{\check{s}}$, as shown by (5.5).
- $\left(\overline{D}_{\check{s}}^{(t)}(k)\right)^2$ is strictly equivalent to the average SRP-PHAT over $\mathbb{S}_{\check{s}}$, as shown by (B.14).

Practical Implementation: In general it is not possible to derive an analytical solution for (5.10). It is therefore approximated with a discrete summation:

$$\overline{D}_{\check{s}}^{(t)}(k) \approx \hat{\overline{D}}_{\check{s}}^{(t)}(k) \quad \text{where} \quad \hat{\overline{D}}_{\check{s}}^{(t)}(k) \stackrel{\text{def}}{=} \sqrt{\frac{1}{N_{\mathbf{v}}}\sum_{n=1}^{N_{\mathbf{v}}}\left[d\left(\mathbf{u}^{(t)}(k), \mathbf{u^{th}}(k, \mathbf{v}_{\check{s},n})\right)\right]^2} \tag{5.11}$$

where $\{\mathbf{v}\} \stackrel{\text{def}}{=} \{\mathbf{v}_{\check{s},1}, \ldots, \mathbf{v}_{\check{s},n}, \ldots, \mathbf{v}_{\check{s},N_{\mathbf{v}}}\}$ is a set of $N_{\mathbf{v}}$ locations in space ($\mathbb{R}^3$) drawn from the prior distribution $P_{\check{s}}(\mathbf{v})$, and $N_{\mathbf{v}}$ is the number of locations used to approximate the continuous distribution $P_{\check{s}}(\mathbf{v})$. The sampling is not necessarily random, e.g. a regular grid for a uniform distribution (Figure 5.3). The rest of this subsection separates the observed phases from the averaging operation. This permits a low-cost implementation.

First, let us expand (5.11) into:

$$\left(\hat{\overline{D}}_{\check{s}}^{(t)}(k)\right)^2 = \frac{1}{N_{\mathbf{v}}}\sum_{n=1}^{N_{\mathbf{v}}}\frac{1}{N_q}\sum_{q=1}^{N_q}\sin^2\left(\frac{\hat{u}_q^{(t)}(k) - u_q^{\text{th}}(k, \mathbf{v}_{\check{s},n})}{2}\right) \tag{5.12}$$

Using the relation $\sin^2 u = \frac{1}{2}(1 - \cos 2u)$ we can write:

$$\begin{aligned}
\left(\hat{\overline{D}}_{\check{s}}^{(t)}(k)\right)^2 &= \frac{1}{2N_q}\sum_{q=1}^{N_q}\left\{1 - \frac{1}{N_{\mathbf{v}}}\sum_{n=1}^{N_{\mathbf{v}}}\cos\left(u_q^{(t)}(k) - u_q^{\text{th}}(k, \mathbf{v}_{\check{s},n})\right)\right\} && (5.13)\\
&= \frac{1}{2N_q}\sum_{q=1}^{N_q}\left\{1 - \Re\left[\frac{1}{N_{\mathbf{v}}}\sum_{n=1}^{N_{\mathbf{v}}} e^{j\left(u_q^{(t)}(k) - u_q^{\text{th}}(k, \mathbf{v}_{\check{s},n})\right)}\right]\right\} && (5.14)\\
&= \frac{1}{2N_q}\sum_{q=1}^{N_q}\left\{1 - \Re\left[e^{ju_q^{(t)}(k)}\frac{1}{N_{\mathbf{v}}}\sum_{n=1}^{N_{\mathbf{v}}} e^{-ju_q^{\text{th}}(k, \mathbf{v}_{\check{s},n})}\right]\right\} && (5.15)\\
&= \frac{1}{2N_q}\sum_{q=1}^{N_q}\left\{1 - \Re\left[e^{ju_q^{(t)}(k)}\, Z_{\check{s},p}^{*}(k, \{\mathbf{v}\})\right]\right\} && (5.16)
\end{aligned}$$

$$\boxed{\left(\hat{\overline{D}}_{\check{s}}^{(t)}(k)\right)^2 = \frac{1}{2N_q}\sum_{q=1}^{N_q}\left\{1-\left|Z_{\check{s},p}(k,\{\mathbf{v}\})\right|\cdot\cos\left[u_q^{(t)}(k)-\angle Z_{\check{s},p}(k,\{\mathbf{v}\})\right]\right\}} \tag{5.17}$$

where $Z_{\check{s},p}(k,\{\mathbf{v}\}) \in \mathbb{C}$ does not depend on the observed phases:

$$Z_{\check{s},p}(k,\{\mathbf{v}\}) \stackrel{\mathrm{def}}{=} \frac{1}{N_{\mathbf{v}}}\sum_{n=1}^{N_{\mathbf{v}}} e^{j u_q^{\mathrm{th}}(k,\mathbf{v}_{\check{s},n})} \tag{5.18}$$

Hence, the approximation is wholly contained in the Z parameters, which need to be computed only once. Any large number $N_{\mathbf{v}}$ can be used[2], so the approximation $\hat{\overline{D}}_{\check{s}}^{(t)}(k)$ can be as close to $\overline{D}_{\check{s}}^{(t)}(k)$ as desired. During runtime, the cost of computing the sector-based average $\hat{\overline{D}}_{\check{s}}^{(t)}(k)$ does not depend on $N_{\mathbf{v}}$: it is directly proportional to N_q, which is the same cost as for computing the point-based $d(\cdot,\cdot)$. Thus, the proposed approach ($\overline{D}_{\check{s}}^{(t)}(k)$) does not suffer from its practical implementation ($\hat{\overline{D}}_{\check{s}}^{(t)}(k)$), concerning both numerical precision and computational complexity. Note that each $Z_{\check{s},p}(k,\{\mathbf{v}\})$ value is nothing but a component of the *average* theoretical cross-correlation matrix over all points $\mathbf{v}_{\check{s},n}$ for $n \in \{1,\cdots,N_{\mathbf{v}}\}$.

5.2.2 Comparing Sectors: the Sparsity Assumption

The previous subsection described the estimation of $\overline{D}_{\check{s}}^{(t)}(k)$, the RMS of the PDM, in a given sector $\mathbb{S}_{\check{s}}$, at a given time frame t. Moreover, it was shown to be roughly equivalent to the *average* delay-sum energy over the sector $\mathbb{S}_{\check{s}}$. The directivity of the delay-sum beamforming is low for the lower frequencies, therefore we can expect "spatial leakage". Spatial leakage means that whenever there is a source in sector $\mathbb{S}_{\check{s}}$ that is active at discrete frequency k, there will be an increase of the sector-based average delay-sum power, not only for the "correct" sector $\mathbb{S}_{\check{s}}$, but also for "wrong" neighbouring sectors, such as $\mathbb{S}_{\check{s}-1}$ and $\mathbb{S}_{\check{s}+1}$. The spatial leakage induces a corresponding increase phenomenon for the $\overline{D}_{\check{s}}^{(t)}(k)$ values, at least at low frequencies.

Spatial leakage thus prevents from using $\overline{D}_{\check{s}}^{(t)}(k)$ as a direct measure of acoustic activity in a sector $\mathbb{S}_{\check{s}}$, at a discrete frequency k. On the other hand, statistical observations of human multi-party speech (Roweis, 2003) show that within a discrete frequency, only one speech source can be assumed dominant in terms of magnitude, and other sources can be neglected. Therefore, we

[2] In all meeting room experiments reported in this thesis (setup described in Fig. 5.3a), we used $N_{\mathbf{v}} = 80^3 = 512000$ points in the approximation, for each sector (all possible combinations of 80 azimuth values, 80 elevation values and 80 radius values).

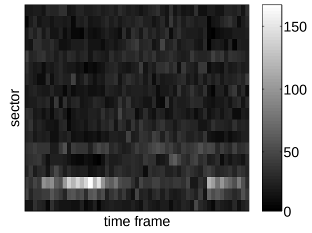


Figure 5.4. Example of sector-based activeness pattern (part of `seq01-1p-0000`). For each sector $\mathbb{S}_{\check{s}}$ and each time frame t, a sector-based activeness value $\zeta_{\check{s},t} \geq 0$ is represented, with larger values in white.

propose to determine, for each discrete frequency k, the sector to which the observed phase vector is the closest:

$$\check{s}_{\min}(k) \quad \stackrel{\text{def}}{=} \quad \arg\min_{\check{s}} \overline{D}_{\check{s}}^{(t)}(k) \tag{5.19}$$

This decision does not require any threshold. Finally, the posterior probability of having at least one active source in sector $\mathbb{S}_{\check{s}_{\min}(k)}$ and at frequency k is modeled with:

$$\boxed{P\left(\text{sector } \mathbb{S}_{\check{s}} \text{ active at discrete frequency } k \;\middle|\; \mathbf{u}^{(t)}(k)\right) \quad = \quad \delta_{Kr}\left(\check{s} - \check{s}_{\min}(k)\right)} \tag{5.20}$$

where $\delta_{Kr}(\xi)$ is the Kronecker function, equal to 1 iff $\xi = 0$, and zero otherwise. This is a sparsity assumption, similar to the one in (Roweis, 2003), which implies that all other sectors $\mathbb{S}_{\check{s}}$, $\check{s} \neq \check{s}_{\min}(k)$ are attributed a zero posterior probability of containing acoustic activity at the discrete frequency k.

5.2.3 Sector-Based Activeness: SAM-SPARSE-MEAN

To measure the wideband acoustic activity within each sector of space, we propose the following activeness measure, called SAM-SPARSE-MEAN. For a given sector $\mathbb{S}_{\check{s}}$ and a given time frame t, it is the number of strictly positive discrete frequencies where the sector is dominant:

$$\zeta_{\check{s},t} \quad \stackrel{\text{def}}{=} \quad \sum_{k=2}^{N_{\mathrm{F}}+1} P\left(\text{sector } \mathbb{S}_{\check{s}} \text{ active at discrete frequency } k \;\middle|\; \mathbf{u}^{(t)}(k)\right) \tag{5.21}$$

$$\boxed{\zeta_{\check{s},t} \quad = \quad \sum_{k=2}^{N_{\mathrm{F}}+1} \delta_{Kr}\left(\check{s} - \check{s}_{\min}(k)\right)} \tag{5.22}$$

A sector containing a wideband acoustic source such as speech is expected to be dominant over other sectors in many discrete frequencies, thus leading to a larger $\zeta_{\check{s},t}$ value. Figure 5.4 shows an example with a single speaker uttering two words from a given location. Note that $\zeta_{\check{s},t} \in \{0, 1, \cdots, N_{\mathrm{F}}\}$, and that for any time frame t, the activeness values of all sectors sum to a constant:

$$\forall t \quad \sum_{\check{s}=1}^{N_{\check{s}}} \zeta_{\check{s},t} \quad = \quad N_{\mathrm{F}} \tag{5.23}$$

5.2.4 Experiments

The task is to evaluate the wideband activeness $\zeta_{\check{s},t}$ in each (sector $\mathbb{S}_{\check{s}}$, time frame t) (Figure 5.4). Ideally, $\zeta_{\check{s},t}$ is large in a (sector, time frame) annotated as "active", and small in an "inactive" one.

Methods: We oppose point-based activeness and sector-based activeness. Point-based activeness is evaluated *at one point* in the middle of each sector (Kellermann, 1991). Sector-based activeness is evaluated *over the volume* $\mathbb{S}_{\check{s}} \subset \mathbb{R}^3$.

For point-based activeness, we selected two methods:

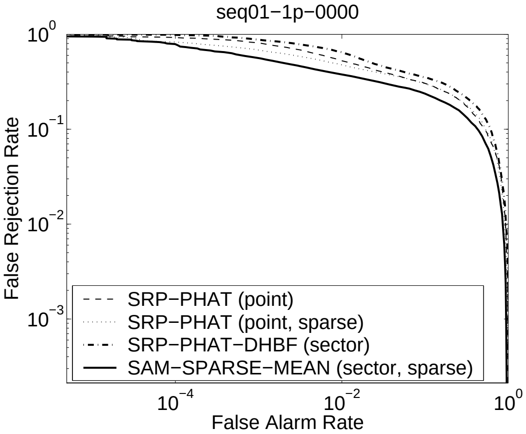
- SRP-PHAT (3.13), which is equivalent to a delay-sum after PHAT. For localization purposes, (DiBiase, 2000) shows that it is superior to both SRP (classical delay-sum), and GCC-PHAT.
- SRP-PHAT (3.13), modified with a sparsity assumption similar to (5.20). For each point, $\mathrm{P}_{\text{SRP-PHAT}}$ is obtained by summing only over discrete frequencies where this point is dominant over other points.

For sector-based activeness, we selected two methods:

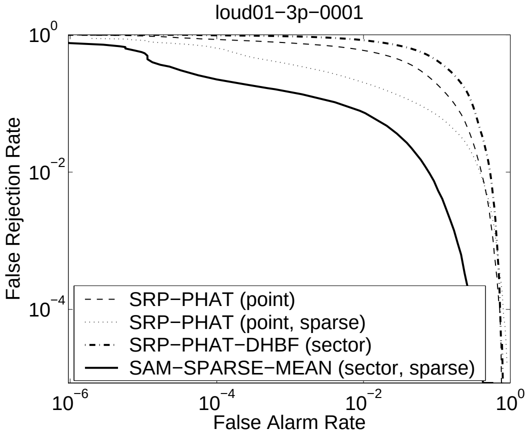
- SRP-PHAT-DHBF (Zotkin and Duraiswami, 2004), which evaluates the activeness within a sector by jointly restricting the volume of space and the bandwidth, based on an imprecision heuristic. DHBF stands for Double Hierarchical BeamForming[3].
- SAM-SPARSE-MEAN: The proposed activeness measure, as in (5.22).

Evaluation: For each method and each recording, we compute a Receiver Operating Characteristic (ROC) curve by comparing the activeness values $\zeta_{\check{s},t}$ to several threshold values. To each

[3]We tried to restrict the evaluation to sectors having a local maximum of SRP-PHAT-DHBF, but it brought a major degradation to the results. Therefore results are reported using the raw SRP-PHAT-DHBF activeness.



(a) one human speaker



(b) three loudspeakers

Figure 5.5. Examples of logarithmic ROC curves for the sector-based activeness $\zeta_{\tilde{s},t}$.

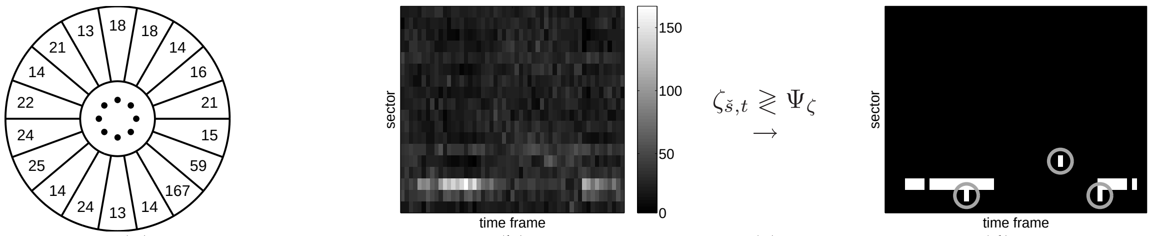
Recording	seq01	seq37	loud01	loud02	loud03
SRP-PHAT (point)	0.3957	0.5782	0.3739	0.4222	0.0735
SRP-PHAT (point, sparse)	0.3781	0.4806	0.1248	0.1169	0.1122
SRP-PHAT-DHBF (sector)	0.4571	0.6649	0.6248	0.6375	0.1217
SAM-SPARSE-MEAN (sector, sparse)	**0.3005**	**0.4199**	**0.0322**	**0.0672**	**0.0323**

Table 5.1. Average FRR for FAR in $[0, 0.1]$ (the lower, the better). Bold face indicates the best result in each column. The FRR values are larger in the case of humans (seq01 & seq37), because of the many short silences between words and syllables that are marked as "speech" in the ground-truth segmentation (see the discussion in Section 5.3.4). On the other hand, the ground-truth segmentation is exact in the case of loudspeakers (loud01, loud02 and loud03).

possible threshold value corresponds a (False Alarm Rate, False Rejection Rate[4]) point on the ROC curve (Figure 5.5). From the ROC curve, we calculate the average FRR over a practical interval of FAR (up to 0.1). FAR and FRR take values between 0 (best value) and 1 (worst value), as formally defined in Appendix A.

Results: We ran the four approaches on five annotated recordings of the AV16.3 Corpus, including a single speaker (seq01) and 3 simultaneous speakers (seq37, loud01, loud02, loud03). Numerical results are reported in Table 5.1, with two illustrative ROC curves in Figures 5.5a and 5.5b. The two SRP-PHAT results clearly show the advantage of the sparsity assumption. Comparing SAM-SPARSE-MEAN with the sparse SRP-PHAT, it is clear that averaging the SRP-PHAT over a sector of space permits better sector-based detection-localization than measuring SRP-PHAT at a single point only. The existing sector-based approach SRP-PHAT-DHBF provides a much lower performance than all others. A possible reason is that the imprecision heuristic used in SRP-PHAT-DHBF

[4]FAR and FRR are formally defined in Appendix A. A False Alarm means that a truly inactive (sector, time frame) is wrongly classified as active. A False Rejection means that a truly active (sector, time frame) is wrongly classified as inactive.



(a) Activeness values $(\zeta_{1,t}, \cdots, \zeta_{\check{s},t})$ for one time frame t

(b) Activeness values $\{\zeta_{\check{s},t}\}$ for all time frames (5.24)

(c)

(d) Final decisions $\left\{\hat{B}_{\check{s},t}\right\}$ (0 or 1). A gray circle is a false alarm.

Figure 5.6. Sector-based detection-localization (20-degree sectors): multichannel waveforms from a microphone array (dots in (a)) are transformed into "activeness" values (a,b), as in (5.22), which are thresholded to obtain the final decision (c). A false alarm happens when the ground-truth is $B_{\check{s},t} = 0$ and the final decision is $\hat{B}_{\check{s},t} = 1$.

is based on a single microphone pair, which does not use the potential of multiple microphone pairs to eliminate part of the spatial aliasing issues encountered at the higher frequencies. Moreover, SRP-PHAT-DHBF does not make a sparsity assumption, which SAM-SPARSE-MEAN does. In all cases, the proposed approach SAM-SPARSE-MEAN yields the best results. We will therefore use SAM-SPARSE-MEAN in the following, whenever sector-based detection-localization is needed.

Implementation: An optimized C implementation of the extraction of the SAM-SPARSE-MEAN values $\zeta_{\check{s},t}$ from the multichannel signals $\{x_1(t), \cdots, x_{N_m}(t)\}$ is available at:
`http://mmm.idiap.ch/Lathoud/2005-SAM-SPARSE-MEAN/`

5.3 Threshold Selection for Sector-Based Detection-Localization

The proposed SAM-SPARSE-MEAN activeness $\zeta_{\check{s},t} \in \{0, 1, \cdots, N_\mathrm{F}\}$ estimates how much wideband acoustic activity is contained in a given sector $\mathbb{S}_{\check{s}} \subset \mathbb{R}^3$ and a given time frame t (Figure 5.6a). Repeating the estimation over time frames $t \in \{t_1, \cdots, t_n, \cdots, t_{N_t}\}$ yields a spatio-temporal pattern of activeness (Figure 5.6b), where $t_n \in \mathbb{N}$ is the center time[5] (in samples) of the n-th time frame, and N_t is the number of time frames. For a given recording, the whole set of observed activeness values is denoted:

$$\{\zeta_{\check{s},t}\} \stackrel{\text{def}}{=} \{\zeta_{\check{s},t} \mid 1 \leq \check{s} \leq N_{\check{s}} \text{ and } t \in \{t_1, \cdots, t_n, \cdots, t_{N_t}\}\} \tag{5.24}$$

[5] Time frames are usually defined by a frame shift (e.g. 100 samples) and a frame length (e.g. 512 samples), as in $t_n = \frac{512}{2} + (n-1) \cdot 100$.

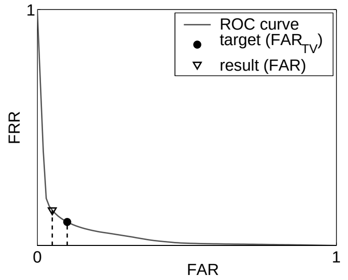


Figure 5.7. ROC curve. The task is to select a threshold Ψ_ζ such that the obtained FAR (triangle) is as close as possible to the target FAR_T (dot). Ideally $\mathrm{FAR} = \mathrm{FAR}_\mathrm{T}$.

The Task: Sector-based detection-localization requires one more step: to take a binary decision $\hat{B}_{\breve{s},t} \in \{0,1\}$ (Figure 5.6d), for example by comparing each activeness value $\zeta_{\breve{s},t}$ to a threshold Ψ_ζ: $\zeta_{\breve{s},t} \gtreqless \Psi_\zeta$ (Figure 5.6c). Errors can be made such as False Alarms (Figure 5.6d) and False Rejections, as formally defined in Appendix A. An optimal value of the threshold Ψ_ζ should be selected, to meet a user-defined performance target – for example $\mathrm{FAR} = \mathrm{FAR}_\mathrm{T}$ (Figure 5.7). The goal is *not* to improve the ROC curve. On the contrary, for a given recording, and for a given test – for example $\zeta_{\breve{s},t} \gtreqless \Psi_\zeta$, the goal is to choose a point on the ROC curve (triangle in Figure 5.7) that is as close as possible to the user-specified target (dot in Figure 5.7). We thus propose to view **sector-based detection-localization as an automatic threshold selection task.** The interest of the approaches proposed further below is that the threshold value Ψ_ζ is chosen *without knowing the true ROC curve.*

The optimal threshold value depends on environmental variations of the distribution of the $\zeta_{\breve{s},t}$ values: there could be one or multiple speakers, clean signals or background noise, etc. Thus, the classical "training" approaches, where Ψ_ζ is set on some training data, then kept fixed afterwards, may well be inadequate when "training" and "testing" environmental conditions differ widely. This section thus focuses on approaches that do not require training data. *For each environmental condition* – for example each recording, the optimal value of the threshold Ψ_ζ is estimated *automatically*, so that the performance metric is as close as possible to a user-specified constant.

Let us now assume that the structure of the observed data $\{\zeta_{\breve{s},t}\}$ is *perfectly known*, in the form of a probabilistic model $\mathcal{M}$ with *known* structure (pdf types etc.), but *unknown* parameters $\mathbf{\Lambda}(\mathcal{M})$. In such a case, the model parameters $\mathbf{\Lambda}(\mathcal{M})$, the type of test and the threshold value can be optimally selected with the Neyman-Pearson and the competitive Neyman-Pearson approaches (Levitan and Merhav, 2002). The "model+data" approach described below addresses a complementary issue, where the types of pdfs are *imperfectly* known – as is often the case in practice. A correction mechanism is proposed that is based on posterior probabilities.

The rest of Section 5.3 if organized as follows:

1. **With Training Data:** Section 5.3.1 describes the usual training/testing approach, where the Ψ_ζ value is set once, on a training set of data, and kept fixed afterwards.

2. **Proposed Approaches: Without Training Data.** As mentioned above, the optimal threshold value very often depends on environmental variations. Thus, Section 5.3.2 investigates alternative methods, where training data is *not* required. *For each environmental condition*, a probabilistic model $\mathcal{M}$ is fitted on the observed values $\{\zeta_{\tilde{s},t}\}$ using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The optimal threshold value is then *estimated* from $\mathcal{M}$. In cases where $\mathcal{M}$ does not properly fit the observed data, a posterior-based correction mechanism is proposed.

3. **Experiments:** Section 5.3.3 reports comparative experiments on real multichannel recordings, where Ψ_ζ is selected with or without training data.

4. **Openings:** Section 5.3.4 tests the proposed approaches ("without training data") with the False Rejection Rate (FRR) metric. Moreover, theoretical investigations extend the above-mentioned "correction mechanism" to multi-class classification tasks.

The proposed automatic threshold selection approaches, including the correction mechanism, are fully generic (items 2. and 4. above). They can thus be applied to any detection task, as long as a probabilistic model is available.

Data: Five real 16kHz audio sequences were taken from the AV16.3 Corpus (Chapter 4), recorded with a horizontal circular 8-microphone array (10 cm radius) set on a table (Figure 1.1a). `loud01` to `loud03` were recorded with either 2 or 3 simultaneously active loudspeakers, at various locations. `seq01` has a single human speaker at various locations. `seq37` has multiple concurrent human speakers. The total duration exceeds 1 hour. Activeness values $\{\zeta_{\tilde{s},t}\}$ are extracted as in (5.22). Time frames are 32 ms long, half-overlapping (one frame every 16 ms).

5.3.1 "Training": Threshold Selection with Training Data

One of the simplest detection strategies would be to compare the $\zeta_{\check{s},t}$ values to a fixed threshold Ψ_ζ:

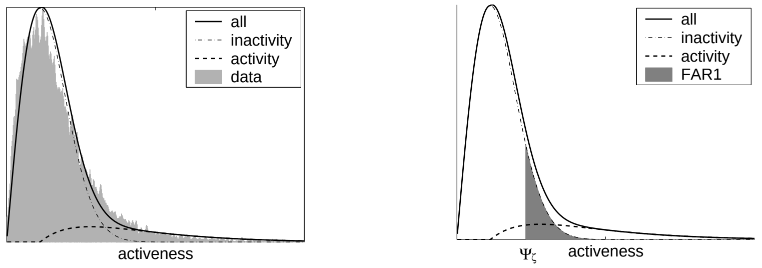
$$\zeta_{\check{s},t} \quad \gtrless \quad \Psi_\zeta \tag{5.25}$$

where Ψ_ζ is typically a preset, unique value, independent of the sector $\mathbb{S}_{\check{s}}$ or time frame t. In order to have an autonomous detection system, this strategy requires to set the threshold *in advance*, that is to train it on previously seen data. A classical approach is to use "training" data where the ground-truth $\{B_{\check{s},t}\}$ is known, and to select a threshold Ψ_ζ such that $\text{FAR}(\Psi_\zeta) = \text{FAR}_\text{T}$. The threshold Ψ_ζ is then kept fixed and applied to any unseen "test" data. For "training" we used the first 3 minutes of `loud01`, for "test" the remaining part of `loud01`, and the complete `loud02`, `loud03`, `seq01` and `seq37`. The "training" approach will fail when there is a mismatch between training conditions and testing conditions, as confirmed by the experiments reported in Section 5.3.3. Although cross-validation procedures can partially reduce the failure in mismatched conditions, their performance is still intrinsically limited by the amount and the *variety* of the training data.

5.3.2 Threshold Selection without Training Data

Training data is not used, and a threshold value Ψ_ζ is selected on *each* recording separately. The hope is to better accommodate the environmental variations across recordings. A probabilistic model $\mathcal{M}$ is fitted on unseen test data using the EM algorithm (Dempster et al., 1977). The threshold value Ψ_ζ is derived from $\mathcal{M}$, such that an estimate $\widehat{\text{FAR}}(\mathcal{M}, \Psi_\zeta)$ is equal to the target value FAR_T ("model-only" approach below). The risk of this approach is to end up relying on a model $\mathcal{M}$ that poorly fits the observed data. This issue is addressed by the "model+data" approach below.

Unsupervised Fit of a Model on the Test Data: For a given recording, the data set $\{\zeta_{\check{s},t}\}$ is collected into 1 dimension, irrespective of $\check{s}$ or t (gray histogram in Figure 5.8a). As detailed in Appendix C.1, this 1-D data can be fitted with a sensible probabilistic model with two components f_0 ("inactivity") and f_1 ("activity"). Each component is assumed to follow a Rice distribution (Rice, 1944, 1945). No manual tuning is needed and the EM cost is very small (Appendix C.1.2). The three curves in Figure 5.8a show an example of fit.



(a) 2-mixture model fitted on test data $\{\zeta_{\check{s},t}\}$

$p\left(\zeta_{\check{s},t} \mid \mathcal{M}\right) = w_0 \cdot f_0(\zeta_{\check{s},t}) + w_1 \cdot f_1(\zeta_{\check{s},t})$

(b) model-only estimation of FAR

$\widehat{\mathrm{FAR}}_1(\mathcal{M}, \Psi_\zeta) = \mathrm{FAR_T}$

Figure 5.8. (a) Unsupervised fit of a 2-mixture model $\mathcal{M}$ with parameters $\mathbf{\Lambda}(\mathcal{M}) = \{w_0, w_1, f_0, f_1\}$. The histogram in (a) is a 1-D view of all data $\{\zeta_{\check{s},t}\}$, irrespective of sector $\mathbb{S}_{\check{s}}$ or time t. w_0 and w_1 are the priors of inactivity and activity, respectively. (b) "Model-only" threshold selection, using the model $\mathcal{M}$ to match the target $\mathrm{FAR_T}$.

"Model-only" Threshold Selection: Once the model $\mathcal{M}$ is fitted on the test data, an estimate Ψ_ζ of the optimal threshold value is determined using the model $\mathcal{M}$ alone (Figure 5.8b), so that $\widehat{\mathrm{FAR}}_1(\mathcal{M}, \Psi_\zeta) = \mathrm{FAR_T}$, where:

$$\widehat{\mathrm{FAR}}_1(\mathcal{M}, \Psi_\zeta) \stackrel{\mathrm{def}}{=} \int_{\Psi_\zeta}^{+\infty} f_0(x) \cdot dx \tag{5.26}$$

Since a model is always a simplification of reality, in some cases it may not fit well the data $\{\zeta_{\check{s},t}\}$, and the $\widehat{\mathrm{FAR}}_1$ estimate will be very different from the actual FAR. The selected threshold Ψ_ζ would then lead to a FAR performance very different from the desired $\mathrm{FAR_T}$.

"Model+data" Threshold Selection: We propose to correct a possible bad fit of the model $\mathcal{M}$ by using the test data itself. Consider the definition of FAR (formally given in Appendix A):

$$\mathrm{FAR} \stackrel{\mathrm{def}}{=} \frac{\text{Number of false alarms}}{\text{Number of silent samples}} \tag{5.27}$$

Numerator and denominator can be approximated with their respective conditional expectations, using posterior probabilities, as described in the following.

Approximation of the numerator: For a given sample $\zeta_{\check{s},t}$, a false alarm happens when the detection decision is $\hat{B}_{\check{s},t} = 1$ and the ground-truth is $B_{\check{s},t} = 0$. Since the ground-truth $B_{\check{s},t}$ is unknown, we estimate the probability of having a false alarm for sample $\zeta_{\check{s},t}$ with a *posterior* probability:

$$\begin{array}{rlcl} & P\left(\hat{B}_{\check{s},t}=1,\ \underline{B}_{\check{s},t}=0 \,\middle|\, \zeta_{\check{s},t},\ \mathcal{M},\ \Psi_\zeta\right) & & \\ = & P\left(\zeta_{\check{s},t}>\Psi_\zeta,\ \underline{B}_{\check{s},t}=0 \,\middle|\, \zeta_{\check{s},t},\ \mathcal{M},\ \Psi_\zeta\right) & & \\ = & P\left(\zeta_{\check{s},t}>\Psi_\zeta \mid \underline{B}_{\check{s},t}=0,\ \zeta_{\check{s},t},\ \mathcal{M},\ \Psi_\zeta\right) & \cdot & P\left(\underline{B}_{\check{s},t}=0 \,\middle|\, \zeta_{\check{s},t},\ \mathcal{M},\ \Psi_\zeta\right) \\ = & \mathbf{1}_{\zeta_{\check{s},t}>\Psi_\zeta} & \cdot & P^{(0)}_{\check{s},t} \end{array} \tag{5.28}$$

where $P^{(0)}_{\check{s},t}$ is the posterior probability of inactivity, for sample $\zeta_{\check{s},t}$, as derived from the Bayes rule:

$$P^{(0)}_{\check{s},t} \stackrel{\text{def}}{=} P\left(\underline{B}_{\check{s},t}=0 \,\middle|\, \zeta_{\check{s},t},\mathcal{M}\right) = \frac{w_0\cdot f_0(\zeta_{\check{s},t})}{w_0\cdot f_0(\zeta_{\check{s},t})+w_1\cdot f_1(\zeta_{\check{s},t})} \tag{5.29}$$

From (5.28), the *expected* number of false alarms is:

$$\sum_{\check{s},t} P\left(\hat{B}_{\check{s},t}=1,\ \underline{B}_{\check{s},t}=0 \,\middle|\, \zeta_{\check{s},t},\ \mathcal{M},\ \Psi_\zeta\right) = \sum_{\substack{\check{s},t\\ \zeta_{\check{s},t}>\Psi_\zeta}} P^{(0)}_{\check{s},t} \tag{5.30}$$

Approximation of the denominator: the *expected* number of silent samples (that is $(\check{s},t)$ such that $\underline{B}_{\check{s},t}=0$) is $\sum_{\check{s},t} P^{(0)}_{\check{s},t}$.

Approximation of the FAR:

$$\widehat{\mathrm{FAR}}_2(\{\zeta_{\check{s},t}\},\mathcal{M},\Psi_\zeta) \stackrel{\text{def}}{=} \sum_{\substack{\check{s},t\\ \zeta_{\check{s},t}>\Psi_\zeta}} P^{(0)}_{\check{s},t} \Big/ \sum_{\check{s},t} P^{(0)}_{\check{s},t} \tag{5.31}$$

Implementation: Ψ_ζ can be determined in an efficient manner, by first ordering samples $\{\zeta_{\check{s},t}\}$ by decreasing value, irrespective of $\check{s}$ or t, and second computing cumulative sums of posteriors $P^{(0)}_{\check{s},t}$. The computational cost can be drastically decreased by ordering and reducing the data to a fixed, small number of samples (e.g. 100), similarly to the data reduction described in Appendix C.1.2.

"Model+data (N-D)" Threshold Selection for Multidimensional Models: So far, the observed values $\{\zeta_{\check{s},t}\}$ were stacked in one dimension, and the same threshold value Ψ_ζ was used for all sectors and time frames, as in the test $\zeta_{\check{s},t} \gtrless \Psi_\zeta$. However, we can use the prior knowledge of (5.23): at any time frame t, the activeness values of the various sectors sum to a constant N_F. This prior knowledge justifies a multidimensional model, where the activeness val-

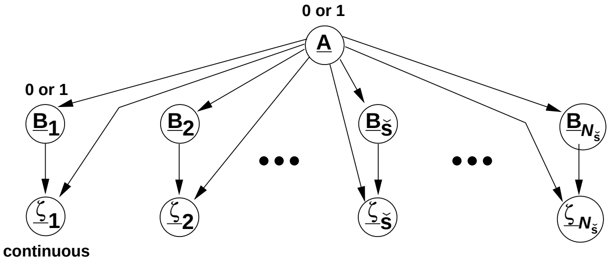


Figure 5.9. Graphical model for the independence assumptions used in the multidimensional model. The r.v. $\underline{A}$ is the frame state (inactive or active) and the r.v. $\underline{B}_{\check{s}}$ is the state (inactive or active) of a given sector $\mathbb{S}_{\check{s}}$. The r.v. $\underline{\zeta}_{\check{s}} \geq 0$ is the activeness of sector $\mathbb{S}_{\check{s}}$. On an active frame ($\underline{A}= 1$) at least one sector is active ($\exists \check{s} \quad \underline{B}_{\check{s}} = 1$).

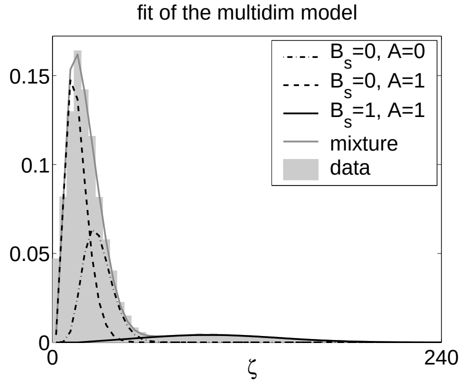


Figure 5.10. Example of fit of the three main distributions used to define the multidimensional model.

ues of the various sectors $\{\zeta_{1,t}, \cdots, \zeta_{\check{s},t}, \cdots, \zeta_{N_{\check{s}},t}\}$ are modelled jointly, by the random variables $\left\{\underline{\zeta}_1, \cdots, \underline{\zeta}_{\check{s}}, \cdots, \underline{\zeta}_{N_{\check{s}}}\right\}$. The model is illustrated in Figure 5.9, and fully described in Appendix C.2, along with its EM derivation. This model describes whether a time frame contains at least one active sector or not ($\underline{A} = 0$ or 1). In the case $\underline{A} = 0$, all sectors are inactive by definition ($\forall \check{s}\ \underline{B}_{\check{s}} = 0$). In the case $\underline{A} = 1$, each sector $\mathbb{S}_{\check{s}}$ can be active or not ($\underline{B}_{\check{s}} = 0$ or 1), and at least one sector is active.

This description results in three possible combinations, each modelled with a different pdf:

- $\underline{B}_{\check{s}} = 0$ and $\underline{A} = 0$: Gamma pdf $\mathcal{G}_{\gamma_{00},\beta_{00}}$ with parameters γ_{00} and β_{00}.
- $\underline{B}_{\check{s}} = 0$ and $\underline{A} = 1$: Gamma pdf $\mathcal{G}_{\gamma_{01},\beta_{01}}$ with parameters γ_{01} and β_{01}.
- $\underline{B}_{\check{s}} = 1$ and $\underline{A} = 1$: Shifted Rice pdf, where the shift is equal to the first moment $\gamma_{01} \cdot \beta_{01}$ of the Gamma pdf $\mathcal{G}_{\gamma_{01},\beta_{01}}$.

An example of fit of the three distributions after convergence of the EM algorithm is depicted in Figure 5.10. The hope is that this multidimensional model will better fit the observed data than the above-described 1-D model, thanks to a higher capacity *and* the incorporation of a prior knowledge.

In a multidimensional space, the 1-D thresholding strategy $\zeta_{\check{s},t} \gtrless \Psi_\zeta$ cannot be used anymore. We propose to modify the above-described "model+data" approach to accommodate multidimensional models, by taking the final binary decision based on posterior probabilities:

$$P^{(1)}_{\check{s},t} \quad \gtrless \quad \Psi_P \tag{5.32}$$

where Ψ_P is an estimate of the optimal threshold on posteriors, such that $\widehat{\mathrm{FAR}}_3(\{\zeta_{\check{s},t}\}, \mathcal{M}, \Psi_P) = \mathrm{FAR_T}$, where:

$$\widehat{\mathrm{FAR}}_3(\{\zeta_{\check{s},t}\}, \mathcal{M}, \Psi_P) \quad \stackrel{\mathrm{def}}{=} \quad \sum_{\substack{\check{s},t \\ P^{(1)}_{\check{s},t} > \Psi_P}} P^{(0)}_{\check{s},t} \;/\; \sum_{\check{s},t} P^{(0)}_{\check{s},t} \tag{5.33}$$

5.3.3 Experiments

The graphical results are given in Figure 5.11, in the form of "FAR curves" that compare the target $\mathrm{FAR_T} \in [0, 0.1]$, and the obtained FAR, where the ideal curve would be $Y = X$. The corresponding numerical statistics are given in Table 5.2 for a practical range of small $\mathrm{FAR_T}$ values, between 0.001 and 0.05. Each numerical statistic is a Root Mean Square (RMS) error:

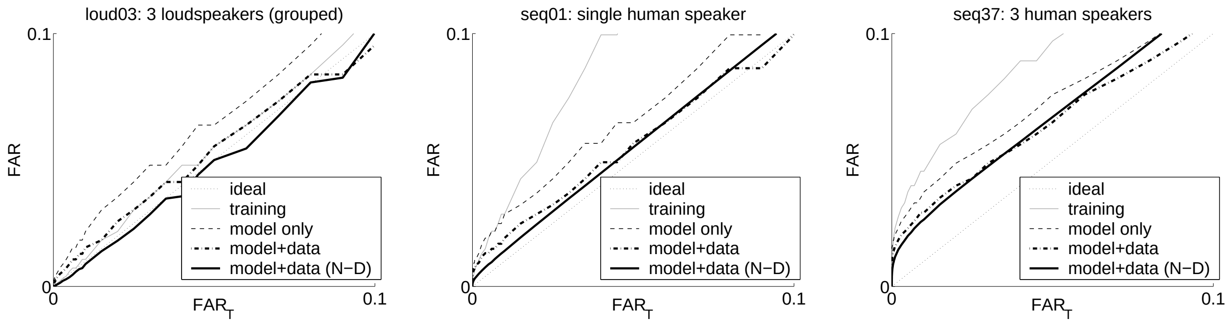
$$\left[\left\langle \left(\frac{\mathrm{FAR}}{\mathrm{FAR_T}} - 1 \right)^2 \right\rangle_{\mathrm{FAR_T} \in [0.001,\ 0.05]} \right]^{\frac{1}{2}} \tag{5.34}$$

This numerical statistic was chosen to normalize results that have varying orders of magnitude (from 0.001 to 0.05). Ideally it is equal to zero.

With Training Data: The training/testing process was repeated for various target values $\mathrm{FAR_T}$. In Figure 5.11, FAR curves compare the target $\mathrm{FAR_T}$ and the obtained FAR. The FAR curve is close to ideal ($Y = X$) on loudspeaker data, but quite far from ideal on human data. Both can be explained by the big difference between the "human" condition (real speech from humans) and the "loudspeaker" condition used during training (synthetic speech from loudspeakers). The threshold Ψ_ζ selected on the training condition *does not generalize* to the test condition.

With the 1-D Model: Two observations can be made about the RMS error:

- Compared to the "training" result, both "model only" and "model+data" approaches yield a degradation on loudspeaker data, and an improvement on human data. A possible explana-



(a) `loud03` (`loud01` & `loud02` are very similar)

(b) `seq01`

(c) `seq37`

Figure 5.11. FAR curves: comparison between target $\mathrm{FAR_T}$ & obtained FAR. In `seq37`, the positive bias is due to body noises (breathing, stomps, shuffling paper) marked as "inactivity" in the ground-truth, since their locations are unknown.

	3 loudspeakers			1 human	3 humans
Recording	(a)	(b)	(c)	(d)	(e)
training	**0.109**	0.142	0.154	1.898	3.929
model only	0.576	1.022	0.977	1.780	3.119
model+data	0.217	0.494	0.443	1.121	2.344
model+data (N-D)	0.117	**0.078**	**0.121**	**0.452**	**1.846**

Table 5.2. RMS error, as defined by (5.34), for $\mathrm{FAR_T} \in [0.001, 0.05]$. This is the RMS of $(\mathrm{FAR}/\mathrm{FAR_T} - 1)$: the lower, the better. The best result for each recording is indicated in boldface.

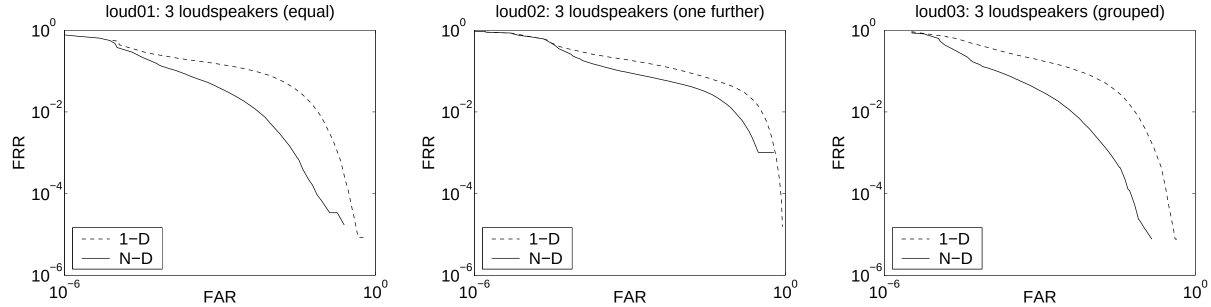
tion is that no condition-specific tuning is made in the model-based approaches, while in the "training" case, tuning was done on loudspeaker data.

- The "model+data" approach systematically reduces the RMS error, as compared with the "model only" approach. This is also visible on the FAR curves.

Both points confirm previous expectations. It is important to bear in mind that all three approaches "training", "model only" and "model+data" have the exact same ROC curve (FRR as a function of FAR), since the decision process is the same: $\zeta_{\breve{s},t} \gtrless \Psi_{\zeta}$.

Overall, although there is a major improvement over the "training" approach in terms of robustness across conditions, especially visible in Figures 5.11b and 5.11c, we can see that the results are sometimes suboptimal (loudspeaker data).

With the Multi-Dimensional Model: In all recordings, for larger values $\mathrm{FAR_T} > 0.05$, the results are similar to those of the 1-D "model+data" approach. For lower values $\mathrm{FAR_T} < 0.05$, in all recordings a systematic improvement is seen over the 1-D "model+data" approach. On the `loud01` recording, results are similar to those of best one: "training", which itself was tuned on part



(a) `loud01` (b) `loud02` (c) `loud03`

Figure 5.12. Logarithmic ROC curves on the loudspeaker recordings, for the 1-D approaches ("training", "model only", "model+data"), and for the multidimensional approach ("model+data (N-D)").

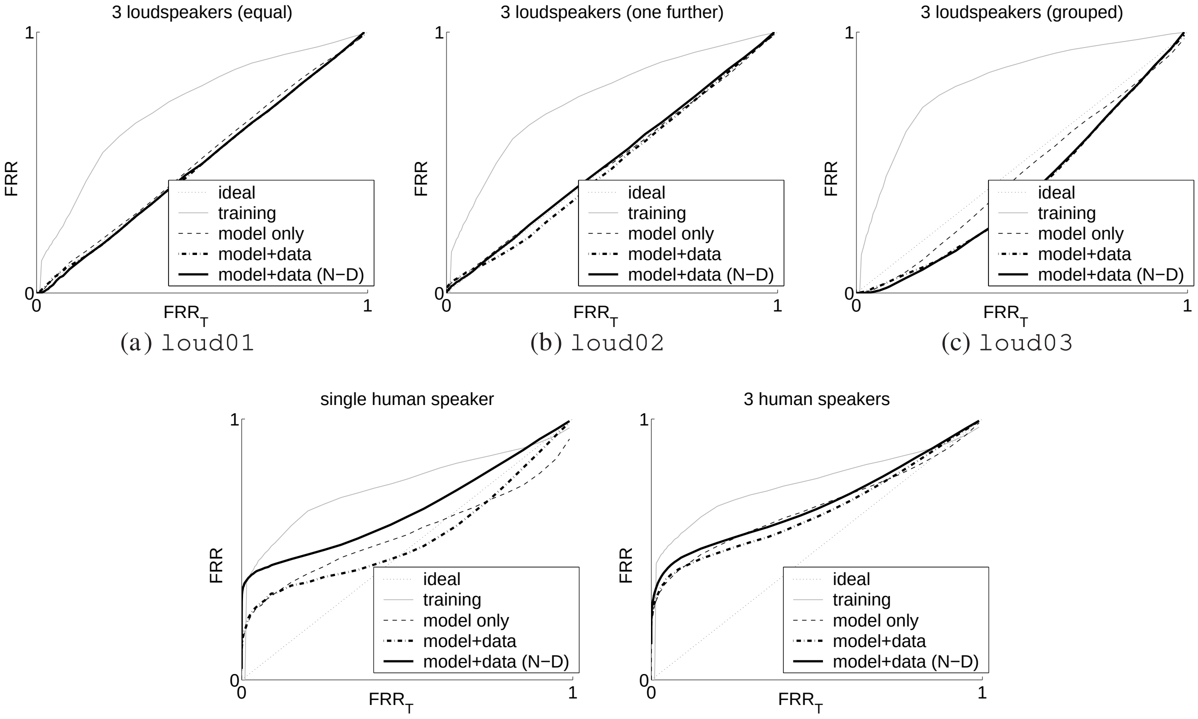
of `loud01`. This result is quite interesting given that the multidimensional approach does not use any training data. On recordings `loud02`, `loud03`, `seq01` and `seq37` the multidimensional approach yields the best results of all approaches.

However, Figure 5.11 and Table 5.2 only involve the FAR prediction performance. For the sake of completeness, we also looked at the ROC curves. As explained earlier, the three 1-D methods share the same ROC curve. On the contrary, the ROC curve of the multidimensional approach is different, because the test (5.32) is different. In the case of loudspeaker recordings a systematic improvement is seen as compared to the 1-D approaches (Figure 5.12). ROC curves on human recordings are not reliable, because of ground-truthing issues related to the FRR, as explained in Section 5.3.4.

Comparison with Frame-Level Detection Features: In Section 3.1.3, a preliminary experiment on frame-level detection for localization led us to reject traditional detection features, such as frame energy. On a frame-level detection task, Appendix D compares the proposed approach (SAM-SPARSE-MEAN + multidimensional model) with (1) frame energy, (2) a multimicrophone estimate of frame SNR (Chen and Ser, 2000), (3) the maximum SRP-PHAT value. When the detection threshold is made more conservative (increased), the proposed approach leads to a decrease of the localization error. This is not the case of the three other features. In all of the following, we therefore use the (SAM-SPARSE-MEAN + multidimensional model) approach.

5.3.4 Openings

This subsection provides insights about future extensions of the proposed model-based threshold selection approaches, from both practical and theoretical points of view.



(a) loud01 (b) loud02 (c) loud03

(d) seq01 (e) seq37

Figure 5.13. Threshold selection with and without training data, applied to loudspeaker recordings (a,b,c) and human recordings (d,e): comparison between desired target and measured False Rejection Rate. Note that all $\mathrm{FRR_T}$ values are shown (from 0 to 1). (d) and (e) illustrate the ground-truthing issue with human data.

Experiments with FRR: Similarly to $\widehat{\mathrm{FAR}}_1$, a "model only" estimate of FRR can be proposed:

$$\widehat{\mathrm{FRR}}_1(\mathcal{M}, \Psi_\zeta) \quad \stackrel{\mathrm{def}}{=} \quad \int_0^{\Psi_\zeta} f_1(x) \cdot dx \tag{5.35}$$

Similarly to $\widehat{\mathrm{FAR}}_2$, a "model+data" estimate of FRR can be proposed:

$$\widehat{\mathrm{FRR}}_2(\{\zeta_{\check{s},t}\}, \mathcal{M}, \Psi_\zeta) \quad \stackrel{\mathrm{def}}{=} \quad \sum_{\substack{\check{s},t \\ \zeta_{\check{s},t} \leq \Psi_\zeta}} P^{(1)}_{\check{s},t} \;/\; \sum_{\check{s},t} P^{(1)}_{\check{s},t} \tag{5.36}$$

Figure 5.13 shows FRR curves, that depict the resulting FRR, as a function of the target $\mathrm{FRR_T}$. Note that the whole $[0, 1]$ interval is shown. Two observations can be made:

- On loudspeaker recordings (Figures 5.13a,b,c), all the proposed model-based approaches provide a reasonable estimation of FRR, while the "training" approach fails – including on `loud01`, on which "training" was tuned (Figure 5.13a). A possible reason is that FRR is by definition linked to the distribution of "activity" (speech), which may be more variable than "inactiv-

Recording	3 loudspeakers (a)	(b)	(c)	Maximum
training	4.072	5.435	5.623	5.623
model only	0.400	2.356	0.846	2.356
model+data	**0.279**	2.759	**0.783**	2.759
model+data (N-D)	0.728	**0.743**	0.976	**0.976**

Table 5.3. RMS error over the interval $\mathrm{FRR_T} = [0.001, 0.05]$. This is the RMS of $(\mathrm{FRR}/\mathrm{FRR_T} - 1)$: the lower, the better. The best result for each recording is indicated in boldface. The rightmost column shows the maximum over all 3 recordings.

ity" (background noise), hence the "training" results are much worse than in the FAR case.

- On human recordings (Figure 5.13d,e), for all approaches a large bias can be seen in the region of small $\mathrm{FRR_T}$. The reason is most likely an issue with ground-truthing: for each location, speech segments were marked as begin- and end-point, by listening to the signal. Each speech segment very often contains many short silences and low energy speech frames, and therefore possibly many frames with low values of $\zeta_{\tilde{s},t}$. This artificially lifts up the FRR for conservative thresholds (low $\mathrm{FRR_T}$). Thus, in the human case, it is not possible to judge or to compare the FRR prediction curves.

Table 5.3 shows the RMS error for loudspeaker recordings, on the $\mathrm{FRR_T} \in [0.001, 0.05]$ range. We can see that the "model+data" approach always performs best. The "maximum" column shows that the multidimensional approach is the most robust.

Extension to Multiclass Classification: The proposed threshold selection approach can be extended to a multi-class classification context. From an observed data sample ξ and a model $\mathcal{M}$, the Maximum A Posteriori (MAP) choice of a class Q_{MAP} from a set $\{Q_1, \cdots Q_n, \cdots Q_N\}$ is:

$$Q_{\mathrm{MAP}}(\xi, \mathcal{M}) \stackrel{\mathrm{def}}{=} \underset{Q \in \{Q_1, \cdots, Q_N\}}{\arg\max} \; P(Q \mid \xi, \mathcal{M}) \tag{5.37}$$

Intuitively, if all posteriors $(P(Q_1 \mid \xi, \mathcal{M}), \cdots, P(Q_N \mid \xi, \mathcal{M}))$ have comparable values, selecting the maximum is almost equivalent to a random choice. Thus, one may want to determine the *confidence* of the decision $Q_{\mathrm{MAP}}(\xi, \mathcal{M})$. For example, a speaker recognition system would ask the user to speak again, if the maximum posterior is below a threshold:

$$\begin{cases} \text{confident}: & P(Q_{\mathrm{MAP}}(\xi, \mathcal{M}) \mid \xi, \mathcal{M}) > \Psi_P \\ \text{not confident}: & P(Q_{\mathrm{MAP}}(\xi, \mathcal{M}) \mid \xi, \mathcal{M}) \leq \Psi_P \end{cases} \tag{5.38}$$

Then again, the question of selecting the threshold Ψ_P for a given objective criterion ($\text{FAR} = \text{FAR}_\text{T}$ or other) can be addressed. Indeed, (5.38) can be seen as a detection task, and by definition:

$$P\,(\text{correct decision} \mid \xi, \mathcal{M}) \quad = \quad P\,(Q_{\text{MAP}}\,(\xi, \mathcal{M}) \mid \xi, \mathcal{M}) \tag{5.39}$$

From (5.39), the objective criterion can be estimated and the threshold Ψ_P can be selected, exactly as in the "model+data (N-D)" approach. For example, the threshold Ψ_P can be updated in an online manner: in a mobile speaker recognition system, it would be desirable to adapt the threshold to varying environmental conditions (clean, noisy, etc.).

5.4 Point-Based Localization

A fully integrated multisource detection-localization system needs to determine active time frames *and* locations ("Where? When?" question in Figure 1.2b). Moreover, the review in Section 3.1.2 showed that many existing localization methods can benefit from a prior step that reduces the search space to a limited volume of space. Thus, as a first step, we propose to restrict the search in space and time *jointly*, through sector-based detection-localization. Section 5.2 proposed a measure of wideband acoustic activity in a sector of space, called the SAM-SPARSE-MEAN activeness $\zeta_{\check{s},t}$ (Figure 5.1a). Section 5.3 investigated thresholding strategies, such as $\zeta_{\check{s},t} \gtrless \Psi_\zeta$, for sector-based detection-localization (Figure 5.1b). The present section investigates the point-based localization of the active sources, in the active sectors (Figure 5.1c). A "point" means a precise point source location $\boldsymbol{\ell}^{\text{ps}} \in \mathbb{R}^3$, e.g. within a sector $\mathbb{S}_{\check{s}} \subset \mathbb{R}^3$. The rest of this section is organized as follows:

Section 5.4.1 describes the proposed 2-step approach: sectors, then points.

Section 5.4.2 derives the gradient expression for point-based localization.

Section 5.4.3 discusses the computational cost of the gradient descent.

Section 5.4.4 applies the approach to multiple arrays.

Section 5.4.5 details a low-cost implementation.

Section 5.4.6 introduces the evaluation method for detection *and* localization.

Section 5.4.7 describes the test recordings.

Section 5.4.8 provides the results and a discussion.

5.4.1 Proposed Multisource Detection-Localization

Step 1: Sector-Based Detection-Localization: After evaluating various performance targets including FAR and FRR, we selected the following compromise. First, "conservative" detection is realized by selecting a threshold corresponding to the target FAR $= 0.005$. Second, for each active (sector, time frame) detected so far, within the same sector and within a window of time frames, e.g. ± 0.5 sec, a second threshold is applied, as determined by the "less conservative" target FRR $= 0.005$. This compromise amounts to first select utterances that are detected almost for sure (conservative FAR test), then to retrieve as many time frames as possible from each such utterance (less conservative FRR test).

Step 2: Point-Based Localization within the Active Sectors: For each time frame t, within each active sector of space (white squares in Figure 5.6d), produce an estimate $\hat{\boldsymbol{\ell}}^{(t)} \in \mathbb{R}^3$ of the most likely point of origin of the sound. By locating a source within each active sector, we can potentially achieve multisource localization, producing a set of location estimates $\left\{\hat{\boldsymbol{\ell}}_1^{(t)}, \cdots, \hat{\boldsymbol{\ell}}_{N_{\mathrm{loc}}(t)}^{(t)}\right\}$. The value of $N_{\mathrm{loc}}(t)$ can vary over time t: for instance $N_{\mathrm{loc}}(t) = 0$ on silence periods, $N_{\mathrm{loc}}(t) = 1$ on single speaker periods, $N_{\mathrm{loc}}(t) > 1$ on overlapped speech periods. As explained in Section 3.1.2, we chose to use the SRP-PHAT localization approach (DiBiase, 2000), where each location estimate is a local maximum of SRP-PHAT. To be consistent with the PDM interpretation of SRP-PHAT presented in Section 5.1.3, we implemented the search through the minimization of a PDM-based cost function Δ, within each active sector.

The issue of localization methods that try to find parameters that minimize a cost function Δ is the computational complexity of exploring the entire search space $\mathbb{R}^3$. Typically, a gradient descent approach could require many steps to converge, depending on its initialization point. Our approach reduces the cost in two ways: first, the search is limited to the active sectors. Second, the minimization is done through the Scaled Conjugate Gradient (SCG) algorithm (Moller, 1993). The SCG was chosen because of its speed efficiency, relative to other descent methods, due to its efficient approximation of second order information (Hessian) using first order derivatives only (gradient). In our case, the SCG descent only requires a few iterations to converge (typically 5 to 10). Moreover, although SCG does have numerical parameters, in practice they do not require tuning. The paper (Moller, 1993) contains very clear step-by-step instructions describing its implementation.

The critical point is to express the gradient of the cost function Δ with respect to location parameters. We are using a UCA (Figure 1.1a), which is known to realize most spatial discrimination in terms of direction, especially azimuth (direction angle in the horizontal plane), while having very poor resolution in terms of radius. Therefore, spherical coordinates $[\theta, \varphi, \rho]^{\mathrm{T}}$ are preferred. Moreover, in order to enforce the $\rho > 0$ constraint without adding any specific constraint to the gradient descent framework, we introduce "logspherical" coordinates $[\,\theta,\ \varphi,\ \mathcal{L}_\rho\,]^{\mathrm{T}} \in \mathbb{R}^3$ where:

- θ is the azimuth angle in radians,
- φ is the elevation angle in radians,
- $\mathcal{L}_\rho \stackrel{\text{def}}{=} \log \rho$, and $\rho > 0$ is the radius in meters,

of the hypothesized source location, relative to the center of the microphone array, whose geometry is known. On the other hand, we will see that the expression of the PDM cost Δ and its gradient involve the expression and the derivation of Euclidean distances. We therefore express the gradient of Δ in Euclidean space $[\mathcal{X},\ \mathcal{Y},\ \mathcal{Z}]^{\mathrm{T}}$, then convert it to logspherical space using the following formula:

$$\begin{bmatrix} \frac{\partial \Delta}{\partial \theta} \\ \frac{\partial \Delta}{\partial \varphi} \\ \frac{\partial \Delta}{\partial \mathcal{L}_\rho} \end{bmatrix} = \begin{bmatrix} -\mathcal{Y} & \mathcal{X} & 0 \\ -\mathcal{Z}\cos\theta & -\mathcal{Z}\sin\theta & \rho\cos\varphi \\ \mathcal{X} & \mathcal{Y} & \mathcal{Z} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial \Delta}{\partial \mathcal{X}} \\ \frac{\partial \Delta}{\partial \mathcal{Y}} \\ \frac{\partial \Delta}{\partial \mathcal{Z}} \end{bmatrix} \tag{5.40}$$

which is obtained from the decomposition:

$$\frac{\partial \Delta}{\partial \theta} = \frac{\partial \Delta}{\partial \mathcal{X}}\frac{\partial \mathcal{X}}{\partial \theta} + \frac{\partial \Delta}{\partial \mathcal{Y}}\frac{\partial \mathcal{Y}}{\partial \theta} + \frac{\partial \Delta}{\partial \mathcal{Z}}\frac{\partial \mathcal{Z}}{\partial \theta} \tag{5.41}$$

and similar decompositions for φ and $\mathcal{L}_\rho$. From (5.40), we can see that the additional computational complexity of using logspherical coordinates is very small, as compared to Euclidean coordinates. Moreover, in a sanity check experiment, we verified that the SCG takes less iterations to converge in logspherical coordinates, than in Euclidean coordinates.

5.4.2 The Cost Function and its Gradient in Euclidean coordinates

This subsection gives the mathematical definition of Δ, then expresses its gradient in Euclidean coordinates. It is presented in a general case where only a subset of the strictly positive discrete frequencies $\{2, \cdots, N_{\mathrm{F}}+1\}$ is used to define Δ. Note that microphone pairs $\left(\boldsymbol{\ell}_{a_q}, \boldsymbol{\ell}_{b_q}\right)$, where $q \in \{1, \cdots, N_q\}$, can be placed anywhere in the room. In particular, this allows for the use of multiple microphone arrays with exactly the same mathematical development, as long as the distances involved are reasonable.

For a given time frame t:

- Let $\hat{\boldsymbol{\ell}}_n^{(t)} \stackrel{\mathrm{def}}{=} \left[\hat{\mathcal{X}}_n^{(t)},\ \hat{\mathcal{Y}}_n^{(t)},\ \hat{\mathcal{Z}}_n^{(t)}\right]^{\mathrm{T}}$ be an instantaneous location estimate in an active sector, in 3-D Euclidean coordinates.
- Let $\left\{\mathbf{u}^{(t)}(k)\right\} \stackrel{\mathrm{def}}{=} \left\{\mathbf{u}^{(t)}(2), \cdots, \mathbf{u}^{(t)}(k), \cdots, \mathbf{u}^{(t)}(N_{\mathrm{F}}+1)\right\}$ be the set of vectors of observed phase values, one vector $\mathbf{u}^{(t)}(k) \in \mathbb{R}^{N_q}$ for each strictly positive discrete frequency $k \in \{2, \cdots, N_{\mathrm{F}}+1\}$. Each vector $\mathbf{u}^{(t)}(k) \in \mathbb{R}^{N_q}$ is defined by (5.3) and (5.6).
- Let $\mathbf{u^{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right) \in \mathbb{R}^{N_q}$ be a vector of theoretical phase values for location $\hat{\boldsymbol{\ell}}_n^{(t)}$ and discrete frequency k. $\mathbf{u^{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right) \in \mathbb{R}^{N_q}$ is defined by (5.4) and (5.7).
- Let $\Upsilon \subset \{2, \cdots, N_{\mathrm{F}}+1\}$ be a subset of the strictly positive discrete frequencies, of cardinal N_Υ ($N_\Upsilon \leq N_{\mathrm{F}}$).

We define the cost function Δ, to be minimized with respect to $\hat{\boldsymbol{\ell}}_n^{(t)}$:

$$\boxed{\begin{aligned} \Delta\left(\left\{\mathbf{u}^{(t)}(k)\right\}, \Upsilon, \hat{\boldsymbol{\ell}}_n^{(t)}\right) &\stackrel{\mathrm{def}}{=} \frac{1}{N_\Upsilon} \sum_{k \in \Upsilon} d^2\left[\mathbf{u}^{(t)}(k), \mathbf{u^{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right)\right] \\ &= \frac{1}{N_\Upsilon} \sum_{k \in \Upsilon} \frac{1}{N_q} \sum_{q=1}^{N_q} \sin^2\left[\frac{u_q^{(t)}(k) - u_q^{\mathrm{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right)}{2}\right] \end{aligned}} \tag{5.42}$$

Using the $\sin^2 u = \frac{1}{2}(1 - \cos 2u)$ equality, we obtain:

$$\Delta = \frac{1}{2} - \frac{1}{2\, N_\Upsilon\, N_q} \sum_{k \in \Upsilon} \sum_{q=1}^{N_q} \Delta_{k,q} \tag{5.43}$$

where:

$$\Delta_{k,q} \stackrel{\mathrm{def}}{=} \cos\left[u_q^{(t)}(k) - u_q^{\mathrm{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right)\right] \tag{5.44}$$

We now express the derivative of Δ with respect to one parameter $\hat{\mathcal{X}}_n^{(t)}$:

$$\frac{\partial \Delta}{\partial \hat{\mathcal{X}}_n^{(t)}} = -\frac{1}{2\, N_\Upsilon\, N_q} \sum_{k\in\Upsilon} \sum_{q=1}^{N_q} \frac{\partial \Delta_{k,q}}{\partial \hat{\mathcal{X}}_n^{(t)}} \tag{5.45}$$

Since $\frac{\partial}{\partial u}(\cos u) = -\sin u$, each term of the sum in (5.45) develops into:

$$\frac{\partial \Delta_{k,q}}{\partial \hat{\mathcal{X}}_n^{(t)}} = \frac{\partial}{\partial \hat{\mathcal{X}}_n^{(t)}} \left[u_q^{\mathrm{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right)\right] \cdot \sin\left[u_q^{(t)}(k) - u_q^{\mathrm{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right)\right] \tag{5.46}$$

From the definitions (5.4), (3.7) and (3.5) we obtain:

$$\frac{\partial}{\partial \hat{\mathcal{X}}_n^{(t)}} \left[u_q^{\mathrm{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right)\right] = -\pi \cdot \frac{k-1}{N_{\mathrm{F}}} \cdot \frac{f_s}{c} \cdot \frac{\partial}{\partial \hat{\mathcal{X}}_n^{(t)}} \left[\|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_{a_q}\| - \|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_{b_q}\| \right] \tag{5.47}$$

Using the relation $\frac{\partial b}{\partial a} = \frac{1}{2b}\frac{\partial}{\partial a}\left[b^2\right]$, we can write, for each microphone index $m = a_q$ or b_q:

$$\frac{\partial}{\partial \hat{\mathcal{X}}_n^{(t)}} \left[\|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_m\| \right] = \frac{1}{2\|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_m\|} \cdot \frac{\partial}{\partial \hat{\mathcal{X}}_n^{(t)}} \left[\|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_m\|^2 \right] = \frac{\hat{\mathcal{X}}_n^{(t)} - \mathcal{X}_m}{\|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_m\|} \tag{5.48}$$

For each point source location estimate $\hat{\boldsymbol{\ell}}_n^{(t)}$ $(1 \le n \le N_{\mathrm{loc}}(t))$, (5.45) thus becomes:

$$\boxed{\begin{aligned} \frac{\partial \Delta}{\partial \hat{\mathcal{X}}_n^{(t)}} = {} & \frac{1}{2\, N_\Upsilon\, N_q} \cdot \frac{\pi}{N_{\mathrm{F}}} \cdot \frac{f_s}{c} \\ & \cdot \sum_{k\in\Upsilon} \sum_{q=1}^{N_q} \left\{ (k-1) \cdot \sin\left[u_q^{(t)}(k) - u_q^{\mathrm{th}}\left(k, \hat{\boldsymbol{\ell}}_n^{(t)}\right)\right] \cdot \left[\frac{\hat{\mathcal{X}}_n^{(t)} - \mathcal{X}_{a_q}}{\|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_{a_q}\|} - \frac{\hat{\mathcal{X}}_n^{(t)} - \mathcal{X}_{b_q}}{\|\hat{\boldsymbol{\ell}}_n^{(t)} - \boldsymbol{\ell}_{b_q}\|} \right] \right\} \end{aligned}} \tag{5.49}$$

The exact same derivation can be conducted with $\hat{\mathcal{Y}}_n^{(t)}$ and $\hat{\mathcal{Z}}_n^{(t)}$.

Comparison with SRP-PHAT: Appendix B.3 shows that in the case where all strictly positive discrete frequencies are used ($\Upsilon = \{2, \cdots, N_{\mathrm{F}}+1\}$), minimizing Δ is strictly equivalent to maximizing $P_{\text{SRP-PHAT}}$. For both $P_{\text{SRP-PHAT}}$ and Δ, most of the computational complexity of the gradient descent lies in the gradient expression. Indeed, we derived the gradient of $P_{\text{SRP-PHAT}}$, and obtained the exact same pairwise comparisons (sine term in (5.49)). So the computational complexity of the SRP-PHAT gradient is also proportional to N_q, and there is little difference between the two complexities.

Choice of the frequency subset Υ: The sparsity assumption (5.20) led to successful sector-based detection-localization results. However, once an active sector is detected, using a similar

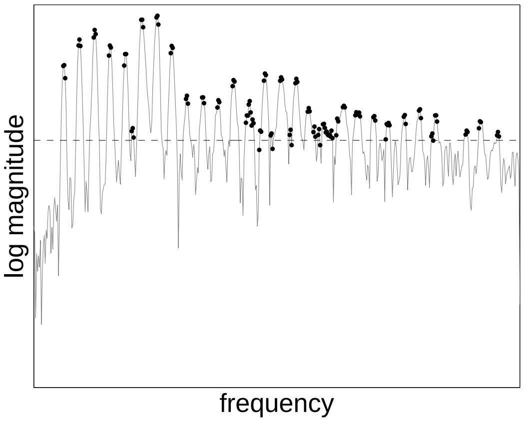


Figure 5.14. Example of frequency selection (dots): we select only the discrete frequencies with magnitude above the geometric mean (horizontal dashed line), and at or next to a magnitude peak.

assumption to define Υ tends to bias the point-based location estimate towards the middle of the sector. An alternative approach is proposed in Section 5.4.3.

5.4.3 Optimization of the Computational Complexity

The computational complexity of computing Δ (5.42) and all coordinates of its gradient (5.49) is directly proportional to the product $N_\Upsilon \cdot N_q$. In this subsection, we examine each of these factors.

$\underline{N_\Upsilon}$: We propose to only use "active" frequencies – typically spectral peaks – to define the subset Υ. Inspired from the Unsupervised Spectral Subtraction (USS) approach presented in Section 8.2, we propose the following restriction (illustrated in Figure 5.14). For each time frame t, for each discrete frequency $k \in \{2, \cdots, N_\mathrm{F} + 1\}$, the geometrical mean $M^{(t)}(k)$ of the magnitudes is computed, across all microphones of an array: $M^{(t)}(k) \stackrel{\mathrm{def}}{=} \exp\left\langle \log M_m^{(t)}(k) \right\rangle_m$. The proposed subset Υ contains only the discrete frequencies k that verify the following two conditions:

- k is at a peak of magnitude ($M^{(t)}(k) > \max\left(M^{(t)}(k-1),\ M^{(t)}(k+1)\right)$), or right next to a peak ($k-1,\ k+1$).
- k has its magnitude $M^{(t)}(k)$ above the geometrical mean $\exp\left\langle \log M^{(t)}(k) \right\rangle_k$.

$\underline{N_q}$: since we need to work on a single short time frame (e.g. 32 ms), where both speech signal and location can be assumed stationary, a large enough number of microphone pairs is required to achieve a decent spatial resolution. This is well explained in (DiBiase, 2000, Section 6.6). We verified this in practice, by reducing the number of microphones in the array (and thus the total number of pairs). In our setup, using less than 6 microphones did not provide usable localization results. All results presented in this thesis use, for each microphone array, all microphones ($N_m = 8$), and the full number of microphone pairs $N_q = N_m (N_m - 1)/2$.

Active sectors: Finally, it is also possible to limit the maximum number of active sectors to be searched, to a "reasonable" value. For example, with 20-degree azimuth sectors, the whole 360-degree range is spanned with 18 sectors. Thus, one can limit the search to the $N_{\max} = 6$ most active sectors, for example according to the posteriors $P^{(1)}_{\check{s},t}$.

SCG iterations: In practice, we found that 5 to 6 iterations are enough for the SCG descent to converge, when using the proposed cost function Δ, in logspherical coordinates.

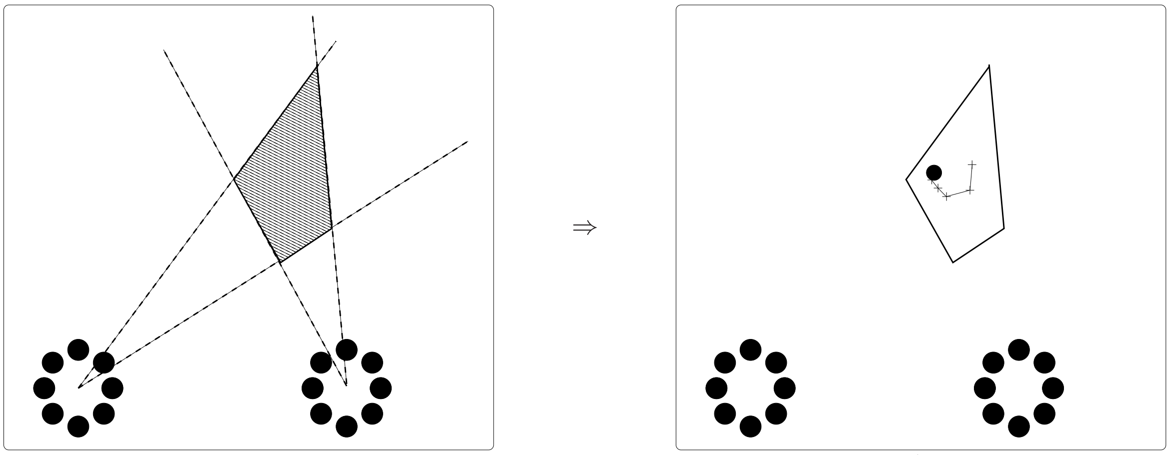
5.4.4 Multiple Microphone Arrays

The PDM cost function Δ defined by (5.42) puts no constraint on the placement of microphone pairs. Thus, the SCG descent can be applied to multiple microphone arrays. Spatial resolution is then much finer than with one array (at least in the near-field area defined by the multiple microphone arrays), so Δ and its gradient are better expressed using Euclidean coordinates $[\mathcal{X}, \mathcal{Y}, \mathcal{Z}]^{\mathrm{T}} \in \mathbb{R}^3$. However, without prior information, a complete search through the entire $\mathbb{R}^3$ space would be intractable in real-time. We thus propose to apply the same 2-step approach as in Section 5.4.1 to the case of multiple microphone arrays. In the first step, for each array independently, each sector is determined to be active or inactive (Figure 5.15a). The intersections of active sectors are then limited volumes of space in which to search for the location(s) of the source(s), through SCG descent in terms of 3-D location (Figure 5.15b)[6].

One advantage of this 2-step, sector-based approach is that these volumes, determined by the intersection of any pair of sectors, can be precomputed once for all, and stored in memory. As for the second step (SCG descent), the implementation for multiple microphone arrays is exactly the same as for a single microphone array, as already mentioned in Section 5.4.2.

We implemented the proposed technique for two arrays, and tested it on `seq01` of the AV16.3 Corpus. The two arrays are depicted in Figure 4.1. Whenever the speaker was in the near-field defined

[6] In the case that the active sectors do not intersect, we run the SCG descent on each array separately, as previously described.



(a)
Sector-based
detection-localization

(b)
Gradient descent
(SCG)

Figure 5.15. Proposed 2-step approach, with two microphone arrays.

by the two arrays, we obtained a localization error of about 10 cm in the horizontal plane $[\mathcal{X}, \mathcal{Y}]$. Moreover, a majority of location estimates were produced by the combination of the two arrays (when the active sectors actually intersect, as in Figure 5.15a). We concluded that the approach is working decently. An extensive evaluation would require additional data, with slightly more distributed geometries, so that the whole room is in the "near-field" of the two arrays. Therefore, all results presented below use one microphone array only.

5.4.5 "FULL", "FAST" and "FASTTDE" Implementations

This subsection describes three different implementations of the 2-step approach for detection-localization presented in Section 5.4.1. The goal is to compare the performance of a full-search implementation with that of low-cost, near real-time implementations. As previously mentioned, for the SCG descent we use only one location per active sector. All implementations are done in a fully online manner: the data is processed by blocks (e.g. 10 seconds), and model parameters for sector-based detection-localization (Appendix C) are updated *at the end* of each block. Finally, concerning sector-based detection-localization, the Shifted Rice (Appendix C) used in offline analysis was replaced with with a Shifted Erlang (Section 8.2) for greater stability, because each block contains less data than tested previously in the offline case (Section 5.3.3).

"FULL" implementation: In the following, the abbreviation "FULL" refers to the original, unconstrained implementation described in Sections 5.4.1 and 5.4.2, that is:

- SCG is applied within all active sectors.
- For each SCG descent, at most 30 iterations. The search is initialized in the middle of the corresponding active sector.
- All strictly positive discrete frequencies are used: $\Upsilon = \{2, \cdots, N_{\mathrm{F}} + 1\}$.
- The frame shift is 10 ms, the frame length is 32 ms.

"FAST" implementation: On the contrary, in the following, "FAST" refers to a low-cost implementation of Sections 5.4.1 and 5.4.2, using the following constraints:

- SCG is applied within, at most, the $N_{\max} = 6$ most active sectors, according to the posterior probability of activity $P^{(1)}_{\check{s},t}$.
- For each SCG descent, at most 10 iterations.
- The subset Υ of the strictly positive discrete frequencies is defined as in Section 5.4.3.
- The frame shift is 16 ms, the frame length is 32 ms.

"FASTTDE": Finally, we also implemented a variant called "FASTTDE", where the SCG descent in "FAST" is replaced with a (very fast) direct method based on time-delay estimation:

- From the 8-microphone circular array, two square subarrays are defined: $\{\boldsymbol{\ell}_1, \boldsymbol{\ell}_3, \boldsymbol{\ell}_5, \boldsymbol{\ell}_7\}$ and $\{\boldsymbol{\ell}_2, \boldsymbol{\ell}_4, \boldsymbol{\ell}_6, \boldsymbol{\ell}_8\}$ in Figure 2.1b.
- The time domain GCC-PHAT function (3.10) is estimated for the two diagonal pairs of each subarray: $(\boldsymbol{\ell}_1, \boldsymbol{\ell}_5)$, $(\boldsymbol{\ell}_3, \boldsymbol{\ell}_7)$, $(\boldsymbol{\ell}_2, \boldsymbol{\ell}_4)$ and $(\boldsymbol{\ell}_6, \boldsymbol{\ell}_8)$ in Figure 2.1b.
- For each active sector of space, and for each microphone pair, the Time-Delay Estimation (TDE) is implemented by finding the maximum (3.11) of the time domain GCC-PHAT function *within the range of time-delays corresponding to this sector*.[7]
- For each active sector of space and for each subarray, the direction of the source is estimated as an azimuth[8], from the two time-delays, as in (Brandstein, 1995, Section 7.2). It is considered

[7] This range can be estimated once, offline, from the geometry of the array and the sectors. It is then stored for all further computations.

[8] Note that elevation is also estimated during this process, but it is not very precise with UCAs.

as valid only if within the active sector. Since there are two subarrays, there may be two valid direction estimates, in which case we average them. We thus end up with zero or one azimuth direction estimate per active sector.

In FASTTDE, the sector-dependent range of permitted time-delays implicitly allows to have different time-delay values for different sectors, for the same pair of microphones. It can be seen as a principled way to apply the single-source GCC-PHAT method to a multisource problem.

Code optimization: all approaches are available in a Matlab implementation that includes C functions for GCC-PHAT, SAM-SPARSE-MEAN, SCG descent and TDE, through the MEX interface: `http://mmm.idiap.ch/Lathoud/2006-multidetloc`

5.4.6 Evaluation Method

We proposed an integrated multisource detection-localization system, so we need to jointly evaluate:

- Localization: The spatial precision resulting from the 2-step approach. The goal is to check whether the system is providing decent localization or not, across various test cases.
- Detection: The number of *correctly* localized speakers at each time frame. The goal is to check whether the sector-based detection-localization step is able to (1) detect when nobody can be correctly localized (e.g. on silences), (2) detect when an active speaker can be correctly localized (as exposed in Section 3.1.3 and Appendix D, this task differs from the single channel speech/silence discrimination task), (3) detect multiple speakers active at the same time.

While spatial precision can be defined in terms of bias and standard deviation, it is not clear what "*correctly* localized" means. Usually a threshold is arbitrarily defined on the localization error (e.g. 5 degrees). This is subject to caution, since the localization error may vary across test cases, so defining a single threshold for all test cases may not be the best choice. On the other hand, defining a separate threshold for each test case does not permit to compare results between test cases.

Thus, we propose to avoid the use of a threshold, replacing it with a statistical approach. Instead of first estimating the localization precision, then estimating the number of correctly localized speakers, both are estimated jointly, as mathematical expectation quantities, based on a simple Gaussian + Uniform model ($\mathcal{N}+\mathcal{U}$), which can be fitted on the localization error θ^{ERR} using the EM algorithm (Dempster et al., 1977). An example of fit of this model is shown in Figure 5.16.

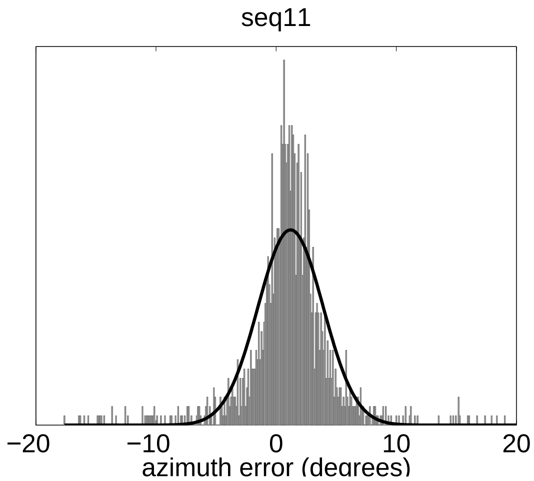


Figure 5.16. Example of fit of the Gaussian + Uniform model $(\mathcal{N}+\mathcal{U})$ on the localization error θ^{ERR}. $(\mathcal{N}+\mathcal{U})$ is used for evaluation of the detection-localization. `seq11` is a recording with a single moving speaker, from the AV16.3 Corpus (Chapter 4). The gray histogram represents the distribution of localization errors. The dark curve represents the Gaussian pdf, modelling "correct" location estimates. The uniform pdf is not represented.

Formally, let θ^{ERR} be an azimuth error in degrees, that is the difference between a given result location estimate and its closest location in the ground-truth, between -180 and +180 degrees. The probability density function (pdf) of the $(\mathcal{N}+\mathcal{U})$ model is defined as a mixture of two components:

$$p\left(\theta^{\mathrm{ERR}} \mid (\mathcal{N}+\mathcal{U})\right) \quad \stackrel{\mathrm{def}}{=} \quad P_{\mathcal{N}} \cdot f_{\mathcal{N}}\left(\theta^{\mathrm{ERR}}\right) + \frac{P_{\mathcal{U}}}{360} \tag{5.50}$$

where $P_{\mathcal{N}}$ and $P_{\mathcal{U}}$ are the priors of "correctly localized" and "incorrectly localized", and $f_{\mathcal{N}}$ is assumed to be a Gaussian pdf with parameters $\mu_{\mathcal{N}}$ and $\sigma_{\mathcal{N}}$:

$$f_{\mathcal{N}}\left(\theta^{\mathrm{ERR}}\right) \quad \stackrel{\mathrm{def}}{=} \quad \mathcal{N}_{\mu_{\mathcal{N}},\sigma_{\mathcal{N}}}\left(\theta^{\mathrm{ERR}}\right) \tag{5.51}$$

Then the desired quantities for performance evaluation of localization and detection can all be directly estimated from the $(\mathcal{N}+\mathcal{U})$ model:

- $\mu_{\mathcal{N}}$ and $\sigma_{\mathcal{N}}$ are the proposed estimates of the bias and standard deviation of the localization error, for an audio source that was detected and "correctly localized".
- $P_{\mathcal{N}}$ is the proportion of location estimates that are correct.
- In a given time frame t, the number of correctly located speakers $\hat{n}_{\mathrm{C}}(t)$ is estimated as a

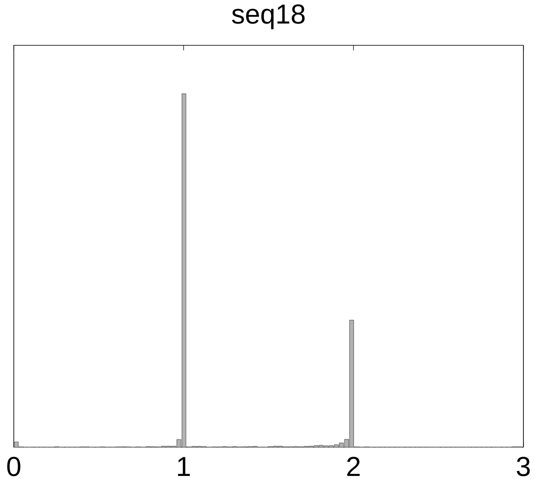


Figure 5.17. Histogram of $\hat{n}_{\mathrm{C}}(t)$, the estimated number of correctly localized speakers in the 2-speaker sequence `seq18`. Note that $\hat{n}_{\mathrm{C}}(t) \in \mathbb{R}$ can take non-integer values.

conditional expectation:

$$\begin{aligned}
\hat{n}_{\mathrm{C}}(t) &\stackrel{\mathrm{def}}{=} \mathrm{E}\left\{ \underline{n}_{\mathrm{C}} \;\middle|\; (\mathcal{N}+\mathcal{U}), \left\{\theta_1^{\mathrm{ERR}}(t), \cdots, \theta_n^{\mathrm{ERR}}(t), \cdots, \theta_{N_{\mathrm{loc}}(t)}^{\mathrm{ERR}}(t)\right\}\right\} && (5.52) \\
&= \sum_{n=1}^{N_{\mathrm{loc}}(t)} P\left(\text{correctly localized} \mid (\mathcal{N}+\mathcal{U}), \theta_n^{\mathrm{ERR}}(t)\right) && (5.53) \\
&= \sum_{n=1}^{N_{\mathrm{loc}}(t)} \frac{P_{\mathcal{N}} \cdot f_{\mathcal{N}}\left(\theta_n^{\mathrm{ERR}}(t)\right)}{P_{\mathcal{N}} \cdot f_{\mathcal{N}}\left(\theta_n^{\mathrm{ERR}}(t)\right) + \frac{P_{\mathcal{U}}}{360}} && (5.54)
\end{aligned}$$

where $N_{\mathrm{loc}}(t)$ is the number of location estimates at time frame t, given by the detection-localization system (0, 1 or more), $\left\{\theta_1^{\mathrm{ERR}}(t), \cdots, \theta_n^{\mathrm{ERR}}(t), \cdots, \theta_{N_{\mathrm{loc}}(t)}^{\mathrm{ERR}}(t)\right\}$ the corresponding azimuth errors, assumed independent. An example of histogram of all $\hat{n}_{\mathrm{C}}(t)$ values for all time frames t is shown in Figure 5.17. To summarize this result, we discretize $\hat{n}_{\mathrm{C}}(t)$ into integer bins, counting the values $\hat{n}_{\mathrm{C}}(t)$ such that: $0 \leq \hat{n}_{\mathrm{C}}(t) < 0.5$, $0.5 \leq \hat{n}_{\mathrm{C}}(t) < 1.5$, $1.5 \leq \hat{n}_{\mathrm{C}}(t) < 2.5$, etc.

A complete Matlab code to conduct the evaluation, along with examples, is available at:

`http://mmm.idiap.ch/Lathoud/2006-multidetloc`

5.4.7 Experimental Protocol

The three implementations FULL, FAST and FASTTDE were run on 8 recordings of the AV16.3 Corpus. We present azimuth results obtained with one 8-microphone circular array[9]. 3 cameras were used to reconstruct the "true" 3-D mouth location of each speaker, relative to the microphones, as described in Section 4.3.2 (with an error less than 1.2 cm). We summarize here the contents of the 8 recordings (Section 4.2.3).

Two recordings contain mostly static speakers:

- `seq01`: 1 speaker at 16 different locations, facing the arrays,
- `seq37`: 3 simultaneous speakers, 2 seated and 1 standing at 5 different locations, facing the arrays.

Six recordings contain mostly moving speakers:

- `seq11`: 1 moving speaker, speaking continuously, facing the arrays.
- `seq15`: 1 moving speaker, speaking discontinuously, with long silences.
- `seq18`: separation test: 2 moving speakers getting as close to each other as possible, facing the arrays.
- `seq24`: crossing test: 2 moving speakers passing in front of each other, facing the arrays.
- `seq40`: partial occlusion: test similar to `seq37`, except that the standing speaker is continuously moving.
- `seq45`: motion & full occlusion: 3 moving speakers, walking around while speaking continuously.

In all cases, after running multisource detection-localization (FULL, FAST or FASTTDE), we removed noisy location estimates, using short-term clustering (Chapter 6) followed by the cheap SNSLOW Speech/Non-Speech discrimination (Section 5.5.2). The focus here is the quality of the detection-localization ; please refer to Sections 5.5 and 6.4.2 for more details on SNS.

[9] This is "array one" in the AV16.3 Corpus.

5.4.8 Results and Discussion

Figure 5.18 depicts examples of localization results, along with the ground-truth locations of the various speakers. Table 5.4 presents localization results on all 8 recordings, in terms of bias, standard deviation and percentage correct, as given by the $(\mathcal{N}+\mathcal{U})$ evaluation (Section 5.4.6). As for detection, Table 5.5 presents the distribution of $\hat{n}_{\mathrm{C}}(t)$, the number of speakers correctly detected *and* located, estimated as described in Section 5.4.6. Table 5.6 shows the effective computational complexity.

Global results: In all cases, speakers appear to be detected and located in a seamless manner, while they move from one sector to the next. For all eight recordings except `seq40`, visual inspection of the location estimates against the ground-truth confirms that FULL, FAST and FASTTDE (1) effectively detect and locate multiple sources, (2) exhibit a low number of spurious location estimates. (1) is confirmed by the distributions shown in Table 5.5, which have significant components with 2 or more speakers. (2) is confirmed by the "percentage correct" in Table 5.4, which is often between 95 % and 100 %. The failure on `seq40` (Figure 5.18) is reflected by the high standard deviation values in Table 5.4. This may have two possible (non-exclusive) explanations: (1) strong interference between the three speech signals, due to *partial* occlusions, (2) low power received at the microphone array, due to the downward orientation of the speakers'heads (Chu and Warnock, 2002), because they are reading books aloud. It contrasts with the success on `seq45` (Figure 5.18), which also contains three simultaneous speakers, but *full* occlusions. For all three methods, the standard deviation on `seq45` is within the range of other, "easier" recordings.

Comparison between FULL and FAST: Looking at the averages in Table 5.4, two observations can be made. The FAST implementation exhibits a localization precision similar to that of the FULL implementation, but a higher percentage of correct location estimates. This is possibly a positive impact of the discrete frequency selection strategy used in FAST (Figure 5.14). The total duration of correct location estimates (not reported) is lower for FAST than for FULL, by 4.2 %. This can be explained by the limit on the number of active sectors, in the case of FAST. Finally, the computational complexity of FAST is much lower, in fact close to real-time speed (Table 5.6), although part of the implementation is still in Matlab. Overall, we can state that FAST is an effective multisource implementation of a parametric search for *multiple* local maxima of SRP-PHAT. Various possibilities arise to achieve real-time implementation, including:

- Parallelization of the SCG descent over several dedicated CPUs (one per active sector).
- Integration within a tracking framework (use past knowledge through an update mechanism).
- Implementation in C of the remaining Matlab code (see the “Input” column in Table 5.6).

Comparison between FAST and FASTTDE: In the case of FASTTDE, the localization cost becomes negligible (“TDE” column in Table 5.6), but the localization precision is degraded (standard deviation in Table 5.4). This degradation confirms the extensive study in (DiBiase, 2000), that compares GCC-PHAT and SRP-PHAT. Also, some noisy location estimates appear, as shown by the “2” column in Table 5.5, for the single-speaker sequences `seq01`, `seq11` and `seq15`.

Specific multispeaker case: `seq37` is a case with only simultaneous speakers (2 or 3), for extended periods of time (about one minute for each combination of speakers and locations). We compared visually the location estimates with the ground-truth. On the positive side, the 2 or 3 speakers are correctly detected and located over a long run (e.g. a minute), *as long as they are at comparable distances from the array*. On the other hand, whenever one speaker was standing about two times further from the array than the other two (seated) speakers, he was nearly completely missed[10]. Further analysis showed that this is because the “missed” speaker is dominant in fewer discrete frequencies, because of his/her further distance from the array. This induces a low SAM-SPARSE-MEAN value $\zeta_{\check{s},t}$, and therefore a low posterior probability of activity $P^{(1)}_{\check{s},t}$.

To conclude, we can state that the proposed detection-localization system was proved able to detect and correctly locate up to 3 simultaneous speakers, including on a highly dynamic 3-speaker case (e.g. FAST on `seq45`). On the other hand, the failure observed on the “partial occlusion” case `seq40` suggests that to investigate joint spectrum-location estimation. Work going in that direction includes a signal subspace microphone array approach relying on a frequency-dependent model for each source (Grenier, 1994), a joint separation of the spectra from known locations (Sekiya and Kobayashi, 2004), and a data-driven approach for binaural localization (Nix and Hohmann, 2006).

The computational complexity of the gradient descent was effectively reduced by using only the selected set Υ of discrete frequencies (Figure 5.14). A further reduction of the computational complexity may be obtained by reducing the number of discrete frequencies in a uniform manner. Indeed, within each active sector, one could combine several consecutive discrete frequencies into a single band, through array interpolation, similarly to (Friedlander and Weiss, 1993).

[10] A similar phenomenon appears in `loud02`, in the loudspeaker experiments reported in Section 5.3.3.

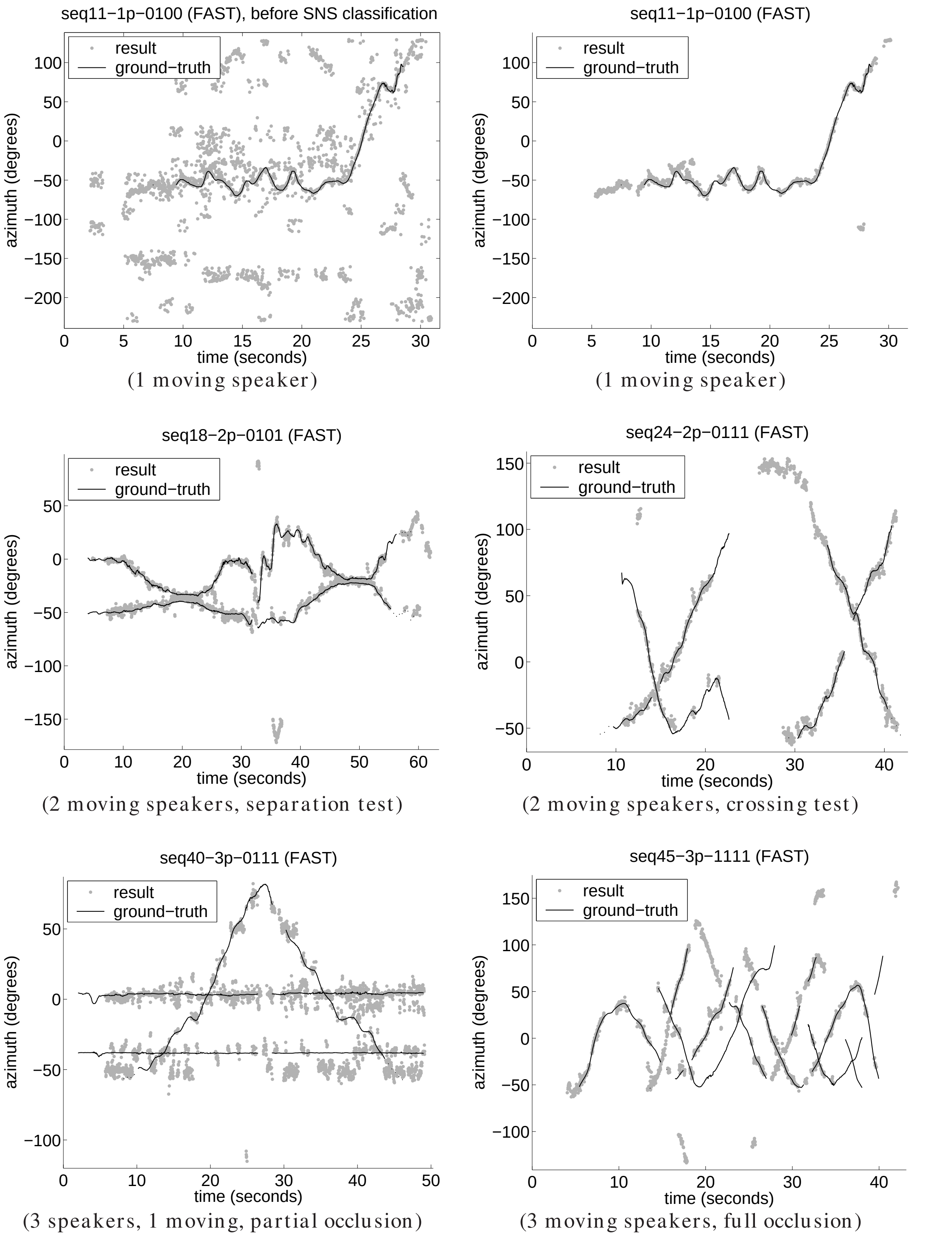


Figure 5.18. Result (red dots) of the detection-localization ("FAST" implementation, followed by short-term clustering and SNSLOW). The ground-truth (black curves) is derived from the cameras, including on silences. Gaps are due to the mouth of a person being occluded on at least one camera (*gaps are not related to silences*).

Recording	FULL (all active sectors) $\mu_{\mathcal{N}}$ Bias	$\sigma_{\mathcal{N}}$ Std dev.	$P_{\mathcal{N}}$ % corr.	FAST (up to 6 sectors) $\mu_{\mathcal{N}}$ Bias	$\sigma_{\mathcal{N}}$ Std dev.	$P_{\mathcal{N}}$ % corr.	FASTTDE (up to 6 sectors) $\mu_{\mathcal{N}}$ Bias	$\sigma_{\mathcal{N}}$ Std dev.	$P_{\mathcal{N}}$ % corr.
seq01 (1, static)	-0.47	2.65	96.4	-0.33	2.60	97.6	0.38	3.46	98.7
seq37 (3, static)	-0.05	2.63	90.3	0.63	2.68	95.8	2.75	6.57	97.4
seq11 (1, moving)	1.18	2.78	87.3	1.29	2.67	92.6	2.36	5.69	97.3
seq15 (1, moving)	0.30	1.76	79.1	0.17	1.77	89.3	1.19	5.30	88.0
seq18 (2, moving, separation test)	0.32	2.09	93.4	0.39	2.06	96.2	0.61	3.18	98.1
seq24 (2, moving, crossing test)	0.16	2.99	90.4	0.22	2.99	96.3	-0.00	4.04	98.6
seq40 (3, moving, partial occlusion)	-1.31	**5.37**	100	-1.94	**6.02**	99.7	-0.16	**6.44**	100
seq45 (3, moving, full occlusion)	0.36	**3.30**	91.3	0.38	**2.46**	88.3	0.16	**3.65**	93.7
Average	0.06	2.95	91.0	**0.10**	**2.91**	**94.5**	0.91	4.79	96.5

Table 5.4. Localization precision, in degrees, along with the percentage of correct location estimates.

Recording	FULL 0	1	2	3	4	FAST 0	1	2	3	4	FASTTDE 0	1	2	3	4
seq01-1p-0000	1.4	98.3	0.4	0.0	0.0	1.2	98.5	0.3	0.0	0.0	0.1	89.3	10.5	0.0	0.0
seq37-3p-0001	0.6	62.9	**35.2**	**1.3**	0.0	0.4	68.0	**30.4**	**1.3**	0.0	0.2	50.7	**40.7**	**8.2**	0.1
seq11-1p-0100	3.7	95.3	1.1	0.0	0.0	2.1	97.1	0.8	0.0	0.0	0.6	82.2	17.2	0.0	0.0
seq15-1p-0100	2.4	97.1	0.5	0.0	0.0	2.1	97.7	0.2	0.0	0.0	2.2	78.9	18.3	0.6	0.0
seq18-2p-0101	0.6	65.7	**33.5**	0.2	0.0	0.6	72.2	**27.1**	0.1	0.0	0.5	56.2	**34.2**	9.1	0.0
seq24-2p-0111	1.6	73.4	**24.1**	0.9	0.0	0.5	76.9	**21.9**	0.7	0.0	0.3	76.7	**21.1**	1.9	0.0
seq40-3p-0111	0.0	51.7	**40.4**	**7.5**	0.5	0.0	54.5	**38.7**	**6.4**	0.5	0.0	41.1	**41.8**	**15.6**	1.5
seq45-3p-1111	0.3	57.2	**30.4**	**12.1**	0.0	0.6	66.9	**28.7**	**3.9**	0.0	0.0	74.7	**19.0**	**5.2**	1.1

Table 5.5. Distribution of the number of correct simultaneous location estimates (percentage of frames). For recordings with multiple simultaneous speakers, the multispeaker cases are in bold face.

Recording	FULL Input	SCG	Total	FAST Input	SCG	Total	FASTTDE Input	TDE	Total
seq01-1p-0000	0.70	9.56	13.29	0.28	0.42	1.54	0.43	0.07	1.29
seq37-3p-0001	0.68	20.93	26.93	0.27	0.91	2.84	0.43	0.12	2.15
seq11-1p-0100	0.69	19.04	24.39	0.27	0.73	2.21	0.41	0.10	1.76
seq15-1p-0100	0.69	9.68	14.10	0.27	0.43	1.65	0.42	0.06	1.46
seq18-2p-0101	0.70	25.50	31.52	0.25	1.02	2.82	0.42	0.12	1.94
seq24-2p-0111	0.69	19.02	24.43	0.28	0.76	2.37	0.42	0.11	1.78
seq40-3p-0111	0.67	27.19	33.55	0.28	0.97	2.76	0.43	0.11	1.96
seq45-3p-1111	0.67	22.71	28.71	0.28	0.85	2.53	0.43	0.11	1.91
Average	0.69	19.20	24.62	0.27	0.76	2.34	0.42	0.10	1.78

Table 5.6. Effective computational complexity: computation duration divided by recording duration *(real time = 1)*. We used a Matlab/C implementation on a Pentium 4, with 3.2 GHz CPU speed and 1 GB of RAM. "SCG" is the time spent doing SCG descent only. "TDE" is the time spent doing TDE-based localization only. "Input" is the time spent reading and buffering wave files (variations due to Matlab). The cost of FFT and GCC-PHAT is very small (around 0.003 real time duration).

5.5 Speech/Non-Speech (SNS) Classification

In an environment with human sound sources only, relying on the SAM-SPARSE-MEAN detection-localization approach, as in the above sections, may be sufficient to discriminate speech from silence, as illustrated by the experiments reported above. However, the type of sound source may well be less constrained, even indoors: for example, machines such as a beamer and laptops may be used in a meeting. In such a case, a "wideband source = speech source" assumption is not enough to discriminate between speech and non-speech.

This section thus proposes a further extension of SAM-SPARSE-MEAN to the Speech/Non-Speech (SNS) classification task. The spectrum is filtered in a location-dependent manner, thus producing "Sector-Based MFCCs". Two SNS classifiers are proposed: SNSLOW and SNSGMM. SNSLOW uses a fixed threshold on the Sector-Based MFCC_0, for virtually no cost. SNSGMM models Sector-Based MFCCs in an unsupervised manner, with a full covariance matrix GMM. This model is then splitted in two, to discriminate between speech activity and machine activity. Chapter 6 shows that it is highly beneficial to integrate SNSLOW or SNSGMM within a *dynamical* analysis, therefore all comparative SNS experiments are included in Chapter 6.

5.5.1 Sector-Based MFCCs

We propose to filter the spectrum of a single microphone $\boldsymbol{\ell}_m$, prior to MFCC extraction, in a location-dependent manner. Based on the sparsity assumption defined in (5.20), for each sector, the frequency spectrum is filtered in a binary manner, setting to zero the magnitude at a discrete frequency if a sector is not dominant in that frequency, as in the expression $M_m^{(t)}(k) \cdot \delta_{Kr}(\check{s} - \check{s}_{\min}(k))$. However, since setting magnitude to zero may introduce artificial dynamics in the cepstrum domain, the spectrum is floored to a non-zero value σ corresponding to the average background noise level, as in USS (Section 8.2). We define the sector-based magnitude spectrum as:

$$M_m^{(t)}(k, \check{s}) \stackrel{\text{def}}{=} \max\left(1, \frac{M_m^{(t)}(k)}{\sigma} \cdot \delta_{Kr}(\check{s} - \check{s}_{\min}(k))\right) \tag{5.55}$$

where σ is the Rayleigh parameter in the Rayleigh + Shifted Erlang model described in Section 8.2. Sector-based MFCC coefficients can then be extracted from $M_m^{(t)}(k, \check{s})$. This way of "separating" spectra from different sources is very approximate for two reasons. First, it does not use the precise spatial location of each source, but only their sector index. Second, the magnitude spectrum of only one microphone is used. Nevertheless, it is sufficient to build an unsupervised speech/non-speech classifier, as described in Section 5.5.2. The effective computational complexity is negligible. A complete Matlab implementation of sector-based MFCC extraction is available at:

`http://mmm.idiap.ch/Lathoud/2006-multidetloc`

5.5.2 Low-Cost Speech/Non-Speech Classifier (SNSLOW)

As a baseline system, we propose to use measures of non-stationarity and wideband activeness: the Sector-Based MFCC_0 coefficient and SAM-SPARSE-MEAN, respectively. All non-speech segments are discarded, only speech segments are kept. A speech segment is defined as a segment that has:

$$\text{standard deviation of Sector} - \text{Based MFCC}_0 \quad > \quad 0.6 \tag{5.56}$$

and at least two "wideband (sector, frame)". A wideband (sector, frame) $(\mathbb{S}_{\breve{s}}, t)$ must verify:

$$P^{(1)}_{\breve{s},t} \quad > \quad \Psi_P\,(\mathrm{FAR} = 0.001) \tag{5.57}$$

The threshold $\Psi_P\,(\mathrm{FAR} = 0.001)$ is determined in an automatic manner, without training data, following (5.33). It corresponds to a target FAR_T of 0.001. The fixed threshold value of 0.6 on the sector-based MFCC_0 coefficient is justified by the fact that sector-based MFCCs are derived from a normalized spectrum (5.55).

5.5.3 Full Covariance GMM Speech/Non-Speech Classifier (SNSGMM)

One approach for Speech/Non-Speech (SNS) classification is to train a model on a set of recordings, against a known ground-truth segmentation, and to test on another set of recordings, where the unknown segmentation is to be estimated (Lu and Zhang, 2002). However, this type of approach runs the risk of "overfitting" the data seen during training, thus not performing well on unseen data that widely differs from the training data, as for example, different microphone characteristics, different types of noises, different types of room reverberation, etc. We thus opted for an approach that does not need training data. This approach relies on the following assumption: in spite of the Discrete Cosine Transform, the various dimensions of the MFCC features of speech signals are still correlated, while for machines such as a projector, there is much less correlation (in particular, if the machine noise is stationary). This leads to the following approach. For a given recording:

- Fit a full covariance matrix GMM on the sector-based MFCC features of active sectors, using the EM algorithm (Dempster et al., 1977).
- During EM, some of the components of the GMM may need to be constrained. This typically happens when a full covariance matrix becomes badly conditioned, thus not invertible within the numerical limits of the computer. In such a case, we chose to constrain the covariance matrix to be diagonal, and the EM algorithm continues with the newly constrained model.
- The number of GMM components is chosen by trying various numbers, from 2 to 10, and picking the fitted GMM with the maximum Bayesian Information Criterion (Schwartz, 1978).
- The selected GMM is then separated into two GMMs: diagonal components in one GMM, to model noise sources, non-diagonal components in the other GMM, to model speech sources.

Final decision: All non-speech segments are discarded, only speech segments are kept. A speech segment is defined as a segment that has:

$$\text{standard deviation of Sector} - \text{Based MFCC}_0 \quad > \quad 0.6 \tag{5.58}$$

and at least two "speech (sector, frame)". A speech (sector, frame) $(\mathbb{S}_{\check{s}}, t)$ must verify:

$$P\left(\text{wideband and non} - \text{noisy}\right) \quad > \quad \Psi_P\left(\text{FAR} = 0.01\right) \tag{5.59}$$

where a simplifying independence assumption gives the posterior probability:

$$P\left(\text{wideband and non} - \text{noisy}\right) \quad = \quad P^{(1)}_{\check{s},t} \cdot P\left(\text{non} - \text{noisy}\right) \tag{5.60}$$

and $\Psi_P\left(\text{FAR} = 0.01\right)$ is a threshold on the posteriors. $\Psi_P\left(\text{FAR} = 0.01\right)$ is determined automatically, without training data, to correspond to a target FAR_T of 0.01, following (5.33). $P\left(\text{non} - \text{noisy}\right)$ is the posterior probability of "non-noisiness", as given by the two-GMM SNS model. A Matlab implementation of the final SNS decision is available at:

`http://mmm.idiap.ch/Lathoud/2006-multidetloc`

Both SNSLOW and SNSGMM are tested in Chapter 6, using a short-term clustering algorithm to produce the candidate speech and non-speech segments.

5.6 Conclusion

Section 3.1.3 presented a preliminary experiment, which suggested to avoid traditional single-channel features, in the context of detection for localization. Following this suggestion, the present chapter addressed the multisource detection-localization task, which amounts to detect and locate multiple acoustic sources, often moving, from a single time frame of multichannel signals. A 2-step approach has been proposed and evaluated. The first step is a fast search space reduction, called sector-based joint detection-localization. The second step is a gradient descent localization within each active sector. Both steps rely on a topological interpretation of SRP-PHAT. Both steps were integrated into a multisource detection-localization system, with a fully available, near real-time implementation. Experiments on real indoor recordings showed that the integrated system is able to detect and locate three simultaneously speaking, moving speakers. Finally, location-based Speech/Non-Speech classifiers were proposed, that do not need training data. For future work on multisource detection-localization, joint spectrum-location estimation may be a relevant direction, for example based on (Grenier, 1994; Sekiya and Kobayashi, 2004; Nix and Hohmann, 2006). One contribution of the present chapter that goes beyond microphone arrays is the automatic selection of a detection threshold, *without training data*.

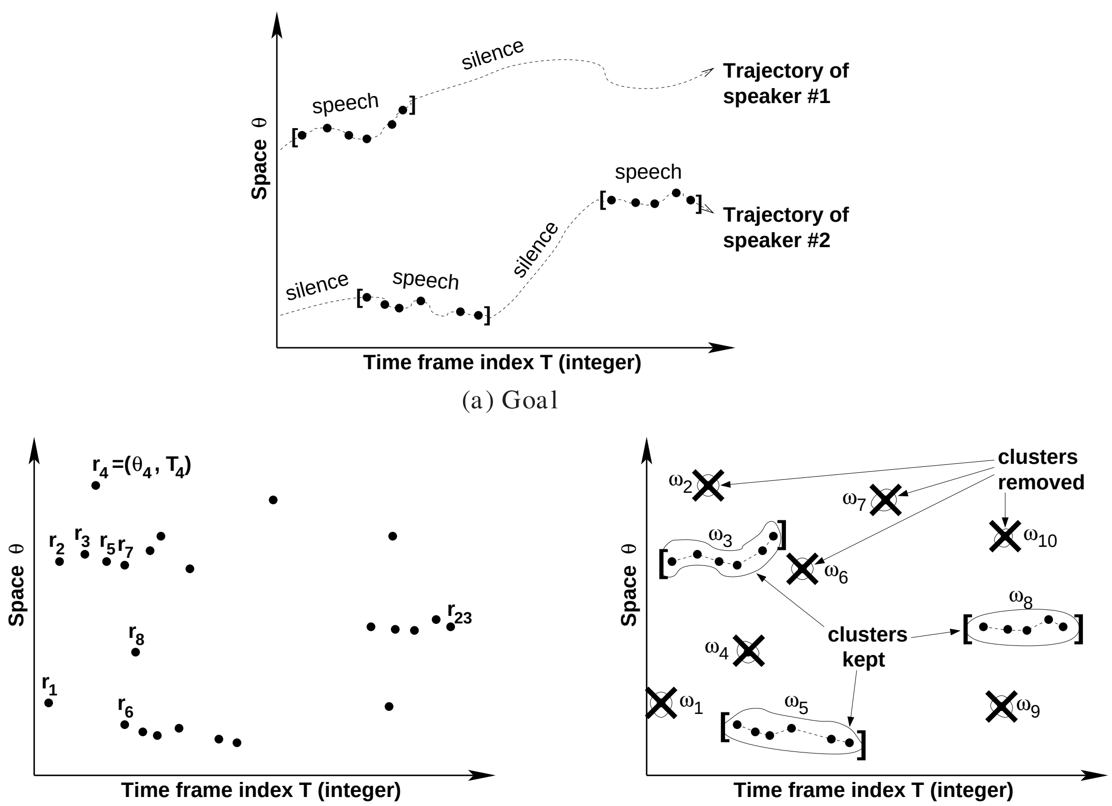
To conclude, the present chapter has investigated the *static* analysis of each recording, where each time frame is processed independently. Although a global model and a strategy to select the detection threshold were introduced, they did not take into account the correlation between consecutive time frames. This is the object of Chapter 6, where a *dynamical* model is proposed.

Chapter 6

Short-Term Spatio-Temporal Clustering

Distant microphones permit to process spontaneous multi-party speech with very little constraints on speakers, as opposed to close-talking microphones. Minimizing the constraints on speakers permits a large diversity of applications, including meeting summarization and browsing, surveillance, hearing aids, and more natural human-machine interaction (Section 1.1). When using distant microphones, a basic requirement of such applications is to determine where and when the speakers are talking (Figure 6.1a). This is inherently a multisource problem, because of background noise sources, as well as the natural tendency of the multiple speakers to talk over each other (Section 3.2.1). Chapter 5 investigated the *instantaneous* localization of multiple speech sources, that is from a single time frame (Figure 6.1b). The present chapter investigates the *dynamical* analysis of the resulting location estimates, across consecutive time frames (Figure 6.1c).

As exposed in Section 3.1.4, spontaneous speech utterances are highly discontinuous, which makes difficult to track the multiple speakers with classical filtering approaches, such as Kalman filtering or Particle Filters. As an alternative, this chapter proposes a probabilistic framework to determine the trajectories of multiple moving speakers in the short-term only – that is only while they speak. Instantaneous location estimates (dots in Figure 6.1b) that are close in space *and* time are grouped into "short-term clusters" (round lines in Figure 6.1c) in a threshold-free, princi-



(a) Goal

(b) Result of Chapter 5: multiple source location estimates

(c) The present chapter: short-term clustering and Speech/Non-Speech classif.

Figure 6.1. The goal (a) and the proposed approach (b)(c). Dots depicts instantaneous location estimates $r_i \stackrel{\text{def}}{=} (\theta_i, T_i)$. Dashed lines depict trajectories of the sources (true in (a), estimated in (c)). Square brackets depict the beginning and the end of each speech utterance. Round, continuous lines depict the short-term clusters $\omega_1, \cdots, \omega_{10}$.

pled manner. As a by-product, the start and end times of each utterance are precisely determined (square brackets in Figure 6.1c). Contrastive experiments clearly show the benefit of using short-term clustering, on real indoor recordings with seated speakers in meetings, as well as multiple moving speakers.

The rest of this chapter is organized as follows:

- Section 6.1 presents an assumption on the local dynamics of the location estimates – over time periods of about 250 ms – and derives a principled, threshold-free *Short-Term spatio-temporal Clustering* (STC) framework.
- Section 6.2 presents online and offline optimization algorithms to implement STC.
- Section 6.3 illustrates the flexibility of the STC framework with experiments on synthetic

data, showing that STC permits to detect trajectory crossings in a threshold-free manner.

- Section 6.4 applies STC to Speech/Non-Speech (SNS) classification in the case of multiple moving, overlapping speakers. An experimental evaluation is conducted on the AV16.3 Corpus, comparing frame-level SNS classification and cluster-level SNS classification.
- Section 6.5 applies STC to a more static case, where a speech/silence time segmentation is produced for each speaker seated in a meeting.
- Section 6.6 concludes.

As announced in Section 1.1, our underlying aim is to have a single system that copes with both moving and seated speakers. Therefore, both dynamic and static experimental evaluations (Sections 6.4 and 6.5, respectively), use as input the instantaneous multisource location estimates of Chapter 5. Similarly, both use the same STC implementation (Section 6.2). Dr Jean-Marc Odobez from IDIAP contributed to the simulated annealing part (Section 6.2.2).

This chapter does *not* attempt to determine the identities of the speakers, but only where and when they are active in the form of segments of trajectories (brackets and dashed lines in Figure 6.1c), as defined by each short-term cluster (round lines in Figure 6.1c). This implies to exclude non-speech clusters (crosses in Figure 6.1c). The resulting speech segments are used as a starting point for speaker clustering in Chapter 7.

6.1 Short-Term Spatio-Temporal Clustering

This section presents the proposed short-term spatio-temporal clustering approach. The context is multiple moving sources: for each source and for each time frame, an *instantaneous* location estimate $r_i \stackrel{\text{def}}{=} (\theta_i, T_i)$ may or may not be available, where i is an integer index, θ_i denotes the spatial location of the source, and $T_i \in \mathbb{N} \setminus \{0\}$ denotes a time frame index[1]. For each possible time frame index $T \in \mathbb{N} \setminus \{0\}$, there can be zero, one or multiple location estimates $r_i = (\theta_i, T_i)$, such that $T_i = T$. The proposed approach relies on a threshold-free criterion to cluster these location estimates into short-term trajectories.

[1] Following the frame-based analysis presented in Chapter 5, the present chapter uses integer time frame indices $T_i \in \mathbb{N} \setminus \{0\}$ (Figure 6.1b). However, the short-term clustering approach defined in the present chapter could be formulated very similarly without time frames, using instead time values t_i expressed in sampling periods.

Although the approach is fully generic, throughout this chapter the practical context will be one 8-microphone 10 cm-radius Uniform Circular Array (UCA) of microphones on a table (Figure 1.1a), recording multi-party speech in a meeting room (Figure 6.9). It is used to provide *instantaneous* location estimates of the multiple audio source, as described in Chapter 5. The spatial location θ_i is for example an azimuth value in degrees. Our ultimate goal in this chapter is to cluster the correct location estimates into speech utterances, and to discard the incorrect location estimates.

6.1.1 Assumptions on Local Dynamics

Let $r_i = (\theta_i, T_i)$ for $i \in \{1, \ldots, N_r\}$ be all instantaneous location estimates of events emitted by the various sources, over an entire recording. This includes the desired events (speech sounds) as well as noise. $\theta_i \in \mathbb{R}^D$ is a location in space, while $T_i \in \mathbb{N} \setminus \{0\}$ is a time frame index: $T_i \in \{1, 2, 3, \ldots\}$. The notation $r_{1:N_r}$ designates the set of all location estimates: $r_{1:N_r} \stackrel{\text{def}}{=} \{r_1, r_2, \cdots, r_{N_r}\}$. For convenience, without loss of generality, we assume the location estimates ordered in time (Figure 6.1b):

$$T_1 \leq T_2 \leq \cdots \leq T_{N_r} \tag{6.1}$$

Note that there can be multiple location estimates per time frame, i.e. $T_i = T_{i+1}$.

For any pair of location estimates (r_i, r_j), we define the two hypotheses:

- $H_0(i,j) \stackrel{\text{def}}{=}$ "r_i and r_j correspond to **different** sources"
- $H_1(i,j) \stackrel{\text{def}}{=}$ "r_i and r_j correspond to the **same** source"

The two hypotheses are complementary: $H_1(i,j) = \overline{H_0(i,j)}$. As a preliminary experiment, we ran instantaneous audio source localization with a UCA on real data (`seq01` from the AV16.3 Corpus), using the SRP-PHAT approach (DiBiase, 2000). For each location estimate $r_i = (\theta_i, T_i)$, θ_i is an estimate of the direction of an active acoustic source (azimuth angle in the horizontal plane). We observed the values of the difference $\theta_i - \theta_j$ for short delays $|T_i - T_j|$ up to T_{short}, where T_{short} is a small number of time frames (e.g. 2). Figure 6.2 displays a typical histogram of location variations $\theta_i - \theta_j$ (in gray). Our interpretation is as follows: two location estimates r_i and r_j either correspond to the same source or not. In the first case (H_1) the difference $\theta_i - \theta_j$ is small: a source does not move a lot during a short time period. Hence the zero-mean central peak in the histogram. In the

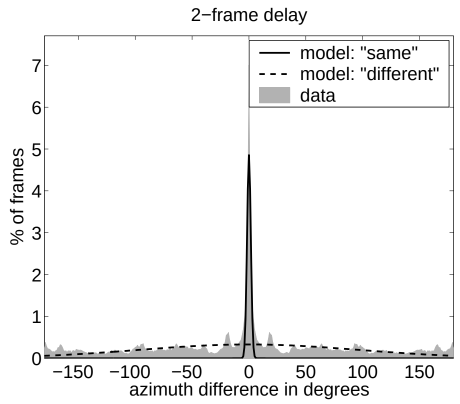


Figure 6.2. Histogram of azimuth angle variations $\theta_i - \theta_j$ over a 2-frame delay ($|T_i - T_j| = 2$), on real data (recording `seq01` from the AV16.3 Corpus, see Chapter 4). The super-imposed curves depict the bi-Gaussian mixture model obtained through EM training.

second case (H_0) the difference $\theta_i - \theta_j$ is random: the trajectories of two sources are independent, at least in the short-term. We therefore propose the following model for local dynamics, i.e. for $|T_i - T_j| \leq T_{\text{short}}$:

$$\begin{cases} p\left(\theta_i - \theta_j \mid H_0(i,j)\right) & = & \mathcal{N}_{0,\sigma^{\text{diff}}_{|T_i - T_j|}}\left(\theta_i - \theta_j\right) \\ \\ p\left(\theta_i - \theta_j \mid H_1(i,j)\right) & = & \mathcal{N}_{0,\sigma^{\text{same}}_{|T_i - T_j|}}\left(\theta_i - \theta_j\right) \end{cases} \tag{6.2}$$

where $\forall T \leq T_{\text{short}} \quad \sigma_T^{\text{same}} < \sigma_T^{\text{diff}}$, and $\mathcal{N}_{\mu,\sigma}(\cdot)$ denotes the Gaussian pdf with mean μ and standard deviation σ. Although an intuitive choice in the case of H_0 would be a uniform distribution, we opted for a Gaussian in order to capture the dependency of σ_T^{diff} on the delay T. This dependency was observed on real data, examples can be found in (Lathoud et al., 2004).

The standard deviation σ_T^{same} accounts for short-term variations of location estimates due to both local motion and measurement imprecision. We argue that there is no need to distinguish between the two, as long as the analysis is restricted to short delays $T \leq T_{\text{short}}$. For each delay $T \leq T_{\text{short}}$, σ_T^{same} and σ_T^{diff} can be estimated simply, through EM training (Dempster et al., 1977) of a bi-Gaussian mixture model, either on the entire data $\{\theta_i - \theta_j\}$ such that $|T_i - T_j| = T$, or in a blockwise fashion when the data is processed online, as in Section 6.2.1. The mean of each Gaussian is fixed to zero. Although the weights are also trained during EM, they are not used in the rest of the process. See Figure 6.2 for an example of bi-Gaussian mixture model.

(a) Short-term cluster

(b) Long-term cluster

Figure 6.3. The two types of clusters. This chapter focuses on short-term clusters (a), obtained with location cues only. Long-term clustering (b) requires additional cues, as investigated in Chapter 7.

We note that the model allows for location differences $\theta_i - \theta_j$ to be close to zero, while r_i and r_j belong to two different sources: $H_0(i,j)$. Such a situation may happen in reality, whenever two sources' trajectories cross each other. Section 6.3 provides a discussion on this topic.

The present chapter reports tests in 1-D space (azimuth angle). For higher dimensions, e.g. in spherical or Euclidean coordinates, one could simply replace σ^{same} and σ^{diff} with covariance matrices (diagonal should be sufficient). The rest of the approach presented below is unaffected by such a modification, because it relies on probabilities only (6.2).

6.1.2 Short-Term Clustering (STC)

Given a value of T_{short}, a cluster $\omega \subset r_{1:N_r}$ is "short-term" iff it has "time gaps" of at most T_{short} (Figure 6.3a). All other clusters are called "long-term clusters" (Figure 6.3b). This subsection formally defines a short-term cluster of location estimates, as well as a short-term partition of $r_{1:N_r}$.

Formally, a cluster ω is "short-term" iff:

$$\forall T \in \left[\min_{r_i \in \omega} T_i,\ \max_{r_i \in \omega} T_i\right] \quad \exists r_j \in \omega \text{ s.t. } |T_j - T| \leq \frac{T_{\text{short}}}{2} \tag{6.3}$$

A partition $\Omega = \{\omega_1, \cdots, \omega_n, \cdots, \omega_{N_\Omega}\}$ of the data $r_{1:N_r}$ is then "short-term" iff all clusters $\omega_n \in \Omega$ are "short-term". We denote this property with:

$$\Omega \in O_{\text{ST}} \tag{6.4}$$

where O_{ST} is the set of all possible short-term partitions Ω of the data $r_{1:N_r}$, for a given value of T_{short}.

6.1.3 Threshold-Free Maximum Likelihood Clustering

Given the local dynamics (6.2), we propose to detect and track events as follows: find a partition Ω of $r_{1:N_r}$ (round lines in Figure 6.1c) that maximizes the likelihood of the observed data $r_{1:N_r}$:

$$\Omega^{\mathrm{ML}} \stackrel{\mathrm{def}}{=} \arg\max_{\Omega \in O_{\mathrm{ST}}} p\left(r_{1:N_r} \mid \Omega\right) \tag{6.5}$$

Note that the number of clusters N_Ω has to be estimated as well. Each cluster $\omega_k \subset r_{1:N_r}$ contains locations for one event, e.g. a speech utterance. We are *not* trying to produce a single trajectory per source, but rather an oversplitted solution where $N_\Omega >> 1$ is the number of individual events, for example speech utterances. The exact value of N_Ω is thus not important: we rather want to be *sure* that all location estimates within each cluster ω_k correspond to the *same* source. Defining one cluster per location estimate obviously fulfills this constraint, although it is of little practical interest. Therefore, within each cluster, we would also like to have as many location estimates as possible, that belong to the same source. In other words, a criterion should be derived from the data-driven dynamical constraints 6.2, that also minimizes N_Ω as much as possible.

Over time, a source may move while being unobservable (e.g. silent, moving speaker). Using location cues alone, it is impossible to determine whether location estimates before and after the "silence" period belong to the same source. Thus, we can relate location cues in the short-term only (Figure 6.3a). We therefore propose to maximize the following "short-term criterion", using a simplifying independence assumption between all pairwise differences $\theta_i - \theta_j$:

$$\boxed{p_{\mathrm{ST}}\left(r_{1:N_r} \mid \Omega\right) \propto \prod_{\substack{0 \leq i < j \leq N_r \\ 0 \leq |T_i - T_j| \leq T_{\mathrm{short}}}} p\left(\theta_i - \theta_j \mid H^{\Omega}(i,j)\right)} \tag{6.6}$$

where $H^{\Omega}(i,j)$ is either $H_0(i,j)$ or $H_1(i,j)$, depending on whether or not r_i and r_j belong to the same cluster ω_n in the candidate partition Ω, as depicted by Figure 6.4. Each term of the product is expressed using (6.2). One important characteristic of this approach is that it does *not* need to explicitly model the true number of sources whose events are observed. Therefore, complex dynamical constraints and associated birth/death rules are not needed.

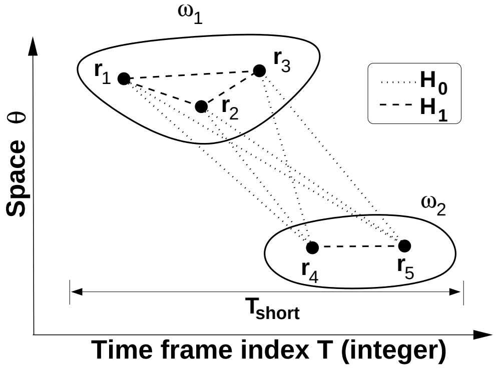


Figure 6.4. This two-cluster partition $\Omega = \{\omega_1, \omega_2\}$ of the set of location estimates $r_{1:5}$ (dots) is equivalently defined by six local decisions $H_0\,(i,j)$ (dotted lines) and four local decisions $H_1\,(i,j)$ (dashed lines). In this particular case, all location estimates (dots) are within T_{short} time frames of each other.

The proposed task, to cluster observations, fundamentally differs from Kalman filtering or Particle Filtering, which estimate a hidden state variable from the observations. In addition, filtering usually relies on a conditional independence assumption between consecutive observations, given the state values (Odobez et al., 2006). On the contrary, the proposed STC precisely consists in modelling dependencies between several consecutive observations, up to the order T_{short}.

6.2 Optimization Algorithms

The goal is to find a short-term partition Ω of the observed location estimates $r_{1:N_r}$ that maximizes the criterion (6.6). Even short recordings contain thousands of location estimates: $N_r >> 1$, for example $N_r = 50000$ for a 5-minute recording. It is thus untractable to try all possible short-term partitions $\Omega \in O_{\text{ST}}$. Sections 6.2.1 and 6.2.2 propose tractable, suboptimal implementations (online and offline).

6.2.1 Online: Sliding Window (SW)

We propose to find a suboptimal solution $\hat{\Omega}^{\text{ML}}$ by using a sliding analysis window, shifted at each iteration by N_{future} location estimates, where location estimates $r_{1:N_r}$ are ordered by increasing times:

$$T_1 \leq T_2 \leq \cdots \leq T_{N_r} \tag{6.7}$$

(Step 1) Initialization: For $T \in \{0, \cdots, T_{\text{short}}\}$, initialize standard deviations σ_T^{same} and σ_T^{diff}, with unsupervised EM training on the beginning or all of $r_{1:N_r}$. $n \leftarrow 1$. (Step 2) $F \leftarrow r_{n:n+N_{\text{future}}-1}$. Define all possible partitions of location estimates in F. Choose the most likely partition $\hat{\Omega}_F^{\text{ML}}$. (Step 3) $P \leftarrow \{r_i = (\theta_i, T_i) \mid T_n - T_{\text{short}} \leq T_i < T_n\}$. Define all possible merges between $\hat{\Omega}_P^{\text{ML}}$ and $\hat{\Omega}_F^{\text{ML}}$. Choose the most likely merged partition and update $\hat{\Omega}_{P\cup F}^{\text{ML}}$. (Step 4) Optionally, update σ_T^{same} and σ_T^{diff} using recently seen data (EM training, as in Step 1). (Step 5) $n \leftarrow n + N_{\text{future}}$ and loop to Step 2.

Table 6.1. The online Sliding Window (SW) maximum likelihood algorithm. The likelihood of a partition is estimated with (6.6). Location estimates are ordered by increasing times ($\forall n \quad T_n \leq T_{n+1}$).

Number of elements N_{future}	Number of possible partitions: Bell number (Weisstein, 2006a)
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4 140
9	21 147
10	115 975
11	678 570
>11	prohibitive

Table 6.2. SW algorithm: Number of possible partitions, for each possible number of elements (Step 2 in Table 6.1).

Table 6.1 describes the algorithm. Step 1 is the initialization: for each delay $T \leq T_{\text{short}}$, a bi-Gaussian model is fitted on azimuth differences $\{\theta_i - \theta_j\}$ such that $|T_i - T_j| = T$, as in Section 6.1.1. Steps 2, 3, and the optional Step 4 constitute one iteration of the algorithm. Step 2 selects the Maximum Likelihood (ML) partition of the N_{future} location estimates in the future set F, independently of all other data. The future set F has a fixed size (N_{future}), given by the user. Step 3 merges some clusters of the partition of F selected at Step 2, with some clusters in the past set P, again maximizing the likelihood (6.6). P contains all location estimates within T_{short} time frames in the past. There can be a variable number of location estimates for each time frame, therefore the set P has a variable size. The optional Step 4 updates the bi-Gaussian models with recently seen data. The result of this algorithm is an estimate $\hat{\Omega}^{\text{ML}} \in O_{\text{ST}}$ of the ML short-term partition $\Omega^{\text{ML}} \in O_{\text{ST}}$ of all observed data $r_{1:N_r}$. The entire process is online, threshold-free and can be fully deterministic[2]. As discussed in Section 6.1.3, this process fundamentally differs from Kalman filtering or Particle Filtering. In particular, the proposed approach models observation dependencies (6.2) up to the order T_{short}, even in the case $N_{\text{future}} = 1$.

One of the advantages of this approach is the bounded computational load. Indeed, evaluating

[2]A deterministic initialization of the EM training of the bi-Gaussian model can be used, similarly to Appendix C.1.2.

a candidate partition (Step 2) or a candidate merge (Step 3) following (6.6) is easily implemented through a sum in the log domain over location estimates within F (Step 2) or $P \cup F$ (Step 3). In both cases, to determine the computational load, we need to determine the maximum number of partitions that are evaluated.

The total number of partitions evaluated at Step 2 is shown in Table 6.2. For $N_{\text{future}} \leq 7$, there are at most 877 such partitions. As for Step 3, the worst case computational complexity was investigated in (Lathoud et al., 2004), in the special case where there is only one location estimate per time frame: for $T_{\text{short}} = 6$, there are at most 13 327 possible merges. However, in the general case investigated here, there can be multiple location estimates per time frame, thus many more possible merges. In practice, the worst case situation is rarely encountered, as it corresponds to a case where most location estimates in $P \cup F$ are unrelated to each other. Two practical solutions can be used, favoring oversplitting. First, one could set a hard limit on the number of partitions that are constructed at Step 3 (e.g. 10 000), always including *at least* the case without any merge. Second, a heuristic can be used to prune out most of the "unlikely" merges, by forbidding short-term partitions Ω of the analysis window, that include "new" decisions $H^{\Omega}(i,j) = H_1(i,j)$ whenever

$$\frac{p\left(\theta_i - \theta_j \mid H_1(i,j)\right)}{p\left(\theta_i - \theta_j \mid H_0(i,j)\right)} \leq \epsilon \tag{6.8}$$

where ϵ is a small value, e.g. 10^{-10}. On tests with synthetic data (Section 6.3.2) we obtained exactly the same results with pruning or without pruning.

In the following, SW-N_{future} denotes the Sliding Window algorithm with a particular N_{future} value: for example SW-1 or SW-7.

6.2.2 Offline: Simulated Annealing Optimization (SA)

Alternatively, the proposed modeling can be cast into a Markov Random Field (MRF) framework (Li, 1995), by defining a label field $E = \{E_i,\ i = 1 \ldots N_r\}$, where E_i is the label associated with the observed location estimate r_i, and $\underline{E}_i$ is the r.v. associated with E_i. The actual label values are not important and can be e.g. integers. We define a graph $\langle E,\ \mathcal{G}\rangle$, where E represents the set of nodes, and $\mathcal{G}$ denotes the neighborhood system. $\underline{E}_i$ is a neighbor of $\underline{E}_j$ iff $|T_i - T_j| \leq T_{\text{short}}$. A graph $\langle E,\ \mathcal{G}\rangle$ uniquely defines a short-term partition $\Omega \in O_{\text{ST}}$. Given the observations $r_{1:N_r}$, the goal is to

1) Initialization: Temperature: $\eta \leftarrow \eta_0$ Label field: $E \leftarrow E_{init}$ 2) Do $E \leftarrow SA(E, \eta)$ $\eta \leftarrow \eta * \lambda$ While $\eta > \eta_{end}$ 3) Iterated Conditional Mode (ICM) optimization steps until there is no label change: Do $E' \leftarrow E$ $E \leftarrow SA(E', 0)$ While $E \neq E'$

Table 6.3. SA algorithm: The MRF optimization (in practice $\lambda = 0.97$). $SA(E, \eta)$ is described in Table 6.4.

1) Initialization: $I \leftarrow \{1, ..., N_r\}$

2) While $I \neq \emptyset$

- sample $i \in I$ uniformly.
- define candidate labels $L_i \leftarrow E_{\mathcal{G}_i} \cup \{NewLabel\}$ where $E_{\mathcal{G}_i}$ is the set of current labels from the $\underline{E}_j$'s in the neighborhood of $\underline{E}_i$.
- compute the posterior probabilities $P_{i,n} \stackrel{\text{def}}{=} P\left(E_i = l_n \mid E \setminus E_i\right)$ over the candidate labels $l_n \in L_i$:

$$P_{i,n} \propto \exp\left[-\frac{1}{\eta} \sum_{\substack{j \\ \langle i,j \rangle \in \mathcal{C}}} \beta_{ij}^{\text{potts}} \cdot \delta_{Kr}\left(l_n - E_j\right)\right]$$

- sample $\underline{E}_i \sim \text{Multinomial}(P_{i,n})$.
- remove i from I.

Table 6.4. SA algorithm: One simulated annealing step $SA(E, \eta)$.

estimate the label field E that maximizes the ML criterion (6.6) or, equivalently, that maximizes the following Potts field (Geman and Geman, 1984):

$$P_{\text{potts}}(E) \stackrel{\text{def}}{=} \frac{1}{Z} \cdot e^{-U(E)} \tag{6.9}$$

where $U(E)$ is the following energy function:

$$U(E) \stackrel{\text{def}}{=} \sum_{\langle i,j \rangle \in \mathcal{C}} V_{ij}(E) \stackrel{\text{def}}{=} \sum_{\langle i,j \rangle \in \mathcal{C}} \beta_{ij}^{\text{potts}} \cdot \delta_{Kr}\left(E_i - E_j\right) \tag{6.10}$$

where $\mathcal{C}$ is the set of pairwise cliques of the neighborhood system $\mathcal{G}$, Z is the partition function (normalization factor), and $\delta_{Kr}(\xi)$ is the Kronecker function, the value of which is 1 when $\xi = 0$, and 0 otherwise. The $\beta_{ij}^{\text{potts}}$ are called the Potts coefficients, the values of which depend on the observations and can be derived from (6.2) and (6.6):

$$\beta_{ij}^{\text{potts}} = \log\left[\frac{p\left(\theta_i - \theta_j \mid H_0(i,j)\right)}{p\left(\theta_i - \theta_j \mid H_1(i,j)\right)}\right] \tag{6.11}$$

The maximization of the probability $P_{\mathrm{potts}}(E)$ with respect to the label field E is equivalent to the minimization of the energy function $U(E)$ and can be conducted using standard techniques. We adopted a simulated annealing approach (Geman and Geman, 1984; van Laarhoven and Aarts, 1987), with Gibbs sampling and an exponentially decaying temperature, followed by an Iterated Conditional Mode (ICM) procedure (Geman and Geman, 1984; van Laarhoven and Aarts, 1987), as described in Tables 6.3 and 6.4. Note that any short-term partition configuration ($\Omega \in O_{\mathrm{ST}}$) can be reached with a non-zero probability, which is a requirement of simulated annealing.

We consider three different alternatives for the label field initialization E_{init}:

- SA(1): The initial label field E_{init} has a single label shared by all nodes.
- SA(N_r): The initial label field E_{init} has one different label per node.
- SA(SW-1): The initial label field E_{init} is constructed in a sequential and causal fashion: for each new observation, we select the label that minimizes $U(E)$ given all previous observations. This is strictly equivalent to SW-1: the SW algorithm with $N_{\mathrm{future}} = 1$.

Section 6.5.5 discusses the outcome of these alternatives and on their impact on the criterion (6.6), the number of short-term clusters N_Ω, and the performance.

6.3 Application: Threshold-Free Detection of Trajectory Crossings

This section defines a confidence measure for each possible individual decision $H_a(i,j)$ ($a \in \{0,1\}$), and uses this confidence measure to detect and deal with low confidence situations such as trajectory crossings. The goal is to illustrate the flexibility of the proposed probabilistic framework (6.6). It is relevant to contexts where the events emitted by the various sources are somewhat "continuous", such as acoustic signals from vehicles (Pham and Fong, 1997). In the present section, we investigate the extraction of trajectory segments, where each segment belongs *for sure* to a single source. Achieving this task would be useful as a first step, for instance prior to agglomerating the trajectory segments that belong to the same source, as in Bayesian Network tracking (Jorge et al., 2004).

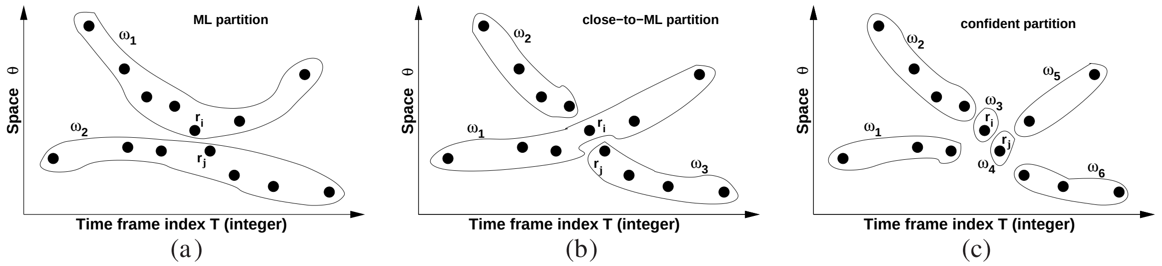


Figure 6.5. Example of low confidence decision $H_0(i,j)$ at a trajectory crossing. Each dot is a location estimate. A continuous line depicts each short-term cluster ω_n.

For $a \in \{0,1\}$, we propose to use the posterior probability $P\left(H_a(i,j) \mid r_{1:N_r}\right)$ as a confidence measure for a given local decision $H_a(i,j)$. Assuming equal priors for all possible short-term partitions $\Omega \in O_{\mathrm{ST}}$ of the observed data $r_{1:N_r}$, the posterior probability of the local decision can be expressed as follows, for $a \in \{0,1\}$:

$$P\left(H_a(i,j) \mid r_{1:N_r}\right) \propto \sum_{\substack{\Omega \in O_{\mathrm{ST}} \\ H^{\Omega}(i,j) = H_a(i,j)}} p\left(r_{1:N_r} \mid \Omega\right) \tag{6.12}$$

where $P\left(H_0\left(i,j\right) \mid r_{1:N_r}\right) + P\left(H_1\left(i,j\right) \mid r_{1:N_r}\right) = 1$. Section 6.3.1 proposes to use this confidence measure in order to modify the ML optimization procedure.

6.3.1 Threshold-Free Confident Clustering

We would like to determine when trajectories cross, and to split short-term clusters accordingly. Figure 6.5a gives an example of ML partition. r_i and r_j are very close, it is thus not clear which short-term cluster, r_i and r_j should ideally belong to. In such a case, there may exist a different partition with a close-to-optimal likelihood (Figure 6.5b). We propose here to break each short-term cluster that contains r_i or r_j into two "confident" parts, and to create two separate one-element clusters $\{r_i\}$ and $\{r_j\}$ (Figure 6.5c).

Let us assume that the ML criterion (6.6) leads to the decision $H^{\hat{\Omega}^{\mathrm{ML}}}\left(i,j\right) = H_0\left(i,j\right)$, as illustrated in Figure 6.5a. Let us assume a low "confidence" in the decision $H_0\left(i,j\right)$, in the sense that there is at least one close-to-ML partition Ω' with $H^{\Omega'}\left(i,j\right) = H_1\left(i,j\right)$, as illustrated in Figure 6.5b. This implies not only a low posterior probability (6.12) for $H_0\left(i,j\right)$, but also that in Ω', several other decisions H_0 and H_1 involving r_i or r_j are the opposite of the corresponding ML decisions. This in turn implies that several ML decisions H_0 and H_1 involving r_i and/or r_j have a low posterior probability (6.12).

We propose to detect "low confidence" in a ML decision $H_0(i,j)$, by comparing it to all ML decisions $H_1(n_1,n_2)$ in the same analysis window $W \subset r_{1:N_r}$. Formally, a "low confidence" $H_0\left(i,j\right)$ decision is defined as verifying:

$$P\left(H_0\left(i,j\right) \mid r_{1:N_r}\right) \quad < \quad M_1\left(\hat{\Omega}_W^{\mathrm{ML}}\right) \tag{6.13}$$

where:

$$M_1\left(\hat{\Omega}_W^{\mathrm{ML}}\right) \stackrel{\mathrm{def}}{=} \max_{\substack{n_1 < n_2 \\ (r_{n_1}, r_{n_2}) \in W \times W \\ H^{\hat{\Omega}_W^{\mathrm{ML}}}(n_1,n_2) = H_1(n_1,n_2)}} P\left(H_1\left(n_1,n_2\right) \mid r_{1:N_r}\right) \tag{6.14}$$

For the Sliding Window (SW) algorithm, "confident clustering" is implemented by modifying Steps 2 and 3 in Table 6.1 as follows:

- For all (r_i, r_j) in the analysis window $W = F$ (Step 2) or $W = P \cup F$ (Step 3), estimate $P\left(H^{\hat{\Omega}_W^{\mathrm{ML}}}(i,j) \,\middle|\, W\right)$ using (6.12). For O_{ST}, we use the set of all candidate partitions in W.
- Step 2: whenever a decision $H_0(i,j)$ given by the ML algorithm has "low confidence" (6.13), split in two parts the short-term cluster containing r_i, at time T_i. Idem for r_j. Additional one-element clusters $\{r_i\}$ and $\{r_j\}$ are created (Figure 6.5c).
- Step 3: whenever a decision $H_0(i,j)$ given by the ML algorithm has "low confidence" (6.13), forbid any merge between each of the two short-term clusters containing r_i (resp. r_j), and any other short-term cluster.

Confident clustering requires $N_{\mathrm{future}} > 1$. Indeed, with $N_{\mathrm{future}} = 1$, cancellation of a single ML merge (Step 3) is most often replicated in the future, resulting in an unnecessarily long series of one-element clusters. This was verified on the same synthetic data as the one used in Section 6.3.2.

In the case of simulated annealing (SA), only some of the partitions are explored, therefore a different implementation may be needed to detect trajectory crossings.

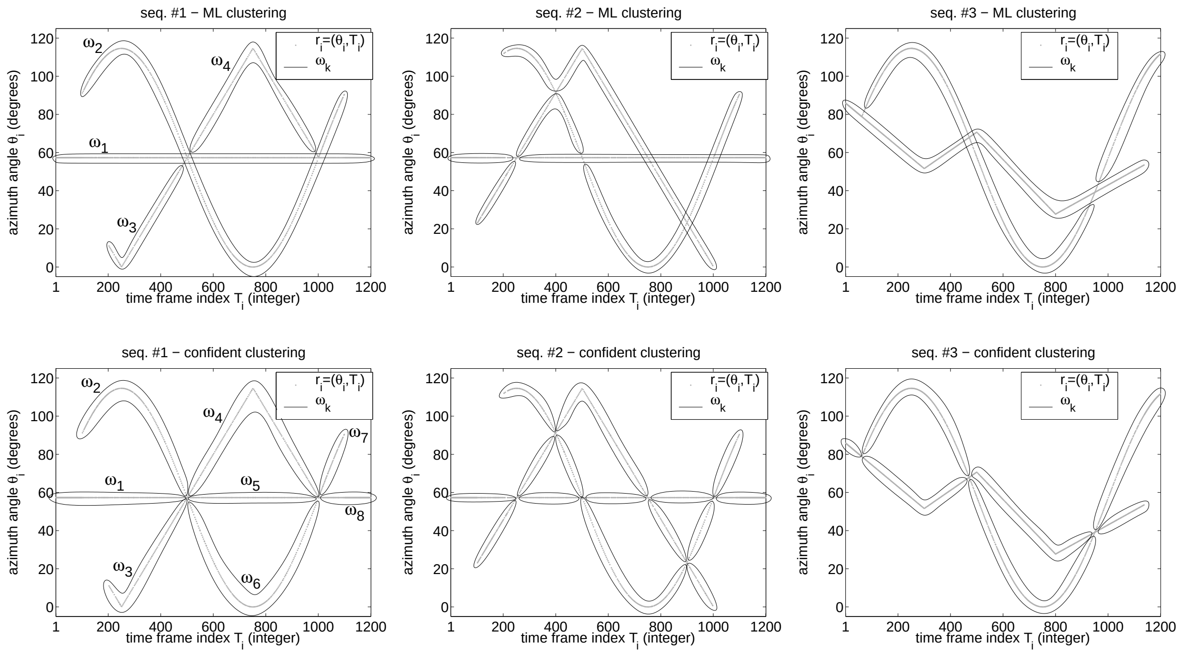


Figure 6.6. Comparison ML clustering / confident clustering on multiple source cases, where the number of active sources varies over time. Gray dots: location estimates $r_i = (\theta_i, T_i)$. Black lines: clusters ω_k. The ML clustering algorithm takes arbitrary decisions at trajectory crossings. On the contrary, the confident clustering correctly splits the short-term clusters at each trajectory crossing.

6.3.2 Multi-Source Tracking Examples

To compare the ML clustering with the confident clustering, we generated data that simulates "sporadic" and "concurrent" events by restricting $r_{1:N_r}$ to have at most only one location estimate per time frame ($\forall i \;\; T_i < T_{i+1}$), yet with trajectories that look continuous enough so that it is still a tracking problem. In all test sequences, the number of active sources varies over time, and trajectories cross several times. The task is twofold:

- Task 1: From instantaneous location estimates $r_{1:N_r}$, build the various trajectories accurately.
- Task 2: Extract pieces of trajectory (clusters), where each piece belongs *for sure* to a single source. This implies that no short-term cluster extends beyond any trajectory crossing.

Figure 6.6 compares the result of ML clustering (SW implementation, with $T_{\text{short}} = 7$ and $N_{\text{future}} = 1$) with the result of the confident clustering described in Section 6.3.1. Although the ML clustering correctly builds the various trajectories (task 1), it produces arbitrary decisions around the points of crossing. On the contrary, confident clustering correctly splits the trajectories at all crossing points (task 2). Thus, confident clustering could be particularly useful to create *reliable* pieces of trajectories, which do not include any crossing point. These pieces of trajectories can then be linked using approaches such as Bayesian Networks (Jorge et al., 2004).

A Matlab implementation of short-term clustering (ML and confident, SW implementation) can be found on the following website, along with 10 synthetic data examples:
`http://mmm.idiap.ch/Lathoud/2006-short-term-clustering/`

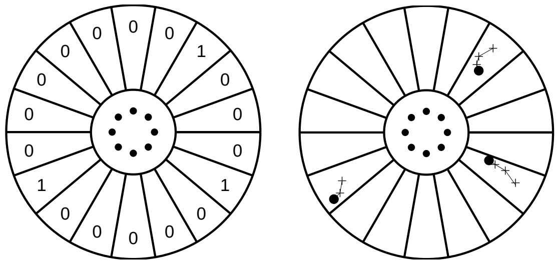
6.4 Application to Detection-Localization of Multiple Speakers

This section presents an integrated system for detection and localization of multiple speakers, along with experimental results on recordings with multiple moving speakers. We show that the use of STC for Speech/Non-Speech (SNS) classification permits to achieve substantial improvements over frame-level approaches. The resulting integrated system is used as a platform for multispeaker segmentation, in Section 6.5.

Since the focus of this chapter is STC, the part of the system that implements *instantaneous* multisource detection-localization is summarized as much as possible (Section 6.4.1). A detailed description of this implementation was given in Chapter 5.

6.4.1 Instantaneous Multisource Detection-Localization

Zero, one or more location estimates $r_i = (\theta_i, T_i)$ are produced at each time frame, where θ_i is the azimuth of an audio source with respect to a microphone array (Figure 1.1a), and $T_i \in \mathbb{N} \setminus \{0\}$ is the time frame index. "Instantaneous" means that each time frame is processed individually. A 2-step approach is used, as illustrated by Figure 6.7. First, the sector-based detection-localization (Figure 6.7a, Section 5.3) limits the search space to zero, one or more sectors of space around the microphone array. Second, the SRP-PHAT local maximum is found within each active sector, through Scaled Conjugate Gradient descent (Moller, 1993), as in Figure 6.7b and Section 5.4. The reader is referred to Chapter 5 for full details, freely available code, and tests on real data that show that this part of the system achieves detection-localization of up to three multiple simultaneous speakers, with near real-time performance (implementation called "FAST" in Chapter 5). We used a 32 ms frame length with 50 % overlap (16 ms frame shift).



(a) (b)

Figure 6.7. 2-step implementation for multisource detection-localization (Chapter 5). The eight dots indicate the locations of the microphones. (a) sector-based detection-localization. (b) gradient descent within each active sector.

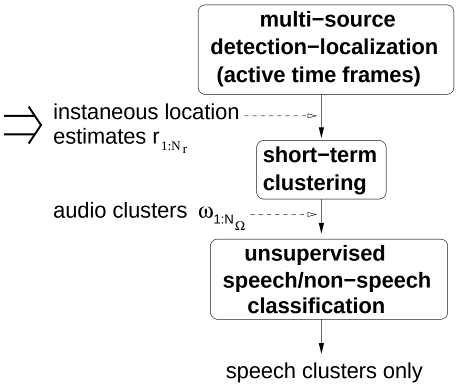


Figure 6.8. Detection-localization of multiple speakers, using microphone arrays (systems SW-1, SW-7 and SA).

6.4.2 Speech/Non-Speech (SNS) Classification

Let us assume that we have a system for instantaneous detection and localization of multiple audio sources, as described above. "Audio sources" include not only human speakers, but also noise sources such as a projector, a laptop and the various reverberations, as shown in Figure 6.10a. But our final task is multi-*speaker* detection-localization, so it is needed to remove the non-speech location estimates (see the result in Figure 6.10b). In other words, each location estimate must be classified as speech or non-speech. In this chapter, two systems are investigated: the SNS decision is taken either at the location estimate level (r_i) – not using the context – or at the short-term cluster level (ω_n) – using the context.

"Individual SNS": SNS decision for each individual location estimate r_i separately: As detailed in Section 5.5.3, we compare to a threshold the posterior probability (5.60) of having a wideband, non-noisy signal emitted by the source at location θ_i and time frame index T_i. The threshold is determined without tuning, as in (5.33), in order to match a user-defined target $\mathrm{FAR_T}$ of detection False Alarm Rate, for example $\mathrm{FAR_T} = 0.01$.

"Cluster SNS": SNS decision per short-term cluster ω_n: As detailed in Section 5.5.3, when a short-term cluster contains more than one location estimate, it is possible to estimate the non-stationarity of the received signal across the whole short-term cluster, based on a location-dependent way of extracting MFCC, as in (5.58). As detailed in Section 5.5.3, a "speech cluster" ω_n must verify the two following conditions:

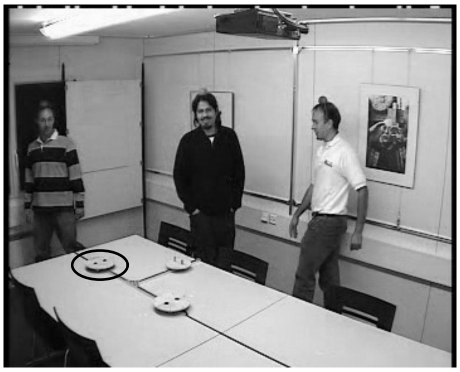


Figure 6.9. Recording `seq45` from the AV16.3 Corpus (Chapter 4), with three moving speakers. The 8-microphone array is marked with an ellipse. The ball markers on the heads were used to construct the ground-truth location of each speaker with respect to the array.

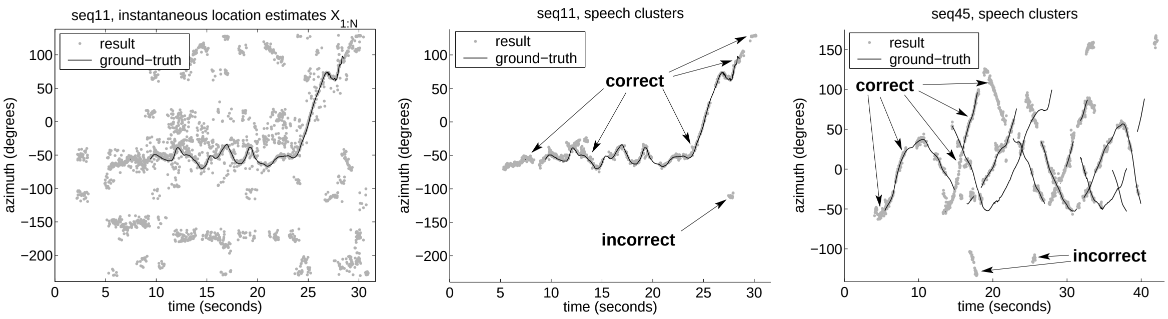
1. At least two location estimates (r_i, r_j, $i \neq j$) corresponding to a wideband, non-noisy source. (Individual SNS decisions.)

2. Non-stationarity above a threshold (global SNS decision). In practice, this threshold is fixed and does not require tuning, due to an underlying spectrum normalization (Section 5.5.3).

Two advantages can be expected from "Cluster SNS" over "Individual SNS". First, *weak but correct* location estimates that do not pass the "Individual SNS" test may still be part of a speech cluster, and thus be correctly detected by the "Cluster SNS" test. Thus, more speech should be retrieved by "Cluster SNS", as verified in Section 6.4.4. Second, the non-stationarity measure allows to exclude machine noise sources such as a projector or a laptop. As shown in Section 6.5.4, this is particularly useful in a meeting environment.

6.4.3 Experimental Protocol

To assess whether STC is beneficial to the detection decision, we compared the two SNS decision granularities (individual vs. cluster) using the same "FAST" underlying instantaneous multisource detection-localization implementation (Section 6.4.1, top block in Figure 6.8). We ran the two systems on eight real indoor recordings from the freely available AV16.3 Corpus (Chapter 4). Multiple simultaneous speakers are moving around a table, with a 8-microphone, 10-cm radius, UCA on its top (Figures 1.1a and 6.9). Three cameras were used to reconstruct the 3-D ground-truth location of each speaker, with an error inferior to 1.2 cm (Chapter 4). In the "Cluster SNS" case, we used the SW algorithm with $N_{\text{future}} = T_{\text{short}} = 7$.



(a) One moving speaker, before SNS classification | (b) One moving speaker, after cluster-level SNS | (c) Three moving speakers, after cluster-level SNS

Figure 6.10. Comparison between audio location estimates (dots) and the ground-truth location(s) obtained with three cameras (black lines, when all three cameras are available: *gaps in the ground-truth do not correspond to silences*), on the AV16.3 Corpus (Chapter 4). (a) Raw azimuth estimates for a single moving speaker, result of the multisource detection-localization, (b) After STC and removal of non-speech clusters. (c) Example with three simultaneous speakers.

SNS decision granularity	Total detected	Correctly localized	Precision (deg.) bias	std
Individual SNS: X_i	**284.9 s**	91.85%	0.335	2.158
Cluster SNS: ω_k (STC with SW-7)	**699.0 s**	92.05%	0.381	2.542

Table 6.5. Comparison between the two types of SNS decision, on the AV16.3 Corpus (Chapter 4), including real recordings with multiple moving speakers, simultaneously speaking. Bias and standard deviation (std) are expressed in degrees.

The focus here is the correct detection-localization of multiple moving speakers. For both systems, speech location estimates are compared with the closest ground truth speaker location. As in Section 5.4.6, we derive the following performance metrics on intervals on which the ground-truth locations of all speakers are known:

- Bias and standard deviation in degrees, to assess the precision of the localization.
- The percentage of detected speech that was correctly localized, i.e. within a small error margin. The margin is derived from the bias and the standard deviation, as detailed in Section 5.4.6.

6.4.4 Results and Discussion

From Figures 6.10a and 6.10b, one can see that the SNS decision using short-term clusters permits to remove most of the incorrect location estimates, while keeping most of the correct location estimates. This is also visible on Figure 6.10c, which presents a 3-speaker case. Note that the gaps in the ground-truth do not mean that a speaker is silent, but simply that the mouth was not visible on a camera – and thus the ground-truth location is unavailable.

Table 6.5 presents the overall detection-localization results. The percentage of correct location estimates is very similar for both "Individual SNS" and "Cluster SNS", but STC clearly retrieves much more speech signal[3]. Indeed, as discussed in Section 6.4.2, each short-term *speech* cluster contains some "weak but correct" location estimates, which would not pass the "Individual SNS" test. This confirms the interest of grouping location estimates *before* rejecting noise. The price to pay is a slight decrease in localization precision, probably due to those "weak" location estimates. This loss of precision can anyway be compensated for by smoothing the trajectory described by each short-term cluster, e.g. using Kalman filtering Kalman (1960); Welch and Bishop (2004) or RTS smoothing Rauch et al. (1965), but this is out of the scope of this thesis. Overall, the proposed "Cluster SNS" approach yields a much larger total amount of detected location estimates (from 284.9 sec to 699.0 sec), but keeps the same proportion of correct ones as in the "Individual SNS" case (from 91.85 % to 92.05 %). The "Cluster SNS" result can thus be seen as a significant improvement over the "Individual SNS" result. The proposed "Cluster SNS" approach could be useful as a prior step to trajectory analysis, as done in (Jorge et al., 2004). The next section uses the proposed approach for meeting segmentation.

6.5 Meeting Segmentation Application

In this section we report experiments conducted on real meeting data recorded with a UCA, the M4 Corpus (McCowan et al., 2005). We use the system described in Section 6.4, with cluster-level SNS classification. A comparison with close-talking lapel microphones is given. These experiments can be seen as a more static counterpart to the moving speaker experiments reported above. We want to determine whether *the same system* can cope with both static and dynamical contexts. In the previous section, the focus was on correct detection for precise localization of multiple moving speakers. In this section, we focus on the speech segmentation task: we have a precise time-domain ground-truth, but an approximate spatial ground-truth.

[3] To obtain the same "Total detected" duration as for the cluster SNS method (699.0 s), the individual SNS method can be made less conservative. The percentage correct then falls to 62.28%, with precision bias 0.375, std 2.869.

"Speech segmentation" means that we are only trying to separate the different speakers in the short-term (where? when?). The target is one short-term cluster per speech utterance. Results reported in Chapter 7 show that the proposed speech segmentation system forms a strong basis for long-term speaker clustering (who?) with distant microphones, where the goal is to obtain only one cluster per speaker. However, this is out of the scope of the present chapter.

The proposed approach departs from a previous work (Ajmera et al., 2004) that essentially extended a MFCC-GMM/HMM speaker clustering approach (Ajmera and Wooters, 2003) to include location information. The approach proposed here exhibits several differences:

- We are focusing on the speech segmentation task only, not on the speaker clustering task.
- We use distant microphones only (no lapel).
- We segment each meeting independently.
- The proposed approach does not rely on a Hidden Markov Model (HMM).

On the contrary to the preliminary results reported in (Lathoud et al., 2004), all systems presented here perform automatic removal of non-speech sources (e.g. projector).

6.5.1 Test Data

The test corpus includes 21 short meetings from the publicly available M4 Corpus[4]. The total amounts to about 2h of multichannel speech data. 3 meetings were used as a development set to tune post-processing parameters (Section 6.5.4), and after that, 18 meetings were used as a test set to evaluate performance metrics.

In the data, people are seated around a table, and sometimes stand up and move to the screen for a presentation using a projector, or to the whiteboard. In all meetings, an independent observer provided a very precise speech/silence segmentation[4]. Because of this high precision, the ground-truth includes many very short segments. Indeed, more than 50% of the speech segments are shorter than 1 second, as depicted in Figure 6.11.

[4] `http://mmm.idiap.ch`

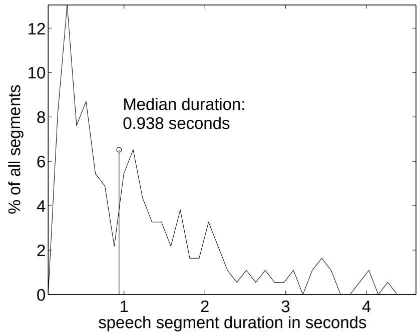


Figure 6.11. Histogram of speech segment durations in the ground-truth (M4 Corpus (McCowan et al., 2005)).

6.5.2 Proposed Systems

We tested several variants of the system described in Section 6.4, corresponding to the different optimization algorithms introduced in Section 6.2. The online systems SW-1 and SW-7 use the sliding window algorithm, both with $T_{\text{short}} = 7$ (maximum interval in which a local decision is defined). As defined in Section 6.2.1, in the case of SW-1 we have $N_{\text{future}} = 1$, which means that the sliding window is shifted by only one location estimate at a time. On the contrary, in the case of SW-7, the sliding window is shifted by steps of $N_{\text{future}} = 7$ location estimates each. $N_{\text{future}} = 7$ was not tuned, it was only chosen to keep the computational cost low (Table 6.2). We also tested the offline systems based on simulated annealing (SA(1), SA(N_r) and SA(SW-1)).

In all cases we use Maximum Likelihood clustering (Section 6.1.3) for this application. The confident clustering described in Section 6.3.1 is not necessary in the case of speech, since trajectory crossings are rarely seen due to the sporadicity of speech. Confident clustering is more relevant to cases where the signals are more continuous in time, such as vehicles (Pham and Fong, 1997).

6.5.3 Baseline System using Lapels

The proposed systems use distant microphones only. We compared them to a lapel-only baseline. The latter is an energy-based technique that selects the lapel with the most energy at each frame, and applies energy thresholding to classify the frame as speech or silence. We tried to use Zero-Crossing Rate (ZCR) as well, but it degrades significantly the segmentation performance. Indeed, ZCR appeared very sensitive to some noises found in meetings, such as writing on a sheet of paper. Therefore, results are reported with energy only. Note that lapels have a SNR around 18.7 dB, while distant microphones have a SNR around 10.7 dB (Table 3.1).

6.5.4 Performance Measures

We evaluated the result of each system as follows. For each of the proposed systems (SW-1, SW-7, SA(1), SA(N_r) and SA(SW-1))[5], for each speech location estimate, the corresponding time frame (32 ms segment) is attributed to the closest human speaker in space (the ground-truth location(s) of each speaker are known). Similarly, in the case of the lapel baseline, for each lapel, each speech time frame is attributed to the speaker wearing the lapel. For each speaker, the resulting speech/silence segmentation is further post-processed with basic morphological operators (Smith, 1999): dilation, erosion, closure and opening, as in (Lathoud et al., 2003). For each system, the post-processing parameters are tuned to maximize the F-measure on the development set (3 meetings). Each system is then applied on the test set (18 meetings). The performance metrics described in the following were evaluated for each meeting separately. Averages across all meetings are reported in Tables 6.6, 6.7, and 6.8. As opposed to previous results (Lathoud et al., 2004), all systems must include automatic removal of non-speech sources such as the projector.

For each meeting, evaluation was performed as follows. For each speaker, the resulting speech/silence segmentation is compared to the ground truth. Following Appendix A, four types of durations (in seconds) are calculated:

- D_{TP}: total duration of all segments in a meeting where a speaker is speaking in both result and ground-truth.
- D_{TN}: total duration of all segments in a meeting where a speaker is silent in both result and ground-truth.
- D_{FP}: total duration of all segments in a meeting where a speaker is speaking in the result, but silent in the ground-truth.
- D_{FN}: total duration of all segments in a meeting where a speaker is silent in the result, but active in the ground-truth.

and similarly to Appendix A, six metrics FAR, FRR, HTER, PRC, RCL, F are defined using the four durations. For example FAR $\stackrel{\mathrm{def}}{=} \dfrac{D_{\mathrm{FP}}}{D_{\mathrm{FP}} + D_{\mathrm{TN}}}$. In the optimal case, FAR, FRR and HTER are all

[5] In order to have a fair comparison between online and offline implementations, in all cases we used the same σ^{diff} and σ^{same} values for each recording, obtained offline, through EM fitting on the whole recording data $r_{1:N_r}$.

equal to 0, and PRC, RCL and F are all equal to 1. The F-measure is a harmonic mean of PRC and RCL, therefore, a large value of F-measure requires a large value for *both* PRC and RCL.

We also report results in terms of Diarization Error Rate (DER), a percentage metric defined by NIST (NIST, 2003). As opposed to the PRC/RCL/F results, DER excludes part of the data from the evaluation: within a collar of 0.25 seconds around each speech segment end-point, results are not evaluated. Moreover, silences of less than 0.300 seconds are removed from both result and ground-truth. Within the remaining segments, the DER is then defined as the percentage of speech that was wrongly attributed: DER = MISS + FA + SPKR, where MISS and FA are the percentages of missed speech and false alarms, respectively, and SPKR is the percentage of speech attributed to the wrong speaker. In the present chapter, the true number of speakers is already known so a low DER only indicates that the estimated speaker segmentation is close to the true speaker segmentation. Full details on the DER can be found in (NIST, 2003).

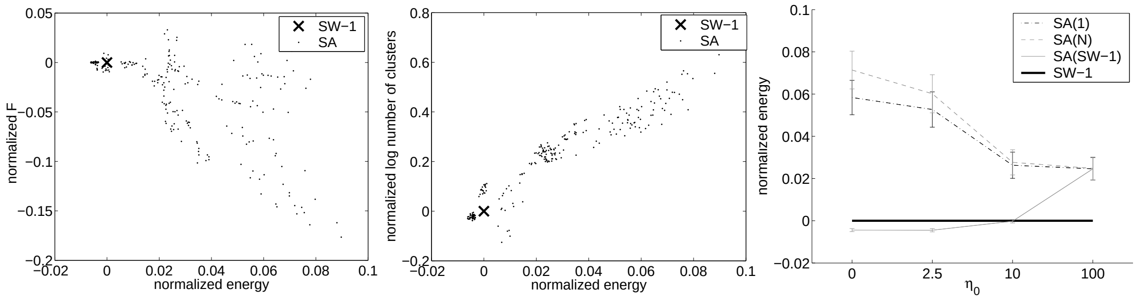
In any case (PRC/RCL/F or DER), it is important to bear in mind that in this chapter we are *only* evaluating the speech segmentation quality (one cluster per utterance). Evaluation of the application of STC to speaker clustering (one cluster per speaker) is found in Chapter 7.

6.5.5 Results and Discussion

Choice of an Optimization Method: Figure 6.12 presents a comparison of various instances of simulated annealing (SA), where different initializations and different values of the initial temperature η_0 are tried. Results are reported in terms of energy $U(E)$ (6.10), final number of clusters $N_{\hat{\Omega}}$, and segmentation performance F. In order to accommodate the various lengths of the meetings, we have normalized all three measures with respect to a reference method (SW-1):

- Normalized energy: for each meeting, $\frac{U(E)-U\left(E^{(SW-1)}\right)}{N_{\text{terms}}}$, where N_{terms} is the number of terms in the sum in (6.10).
- Normalized log number of clusters: for each meeting, $\log N_{\hat{\Omega}} - \log N_{\hat{\Omega}}^{(SW-1)}$.
- Normalized F: for each meeting, $F - F^{(SW-1)}$.

Figure 6.12a shows that the proposed criterion is effectively related to the final segmentation performance: the lower the energy, the higher the performance. All lowest energies lead to very similar



(a) (b) (c)

Figure 6.12. Simulated Annealing (SA): Comparison between different initialization methods (Sections 6.2.2 & 6.5.5). In (a) and (b), each point represents a result for one meeting and one initialization method (SA(1), SA(N_r) or SA(SW-1)). For each meeting, all results are normalized through subtraction with respect to the reference SW-1 (Section 6.5.5). In (c), η_0 is the initial temperature. In the case $\eta_0 = 0$, only the ICM optimization is used, without simulated annealing.

performances. It could be concluded that the dynamics (6.2), in conjunction with the proposed criterion (6.6), constrain the type of solution that can be obtained. Figure 6.12b shows that minimizing $U(E)$ is highly correlated with minimizing $N_{\hat{\Omega}}$, which was one of the objectives announced in Section 6.1.3. Figure 6.12c shows that a high initial temperature η_0 leads to a result independent from the initialization. This is similar to the well-known property of simulated annealing when temperature decreases in a logarithmic fashion (Geman and Geman, 1984).

The diversity of behaviors observed for a lower initial temperature η_0 can be explained as follows: when the initial labeling is rather bad (SA(1) and SA(N_r)), since the local optimization is pointwise and points are visited at random (see Table 6.4), the procedure tends to accept too often the NewLabel tag, which ultimately results in an oversplitted solution. This effect does not appear when using the SA(SW-1) solution, because the labels are much more stable, due to a lower local posterior probability of the NewLabel tag. Overall, results with the lowest energies are obtained using a somewhat low initial temperature η_0, and SA(SW-1). SW-1 alone provides close-to-optimal results, in terms of energy. Thus, in the following, results are reported for SW-1 only.

Comparison with lapels: Table 6.6 gives the segmentation performance on the test set for SW-1 and the lapel baseline. The proposed approach SW-1 compares well with the lapel baseline, both in terms of F-measure and DER. SW-1 also yields a major improvement on overlapped speech. These results are particularly significant, given the high precision of the ground-truth and the fact that we use distant microphones only. Indeed, close-talking lapel signals are about 8 dB cleaner than distant microphone array signals, due to the difference of distance (Table 3.1). The decrease

	Lapel baseline	SW-1	SW-7
PRC	89.3 (67.8)	83.8 (71.8)	83.9 (71.7)
RCL	90.4 (63.8)	90.9 (82.0)	90.6 (81.6)
F	89.8 (64.6)	87.2 (75.7)	87.0 (75.5)
DER	8.2 (34.9)	11.8 (19.7)	12.0 (20.0)

Table 6.6. Segmentation results on the M4 Corpus. SW-1 and SW-7 use distant microphones only. Values are percentages, results on overlaps only are indicated in brackets. PRC, RCL, F: the higher the better. DER: the lower, the better.

in precision may be due to the automatic SNS decision leading to more False Positives (D_{FP}) as compared to lapels, because the decision is taken *without* knowledge of the number of speakers. On the contrary, the number of speakers is implicitly known in the lapel baseline.

Comparison with a previous speaker clustering work: We also compared SW-1 to a HMM-based previous work (Ajmera et al., 2004), on a different task: only 6 meetings are segmented, and the task excludes silences smaller than 2 seconds. Results are reported in Table 6.7. There is a clear improvement. However, the previous work was attacking a wider task: speech segmentation and speaker clustering. This comparison shows that we can obtain a very good speech segmentation with location cues. Chapter 7 builds a speaker clustering approach based on this segmentation.

Window size: In Table 6.6, the two results SW-1 and SW-7 show that N_{future}, the size of the "future" window, has little impact in the framework of the meeting application. However, this may not be the case in other contexts: for example, the confident clustering approach introduced in Section 6.3.1 *requires* $N_{\text{future}} > 1$.

Interest of STC: As in Section 6.4.3, the same segmentation experiments were also conducted with the Speech/Non-Speech decision taken for each location estimate r_i individually – without STC. Results reported in Table 6.8 clearly show that the proposed STC method leads to the best results, and is a lot less dependent on segmentation post-processing. Finally, we observed on the raw results that the non-stationarity test mentioned in Section 6.4.2 effectively removes all short-term clusters belonging to the projector.

Overall, STC permits to fulfill two goals of the segmentation application: to obtain with distant microphones a segmentation performance comparable to that obtained with close-talking microphones, and to handle multiple simultaneous speakers in an appropriate manner. It can serve as a strong starting point for unsupervised speaker clustering with distant microphones only: Chapter 7 reports results superior to that of a state-of-the-art approach.

	SW-1	HMM-based
HTER	4.3	17.3

Table 6.7. Comparison with a previous speaker clustering work: segmentation results on 6 meetings, with a silence minimum duration of 2 seconds. Values are percentages: the lower, the better.

SNS decision granularity	Result without post-processing	Result with post-processing
Individual: r_i	48.1	84.6
Cluster: ω_n (STC with SW-1)	83.1	87.2

Table 6.8. F-measure on the M4 Corpus with SW-1, for two types of speech/non-speech decisions. The segmentation post-processing is detailed in Section 6.5.4.

6.6 Conclusion

Accurate segmentation and tracking of speech in a meeting room is critical for a number of tasks, including speech acquisition and recognition, speaker tracking, and recognition of higher-level events.

In this chapter, we first described a generic, threshold-free scheme for Short-Term Clustering (STC) of sporadic and concurrent events. The motivation behind this approach is that with highly sporadic modalities such as speech, it may not be relevant to try to output a single trajectory for each source over the entire data, which would lead to complex data association issues. We proposed here to track speakers in the short-term only, thus avoiding such issues. The core of our approach is a threshold-free probabilistic criterion. We described an algorithm based on a sliding-window analysis, spanning a context of several time frames at once. It is online, can be fully deterministic and can function in real-time when using reasonable context durations (N_{future}). It is unsupervised: local dynamics are extracted from the data itself, and the STC is threshold-free. We also presented investigations on the problem of trajectory crossings, useful for example in the context of acoustic vehicle tracking (Pham and Fong, 1997) or visual tracking (Jorge et al., 2004).

Second, we described speech specific applications of this algorithm. STC was used to build a multispeaker detection-localization system with microphone arrays, which was then successfully applied to both dynamic and static recordings with multiple simultaneous speakers. In both cases, STC permits to discriminate between speech and non-speech in a much more advantageous manner, as compared to an individual decision for each location estimate. Highly dynamic, non-linear human motions are well handled by the STC algorithm. In particular, a comparison with offline simulated annealing optimization shows that the proposed online implementation is sufficient. In addition, experiments on synthetic data highlighted the benefit of processing several time frames at once ($N_{\mathrm{future}} > 1$).

In terms of final performance, STC leads to a meeting segmentation performance, with distant microphones only, close to that obtained with close-talking microphones. This result can already be considered as a success, since distant microphones are much more noisy than close-talking microphones. Moreover, since multiple speech sources are effectively "tracked in the short-term", a dramatic improvement is observed in the case of overlapped speech, which is often found in spontaneous multi-party speech. These results validate the STC algorithm, as well as the idea of relying on location cues to obtain high precision short-term tracking and speech segmentation of multiple moving speakers. This in turn permits a much wider range of applications than with close-talking microphones, due to the non-intrusive aspect of distant microphones. Investigations on the unsupervised speaker clustering task with distant microphones in Chapter 7 show that STC can serve as a foundation for a speaker clustering approach that has a performance superior to that of a state-of-the-art approach.

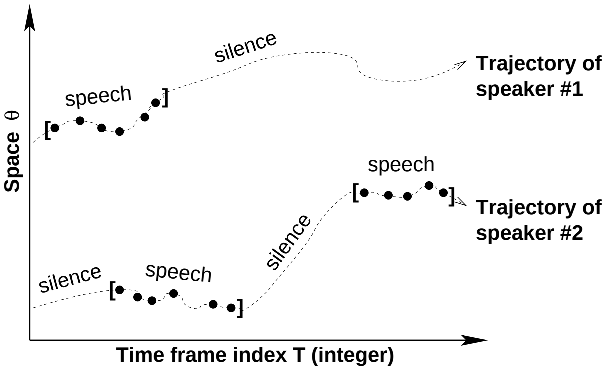
Chapter 7

Speaker Clustering with Distant Microphones

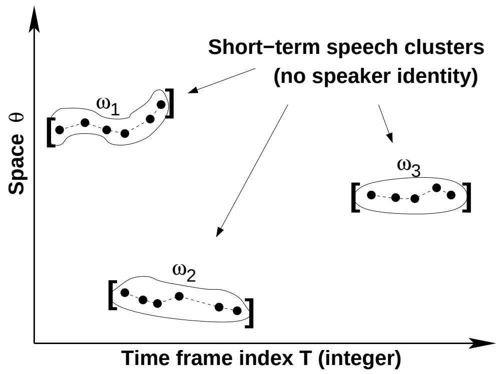
As explained in Section 1.1, the global objective of this thesis is to determine who spoke where and when (Figure 7.1a), in the context of spontaneous speech and varying, *possibly distant*, speakers. Chapter 6 addressed the "Where? When?" questions, producing short-term speech clusters of location estimates (Figure 7.1b). This chapter addresses the remaining question "Who?" in the general case where enrollment data is not available: this is the speaker clustering task. We propose to group the short-term speech clusters to form longer-term speaker clusters (Figure 7.1c).

Section 3.2.1 explained why the speaker clustering task is difficult in the present context, giving three main reasons that we summarize here. First, spontaneous speech utterances are sporadic and can be very short, whereas MFCC/GMM-based speaker modelling requires a minimum amount of data for each speech utterance (2 or 3 seconds). Second, speakers spontaneously tend to talk over each other, so using a minimum duration constraint would lead to incorporate in each segment not only the speech from a given speaker, but also from the speech from overlapping speakers. Third, only distant microphones are used, which have a lower SNR than close-talking microphones (Table 3.1).

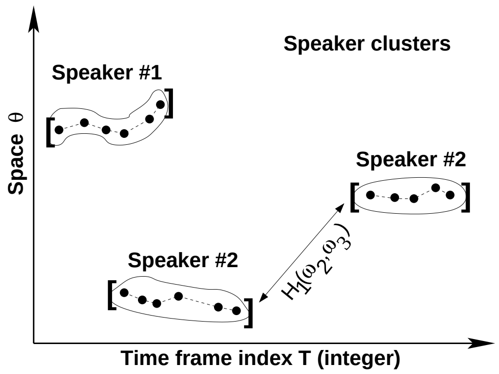
In the context of spontaneous multi-party speech, and if only acoustic information (MFCC) is used, one possible solution is to use post-processing steps (Anguera et al., 2005). On the other



(a) Goal



(b)
Result of Chapter 6:
Short-term speech clusters
(Where? When?)



(c)
The present chapter:
Long-term speaker clustering
(Who?)

Figure 7.1. The goal (a) and the proposed approach (b)(c). Dots depicts instantaneous location estimates $r_i \stackrel{\mathrm{def}}{=} (\theta_i, T_i)$. Dashed lines depict trajectories of the sources (true in (a), estimated in (b) and (c)). Square brackets depict the beginning and the end of each speech utterance. Round, continuous lines depict the short-term clusters $\omega_1, \omega_2, \omega_3$.

hand, Chapter 6 showed that microphone arrays allow for a *precise* speech/silence time segmentation, using Short-Term Clustering (STC) of microphone array-based speaker location estimates (Figure 7.1b). The goal of the present chapter is to exploit the complementarity of the two modalities (location and MFCC) in a principled manner, for speaker clustering with distant microphones:

- Location cues permit excellent short-term discrimination between speakers, but provide no speaker identity information (a speaker can move while silent, then speak again elsewhere).
- Acoustic cues (MFCC) carry long-term speaker identity information, but a minimum duration is needed to build reliable speaker models.

The rest of this chapter is organized as follows:

- Section 7.1 proposes to combine the location and acoustic modalities for speaker clustering, through a Bayesian Information Criterion modified for multiple modalities. Experimental results on the M4 Corpus (McCowan et al., 2005) are provided, showing a performance superior to that of a state-of-the art approach.
- The proposed approach correctly groups in one speaker cluster, the speech utterances from the same speaker located at different azimuth directions. However, it fails when the speaker increases his/her distance from the array. Possible causes include specificities of the human acoustic radiation characteristics (Schwetz et al., 2004).
- When available, the visual modality can help to circumvent the audio shortcomings. However, audio-visual speaker tracking necessitates a relative calibration of microphones and cameras. Section 7.2 proposes two unsupervised audio-visual calibration methods.
- Section 7.3 concludes.

7.1 Speaker Clustering with Audio Modalities

Chapter 6 proposed to estimate the speech segments in both time and space, through Short-Term Clustering (STC), thus answering the "Where? When?" questions (Figure 7.1b). The present section addresses the "Who?" question, where the speaker identity needs to be determined for each speech utterance. More precisely, we investigate the speaker clustering task, where no enrollment data is available. Many works (Sugiyama et al., 1993; Chen and Gopalakrishnan, 1998; Ajmera and Wooters, 2003; Galliano et al., 2005; Valente, 2006) examined the single, close-talking audio channel, in constrained environments such as broadcast news data. In contrast, the present section investigates multichannel recordings of highly dynamic, spontaneous multi-party speech. The main contribution presented here is a method that combines location cues from a microphone array, and acoustic cues, e.g. represented by MFCC parameters from one of the microphones, to obtain a more robust speaker clustering result. This is particularly challenging given that only distant microphones are used. The underlying motivation is to provide an automatic system that does not force participants to wear any device at all (Section 1.1).

This section is structured as follows. Section 7.1.1 introduces the Bayesian Information Criterion (BIC) for the speaker clustering task. Section 7.1.2 proposes a multimodal BIC criterion for speaker clustering, that merges location cues and acoustic cues (MFCC). Section 7.1.3 evaluates the proposed multimodal BIC criterion on the M4 Corpus. Section 7.1.4 contains a discussion and suggestions for future directions.

The author would like to acknowledge the help of Dr Fabio Valente from IDIAP, for the experiments with the "GMM/HMM" method.

7.1.1 The Bayesian Information Criterion (BIC) for Speaker Clustering

This subsection summarizes the BIC approach for speaker clustering, that was initially introduced by (Chen and Gopalakrishnan, 1998). Let us assume an ordered sequence $\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}}$ of $N_{\boldsymbol{\xi}}$ observed samples:

$$\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}} \stackrel{\text{def}}{=} \left(\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n, \cdots, \boldsymbol{\xi}_{N_{\boldsymbol{\xi}}}\right) \tag{7.1}$$

where each observed sample $\boldsymbol{\xi}_n$ can be any measurement giving some indication about the identity of the current speaker, for example a feature vector (MFCC, location, or other). The goal of speaker clustering is to split the data $\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}}$ into N_Ω *long term* clusters, defining a long-term partition Ω:

$$\Omega \stackrel{\text{def}}{=} \{\omega_1, \omega_2, \cdots, \omega_{N_\Omega}\} \tag{7.2}$$

where each **long-term speaker cluster** $\omega_n \subset \boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}}$ can span several segments of data, possibly separated by long "durations". See for example the cluster ω_3 in the four-cluster long-term partition depicted in Figure 7.2. Formally, in the case of location estimates, long-term clusters include the short-term clusters defined in Chapter 6 (which verify (6.3)) as well as clusters "distributed along time", that do not verify (6.3). We thus permit $\Omega \notin O_{\text{ST}}$, as defined by (6.4).

Ideally, the estimated N_Ω is equal to the true number of speakers, and each long-term cluster ω_n contains all the data from one and only one speaker. One solution is to optimize the long-term partition Ω and its number of clusters N_Ω, with respect to a criterion. In this chapter we opted for the Bayesian Information Criterion (BIC), which permits to evaluate a probabilistic representation of the data, so that the likelihood of the data is maximized while keeping the complexity of the model

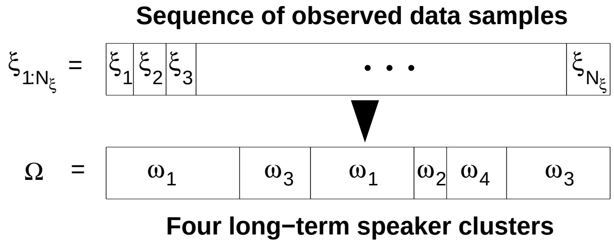


Figure 7.2. Example of speaker clustering: a long-term partition Ω. In this example the data samples $\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}}$ are splitted into $N_\Omega = 4$ *long-term* speaker clusters $\omega_1, \omega_2, \omega_3$ and ω_4.

as low as possible. In practice, there are two cases: either we have built a complete probabilistic model $\mathcal{M}(\Omega)$ of the long-term partition Ω, or we only have local models $(\mathcal{M}(\omega_1), \cdots, \mathcal{M}(\omega_{N_\Omega}))$, one for each cluster ω_n. A model $\mathcal{M}(\omega_i)$ can be a GMM trained on ω_i, for example.

Complete model: For a given long-term partition Ω and its associated model $\mathcal{M}(\Omega)$:

$$\mathrm{BIC}(\Omega) \stackrel{\mathrm{def}}{=} \log p\left(\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}} \,\middle|\, \mathcal{M}(\Omega)\right) - \frac{\lambda}{2} \cdot \kappa(\mathcal{M}(\Omega)) \cdot \log N_{\boldsymbol{\xi}} \tag{7.3}$$

where $p\left(\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}} \,\middle|\, \mathcal{M}(\Omega)\right)$ is the likelihood of the observed data $\boldsymbol{\xi}_{1:N_{\boldsymbol{\xi}}}$, given the model $\mathcal{M}(\Omega)$. λ is an adjusting parameter (in the original BIC definition $\lambda = 1$). BIC allows to compare various models with different number of free parameters $\kappa(\mathcal{M})$, by selecting the model with the maximum BIC value. If the values of the model parameters are selected in an optimal manner, the first term in (7.3) is supposed to increase when the number of free parameters $\kappa(\mathcal{M})$ increases. The second term in (7.3) is often called the "penalty term", as it penalizes models that have too many free parameters.

Local models: In the case of acoustic speaker clustering (Chen and Gopalkrishnan, 1998), a global model $\mathcal{M}(\Omega)$ may not be available. The data in each cluster ω_n is represented by a local model $\mathcal{M}(\omega_n)$ with $\kappa(\mathcal{M}(\omega_n))$ free parameters. For a given partition Ω, a commonly used criterion is a sum of "local" BICs (LBIC):

$$\begin{aligned}\mathrm{LBIC}(\mathcal{M}(\omega_1), \cdots, \mathcal{M}(\omega_{N_\Omega})) &\stackrel{\mathrm{def}}{=} \sum_{n=1}^{N_\Omega} \mathrm{BIC}(\omega_n) \\ &= \sum_{n=1}^{N_\Omega} \left[\log p(\omega_n \mid \mathcal{M}(\omega_n)) - \frac{\lambda}{2} \cdot \kappa(\mathcal{M}(\omega_n)) \cdot \log(\mathrm{card}\,\omega_n)\right]\end{aligned} \tag{7.4}$$

To determine a partition Ω that maximizes LBIC, an iterative merging approach can be used. A typical merging iteration is to test all possible pairs of clusters (ω_i, ω_j), where $1 \le i < j \le N_\Omega$, and to merge the pair of clusters that yields the maximum increase of LBIC, as long as it is positive. Convergence is reached when it is no longer possible to increase LBIC.

To implement iterative merging, only local comparisons are required. In each comparison, the possibility of merging two clusters ω_i and ω_j into a merged cluster $\omega_i \cup \omega_j$ is evaluated by computing the difference $\Delta_{i,j}^{\mathrm{BIC}}$ of the two LBIC scores (before and after merge):

$$\begin{aligned}
\Delta_{i,j}^{\mathrm{BIC}} &\stackrel{\mathrm{def}}{=} \mathrm{BIC}\left(\omega_i \cup \omega_j\right) - \mathrm{BIC}\left(\omega_i\right) - \mathrm{BIC}\left(\omega_j\right) \\
&= \log p\left(\omega_i \cup \omega_j \mid \mathcal{M}\left(\omega_i \cup \omega_j\right)\right) - \log p\left(\omega_i \mid \mathcal{M}\left(\omega_i\right)\right) - \log p\left(\omega_j \mid \mathcal{M}\left(\omega_j\right)\right) \\
&\quad -\frac{\lambda}{2} \cdot \kappa\left(\mathcal{M}\left(\omega_i \cup \omega_j\right)\right) \cdot \log\left(\mathrm{card}\,\omega_i + \mathrm{card}\,\omega_j\right) \\
&\quad +\frac{\lambda}{2} \cdot \left[\kappa\left(\mathcal{M}\left(\omega_i\right)\right) \cdot \log\left(\mathrm{card}\,\omega_i\right) + \kappa\left(\mathcal{M}\left(\omega_j\right)\right) \cdot \log\left(\mathrm{card}\,\omega_j\right)\right]
\end{aligned} \tag{7.5}$$

At a given iteration, if $\max\limits_{i<j} \Delta_{i,j}^{\mathrm{BIC}}$ is positive, then the corresponding merge is applied. If it is negative, then the iterative process is stopped.

7.1.2 Combining Two Modalities: Location Cues and Acoustic Cues

Let us now consider the problem at hand. The goal is to merge the (small) speech segments produced by the STC[1], using two modalities: acoustic cues (MFCC vectors) and location cues (azimuth direction estimates). An iterative merging approach is considered, as described in Section 7.1.1. We start with the result of the STC[1] (Figure 7.1b), that is $N_\Omega = N_\Omega^{\mathrm{init}}$, where N_Ω^{init} is the number of short-term speech clusters produced by the STC[1]. The N_Ω^{init} value can be quite large: for example 3700 for a five-minute meeting. We thus opted for simple modelling approaches for both acoustic and location cues. We first review each modality in more details, then propose the fusion criterion.

Acoustic cues: MFCC vectors

For each *instantaneous* location estimate $r_n = (\theta_n, T_n)$, where θ_i is an azimuth value in radians, and $T_n \in \mathbb{N} \setminus \{0\}$ an integer time frame index[2], we extract an acoustic observation $\boldsymbol{\xi}_n^{(\mathrm{ac})}$, which is a

[1] We use the short-term speech clusters produced by SW-1 in Section 6.5.2, with "Cluster SNS" classification (Section 6.4.2). Since time constraints are not considered in the present chapter, we chose SNSGMM (Section 5.5.3).

[2] Following the frame-based analyses presented in Chapters 5 and 6, the present chapter uses integer time frame indices $T_i \in \mathbb{N} \setminus \{0\}$. However, the speaker clustering approach defined in the present chapter could be formulated very similarly without time frames, using instead time values t_i expressed in sampling periods.

MFCC vector from a single microphone[3] (delay-sum[4] did not bring any improvement, as confirmed by the results reported below). So the total number of acoustic observations $N_{\boldsymbol{\xi}}^{(\mathrm{ac})}$ is equal to the total number of instantaneous location estimates in all short-term speech clusters N_r:

$$N_{\boldsymbol{\xi}}^{(\mathrm{ac})} = N_r \tag{7.6}$$

As mentioned above, we start with one speaker cluster per short-term speech cluster produced by the STC (Figure 7.1b): $N_\Omega = N_\Omega^{\mathrm{init}}$. There can be many of them, so we opted for a simple modelling approach, with one "local" model $\mathcal{M}^{(\mathrm{ac})}(\omega_n)$ for each cluster ω_n. The clusters are merged iteratively until LBIC $\left(\mathcal{M}^{(\mathrm{ac})}(\omega_1), \cdots, \mathcal{M}^{(\mathrm{ac})}(\omega_{N_\Omega})\right)$ reaches a maximum. For each cluster we use a single Gaussian model with a full covariance matrix (instead of a GMM). Indeed, the merge of two single Gaussians into one single Gaussian can be done analytically, as summarized in Appendix E. The cost is thus small, as opposed to the somewhat time-consuming EM fitting of a GMM.

Location cues: azimuth differences

In this case, a global model $\mathcal{M}^{(\mathrm{loc})}$ is available, as defined below. The idea is to use the same criterion as in the Short-Term Clustering (STC, Section 6.1.2). The only difference is that instead of modelling the azimuth difference $\theta_i - \theta_j$ between two individual location estimates, we model here the azimuth difference $\xi_n^{(\mathrm{loc})}$ between two short-term speech clusters of location estimates.

Location observations: $\xi_n^{(\mathrm{loc})}$ is the azimuth *difference* between two short-term clusters produced by the STC (in radians, between $-\pi$ and $+\pi$). This location observation $\xi_n^{(\mathrm{loc})}$ is equal to the difference between the respective average azimuths of the two short-term speech clusters. The total number of such location observations $\xi_n^{(\mathrm{loc})}$ is thus:

$$N_{\xi}^{(\mathrm{loc})} \stackrel{\mathrm{def}}{=} \frac{N_\Omega^{\mathrm{init}} \cdot \left(N_\Omega^{\mathrm{init}} - 1\right)}{2} \tag{7.7}$$

Global location model $\mathcal{M}^{(\mathrm{loc})}$*:* The optimization process amounts to select an optimal coherent graph that links any two short-term speech clusters (vertices) with a relationship "same" or "different" (edges), similarly to STC (Figure 6.4). Initially, all edges are set as "different" ($H_0(\omega_i, \omega_j)$,

[3]In this case (single microphone channel), the MFCC vector is *not* location-dependent, which implies that two location estimates r_i and r_j with the same time $T_i = T_j$ have the same MFCC vector: $\boldsymbol{\xi}_i^{(\mathrm{ac})} = \boldsymbol{\xi}_j^{(\mathrm{ac})}$.

[4]In this case (delay-sum beamforming), the MFCC vector is location-dependent, which implies that two location estimates r_i and r_j with the same time $T_i = T_j$ but different locations $\theta_i \neq \theta_j$ have different MFCC vectors: $\boldsymbol{\xi}_i^{(\mathrm{ac})} \neq \boldsymbol{\xi}_j^{(\mathrm{ac})}$.

Figure 7.1b). "Merging" is done by setting an edge of the graph as "same" ($H_1(\omega_i, \omega_j)$, Figure 7.1c). When a pair of short-term speech clusters is separated by a small time duration (e.g. 5 sec), we model the azimuth difference $\xi_n^{(\mathrm{loc})}$ using a bi-Gaussian model[5], similarly to (6.2) in the STC. Otherwise, we assume a uniform distribution for $\xi_n^{(\mathrm{loc})}$. Thus, $\mathcal{M}^{(\mathrm{loc})}$ has a fixed number of parameters[6] $\kappa\left(\mathcal{M}^{(\mathrm{loc})}\right) = 2$, so that only the likelihood term is relevant:

$$\mathrm{BIC}\left(\mathcal{M}^{(\mathrm{loc})}\right) = \log p\left(\xi^{(\mathrm{loc})}_{1:N_\xi^{(\mathrm{loc})}} \,\middle|\, \mathcal{M}^{(\mathrm{loc})}\right) + \mathrm{const} \tag{7.8}$$

Maximizing $\mathrm{BIC}\left(\mathcal{M}^{(\mathrm{loc})}\right)$ is thus similar to the maximum likelihood scheme in STC (6.6).

Fusion of Acoustic Cues and Location Cues: Multimodal BIC

In practice, location cues are very useful to separate fast changing speaker turns, but carry no reliable long-term information (speakers can move while silent). On the other hand, acoustic cues such as MFCC carry long-term speaker identity information, but a minimum amount of speech is required to build a reliable speaker model (at least 2 or 3 seconds). In spontaneous multi-party speech, many speech segments are short so it is not always possible to build reliable speaker models. Given these considerations, it would be desirable to combine the strengths of both location and acoustic modalities, in order to build a reliable speaker clustering system, for spontaneous multi-party speech. More specifically, the location cues could help to prevent merging two speech clusters that are very close in time but very far in space, as observed in fast-changing speaker turns.

An existing combination approach (Ajmera et al., 2004), which was successful on a limited amount of data (6 meetings from the M4 Corpus (McCowan et al., 2005)), was tested on a larger amount of data (12 meetings from the same corpus). Unfortunately, it failed to produce meaningful speaker clusters, most likely because of the uniform initialization of the clusters (many speakers would fall into the same cluster). Alternatively, we propose here to initialize the clusters using the N_Ω^{init} speech segments provided by the STC (Figure 7.1b and Chapter 6). The clusters are then

[5]One Gaussian for "same" ($H_1(\omega_i, \omega_j)$), the other Gaussian for "different" ($H_0(\omega_i, \omega_j)$). Both are zero-centered, and their standard deviations are the means $\langle\sigma_T^{\mathrm{diff}}\rangle_T$ and $\langle\sigma_T^{\mathrm{same}}\rangle_T$, respectively. σ_T^{diff} and $\sigma_T^{\mathrm{same}}$are defined in (6.2).

[6]The parameters of $\mathcal{M}^{(\mathrm{loc})}$ are the standard deviations of 2 zero-centered Gaussians (one for "same", one for "different").

iteratively merged (as in Section 7.1.1) to maximize the following "multimodal" BIC:

$$\boxed{\frac{\mathrm{LBIC}\left(\mathcal{M}^{(\mathrm{ac})}(\omega_1),\cdots,\mathcal{M}^{(\mathrm{ac})}(\omega_{N_\Omega})\right)}{N_{\boldsymbol{\xi}}^{(\mathrm{ac})}} \quad + \quad \frac{\mathrm{BIC}\left(\mathcal{M}^{(\mathrm{loc})}(\Omega)\right)}{N_{\xi}^{(\mathrm{loc})}}} \tag{7.9}$$

where $\Omega = \{\omega_1, \cdots, \omega_{N_\Omega}\}$ is a candidate partition of the location estimates $r_{1:N_r}$ and their associated MFCC vectors $\boldsymbol{\xi}^{(\mathrm{ac})}_{1:N_{\boldsymbol{\xi}}^{(\mathrm{ac})}}$ ($N_{\boldsymbol{\xi}}^{(\mathrm{ac})} = N_r$). The normalizations by $N_{\boldsymbol{\xi}}^{(\mathrm{ac})}$ and $N_{\xi}^{(\mathrm{loc})}$ factor out the number of terms in the log likelihood term of each BIC score. The idea is to add values that are somewhat comparable. Note that this "multimodal" BIC approach can potentially be used with any number or types of modalities.

Implementation

The initial short-term clusters are very often much shorter than the minimum required to build reliable MFCC speaker models. The first term in (7.9) may thus be very noisy at first. We thus first run a location-only iterative merging, maximizing only the second term of (7.9) until it cannot be increased anymore (order of magnitude for a 5-minute meeting: from 3000 clusters to 300 clusters). Then only, we run the acoustic + location iterative merging, where the complete criterion (7.9) is maximized (order of magnitude for a 5-minute meeting: from 300 clusters to 4 clusters).

7.1.3 Experimental Results

Methods

"GMM/HMM" and "lapelmix-GMM/HMM" start with $N_\Omega = 10$ clusters, and all other methods start with $N_\Omega = N_\Omega^{\mathrm{init}}$ clusters, one per speech cluster given by the STC (Figure 7.1b). Note that we used 24-dimensional MFCC vectors as acoustic cues: $\boldsymbol{\xi}_n^{(\mathrm{ac})} \in \mathbb{R}^{24}$. Five clustering methods use distant microphones only:

"GMM/HMM": Acoustic cues only, where $\boldsymbol{\xi}^{(\mathrm{ac})}_{1:N_{\boldsymbol{\xi}}^{(\mathrm{ac})}}$ are single channel MFCCs from one microphone in the array – not to be confused with Sector-Based MFCCs. The speaker clustering algorithm is described in (Ajmera and Wooters, 2003). For each meeting, we started the merging process with $N_\Omega = 10$ clusters.

"ac only": Acoustic cues only, where $\boldsymbol{\xi}^{(\mathrm{ac})}_{1:N^{(\mathrm{ac})}_{\boldsymbol{\xi}}}$ are single channel MFCCs from one microphone in the array. Each local model $\mathcal{M}^{(\mathrm{ac})}(\omega_n)$ is a single Gaussian with full covariance matrix. The speaker clusters are iteratively merged to maximize LBIC $\left(\mathcal{M}^{(\mathrm{ac})}(\omega_1), \cdots, \mathcal{M}^{(\mathrm{ac})}(\omega_{N_\Omega})\right)$ with a tunable λ, as described in Section 7.1.1.

"ac + loc": Acoustic + location (7.9), also using single channel MFCCs for $\boldsymbol{\xi}^{(\mathrm{ac})}_{1:N^{(\mathrm{ac})}_{\boldsymbol{\xi}}}$.

"ds, ac + loc": Acoustic + location (7.9), where each $\xi^{(\mathrm{ac})}_n$ is a MFCC vector extracted from the delay-sum beamformed signal. The beamforming direction is the azimuth θ_n associated with $\boldsymbol{\xi}^{(\mathrm{ac})}_n$.

"loc. only (cheating)": Location only (7.8), MFCCs are not used. It is "cheating", in the sense that the speaker identity of each speech segment is given by the prior knowledge of the locations of all speakers.

For comparison, two other methods use close-talking microphones only (lapel worn near the throat):

"lapels-seg (cheating)": The multichannel segmentation baseline described in Section 6.5.3. The word "cheating" means that the true number of speakers and the speaker identities are known by construction (one lapel per person).

"lapelmix-GMM/HMM": The single channel iterative speaker clustering scheme described in (Ajmera and Wooters, 2003). For each meeting, the single channel is formed by first adding the four lapel signals. For each meeting, we started the merging process with $N_\Omega = 10$ clusters.

Performance Metrics

For each method, we report the results using the **Diarization Error Rate (DER)** and the **Speaker Activity Detection (SAD)** metrics, both expressed in percentage (the lower, the better), as defined in (NIST, 2003). DER and SAD exclude part of the data from the evaluation: within a collar of 0.25 sec around each speech segment end-point, results are not evaluated. Moreover, short silences are removed from both ground-truth and result. These short silences are defined as less than 0.3 second (as in the NIST specification (NIST, 2003)) or less than 2.0 seconds, thus defining two subtasks called "0.3 task" and "2.0 task", respectively.

Within the remaining segments, the DER is then defined as the percentage of speech that was wrongly attributed (the lower, the better): DER = MISS + FA + SPKR, where MISS and FA are the

Microphones	Method	0.3 task DER		0.3 task SAD	2.0 task DER		2.0 task SAD
Distant (mic. array)	GMM/HMM	37.6	(63.9)	14.4	32.2	(68.6)	4.6
	ac only	51.7	(77.5)	11.4	48.9	(76.7)	4.4
	ds, ac + loc	27.8	(40.4)	11.1	24.8	(35.6)	4.3
	ac + loc	18.1	(37.4)	9.7	15.2	(30.0)	4.0
	loc. only (cheating)	12.5	(19.7)	8.2	10.4	(17.9)	4.1
Close-talking	lapelmix-GMM/HMM	29.7	(66.0)	14.4	24.0	(65.1)	4.5
	lapels-seg (cheating)	8.2	(34.9)	4.7	9.1	(27.2)	2.1

Table 7.1. Speaker clustering results on 18 meetings of the M4 Corpus (McCowan et al., 2005) (the lower, the better). Brackets indicate results on overlapped speech only. "ds" stands for delay-sum beamforming. "GMM/HMM" is the speaker clustering algorithm described in (Ajmera and Wooters, 2003).

percentages of missed speech and false alarms, respectively, and SPKR is the percentage of speech attributed to the wrong speaker. Overall, a low DER implies (1) that the estimated number of speakers is close to the true number of speakers, *and* (2) that the estimated speaker segmentation is close to the true speaker segmentation. The SAD metric (NIST, 2003) is defined similarly to DER, but using only two classes (speech and silence), instead of speaker labels. Full details on the DER can be found in (NIST, 2003).

Test Data and Protocol

Speaker clustering experiments were conducted on the M4 Corpus of meetings (McCowan et al., 2005). We used the 21 short meetings – about 5 minutes each – with ground-truth speech/silence segmentation for each speaker, and 4 participants in each meeting. For a given method, speaker clustering was applied on each meeting separately, then the overall DER and SAD performance metrics were evaluated, as in (NIST, 2003). From the 21 meetings, 3 were used to tune λ and the post-processing parameters, and 18 were used for performance evaluation. In most cases post-processing only meant dropping short silences, for example less than 0.25 sec. We tuned λ and the post-processing parameters in order to achieve equal MISS and FA (NIST, 2003).

Results

From the results reported in Table 7.1, several observations can be made. First, the methods rank in similar orders, whether the task is the "0.3 task" (precise segmentation) or the "2.0 task" (rough segmentation). Second, for all metrics, the proposed combination "ac + loc" significantly improves over both "GMM/HMM" and "ac only" results. The "ac + loc" results are in fact close to the "loc. only (cheating)" results, i.e. the best that can be obtained based on the underlying multispeaker segmentation method. Finally, delay-sum beamforming does not seem to help, as already noted in (Anguera et al., 2005).

When comparing "loc. only (cheating)" with "lapels-seg (cheating)" in terms of DER, it appears that the microphone array-based methods can potentially bring a large improvement on overlapped speech (10 to 15 % absolute) at the cost of a slight overall degradation (3 or 4 % absolute). So they can indeed be used for speaker clustering in meetings. The single channel speaker clustering scheme "lapelmix-GMM/HMM" does not seem to yield effective results. In fact, compensation heuristics such as the "purification" step proposed in (Anguera et al., 2005) are necessary for this scheme to be effective on spontaneous multi-party speech.

7.1.4 Future Directions

Overall, the results presented here show that the proposed "ac + loc" combination brings significant improvement over "ac only". This validates the proposed "multimodal" (location + acoustic) BIC criterion (7.9), in so far as it combines complementary strengths from two modalities. It would be interesting to test a similar fusion of acoustic and location information in the framework of more flexible modelling approaches such as Variational Bayes learning (Valente, 2006). In particular, this would help to model clusters with very heterogeneous lengths.

A closer look at the results may suggest future directions of research. We tested "ac + loc" on a concatenation of three meetings, in which some people appear at different locations: different seats around the microphone array, and some standing locations when doing a presentation (further from the array). The same person seated at different locations tends to be clustered correctly (1 cluster). On the contrary, people who stand up and move away to do a presentation systematically end up being clustered into 2 different clusters, depending on the distance (close or far). We tested low-level signal processing methods that have proved useful to improve MFCC-based Automatic Speech Recognition (ASR) results in such cases, hoping that they would also improve the speaker clustering results. The tested techniques included delay-sum beamforming, dereverberation (Gelbart and Morgan, 2001) and spectral subtraction (as in Section 8.2). Unfortunately, all these techniques resulted in *absolutely no performance improvement*, with respect to the present issue. Therefore, it seems that feature variability between close and distant locations is a bottleneck to speaker clustering with distant microphones, which suggests some basic research. This feature variability at higher distances may be linked to specificities of the human acoustic radiation characteristics (Schwetz et al., 2004).

On the practical side, a speaker recognition study proposed a location-dependent Cepstral Mean Normalization (Wang et al., 2005), which aims to remove the location-dependent transmission channel distortions, without removing speaker-specific characteristics. However, this requires training data with multiple speakers at multiple locations for each different room, which could limit the ease of use for end-users. An alternative direction would be to modify existing, larger speaker clustering systems such as (Anguera et al., 2005), by integrating the use of the joint BIC criterion (7.9).

7.2 Automatic Audio-Visual Calibration

The previous section gave some insights on successes and limits of audio-only speaker clustering, concerning spontaneous multi-party speech within an indoor environment. To compensate shortcomings of the audio-only approach *in its present state*, an additional modality could be used: visual information from cameras. The advantage of the visual modality is that identity cues such as faces are visible at almost all times. Therefore, even if a person moves in silence, Audio-Visual (AV) speaker tracking such as (Gatica-Perez et al., 2006) should permit to determine that it was the same speaker before and after the motion. This visually-inferred knowledge would help to circumvent location-dependent variabilities of the audio channel, such as those characterized in (Schwetz et al., 2004).

However, AV speaker tracking requires some AV calibration information, which relates the locations of the microphones to the locations of the cameras. (Gatica-Perez et al., 2006) includes a review of existing AV tracking works, with respect to the AV calibration task. It appears that so far, there is either automatic calibration of colocated sensors, or manual calibration of non-colocated sensors. For example, a short sequence can be recorded with a single speaker moving around the room, while speaking, and a visual tracker can be run. One practical issue of schemes such as the codebook approach proposed in (Gatica-Perez et al., 2006) is the need for manual initialization of the visual tracking, to start the AV calibration procedure. Moreover, the codebook approach represents a one-to-one mapping, which does not model a variable depth at a given location in the image plane. Sections 7.2.1 and 7.2.2 respectively address each of these two issues, by proposing two alternative schemes that do not require manual initialization. As announced in Section 1.1, the underlying aim is to put the least possible constraints on often non-technical end-users. Therefore, for both proposed schemes, unsupervised calibration experiments are reported that use only 30 seconds of recording, made with a single speaker moving around a room (`seq11` of the AV16.3 Corpus).

7.2.1 Calibration between Discrete Spaces

Both audio and video spaces are discretized, as depicted in Figure 7.3. For each camera, the visual space is discretized into blocks, e.g. 24 blocks vertically and 32 blocks horizontally (Figure 7.3a). The audio space, for example an azimuth direction θ from a microphone array, is discretized into sectors, as in the SAM-SPARSE-MEAN approach (Chapter 5). Figure 7.3b shows an example where 18 sectors, each spanning 20 degrees, cover the entire 360-degree space around a circular microphone array. The discretization of both audio and video spaces permits to estimate the correlation between *all* locations from each modality. For example, by finding peaks of covariance[7], we could determine that having a speaker in a given audio sector is highly correlated with having a speaker in a given video block of a camera (the two fat arrows in Figure 7.3). This type of information can in turn be used for initialization of more complex, model-based parametric AV calibration schemes, that could be inspired from (Svoboda, 2003; Bouguet, 2004).

We tested the proposed AV calibration on the short recording `seq11` from the AV16.3 Corpus (Chapter 4), obtained from an 8-microphone array in the middle of a room, and 3 cameras on the surrounding walls (Figure 7.4, top row). The "audio activity indicator" is the posterior probability $P^{(11)}_{\breve{s},t}$ of having speech activity within a given sector $\mathbb{S}_{\breve{s}}$ (event $\underline{B}_{\breve{s}} = 1$) and a given time frame t (event $\underline{A} = 1$), given the observed wideband activeness values $\{\zeta_{\breve{s},t}\}$ defined in (5.24), and the multidimensional model briefly described in Section 5.3.2, and detailed in Appendix C.2:

$$P^{(11)}_{\breve{s},t} \stackrel{\text{def}}{=} P\left(\underline{B}_{\breve{s}} = 1, \underline{A} = 1 \;\middle|\; \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} = \left[\zeta_{1,t}, \cdots, \zeta_{N_{\breve{s}},t}\right]^{\mathrm{T}}, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{7.10}$$

where the RHS is calculated using (C.65).

The "video motion indicator" is defined as the average $\langle P\left(\text{motion} \mid \text{pixel } x, y, t\right)\rangle_{(x,y)\in\text{block}} \in [0\ 1]$ across all pixels in a given block, at a given time frame t, of the pixel-wise posterior probability $P\left(\text{motion} \mid \text{pixel } x, y, t\right)$ of having motion in all three components R, G, B. Based on a simplifying independence assumption between R, G, and B, for each pixel-time frame (x, y, t):

$$P\left(\text{motion} \mid \text{pixel } x, y, t\right) = P\left(\text{motion} \mid \text{vmf}_{\mathrm{R}}\left(x, y, t\right)\right) \cdot P\left(\text{motion} \mid \text{vmf}_{\mathrm{G}}\left(x, y, t\right)\right) \cdot P\left(\text{motion} \mid \text{vmf}_{\mathrm{B}}\left(x, y, t\right)\right)$$

where each video motion feature $\text{vmf}_{\mathrm{R}}, \text{vmf}_{\mathrm{G}}, \text{vmf}_{\mathrm{B}}$, is defined using 3 consecutive frames $t-1,\ t,\ t+1$:

$$\text{vmf}_{\mathrm{R}}\left(x, y, t\right) \stackrel{\text{def}}{=} \sqrt{\left|\mathrm{c}_{\mathrm{R}}\left(x, y, t+1\right) - \mathrm{c}_{\mathrm{R}}\left(x, y, t\right)\right| \cdot \left|\mathrm{c}_{\mathrm{R}}\left(x, y, t\right) - \mathrm{c}_{\mathrm{R}}\left(x, y, t-1\right)\right|} \tag{7.11}$$

where $\mathrm{c}_{\mathrm{R}}\left(x, y, t\right)$ is the color value (e.g. from 0 to 255) for color R at the pixel-time frame (x, y, t) (similarly for G and B). Each posterior $P\left(\text{motion} \mid \text{vmf}_{\mathrm{R}}\left(x, y, t\right)\right)$ is estimated from a 2-component

[7] We used the covariance in this initial study. Future investigations may use the Mutual Information.

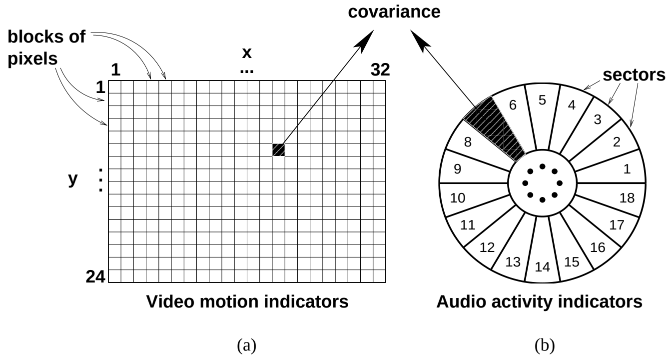


Figure 7.3. Unsupervised AV calibration in discrete space. The covariance is calculated between each video motion indicator (block of pixels) and each audio activity indicator (sector of space around the microphone array). Numbers represent indices of video blocks and audio sectors.

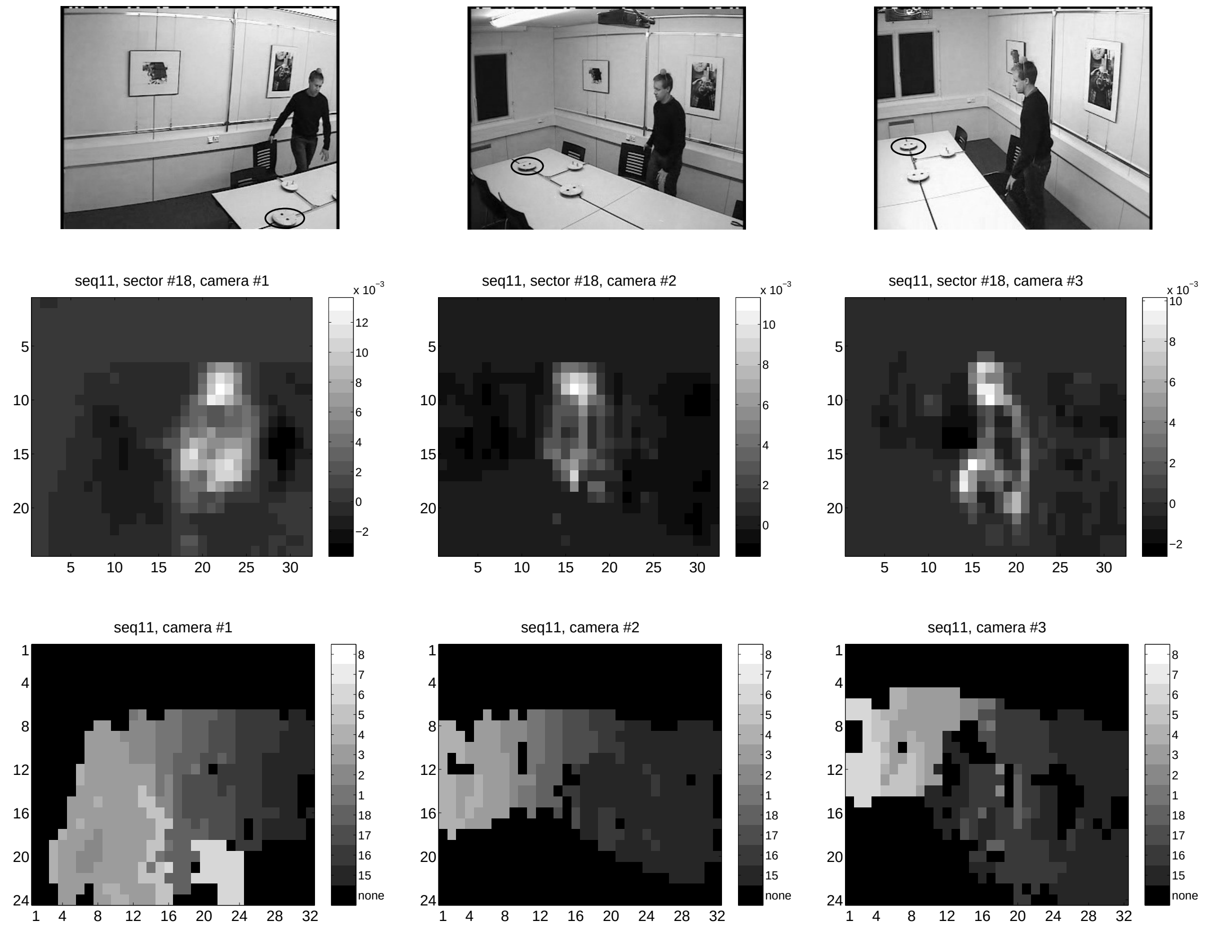


Figure 7.4. Unsupervised AV calibration in discrete space. Top row: snapshot of the `seq11` recording (microphone array indicated by a black ellipse). Middle row: AV covariance between the audio activity of sector $\mathbb{S}_{18}$ and the video motion of each camera. Bottom row: global result of the AV covariance analysis. For each block of pixels, the index (1 to 18) of the audio sector with the highest covariance is represented by a gray level (colorbar on the right side). The black background color ("none") appears whenever all audio sectors have a covariance inferior to e^{-6}.

1-dimensional model similar to the one described in Appendix C.1.2, fitted in an unsupervised manner on $\mathrm{vmf_R}\,(x, y, t)$, using EM, exactly as in Appendix C.1.2. The two components are a Gamma pdf (not moving) and a Shifted Rice pdf (moving), exactly as $\mathcal{G}_{\alpha_{01},\beta_{01}}$ and $\mathcal{R}_{\sigma_{11},V_{11}}$ in Appendix C.2.1.

Covariance analysis for calibration: we compute the covariance over time t between each "audio activity indicator" $P^{(11)}_{\check{s},t}$ and each "video motion indicator" $\langle P\,(\text{motion} \mid \text{pixel } x, y, t)\rangle_{(x,y)\in\text{block}}$, as illustrated in Figure 7.3. This approach is justified as long as the short calibration recording contains a single moving speaker, in a fixed indoor environment (as in `seq11`): there is no data association ambiguity between motion in a region of the video image and speech activity in a sector of space. An appropriate downsampling technique was used to have the same sampling rates for audio and video[8,9]. The middle row of Figure 7.4 depicts an example of AV covariance pattern: between the audio activity of sector $\mathbb{S}_{18}$ and the video motion of each camera. The body of the speaker is clearly marked by higher covariance values. After having computed the AV covariance pattern for each sector, it is possible to determine, for each block of pixels of a camera, which audio sector corresponds the most. The result of this analysis is the one-sector-to-many-pixel-blocks AV mapping depicted in the bottom row of Figure 7.4. Although rough, this AV mapping depicts quite accurately the various locations of the body of the human speaker. This is interesting, given that only 30 seconds of data were used. Note that the same type of covariance analysis could be conducted between cameras.

To conclude, the proposed discretized scheme builds an AV mapping without explicit geometrical model, and without manual intervention. This AV mapping could directly be used within an AV tracking scheme. So far, we only permitted a single depth for each location in the image plane. This limitation is addressed by Section 7.2.2.

7.2.2 Calibration without Discretization

In this section, no discretization is used, and the relationship between audio and video is estimated for each pixel in the image plane, using continuous azimuth audio measurements and pixel-wise video background subtraction. As in the previous subsection, we assume that only one speaker is

[8] The original audio frame rate was 62.5 Hz, more than twice the video frame rate 25 Hz. Frame-level (C.64) and sector-level (C.65) posterior estimates of audio activity were downsampled using the `max` and `mean` operators, respectively.

[9] The covariance needs to be calculated "on speech only". We implemented this by weighting each video frame with the downsampled frame-level posterior estimate (C.64) of audio activity.

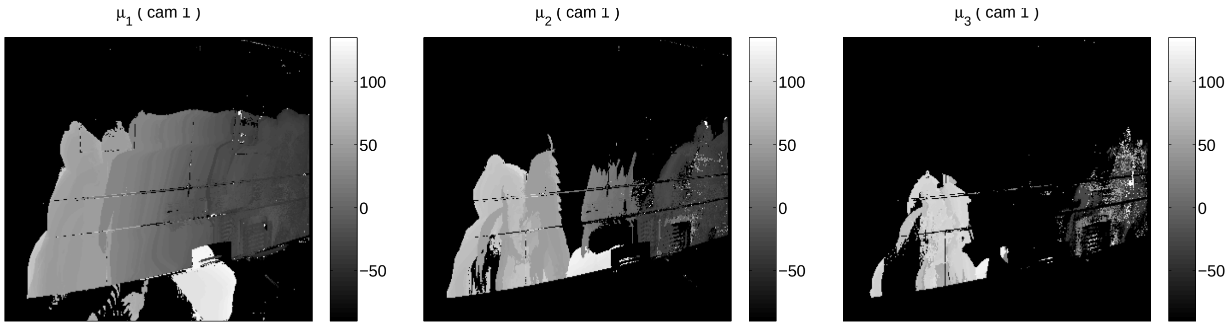
being recorded in the sequence. We propose to turn the video-only adaptive background subtraction approach from (Stauffer and Grimson, 2000), into an AV calibration procedure (step 2 below).

Step 1: Each modality is processed separately to extract measurements:

- The audio speaker detection-localization system described in Chapters 5 and 6 produces zero, one or more *speech* azimuth measurements at each audio time frame. For each video time frame (25 Hz), the corresponding speech azimuth measurements (62.5 Hz) are grouped.
- The video measurements are binary images: each pixel is classified as foreground or background. The binary decision (Figure 7.7a) is taken for each pixel independently, using adaptive background subtraction (Stauffer and Grimson, 2000). For each pixel, it consists in the online update of a GMM in the video color space, denoted "V-GMM", using all past data, then in deciding which Gaussian components in the V-GMM correspond to the foreground, and which Gaussian components correspond to the background. We used $K = 3$ components, $\alpha = 0.01$ as the learning factor, and $T = 0.66$ to discriminate between background and foreground. This choice of parameters was motivated by the possible multimodality of the background color in an indoor environment, where objects such as chairs and laptops can punctually be moved. We did not do any extensive tuning.

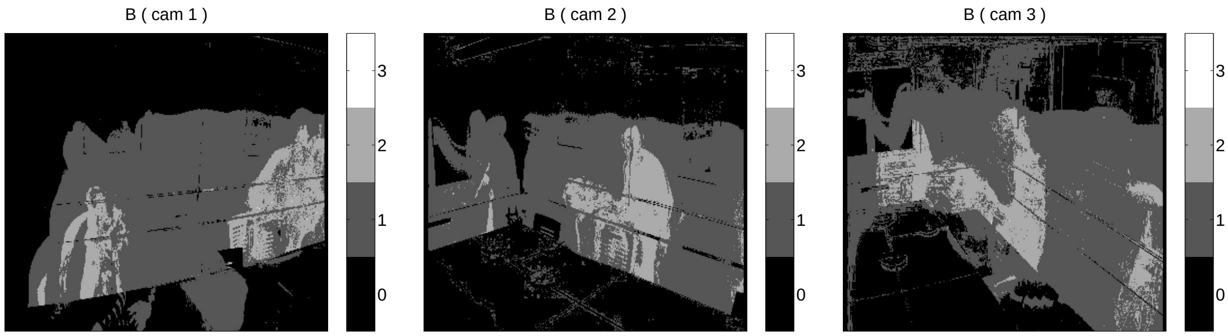
Step 2: Based on the audio and video measurements extracted in Step 1, we estimate the links between audio azimuth locations and video pixel locations. We propose a GMM framework inspired from the one used for video background subtraction (Stauffer and Grimson, 2000). At a given time t, *for each foreground pixel* we update a GMM in the audio azimuth space, denoted "A-GMM", with the current speech azimuth measurements. For each pixel, we then select the "speech" components of the A-GMM that have large weights and small standard deviations, similarly to (Stauffer and Grimson, 2000). The selected A-GMM components define the audio locations associated with this pixel location. Having potentially multiple selected Gaussians for a given pixel permits to model depth indeterminations in the image. These indeterminations are due to the fact that the microphone array and the camera are not colocated.

Calibration experiment: we applied the two steps on `seq11` of the AV16.3 Corpus. The experiment was performed three times: once for each camera. For camera #1, Figures 7.5a,b,c show the means μ_1, μ_2 and μ_3 of the three components in the A-GMM: for each pixel, the mean speech



(a) (b) (c)

Figure 7.5. Unsupervised AV calibration without discretization, camera #1. (a)(b)(c) For each pixel, the means μ_1, μ_2 and μ_3 (in degrees) of the three components of the A-GMM. For each pixel, speech components appear first in (a), then possibly in (b), then possibly in (c), followed by non-speech components in the remaining pictures. For each pixel, the number B of speech components of the A-GMM is shown in Figure 7.6a.

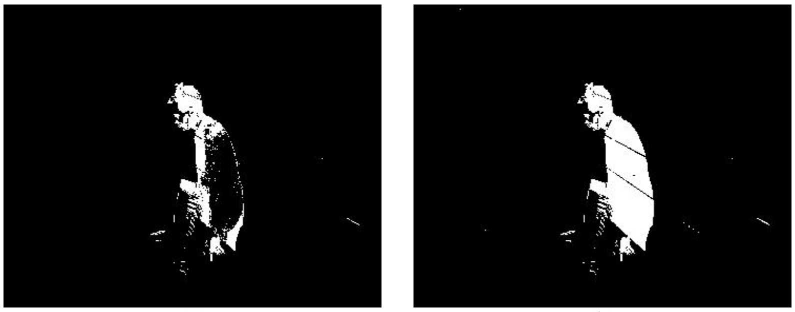


(a) (b) (c)

Figure 7.6. Unsupervised AV calibration without discretization, cameras #1, #2, #3. Each pixel depicts the number B of speech components in the corresponding A-GMM.

azimuth μ_n of one Gaussian component is represented. For each camera #1, #2, #3, Figures 7.6a,b,c show the number B of selected “speech” components in the A-GMM. The parts of the image where $B > 1$ effectively correspond to locations in the image plane, where speech azimuth may vary, depending on the depth. This result is quite interesting, given that only 30 seconds of data were used, without any manual intervention.

Improved Subtraction Experiment: To illustrate the effectiveness of the proposed AV calibration procedure, we generated a second series of “improved” foreground/background decisions. Indeed, using video-only background subtraction, whenever the speaker stops moving, his/her body is eventually absorbed by the background model, after about $1/\alpha$ video frames (4 seconds in our case). This is visible in Figure 7.7a. A partial solution to this issue is to prevent the weight adap-



(a) (b)

Figure 7.7. Unsupervised AV calibration without discretization, camera #3. Example of binary decision on frame #189. Black: background, white: foreground. (a) Video-only adaptive background subtraction (Stauffer and Grimson, 2000). (b) Same, where the weight adaptation is restricted, based on the A-GMM.

tation of the video-only V-GMM, on any pixel-time frame (x, y, t) where a current speech azimuth measurement matches a speech component of the A-GMM[10]. This effectively prevents a speaker that does not move anymore, from being absorbed in the video background, as long as he speaks. Figure 7.7b shows that more of the speaker's body is retained as foreground.

Future work: One could think of joint tracking and calibration, in the AV space. This may require to use more complex models than the one proposed here, for example by adding local continuity constraints between neighbouring pixels. The challenge would be to have a system that is robust to a large variety of data, with possibly several speakers at the same time. The gain would be automatic, joint calibration, of several sensors, *without explicit geometrical models*. At this point of time, from an applicative point of view, it seems reasonable to ask a non-technical user to walk and speak around a meeting room once at the beginning, to initialize the calibration process without any manual intervention.

[10] See (Stauffer and Grimson, 2000) for the definition of a "match" between a measurement and a GMM component.

7.3 Conclusion

Chapters 5 and 6 investigated the "Where? When?" questions, where speech segments are detected and located in both space and time, *without determining the speaker identity*. This chapter examined the remaining question "Who?", in the form of a speaker clustering task, where the previously detected speech segments are iteratively merged into speaker clusters – ideally one cluster per "true" speaker. This task is particularly difficult, (1) because of the spontaneous multi-party speech context, and (2) because only distant microphones are used. A multimodal generalization of BIC was proposed, to exploit the complementarities of long-term acoustic information (MFCCs) and short-term location information (speaker direction). Speaker clustering experiments were carried on the M4 Corpus of meetings. With distant microphones only, the proposed multimodal BIC approach yielded a major improvement over a state-of-the-art acoustic-only approach. The results also show that the speaker clustering performance of the multimodal BIC is close to the optimum that could be obtained with the underlying multispeaker detection-localization system. Results also compare well with those of a close-talking multimicrophone speech segmentation technique.

A closer look at the speaker clustering results revealed one success and one failure. The success is that data from a given speaker, seated at various azimuth angles but constant distance from the microphone array, would be correctly clustered into *one* speaker cluster. The failure is that when a speaker moves a few meters away from the microphone array, an incorrect *second* speaker cluster always appears. Signal normalization methods such as dereverberation and denoising did not provide any improvement. Further basic research is needed to explain the underlying distance-dependent acoustic feature variability.

Finally, audio-visual alternatives were investigated, where visual speaker tracking can help to circumvent the audio-only, location-dependent acoustic feature variabilities. Initial investigations on unsupervised audio-visual calibration were conducted. An approach was proposed that automatically detects the speaker depth variabilities in the image plane, based on the audio measurements. Such unsupervised audio-visual calibration could be useful as part of an audio-visual tracking system. This solution is particularly adapted to non-technical end-users, because the only requirement is to record one sequence where a person walks around the room while speaking.

Chapter 8

Applications to Other Domains

This chapter illustrates the genericity of the techniques developed in this thesis, with two applications outside the meeting room domain.

In the previous chapters, the sector-based *joint* detection-localization was used with UCAs of omnidirectional microphones, permitting reliable detection-localization of multiple speakers, in a meeting room context. Section 8.1 summarizes an in-car hands-free speech enhancement application of sector-based detection-localization, using Uniform Linear Arrays (ULAs) of omnidirectional microphones (Lathoud et al., 2006a)[1]. Although the task and the physical context differ from previously, this chapter shows that sector-based detection-localization permits to reliably control the adaptive filtering process. Experiments are conducted on real recordings made in a Mercedes S320 car, including 100 km/h background noise.

Section 8.2 summarizes a work on single-channel noise-robust ASR task (Lathoud et al., 2006b). In standard spectral subtraction (Boll, 1979; Berouti et al., 1979), the noise energy is estimated, then subtracted from the speech energy, using an independence assumption. This 2-step approach requires tuning parameters such as the spectral floor (Berouti et al., 1979). We propose an alternative, single-step approach that *jointly* estimates both speech and noise powers. As in Appendix C.1, the EM algorithm (Dempster et al., 1977) is used to fit a two-component model (speech and noise) on the observed data, thus determining the spectral subtraction floor in a fully unsupervised manner.

[1]This work was conducted with Julien Bourgeois in the framework of the European HOARSE Research Training Network.

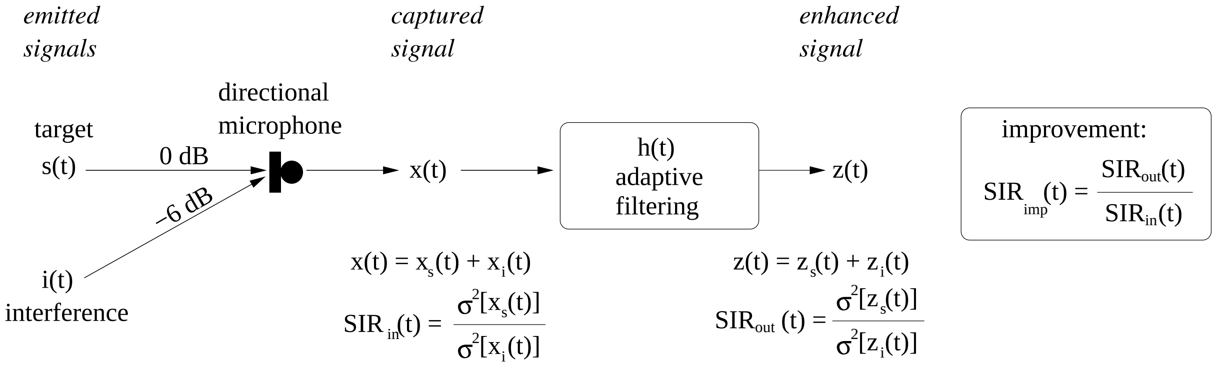


Figure 8.1. Entire acquisition process, from the emitted signals to the enhanced signal (Section 8.1). The focus is on the adaptive filtering block $h(t)$, so that $\mathrm{SIR}_{\mathrm{imp}}(t)$ is maximized when the interference is active (interference cancellation). The s and i subscripts designate contributions of target and interference, respectively. The whole process is supposed to be linear. $\sigma^2[x(t)]$ is the variance or energy of a speech signal $x(t)$, estimated on a short time frame (20 or 30 ms) around t, on which stationarity and ergodicity are assumed.

8.1 Sector-Based Detection for Hands-Free Speech Enhancement in Cars

This section summarizes an application of the sector-based activeness estimation (Chapter 5) to hands-free speech acquisition. The focus is on the separation of the driver's speech (the target) from the codriver's speech (the interference). This work is the result of a collaboration with Julien Bourgeois (while at Daimler-Chrysler), who is responsible for the adaptive filtering part, and the implicit adaptation control. Full details are available in (Lathoud et al., 2006a). An additional analysis on the stability of the implicit adaptation control can be found in (Bourgeois et al., 2005).

As speech-based command interfaces are becoming more and more common in cars, for example in automatic dialog systems for hands-free phone calls and navigation assistance, the ASR performance becomes critical. However, the ASR performance can be greatly hampered by interferences such as speech from a codriver. Moreover, spontaneous multi-party speech contains lots of overlaps between participants (Shriberg et al., 2001). A directional microphone oriented towards the driver provides an immediate *hardware* enhancement, by lowering the energy level of the codriver interference (Figure 8.1). In the Mercedes S320 setup used in the present section, a 6 dB relative difference is achieved. However, an additional *software* improvement is required to fully cancel the codriver's interference, for example with *adaptive* techniques, where a *time-varying* linear filter ($h(t)$ in Figure 8.1) enhances the Signal-to-Interference Ratio (SIR).

(a) Proposed explicit approach

(b) Proposed implicit approach

Figure 8.2. Proposed explicit and implicit adaptation control. $\mathbf{x}(t) = [x_1(t) \cdots x_{N_m}(t)]^{\mathrm{T}}$ are the signals captured by the N_m microphones, and $\mathbf{h}(t) = [\mathbf{h}_1(t) \cdots \mathbf{h}_{N_m}(t)]^{\mathrm{T}}$ are their associated filters. Double arrows denote multiple signals.

As explained in (Lathoud et al., 2006a, Section 1), we selected the Generalized Sidelobe Canceller (GSC) structure. In practice, it is required to adapt *only when the interferer is dominant*, by varying the adaptation speed in a binary manner (explicit control), or in a continuous manner (implicit control). As detailed in (Lathoud et al., 2006a, Section 1), most existing explicit methods rely on prior knowledge of the target location only. There are few implicit methods, such as (Gannot et al., 2001), which varies the adaptation speed based on the input signal itself.

The contribution of (Lathoud et al., 2006a) is twofold, as summarized in the present section. First, an explicit method (Figure 8.2a) is proposed. It relies on a novel input SIR estimate, which itself extends the sector-based detection-localization defined above (5.20). Few works investigated input SIR estimation for non-stationary, wideband signals such as speech. In (Herbordt et al., 2003, 2004), spatial information of the target only is used, represented as a single direction. On the contrary, the proposed approach (1) defines spatial locations in terms of sectors, (2) uses both target's and interference's spatial location information. This is particularly relevant in the car environment, where both driver and co-driver locations are known, but only approximately. In the framework of this thesis, the proposed explicit method illustrates the genericity of the sector-based detection-localization approach proposed in Chapter 5, because it is applied on a different task (Figure 8.1) and different microphone array geometries (Figure 8.3).

The second contribution of (Lathoud et al., 2006a) is an implicit adaptation method (Bourgeois et al., 2005), where the adaptation speed (step-size) is determined from the output signal $z(t)$ (Figure 8.2b), with theoretically-proven robustness to target cancellation issues. Estimation of the input SIR is not needed, and there is no additional computational cost. In the framework of this thesis, the implicit method is used for comparison purposes, thus briefly summarized below.

The rest of this section is organized as follows. Section 8.1.1 defines the two physical setups

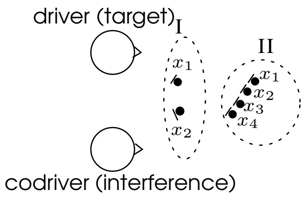


Figure 8.3. Physical setups I (2 mics) and II (4 mics).

as well as the sectors in space. Section 8.1.2 explains how the sector-based detection-localization method can be used to estimate the input SIR, and validates the proposed method with experiments on real in-car data, including in 100 km/h background noise. Section 8.1.3 uses the *estimated* input SIR to explicitly control the step-size of the adaptive filter. Experiments on real in-car data show that the SIR improvement provided by the proposed explicit method is superior to that of a state-of-the-art implicit adaptation control method, including in 100 km/h background noise.

8.1.1 Physical Setups, Recordings and Sector Definition

Two setups are considered for acquisition of the driver's speech in a car. The general problem is to separate the speech of the driver from interferences such as codriver speech.

Physical Setups

Figure 8.3 depicts the two setups, denoted I and II. Setup I has 2 directional microphones on the ceiling, separated by 17 cm. They point to different directions: driver and codriver, respectively. Setup II has 4 directional microphones in the rear-view mirror, placed on the same line with an interval of 5 cm. All of them point towards the driver.

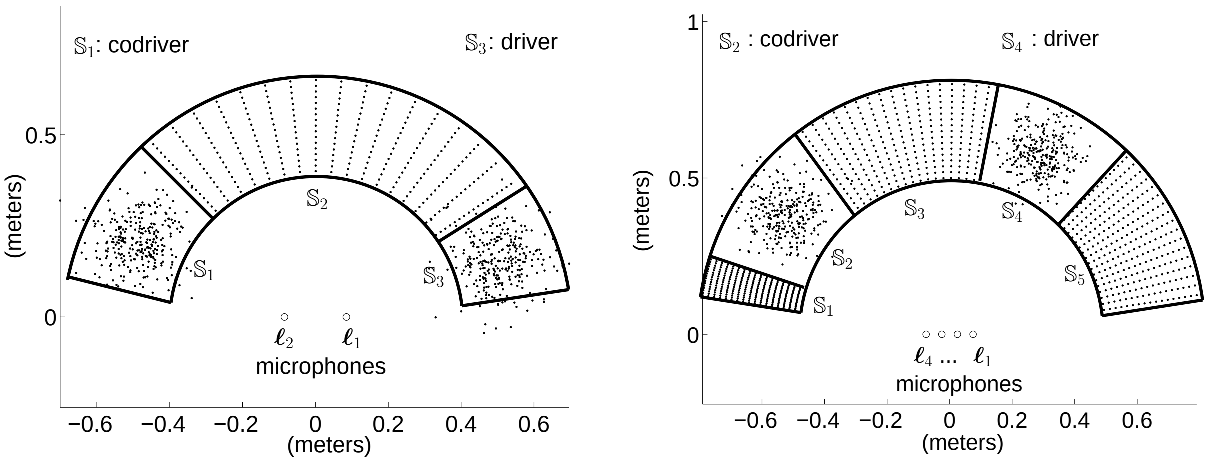
Recordings

Data was not simulated, we opted for real data instead. Three 10-second long recordings sampled at 16 kHz, made in a Mercedes S320 vehicle, are used in the experiments reported below:

`train`: Mannequins playing pre-recorded speech. Parameter values are selected on this data.

`test`: Real human speakers, for testing only: all parameters determined on `train` were "frozen".

`noise`: Both persons silent, the car running at 100 km/h.



(a) Setup I (b) Setup II

Figure 8.4. Sector definition. Each dot corresponds to a $\mathbf{v}_{\check{s},n}$ location, as defined in Section 5.2.1.

For both `train` and `test`, we first recorded the driver, then the codriver, and added the two waveforms. Having separate recordings for driver and codriver permits to compute the *true* input Signal-to-Interference Ratio (SIR) in the microphone signal $x_1(t)$, as the ratio between the instantaneous frame energies of each signal. The true input SIR is the reference for the evaluations presented in Sections 8.1.2 and 8.1.3. The `noise` waveform is then added to repeat speech enhancement experiments in a noisy environment, as reported in Section 8.1.3.

Sector Definition

Figures 8.4a and 8.4b depict the way we defined the sectors for each setup. In both cases, sectors are defined in 2-D space, because the conformation of the array is linear. We used the prior knowledge of the locations of the driver and the codriver with respect to the microphones. The prior distribution $P_{\check{s}}(\mathbf{v})$ (defined in Section 5.2.1) was chosen to be a Gaussian in Euclidean coordinates for the 2 sectors where the people are, and uniform in polar coordinates for the other sectors: $P_{\check{s}}(\mathbf{v}) \propto ||\mathbf{v}||^{-1}$. Each distribution was approximated with $N_{\mathbf{v}}$=400 points. The motivation for using Gaussian distributions is that we know where the people are on average, and we allow a slight motion around the average location. The other sectors have uniform distributions because reverberations may come from any of those directions.

8.1.2 Input SIR Estimation

We describe here a method to estimate the *input* SIR $\mathrm{SIR}_{\mathrm{in}}(t)$, which is the ratio between driver and codriver energies in signal $x_1(t)$ (Figure 8.1). This input SIR estimate relies on the SAM-SPARSE-MEAN sector-based detection-localization (5.20). The proposed input SIR estimate is used by the "explicit" adaptation control method described in Section 8.1.3. As discussed above, the proposed input SIR estimate is a priori well adapted to the car environment, as it uses approximate knowledge of both driver and codriver locations.

Method

For a microphone $\boldsymbol{\ell}_1$, for each time frame t, DFT is applied to the time domain samples $\mathbf{x}_1^{(t)} \in \mathbb{R}^{2N_{\mathrm{F}}}$ to estimate the local spectral representation $\mathbf{X}_1^{(t)} \in \mathbb{C}^{2N_{\mathrm{F}}}$. The energy spectrum for each time frame t is then defined by $E_1^{(t)}(k) = |X_1^{(t)}(k)|^2$, for each discrete frequency k $(1 \le k \le 2N_{\mathrm{F}})$.

In order to estimate the input SIR, we propose to estimate the proportion of the overall frame energy $\sum_k E_1^{(t)}(k)$ that belongs to the driver, and to the codriver, respectively. Then the input SIR is estimated as the ratio between the two. Within the sparsity assumption context of Section 5.2.2, the following two estimates are proposed:

$$\widehat{\mathrm{SIR}}_1 \stackrel{\mathrm{def}}{=} \frac{\sum_f E_1^{(t)}(k) \cdot P\left(\text{sector } \mathbb{S}_{\text{driver}} \text{ active at discrete frequency } k \mid \mathbf{u}^{(t)}(k)\right)}{\sum_f E_1^{(t)}(k) \cdot P\left(\text{sector } \mathbb{S}_{\text{codriver}} \text{ active at discrete frequency } k \mid \mathbf{u}^{(t)}(k)\right)}, \tag{8.1}$$

$$\widehat{\mathrm{SIR}}_2 \stackrel{\mathrm{def}}{=} \frac{\sum_f P\left(\text{sector } \mathbb{S}_{\text{driver}} \text{ active at discrete frequency } k \mid \mathbf{u}^{(t)}(k)\right)}{\sum_f P\left(\text{sector } \mathbb{S}_{\text{codriver}} \text{ active at discrete frequency } k \mid \mathbf{u}^{(t)}(k)\right)}, \tag{8.2}$$

where $\mathbf{u}^{(t)}(k) \in \mathbb{R}^{N_q}$ is the vector of *observed* relative phases between pairs of microphones, at time frame t, as defined in (5.3), and $P\left(\cdot \mid \mathbf{u}^{(t)}(k)\right)$ is the posterior probability given by (5.20). Both $\widehat{\mathrm{SIR}}_1$ and $\widehat{\mathrm{SIR}}_2$ are a ratio between two mathematical expectations over the whole spectrum. $\widehat{\mathrm{SIR}}_1$ weights each frequency with its energy, while $\widehat{\mathrm{SIR}}_2$ weights all frequencies equally. In the case of a speech spectrum, which is wideband but occupies the low frequencies (up to 4 kHz) more often than the high frequencies (above 4 kHz), this means that $\widehat{\mathrm{SIR}}_1$ gives more weights to the low frequencies, while $\widehat{\mathrm{SIR}}_2$ gives equal weights to low and high frequencies. From this point of view, it can be expected that $\widehat{\mathrm{SIR}}_2$ provides better results as long as microphones are close enough to avoid spatial aliasing effects.

Setup	Dynamic range	Method (best on `train`)	Results on `test` (10 seconds) %: RMS error divided by dynamic range (·): correlation true/estimated input SIR	clean	`test+noise`
I	71.6 dB	$\widehat{\mathrm{SIR}}_1$	All frames:	14.0% (0.77)	15.1% (0.73)
			True input SIR > 6 dB:	16.1% (0.25)	17.8% (0.27)
			True input SIR < -6 dB:	12.4% (0.71)	16.3% (0.63)
II	70.2 dB	$\widehat{\mathrm{SIR}}_2$	All frames:	9.3% (0.90)	11.4% (0.84)

Table 8.1. Results on `test` and `test+noise`. Methods and parameters were selected on `train`. The RMS error of the input SIR estimation was calculated in log domain (dB). *Percentages (the lower, the better)* indicate the ratio between the RMS error and the dynamic range of the true input SIR (max - min). *Values in brackets (the higher, the better)* indicate the correlation between the true and the estimated input SIR.

Note that $\widehat{\mathrm{SIR}}_2$ seems less adequate than $\widehat{\mathrm{SIR}}_1$ in theory: it is a ratio of numbers of discrete frequencies, while the quantity to estimate is a ratio of energies. However, in practice it follows the same trend as the input SIR: due to the wideband nature of speech, whenever the target is louder than the interference, there will be more discrete frequencies where it is dominant, and vice-versa. This is supported by experimental evidence in the meeting room domain (Section 5.2.4). To conclude, we can expect a biased relationship between $\widehat{\mathrm{SIR}}_2$ and the true input SIR, that can be compensated with an affine scaling in log domain, as detailed in (Lathoud et al., 2006a). For each setup and for each input SIR estimation method ($\widehat{\mathrm{SIR}}_1$ and $\widehat{\mathrm{SIR}}_2$), the parameters of the scaling are tuned on `train`, then kept fixed and applied on the other recordings `test` and `test+noise`.

Experiments

For each setup I and II, a time plot of the results of the best method is available: Figures 8.5a and 8.5b. The estimate follows the true value very accurately most of the time. Errors happen sometimes when the true input SIR is high. One possible explanation is the directionality of the microphones, which is not exploited by the sector-based detection-localization. Also the sector-based detection-localization gives equal role to all microphones, whereas for setup I we are mostly interested in $x_1(t)$. In spite of these limitations, we can safely state that the obtained SIR curve is very satisfying for triggering the adaptation, as verified in Section 8.1.3.

As it is not sufficient to evaluate results on the same data that was used to select the SIR estimation method and parameters, results on the `test` recording are also reported in Table 8.1 and Figures 8.5c and 8.5d. Overall, all conclusions made on `train` still hold on `test`, which tends to

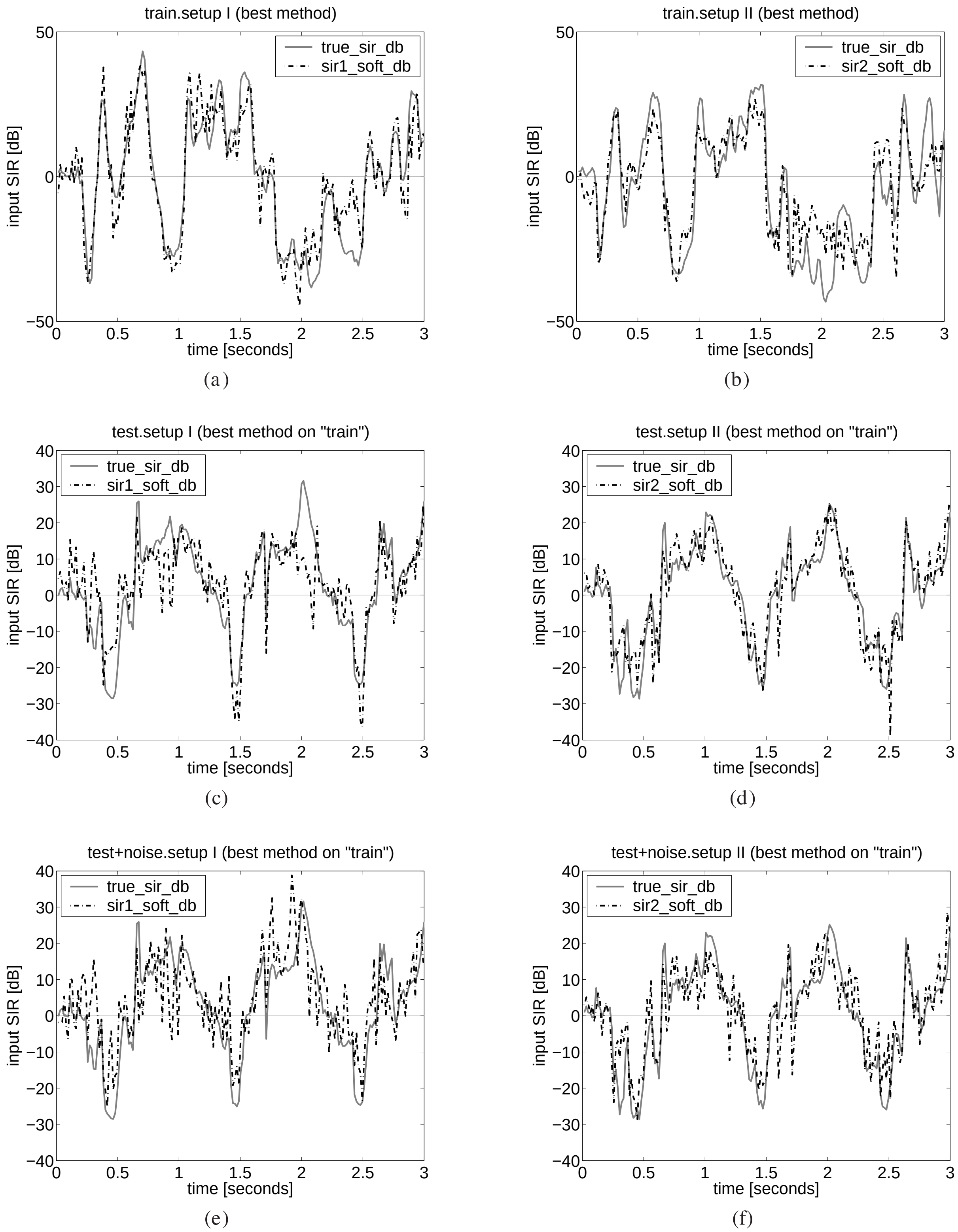


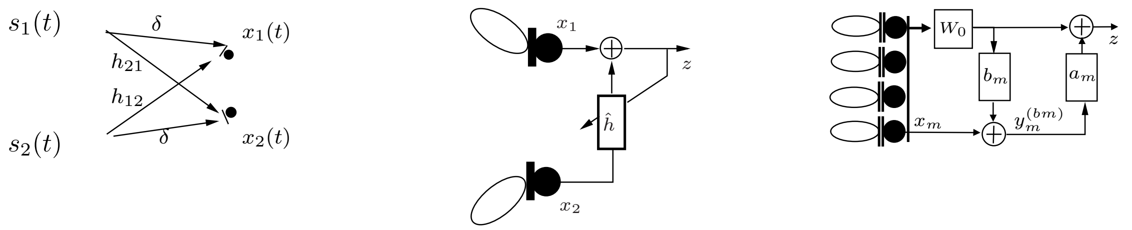
Figure 8.5. Estimation of the input SIR for setups I (left column) and II (right column). Beginning of recordings `train` (top row), `test` (middle row), `test+noise` (bottom row).

prove that the proposed approach is not too dependent on the training data. However, for Setup I, a degradation is observed, mostly on regions with high input SIR, possibly because of the low coherence between the two directional microphones, due to their very different orientations. However, an interference cancellation application with Setup I mostly needs accurate detection of periods of negative input SIR, rather than positive input SIR. On those periods the RMS error is lower (12.4%). Section 8.1.3 confirms the effectiveness of this approach in a speech enhancement application. For Setup II, the results on `test` are quite similar to those on `train`.

Results in 100 km/h noise (`test` + `noise`) are also reported in Table 8.1 and Figures 8.5e and 8.5f. The curves and the relative RMS error values show that the resulting estimate is more noisy, but still follows the true input SIR quite closely in average, and correlation is still high. The estimated ratio still seems accurate enough for adaptation control in noise, as confirmed by Section 8.1.3. This can be contrasted with the fact that the car noise violates the sparsity assumption with respect to speech. A possible explanation is that in both (8.1) and (8.2), numerator and denominator are equally affected, so that the ratio is not biased too much by the presence of noise.

To conclude, the proposed methodology for input SIR estimation gives acceptable results, including in noisy conditions. The estimated input SIR curve follows the true curve accurately enough to detect periods of activity and inactivity of the driver and the codriver[2]. This method is particularly robust since it does not need any thresholding or temporal integration over consecutive frames.

[2]With respect to the speech enhancement application, only one parameter is used (Lathoud et al., 2006a), and the affine scaling in log domain has no impact on the results presented in Section 8.1.3.



(a) Setup I: mixing channels. (b) Setup I: noise canceller. (c) Setup II: GSC.

Figure 8.6. Linear models for the acoustic channels and the adaptive filtering.

8.1.3 Speech Enhancement

Adaptive Interference Cancellation Algorithms

Setup I provides an input SIR of about 6 dB in the driver's microphone signal $x_1(t)$. An estimate of the interference signal is given by $x_2(t)$. Interference removal is attempted with the linear filter $\hat{\mathbf{h}}$ of length L depicted by Figure 8.6b, which is adapted to minimize the output power $\mathrm{E}\left\{z^2(t)\right\}$, using the NLMS algorithm (Widrow and Stearns, 1985) with step size μ:

$$\boxed{\hat{\mathbf{h}}(t+1) = \hat{\mathbf{h}}(t) - \mu \frac{\mathrm{E}\left\{z(t) \cdot \mathbf{x}^{(t)}\right\}}{\left\|\mathbf{x}_2^{(t)}\right\|^2}} \tag{8.3}$$

where $\mathbf{x}_2^{(t)} \in \mathbb{R}^{2N_\mathrm{F}}$ is a time frame of $2N_\mathrm{F}$ samples from microphone ℓ_2, $\hat{\mathbf{h}}(t) = \left[\hat{h}_0(t),\ \hat{h}_1(t),\ \ldots,\ \hat{h}_{L-1}(t)\right]^\mathrm{T}$ is the impulse response of the adaptive filter $\hat{\mathbf{h}}$ at time t, and $\mathrm{E}\{\cdot\}$ denotes expectation, taken over realizations of stochastic processes, as implemented in (Lathoud et al., 2006a).

To prevent instability, adaptation of $\hat{\mathbf{h}}$ must happen only when the interference is active: $\left\|\mathbf{x}_2^{(t)}\right\|^2 \neq 0$, which is assumed true in the rest of this section. In practice, a fixed threshold on the variance of $x_2(t)$ can be used. Moreover, to prevent target cancellation, adaptation of $\hat{\mathbf{h}}$ must happen *only* when the interference is active and dominant.

In setup II, $N_m = 4$ directional microphones are in the rear-view mirror, all pointing at the target. It is therefore not possible to use any of them as an estimate of the codriver interference signal. A suitable approach is the linearly constrained minimum variance beamforming (Griffiths and Jim, 1982) and its robust GSC implementation (Hoshuyama and Sugiyama, 1996). It includes two filters b_m and a_m for each input signal $x_m(t)$, with $m \in \{1, \cdots, N_m\}$, as depicted by Figure 8.6c. Each filter b_m (resp. a_m) is adapted to minimize the output power of $y_m^{(b_m)}(t)$ (resp. $z(t)$), as in (8.3).

To prevent leakage problems, the b_m (resp. a_m) filters must be adapted *only* when the target (resp. interference) is active and dominant.

Implicit and Explicit Adaptation Control

For both setups I and II, an adaptation control is required that slows down or stops the adaptation according to target and interference activity. Two methods are proposed: "implicit" and "explicit". The implicit method introduces a continuous, adaptive step-size $\mu(t)$, whereas the explicit method relies on a binary decision, whether to adapt or not. The reader is referred to (Lathoud et al., 2006a) for the implementation details of both methods.

• **Implicit Method:** We briefly present the method for Setup I. The implicit method also applies to Setup II, as described in (Lathoud et al., 2006a). The goal is to increase the adaptation step-size whenever possible, while not turning (8.3) into an unstable, divergent process. With respect to existing implicit approaches, the novelty is a well-grounded mechanism to prevent instability while using the filtered output. Based on analyses conducted in (Mader et al., 2000) and (Widrow and Stearns, 1985), the theoretical analysis conducted in (Lathoud et al., 2006a) justifies the following "implicit" adaptive step-size control. In (8.3), the constant step-size μ is replaced with a variable, time-dependent step-size $\mu(t)$, as follows:

$$\boxed{\begin{array}{l} \bullet\ \ \mu(t) = \begin{cases} \mu_{\text{impl}}(t) & \text{if} \quad \mu_{\text{impl}}(t) < 2 \quad \text{(stable case)} \\ \mu_0 & \text{otherwise} \qquad\qquad \text{(unstable case)}, \end{cases} \\ \bullet\ \ 0 < \mu_0 \ll 1 \ \text{ is a small constant.} \\ \bullet\ \ \mu_{\text{impl}}(t) \quad \stackrel{\text{def}}{=} \quad \mu_0 \frac{\left\| \mathbf{x}_2^{(t)} \right\|^2}{\left\| \mathbf{z}^{(t)} \right\|^2} \end{array}} \tag{8.4}$$

This effectively reduces the step-size when the current target power estimate is large ($\left\|\mathbf{z}^{(t)}\right\|^2 >> \left\|\mathbf{x}_2^{(t)}\right\|^2$) and conversely it adapts faster in absence of the target ($\left\|\mathbf{z}^{(t)}\right\|^2 << \left\|\mathbf{x}_2^{(t)}\right\|^2$). This was theoretically proved in (Lathoud et al., 2006a), in the case where sources s_1 and s_2 are assumed to be uncorrelated, blockwise stationary white sources. A further theoretical stability analysis of the implicit control method is available in (Bourgeois et al., 2005).

• **Explicit method:** For both setups I and II, the sector-based method described in Section 8.1.2 is used to directly estimate the input SIR at $x_1(t)$. Two thresholds are set to detect when the target (resp. the interference) is dominant, which determines whether or not the fixed step-size adaptation of (8.3) should be applied. The threshold values are tuned on `train`, then kept fixed for performance evaluation on `test` and `test+noise`.

Performance Evaluation

For both setups I and II, we measured the instantaneous SIR improvement on the real 16kHz recordings, with respect to the output when no adaptation is performed. Thus, the reference in Setup I is the true input SIR at microphone x_1, and the reference in Setup II is the SIR at the output of the delay-and-sum beamformer W_0. "Instantaneous" means on half-overlapping, short time-frames – where speech can be safely considered as stationary. We used 32 ms-long time-frames. Section 8.1.1 describes the recordings and the method of computation of the true input SIR.

Five seconds of the `train` recording were used to tune all parameters. Then the entire `test` recording (real human speakers, 10 seconds) was used to test the methods. It contains a significant degree of overlap between the two speakers (56% of speech frames). Three adaptation control methods are tested:

1. **No control (baseline):** Adaptation (8.3) with a fixed step-size μ, on *all* time frames (including silences, noise and overlaps).

2. **Implicit method:** Adaptation (8.4) with a variable, automatically controlled step-size $\mu(t)$.

3. **Explicit method:** Same as "No control", except that the adaptation is applied *only* on frames with an estimated input SIR above a threshold. As detailed in Section 8.1.2, the input SIR is estimated based on the sector-based detection-localization introduced in Chapter 5.

For each method, based on the instantaneous SIR improvement, the segmental SIR improvement[3] is computed in three cases: whether the true input SIR is low, close to 0 dB or high.

[3] "Segmental" means that only frames containing speech from either driver or codriver or both are considered, as detailed in (Lathoud et al., 2006a).

Below is a description of the 3 cases that were evaluated:

1. True input SIR $<$ -6 dB: when the energy of the codriver is dominant in signal x_1. This quantifies how much of the interference signal is canceled during silences of the driver: a significantly positive value. All three methods can be expected to perform well in this case.

2. True input SIR in [-6 +6] dB: when both driver and codriver are comparatively active. This quantifies how much of the interference signal is cancel-led during overlap periods (both persons speaking): a positive value. We can expect a slight degradation in the case of the "No control" baseline method, because of leakage issues.

3. True input SIR $>$ +6 dB: when the energy of the driver is dominant in signal x_1. No improvement is expected here: a value around zero. If this value is markedly negative, it means that a given method is suffering from target cancellation issues – as expected for "No control".

Experiments: clean data

The first 3 seconds of `test` are depicted by Figure 8.7a. The periods where the SIR improvement is consistently close to 0 dB correspond to silences of both speakers. Average SIR improvement over the entire recording is given in Table 8.2a. The result of the "no control" baseline method highlights the target cancellation problem and confirms the necessity of adaptation control. In both setups, both "implicit" and "explicit" methods are robust against this problem, and the explicit method provides the best results. Overall, all above-mentioned expectations are verified.

Experiments with 100 km/h noise

The same experiments were conducted again, after adding the background road noise waveform `noise`. The resulting wave files have an average segmental SNR of 11.6 dB in setup I, and 9.6 dB in setup II. The goal of this experiment is to determine whether the proposed approaches can cope with background noise. It is not obvious, since they do not explicitly model background noise, which may be incoherent, or localized outside of the defined sectors. To that purpose the adaptation step μ_0 was reduced, while keeping all other parameters unchanged (Lathoud et al., 2006a).

The result is given in Figure 8.7b and Table 8.2b. The behaviour in terms of SIR improvement, both over time and in average, is very similar to the clean case. Thus, we can state that both implicit

	Setup I (2 mics, reference: x_1)			Setup II (4 mics, reference: W_0)		
Range of the true input SIR	No control (baseline)	Implicit	Explicit	No control (baseline)	Implicit	Explicit
$<$ -6: (codriver)	6.5	5.9	10.7	10.4	6.1	10.5
$[-6, +6]$: (both)	-0.6	1.2	5.8	0.6	2.3	3.3
$>$ +6: (driver)	-7.7	-0.2	2.6	-10.0	0.0	-0.8

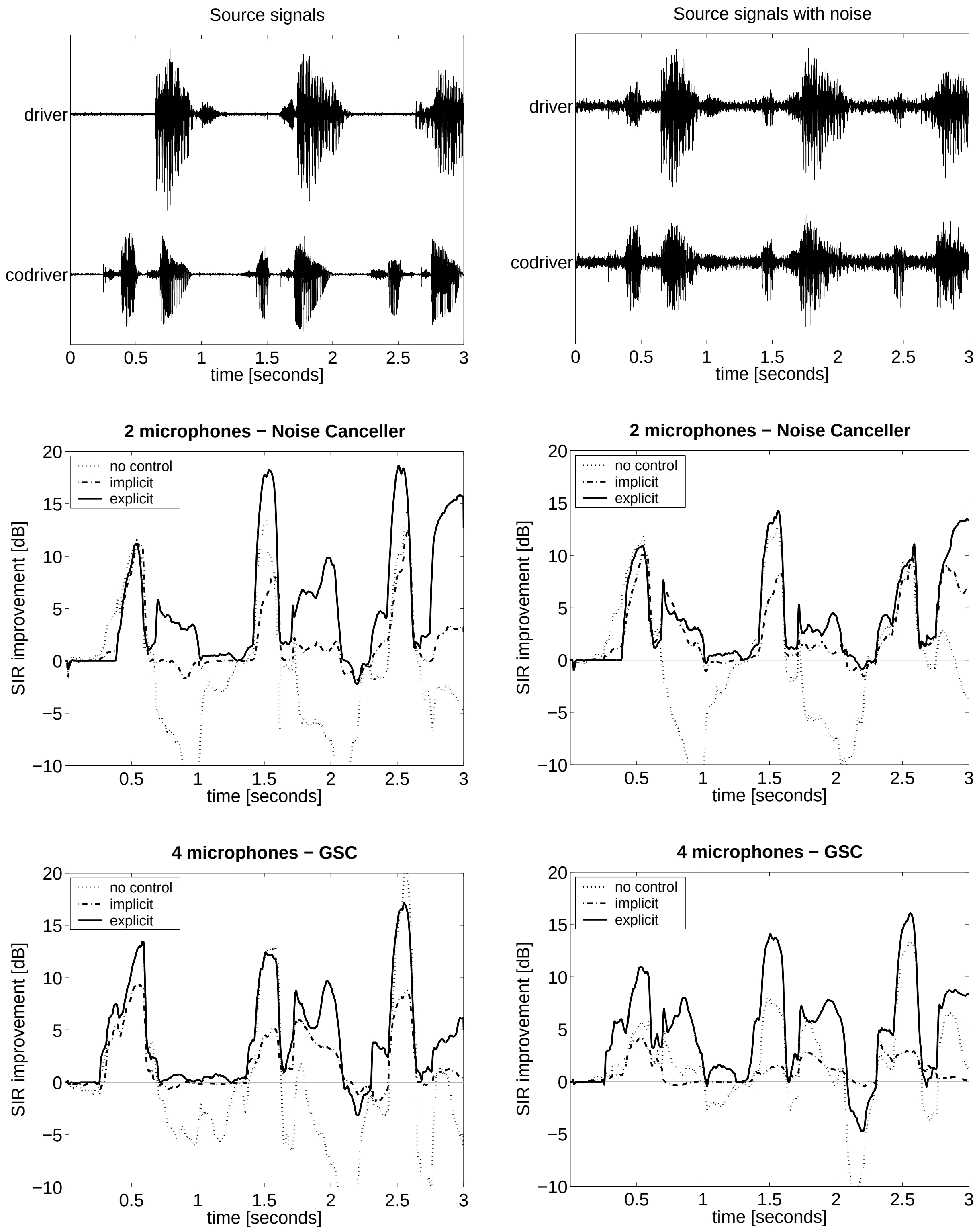
(a) `test` (clean data)

	Setup I (2 mics, reference: x_1)			Setup II (4 mics, reference: W_0)		
Range of the true input SIR	No control (baseline)	Implicit	Explicit	No control (baseline)	Implicit	Explicit
$<$ -6: (codriver)	6.4	7.1	7.4	7.9	3.8	10.3
$[-6, +6]$: (both)	1.0	2.7	3.5	1.2	1.6	3.2
$>$ +6: (driver)	-4.7	0.4	1.9	-6.3	0.2	-2.4

(b) `test+noise`

Table 8.2. Average segmental SIR improvement in dB. In Setup I, the reference is the output x_1 of microphone $\boldsymbol{\ell}_1$. In Setup II, the reference is the output of the delay-sum W_0. (W_0 brings a SIR improvement over x_1 of 0.1, 1.6, 2.2 dB respectively in the "codriver", "both" and "driver" cases.)

and explicit approaches also work in a realistic case of a moving car. The only negative result is "explicit" in the "setup II, driver" case, although it is still no degradation compared to the input SIR of $x_1(t)$. For both setups I and II, the best results on "codriver" and "both" are given by the "explicit" method. This is interesting, given that the thresholds of the "explicit" method were not changed.



(a) Clean conditions.

(b) Noisy conditions.

Figure 8.7. Improvement over input SIR (100 ms moving average, first 3 seconds shown). Column (a) shows results on clean data (`test`), whereas column (b) shows results on noisy data (`test+noise`: 100km/h background road noise).

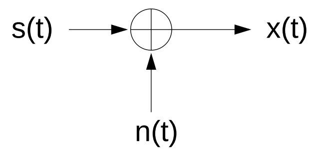
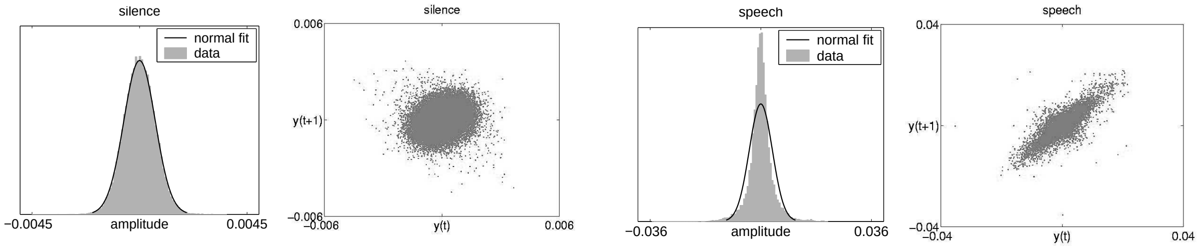


Figure 8.8. Model of the problem: recognize speech from the observed signal $x(t) = s(t) + n(t)$, where $s(t)$ is the clean speech signal and $n(t)$ is the additive acoustic noise signal.

8.2 Noise-Robust ASR: Unsupervised Spectral Subtraction

Let us now assume that we have a *single* channel noisy speech signal $x(t)$. Application domains such as in-car human-machine interaction require noise-robust front-ends in order to cope with the noisy situations encountered in practice, as depicted in Figure 8.8. It is thus desirable to estimate and remove the time-varying ambient noise ($n(t)$ in Figure 8.8), in an online manner. In classical spectral subtraction (Boll, 1979; Berouti et al., 1979), the noise energy is estimated, then subtracted from the speech energy, using an independence assumption. This 2-step approach requires tuning parameters (Berouti et al., 1979). We propose an alternative, single-step approach that *jointly* estimates both speech and noise powers, called Unsupervised Spectral Subtraction (USS). The purpose of the present section is to highlight the versatility of the *joint* modelling proposed in Appendix C.1. In Appendix C.1, a two-component model was used to model sector-based activeness (5.22), thus permitting to classify each sector of space as active or silent. In the present section, a very similar two-component model is used to model the single-channel magnitude spectrogram, thus permitting to determine the noise level for the subsequent spectral subtraction. Experiments on the Aurora 2 setup (Hirsch and Pearce, 2000) show that with only two equations and no parameter tuning, the ASR results are very similar to those of the ETSI Advanced Front-End (AFE) (ETSI, 2003a).

The present section only reports what is related to two-component joint modelling, and its use in USS. Although used in the experiments, the cellphone CHannel Normalization (CHN) is out of scope of this thesis, thus voluntarily de-emphasized. Full details on both USS and CHN can be found in (Lathoud et al., 2005b, 2006b). The rest of the present section is organized as follows. Section 8.2.1 derives a 2-component mixture model from observations on real magnitude spectrograms, (similarly to Appendix C.1). Section 8.2.2 uses the 2-component mixture model to propose a tuning-free USS method. Section 8.2.3 reports noise-robust ASR experiments on the Aurora 2 setup. USS is also used somewhere else in this thesis, to define the sector-based MFCCs (Section 5.5.1).



(a) (b) (c) (d)

Figure 8.9. Observations on real meeting room data (`seq01` in the AV16.3 Corpus, Chapter 4) of a pre-emphasized waveform $y(t) \stackrel{\text{def}}{=} x(t) - 0.97 \cdot x(t-1)$. (a),(c): histograms, (b),(d): phase plots.

This work was done in collaboration with Dr Mathew Magimai.-Doss, Bertrand Mesot and Prof. Hervé Bourlard. Two demos are available on the Internet, with all necessary Matlab code for Unsupervised Spectral Subtraction (USS): `http://mmm.idiap.ch/Lathoud/USS-EXAMPLE` and for CHannel Normalization (CHN): `http://mmm.idiap.ch/Lathoud/2006-CHN-USS`

8.2.1 Proposed 2-Component Mixture Model

In this subsection, the commonly used Rayleigh silence model is justified on real data, and completed with an ad-hoc "activity" model. The main difference with existing, related models such as in (Ephraim and Malah, 1984; Martin and Breithaupt, 2003; Gemello et al., 2004), is that we do not address the complete probabilistic modeling of speech activity, but limit ourselves to large magnitudes only.

Observations on Real Waveforms

Simple observations on silence periods of a pre-emphasized waveform $y(t) \stackrel{\text{def}}{=} x(t) - 0.97 \cdot x(t-1)$ and its covariance matrix, as partially illustrated by Figures 8.9a and 8.9b, show that it is very reasonable to model $\{\underline{y}(t)\}$ as a series of i.i.d, zero-centered Gaussian processes. Under such assumption, the real and imaginary part of the DFT are independent Gaussian distributed variables, as shown in Appendix F. (Note that this derivation is exact and does not rely on asymptotical considerations such as the central limit theorem.) Thus, the magnitude r.v. $\underline{M}^{(t)}(k)$ has a Rayleigh pdf (Rice, 1944, 1945). This type of assumption is used in several existing works (Martin and Breithaupt, 2003; Chen and Loizou, 2005).

On the other hand, speech waveforms are clearly *not* Gaussian distributed, and *not* i.i.d., as shown by Figure 8.9c and 8.9d. As mentioned previously, finding a fully-justified pdf for speech magnitude is still an open research subject. Hence, as shown below, we propose to model large magnitudes of speech only.

Proposed Mixture Model

The proposed pdf for $\underline{M}$ is $f(M) \stackrel{\text{def}}{=} P_{\text{I}} \cdot f_{\text{I}}(M) + P_{\text{A}} \cdot f_{\text{A}}(M)$, where P_{I} and P_{A} are the priors for "silence" and "activity", respectively. Although the model is independent of k and t, slowly time-varying ambient noise can be accommodated through blockwise-processing of the magnitude spectrogram.

f_{I} is the Rayleigh pdf of parameter σ_{I} (Section 2.4) and f_{A} is a pdf that models magnitudes $M > \Psi_M$, where Ψ_M is a threshold defined with respect to f_{I}. Formally, we impose that:

$$\forall M \leq \Psi_M \quad f_{\text{A}}(M) = 0 \tag{8.5}$$

As a starting point, we use $\Psi_M = \sigma_{\text{I}}$, which is the mode of the Rayleigh pdf. The reasoning is that values below the mode of the Rayleigh f_{I} can safely be assumed to be background noise.

Moreover, we constrain f_{A} to fulfill two practical constraints. First, the derivative $f'_{\text{A}}(M)$ of the chosen "activity" pdf should not be zero when M is just above Ψ_M, otherwise the threshold Ψ_M will loose its meaning, as it may be set to an arbitrarily low value. Second, the decay of $f_{\text{A}}(M)$ when M tends towards infinity should be lower than the decay of the Rayleigh, in order to make sure that f_{A} will capture data with large magnitudes, and not f_{I}. A pdf that fulfills the two criteria above is a "shifted Erlang" pdf with h=2 (the Erlang pdf belongs to the Gamma family (Grinstead and Snell, 1997)):

$$f_{\text{A}}(M) \quad \stackrel{\text{def}}{=} \quad \mathbf{1}_{M>\sigma_{\text{I}}} \cdot \lambda_{\text{A}}^2 \cdot (M - \sigma_{\text{I}}) \cdot e^{-\lambda_{\text{A}}(M-\sigma_{\text{I}})} \tag{8.6}$$

where $\mathbf{1}_{M>\sigma_{\text{I}}}$ is equal to 1 if $M > \sigma_{\text{I}}$, and zero otherwise. Note the implicit stationarity assumption: the 4 parameters $\mathbf{\Lambda} = \{P_{\text{I}}, \sigma_{\text{I}}, P_{\text{A}}, \lambda_{\text{A}}\}$ are assumed to be independent of t. Furthermore, independence of k is also assumed; it is justified by the pre-emphasis, which whitens the spectrum.

EM training of $\mathbf{\Lambda}$ (Dempster et al., 1977): Both "E" and "M" steps involve simple mathematical expressions. In the "E" step, posteriors can be estimated as follows:

$$P\left(\text{sil}\middle|M^{(t)}(k),\,\mathbf{\Lambda}\right) = \frac{P_{\text{I}}\cdot f_{\text{I}}(M^{(t)}(k))}{P_{\text{I}}\cdot f_{\text{I}}(M^{(t)}(k))+P_{\text{A}}\cdot f_{\text{A}}(M^{(t)}(k))} \quad (8.7)$$

$$P\left(\text{act}\middle|M^{(t)}(k),\,\mathbf{\Lambda}\right) = 1-P\left(\text{sil}\middle|M^{(t)}(k),\,\mathbf{\Lambda}\right) \quad (8.8)$$

In the "M" step, exact maximization of the likelihood is difficult, so a moment-based approximation is used to update the $(\sigma_{\text{I}}, \lambda_{\text{A}})$ parameters (Lathoud et al., 2005b, 2006b).

Data representation: Similarly to Appendix C.1, one spectrogram is reduced to only 100 representative samples through a deterministic sampling method. Therefore the cost of the EM fitting is very small. An example of fit on one file taken from the OGI Numbers 95 database (Cole et al., 1994) can be seen in Figures 8.10a, 8.10b and 8.10c.

8.2.2 Application to Unsupervised Spectral Subtraction (USS)

A 2-step approach is used:

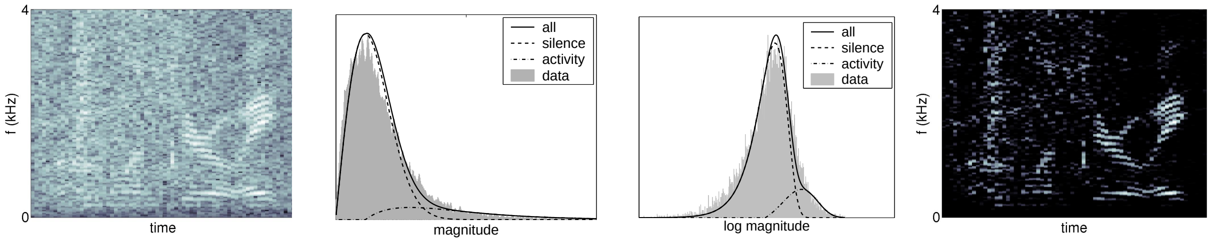
1. EM fitting of the 2-mixture model, as described in Section 8.2.1.

2. Spectral subtraction using the parameter σ_{I} as a floor:

$$M^{(t)}_{\text{USS}}(k) \stackrel{\text{def}}{=} \max\left(1, \frac{M^{(t)}(k)}{\sigma_{\text{I}}}\right) \quad (8.9)$$

Note that the flooring to a non-zero value ($\max(1,\ldots)$) is necessary in the MFCC context. Indeed, leaving zero magnitude values after spectral subtraction would lead to undesirable dynamics in cepstral coefficients. An example of result of (8.9) is shown in Figure 8.10d.

This approach can be compared to previous works. We can note common points "in spirit" with (Van Compernolle, 1989) and to a lesser extent with (Cohen, 1989): adaptation to non-stationary noises is possible both in (Van Compernolle, 1989) and in our approach through block-wise processing. Moreover, on the modeling side, parameters of the "noise" distribution (i.e. silence) are more important than those of the "speech" distribution (i.e. activity) in both approaches, as they are used to floor the magnitude values ($\max$ operator in our approach).



(a) Original spectrogram $M^{(t)}(k)$ | (b) Fit (shown in linear domain) | (c) Fit (shown in log domain) | (d) Spectrogram after subtraction $M_{\mathrm{USS}}^{(t)}(k)$

Figure 8.10. Example of fit of the 2-mixture model on noisy data taken from the OGI Numbers 95 database (Factory 0dB condition). All plots show magnitude data in the frequency domain. On spectrogram plots (a) and (d), the largest magnitudes are white, the smallest magnitudes are black. $f = \frac{k-1}{N_{\mathrm{F}}} \cdot \frac{f_s}{2}$ and $f_s = 8\ kHz$.

However, here the modeling is made directly at the magnitude spectrogram level, with a single model for all frequencies, while in (Van Compernolle, 1989) modeling was made after mel filterbank computation, and a model was defined for each critical band separately. Moreover, the approach proposed here is one-pass, fully unsupervised, without any “feedback” loop, without any tunable threshold, without histogram. In (Van Compernolle, 1989), multiple stages are involved, including histograms, band-specific parameters, short-term and long-term adaptation, a “feedback” loop, tunable parameters that have to be trained and (optionally) injection of an artificial white noise.

Finally, note that all posterior-based filtering approaches previously proposed in (Lathoud et al., 2005a) yielded inferior ASR results, as compared to the simple spectral subtraction approach of (8.9). A consequence is that posterior computation for spectrogram filtering is not necessary anymore, only comparing magnitude values $M^{(t)}(k)$ to the σ_{I} parameter is needed, as in (8.9). The computational cost is therefore reduced.

8.2.3 Noise-Robust ASR Experiments

The Aurora 2 task was designed to evaluate the front-end of ASR systems in noisy conditions (Hirsch and Pearce, 2000). The task is speaker-independent connected digit recognition. The database comprises isolated digits and sequences of up to 7 digits from the TIDigits database (Leonard, 2004) spoken by male and female US-American adults. The original 20 kHz data was downsampled to 8 kHz, in order to obtain a telephone bandpass between 0 and 4 kHz. The resulting data constitutes the clean speech data (clean condition). Noises were then added artificially at different SNR levels

SNR	-5	0	5	10	15	20	clean
MFCC baseline	91.6	85.2	65.6	38.6	17.2	6.1	0.9
CHN-USS	70.0	38.2	16.2	7.1	3.5	2.1	1.1
AFE	69.9	38.1	15.8	7.0	3.4	1.9	0.8

Table 8.3. Word Error Rate results on Aurora 2 (the lower, the better), per SNR level, averaged on the three noisy test sets A, B and C. Training is done on clean signals.

(20 dB to -5 dB). The noises were recorded at different places: suburban train, crowd of people (babble), car, exhibition hall, restaurant, street, airport, and train station. Some noises are fairly stationary, for instance car noise and exhibition noise. Others contain non-stationary segments, as in street noise, babble noise and airport noise. Three test sets A, B, C were built by adding the various noise conditions to the original clean signals (Hirsch and Pearce, 2000). In this section we report averages across the three sets. Individual results can be found in (Lathoud et al., 2006b).

We ran three different feature extractor on the Aurora 2 task:

- MFCC: the baseline ETSI Mel-Cepstrum Front-End (ETSI, 2003b).
- CHN-USS: CHannel Normalization to accommodate various cellphone channels (Lathoud et al., 2006b, Section 4), followed by USS (8.9). Overall, only two equations are needed, as explained in (Lathoud et al., 2006b), and no parameter tuning is involved. We used 1-second blocks to accommodate time-varying background noises.
- AFE: the ETSI Advanced Front-End (AFE) for noise-robust ASR (ETSI, 2003a). The ETSI AFE includes many steps and parameters.

In all three cases, the frame length was 25 ms and the frame shift was 10 ms[4], and the Aurora 2 (Hirsch and Pearce, 2000) training/testing setup was used. Moreover, in all experiments, training was always done using clean data only, because the task is precisely to remove as much noise from the noisy signal as possible. Next, performance evaluation was done on all available noise conditions, and we report the average across the three test sets A, B and C in Table 8.3. It can be seen that the proposed CHN-USS provides results very close to those of the ETSI AFE. CHN-USS can thus be seen as a drastic simplification of the ETSI AFE, and (Lathoud et al., 2006b) suggests that a further improvement can be obtained by merging the two methods.

[4]Further details on feature post-processing can be found in (Lathoud et al., 2006b).

With respect to the present thesis, the CHN-USS results validate the 2-component, joint modelling of background noise and large magnitudes in the noisy speech spectrogram. This shows that the *type* of model used in Appendix C.1 for sector-based activeness, can effectively be used in a very different context.

8.3 Conclusion

The purpose of the present chapter was to highlight the versatility of some of the methods proposed in this thesis. At the signal processing level, Section 8.1 showed that the SAM-SPARSE-MEAN sector-based detection-localization previously used for speaker detection-localization in real meeting room recordings made with a UCA, is also effective for hands-free speech enhancement on real in-car recordings made with a ULA. At the probabilistic modelling level, Section 8.2 showed that the type of 2-component mixture model used to model sector-based activeness, can be used on a very different task: speech+noise modelling of a single-channel, noisy speech magnitude spectrogram. In each application, as opposed to a sequential, 2-step implementation, we focussed on a *joint* implementation:

- Detection *and* localization in terms of sectors (Section 8.1).
- Speech *and* noise modelling in terms of magnitude (Section 8.2).

Both aspects are used to define the sector-based MFCC (Section 5.5.1).

Chapter 9

Conclusion

The subject of this thesis was to determine who spoke where and when, in multi-party spontaneous speech. A wide context of applications motivated this choice: automatic summarization of meetings, surveillance, smart offices and homes, autonomous robots, etc. Means were thus restricted to be non-invasive, that is using distant microphones only. This choice can be opposed to very efficient but invasive methods that require each speaker to wear a close-talking microphone (lapels), as in (Wrigley et al., 2005). Solutions were aimed to be as unsupervised as possible, that is ideally (1) devoid of large training data sets and/or manual threshold tuning procedures that are difficult for non-technical users, (2) adaptive to varying conditions: for example single or multiple speakers, static or moving speakers.

The methodology adopted in this thesis was to build an integrated system in a bottom-to-top approach, starting with "signal-level" tasks such as detection and localization of multiple concurrent speakers from a single time frame, and finishing with long-term speaker clustering, resulting in a high-level annotation in terms of speaker identities and locations. We argue that addressing low-level "signal" issues can strongly benefit the performance of high-level integrated systems, and that the analysis of the high-level output can lead to reveal interesting low-level "signal" issues.

Building such a system led to address several research issues, as summarized below. The central concept is that speaker location information is particularly well adapted to the spontaneous multi-party speech context involved here. Some of the proposed methods – especially concerning detection – are generic and can be applied to other tasks, as illustrated in Chapter 8.

9.1 Data Acquisition

At least one corpus of spontaneous speech with static speakers was already available (McCowan et al., 2005), that was precisely annotated along the time axis, in terms of speaker turns. On the contrary, we could not find a publicly available corpus of audio data that would include *precise 3-D annotation* of the mouths'locations. This is necessary to evaluate the instantaneous detection and localization of multiple speakers. A first contribution of this thesis is thus the AV16.3 Corpus, that contains both static and dynamic scenarios, with multiple (moving) speakers occupying a variety of locations in an indoor environment. The preference was given to real human speakers, since loudspeakers do not necessarily reflect humans' speech radiation characteristics, especially non-linearities (Schwetz et al., 2004). A strong emphasis was given to overlapped speech and non-linear motions. Three calibrated cameras were used to estimate the ground-truth 3-D mouth location of each speaker, with an error inferior to 1.2 cm. To the best of our knowledge, the AV16.3 Corpus is the first publicly available audio-visual corpus that includes precise 3-D mouth location annotation.

9.2 Multispeaker Detection-Localization

From an instantaneous time frame (20 to 30 ms) of a recording made with multiple microphones synchronized in time, and the knowledge of the relative microphones locations, the task is to detect *and* locate multiple simultaneous speakers, as often encountered in spontaneous speech (Shriberg et al., 2001). We selected the framework of Steered Response Power approaches, were the location(s) with maximum beamforming output are located in the space around the microphone array (Krim and Viberg, 1996; DiBiase, 2000). A literature review showed that being able to quickly detect which region(s) of space contain active speakers would greatly speedup the search, which is otherwise prohibitive (entire 3-D space). The key idea of this part of the thesis is that detection for localization has specific requirements, which differ from speech detection for Automatic Speech Recognition, where a threshold on energy-based features can be used. In particular, we investigated whether *joint* detection-localization could be beneficial, as opposed to detection *followed by* localization. The main conclusions of this part are as follows.

- An existing beamforming method for localization, called SRP-PHAT (DiBiase, 2000), was shown to be strictly equivalent to a Phase Domain Metric (PDM). The PDM simply compares frequency-domain phase information between multiple pairs of microphones, *with all the properties of a standard metric*. Based on the PDM, the search space for subsequent multisource localization was greatly reduced, through a virtually costless evaluation of the *average* acoustic activeness in a sector of space. A major improvement was shown, over an existing approach that compares beamforming values at a few *points* in space. Sector-based activeness was also used on a different task (in Chapter 8), to control the adaptation of speech enhancement filters in hands-free, in-car speech acquisition.

- For the sector-based detection-localization step, two-component probabilistic models were proposed that model the average sector activeness with one component for background noise, and another component for audio source activity. They are fitted in a fully unsupervised manner, therefore training data is not necessary. This implicitly permits to adapt the *parameters* of a model to varying conditions. Such an approach is also used in Chapter 8, to separate speech from noise on a completely different task (noise-robust ASR).

- Adapting the parameters is not sufficient, when the chosen *structure* of the probabilistic model (which types of pdfs, how they are tied) is not capable to properly fit the observed data. A generic correction procedure was proposed, where the final detection threshold is selected using not only the fitted model *but also the data again, through posterior probabilities of silence and activity*. Experiments on the sector-based detection-localization task show that the proposed correction procedure yields a dramatic improvement over classical training/testing approaches. This correction procedure can be applied to any other detection task, as long as a probabilistic model is used. Theoretical investigations show that the correction procedure can also be applied to multiclass classification tasks.

- Scaled Conjugate Gradient descent (Moller, 1993) permits fast localization of the individual acoustic sources. The SCG finds local minima of the PDM, within each active sector of space.

- A location-dependent way of extracting MFCCs is proposed, that permits to separate acoustic activity into speech and non-speech. It separates slightly correlated (speech) MFCCs from uncorrelated (noise) MFCCs through full-covariance GMM modelling.

To summarize, this part investigated static analysis, which means analysis of each time frame separately. The result is zero, one or more *audio source* location estimates, for each time frame. Each audio source location estimate is then classified as speech or non-speech, using location-dependent MFCCs.

9.3 Short-Term Clustering

This part investigates dynamical analysis of the location estimates, that is across several time frames. The applicative target is the speech segmentation task in meetings, where a sequence of silence and speech segments has to be marked along the time axis, for each speaker. The main conclusions of this part are as follows.

- Location information can be used to segment speech from each speaker with distant microphones, with a performance comparable to that of close-talking microphones, and a major improvement on overlapped speech.
- To take the Speech/Non-Speech (SNS) decision for each location estimate independently leads to speech segmentation results that are very sensitive to post-processing parameters. In practice, post-processing means grouping the “speech” segments together, and eliminating the small ones.
- On the other hand, tracking a speech source over time and space is difficult, due to the sporadic nature of speech: utterances are short, discontinuous, and interspersed with long silences. We thus proposed an intermediary approach called “short-term clustering” of location estimates, where the location estimates that are close in space *and* time are grouped into clusters. To take the SNS decision in terms of short-term clusters leads to speech segmentation results that are (1) better than with individual SNS decisions, (2) much less sensitive to post-processing parameters.
- Short-term clustering was realized with a novel, principled probabilistic framework, which is threshold-free. It thus requires neither training data nor parameter tuning. A fully deterministic implementation is proposed that processes each recording in an online manner. Short-term clustering can also be used for threshold-free detection of trajectory crossings.

To summarize this part, the integration of the research solutions proposed so far permitted to develop an efficient, non-invasive multispeaker segmentation system. It is particularly efficient on overlapped speech. The result is an answer to the two questions: "where? when?", as a set of speech segments, each segment being defined as a piece of trajectory in *both* time and space.

9.4 Speaker Clustering with Distant Microphones

Based on the previous parts, the remaining task is to determine the identity of the speaker, for each speech segment. More specifically, we address the unsupervised speaker clustering task, where no enrollment data is available. The goal is twofold: (1) to produce a set of labels, with ideally one label per speaker, and (2) to give the correct label to each speech segment. Both goals are met when NIST's Diarization Error Rate (DER) reaches 0%. The main conclusions of this part are as follows.

- The fast-changing speaker turns encountered in spontaneous multi-party speech are difficult to model with the GMM/HMM methods that were developed for broadcast news speech, typically through maximization of the Bayesian Information Criterion (BIC). Indeed, GMMs require each speech segment to have a sufficient length (2-3 seconds), which is often not the case in spontaneous multi-party speech.
- We proposed to complement the MFCC information with location information, in order to accommodate the fast speaker turns. An extension of the BIC was proposed, that fuses information from multiple modalities. A major reduction of the DER was obtained on the M4 Corpus, as compared to a state-of-the-art approach that uses MFCCs alone.
- A closer analysis of the results highlights one success and one failure. First, the success: the proposed multimodal BIC criterion effectively led to give the same label to speech segments from the same speaker, but seated at different places around the microphone array. Second, the failure: when the speaker would stand up and move away to do a presentation, we *always* ended up with two clusters for the same speaker. One cluster contained speech spoken at the seated location, the other cluster contained speech spoken at the distant, standing location.
- Standard linear dereverberation/denoising methods, often used in ASR, did not yield *any* improvement with respect to this issue. Further research is needed to explain the underlying feature variability, possibly extending the study presented in (Schwetz et al., 2004).

- Initial investigations were conducted on audio-visual unsupervised calibration, as a possible alternative to audio-only speaker clustering. An extension of Stauffer's adaptive background subtraction (Stauffer and Grimson, 2000) was proposed to model geometrical ties between non-colocated audio and video sensors.

To summarize this part, an effective speaker clustering scheme was proposed, that uses distant microphones only. The result identifies who spoke where and when. The key concept was to merge MFCC and location information, by means of a multimodal BIC. A clear advantage is obtained over MFCC-only speaker clustering. Analysis of the failures revealed that fundamental signal processing issues need to be addressed to explain the distance-dependent MFCC feature variability.

9.5 Self-Criticism and Future Directions

The various parts of this work suggest several lines of work. One can build new systems on top of the successful parts, but also research the causes of the failures, and *propose* solutions. In my opinion, the failures are in fact quite interesting, because searching for their explanation(s) often leads to new, powerful signal processing methods, that are invariant to a larger set of undesired variabilities in the data.

Speech/Non-Speech classification: We briefly mentioned that the correlation that remains in MFCCs extracted from human speech, in spite of DCT, can be opposed to the uncorrelated MFCCs extracted from fairly stationary machine noises. The main interest is that training data is not used at all. However, we only tested this approach on the M4 Corpus (static speakers) and the AV16.3 Corpus (moving speakers). Although the two environments differ in terms of noise sources, objects in the room and speaker locations/behaviors, the same room (Moore, 2002) was used in both corpora. It is thus highly desirable to test and generalize the proposed approach on other data sets, and to modify it if required.

Speaker Clustering: We did investigate the issue of the often short, fast-changing speaker turns encountered in spontaneous multi-party speech. However, there are also a few long speaker turns – for example monologues such as presentations. Therefore, the speech utterances have very variable lengths. This would require probabilistic models with *adaptive capacity*, as opposed to the full-covariance single Gaussian used in this thesis. Mixtures of Dirichlet processes seem to be an interesting direction to follow, as already shown in Broadcast News speaker clustering experiments (Valente, 2006).

Unsupervised Audio-Visual Calibration: Although many audio-visual *speaker tracking* solutions exists, they often presuppose calibration information to be known either precisely or approximately, as shown by the review in (Gatica-Perez et al., 2006). A non-technical user may not want to go through a possibly complex calibration procedure, which motivated the two approaches proposed in Section 7.2. One could try to merge AV calibration and AV tracking, thus having a continuously updated calibration model. The challenge would be to get the system working on data with multiple moving speakers, as encountered for example in crowded areas (surveillance, guide robots etc.).

Distant Processing of Audio Signals: A distance-dependent variability of acoustic features was identified and characterized, that appears to be detrimental to the use of MFCCs for speaker clustering at variable distances. Further research is needed to explain this variability, possibly extending the research conducted in (Schwetz et al., 2004). One could speculate that invariance to these distance-dependent variabilities would not only help speaker clustering, but also speech recognition, and possibly speaker localization.

Appendix A

Performance Metrics for Detection

For each sector $\mathbb{S}_{\check{s}}$ and each time frame t:

- The ground-truth is $B_{\check{s},t} \in \{0, 1\}$.
- The decision taken by the system is $\hat{B}_{\check{s},t} =\in \{0, 1\}$.

Thus, following (Bengio et al., 2005), four types of cases happen, including correct classifications TP, TN and wrong classifications FP, FN, as defined in Table A.1. The corresponding number of samples $\mathrm{N_{TP}}, \mathrm{N_{TN}}, \mathrm{N_{FP}}, \mathrm{N_{FN}}$ are counted over all (sector, frame) pairs: $\{(\check{s}, t) \mid 1 \leq \check{s} \leq N_{\check{s}} \text{ and } t_1 \leq t \leq t_{N_t}\}$.

The False Alarm Rate (FAR) is defined as follows:

$$\mathrm{FAR} \stackrel{\mathrm{def}}{=} \frac{\mathrm{N_{FP}}}{\mathrm{N_{FP}} + \mathrm{N_{TN}}} \tag{A.1}$$

The False Rejection Rate (FRR) is defined as follows:

$$\mathrm{FRR} \stackrel{\mathrm{def}}{=} \frac{\mathrm{N_{FN}}}{\mathrm{N_{FN}} + \mathrm{N_{TP}}} \tag{A.2}$$

		Ground-truth	
		$B_{\check{s},t} = 0$	$B_{\check{s},t} = 1$
Detection	$\hat{B}_{\check{s},t} = 0$	TN	FN
decision	$\hat{B}_{\check{s},t} = 1$	FP	TP

Table A.1. The four types of results. TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

The Half Total Error Rate (HTER) is the arithmetic mean of FAR and FRR:

$$HTER \stackrel{\text{def}}{=} \frac{\text{FAR} + \text{FRR}}{2} \tag{A.3}$$

All three metrics FAR, FRR and HTER take values between 0 and 1 (the lower, the better).

As an alternative to FAR/FRR/HTER, another triplet of metrics is commonly used.

The Precision (PRC) is defined as follows:

$$\text{PRC} \stackrel{\text{def}}{=} \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{A.4}$$

The Recall (RCL) is defined as follows:

$$\text{RCL} \stackrel{\text{def}}{=} \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{A.5}$$

The F-measure is the harmonic mean of PRC and RCL:

$$\text{F} \stackrel{\text{def}}{=} \frac{2 \times \text{PRC} \times \text{RCL}}{\text{PRC} + \text{RCL}} \tag{A.6}$$

All three metrics PRC, RCL and F-measure take values between 0 and 1 (the higher, the better). Note that maximing F means maximizing *both* PRC and RCL.

Appendix B

Multidimensional Phase Domain Metrics

Section B.1 defines a Phase Domain Metric (PDM). Section B.2 proves that any 1-dimensional PDM can be composed into a multidimensional function which is also a PDM. Section B.3 proves the strict equivalence between SRP-PHAT and the PDM-based cost function Δ.

B.1 Definition of a PDM

Similarly to the classical metric definition, we define a PDM as a function $g(\mathbf{x}, \mathbf{y})$ on $\mathbb{R}^N \times \mathbb{R}^N$ verifying all of the following conditions, for all $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N$:

$$g(\mathbf{x}, \mathbf{y}) \geq 0 \tag{B.1}$$

$$g(\mathbf{x}, \mathbf{y}) = g(\mathbf{y}, \mathbf{x}) \tag{B.2}$$

$$g(\mathbf{x}, \mathbf{y}) = 0 \quad \text{iff} \quad \forall n \in \{1, \cdots, N\} \quad x_n \equiv y_n \tag{B.3}$$

$$g(\mathbf{x}, \mathbf{z}) \leq g(\mathbf{x}, \mathbf{y}) + g(\mathbf{y}, \mathbf{z}) \tag{B.4}$$

It is the same as a classical metric, except for (B.3) which reflects the "modulo 2π" definition of angles.

B.2 Property

Let g_1 be a 1-dimensional PDM (on $\mathbb{R} \times \mathbb{R}$). For $\lambda > 1$ and $N \in \mathbb{N} \setminus \{0\}$, let $G_{\lambda,N}$ be defined as:

$$\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N \qquad G_{\lambda,N}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \left\{ \frac{1}{N} \sum_{n=1}^{N} [g_1(x_n, y_n)]^\lambda \right\}^{\frac{1}{\lambda}} \tag{B.5}$$

The rest of this Section shows that all $G_{\lambda,N}$ functions are also PDMs. (B.1), (B.2) and (B.3) are trivial to demonstrate. (B.4) is demonstrated for $G_{\lambda,N}$ in the following.

Let $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N$. From (B.4) applied to g_1, and (B.5) we obtain:

$$G_{\lambda,N}(\mathbf{x}, \mathbf{z}) \leq \left\{ \frac{1}{N} \sum_{n=1}^{N} [g_1(x_n, y_n) + g_1(y_n, z_n)]^\lambda \right\}^{\frac{1}{\lambda}} \tag{B.6}$$

For any $\lambda > 1$, $\alpha_n > 0$ and $\beta_n > 0$, the Minkowski inequality (Moon and Stirling, 2000) is written:

$$\left[\sum_{n=1}^{N} (\alpha_n + \beta_n)^\lambda \right]^{\frac{1}{\lambda}} \leq \left[\sum_{n=1}^{N} (\alpha_n)^\lambda \right]^{\frac{1}{\lambda}} + \left[\sum_{n=1}^{N} (\beta_n)^\lambda \right]^{\frac{1}{\lambda}} \tag{B.7}$$

We now apply (B.7) to the RHS of (B.6), with $\alpha_n = g_1(x_n, y_n)$ and $\beta_n = g_1(y_n, z_n)$:

$$G_{\lambda,N}(\mathbf{x}, \mathbf{z}) \leq \left\{ \tfrac{1}{N} \textstyle\sum_{n=1}^{N} [g_1(x_n, y_n)]^\lambda \right\}^{\frac{1}{\lambda}} + \left\{ \tfrac{1}{N} \textstyle\sum_{n=1}^{N} [g_1(y_n, z_n)]^\lambda \right\}^{\frac{1}{\lambda}} \tag{B.8}$$

$$G_{\lambda,N}(\mathbf{x}, \mathbf{z}) \leq G_{\lambda,N}(\mathbf{x}, \mathbf{y}) + G_{\lambda,N}(\mathbf{y}, \mathbf{z}) \tag{B.9}$$

Therefore (B.4) is verified by $G_{\lambda,N}$, and $G_{\lambda,N}$ is a PDM[1]. □

Application: In the case $\lambda = 2$, this property proves that the d function defined in (5.8) is a PDM.

B.3 Equivalence Between SRP-PHAT and Δ

The cost function Δ defined by (5.42) is a sum, where each term is written with the squared PDM d^2 defined by (5.8). This section shows that $\mathrm{P}_{\text{SRP-PHAT}}$ and Δ are linearly related. In particular, minimizing Δ is strictly equivalent to maximizing $\mathrm{P}_{\text{SRP-PHAT}}$.

[1] The exact same demonstration is also valid for classical metrics (replace "PDM" with "metric").

Let us write the SRP-PHAT sum defined in (3.13), for a location $\boldsymbol{\ell} \in \mathbb{R}^3$:

$$\mathrm{p}_{\text{SRP-PHAT}}\left(\boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right) = \sum_{k=2}^{N_\mathrm{F}+1} \mathrm{p}_{\text{SRP-PHAT}}\left(k, \boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right) \tag{B.10}$$

where each term of the sum is defined as:

$$\begin{aligned}\mathrm{p}_{\text{SRP-PHAT}}\left(k, \boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right) &\stackrel{\text{def}}{=} \left|\sum_{m=1}^{N_m} \frac{X_m^{(t)}(k)}{\left|X_m^{(t)}(k)\right|} e^{j\pi\frac{k-1}{N_\mathrm{F}}\mathrm{TOF}(\boldsymbol{\ell},\boldsymbol{\ell}_m)}\right|^2 &\text{(B.11)}\\ &= \left|\sum_{m=1}^{N_m} \exp\left[j\left(\angle X_m^{(t)}(k) + \pi\frac{k-1}{N_\mathrm{F}}\mathrm{TOF}(\boldsymbol{\ell},\boldsymbol{\ell}_m)\right)\right]\right|^2 &\text{(B.12)}\end{aligned}$$

Using the $|z|^2 = z \cdot z^*$ decomposition we obtain:

$$\begin{aligned}&\mathrm{p}_{\text{SRP-PHAT}}\left(k, \boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right)\\ &= N_m + \sum_{m=1}^{N_m} \cdot \sum_{\substack{m'\neq m\\ m'=1}}^{N_m} \exp\left\{j\left[\angle X_m^{(t)}(k) - \angle X_{m'}^{(t)}(k) + \pi\frac{k-1}{N_\mathrm{F}}\left(\mathrm{TOF}(\boldsymbol{\ell},\boldsymbol{\ell}_m) - \mathrm{TOF}(\boldsymbol{\ell},\boldsymbol{\ell}_{m'})\right)\right]\right\}\end{aligned}$$

Using the $z + z^* = 2 \cdot \Re(z)$ equality we obtain:

$$\begin{aligned}&\mathrm{p}_{\text{SRP-PHAT}}\left(k, \boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right)\\ &= N_m + 2 \cdot \sum_{m=1}^{N_m} \cdot \sum_{m'=m+1}^{N_m} \cos\left[\angle X_m^{(t)}(k) - \angle X_{m'}^{(t)}(k) + \pi\frac{k-1}{N_\mathrm{F}}\left(\mathrm{TOF}(\boldsymbol{\ell},\boldsymbol{\ell}_m) - \mathrm{TOF}(\boldsymbol{\ell},\boldsymbol{\ell}_{m'})\right)\right]\\ &= N_m + 2 \cdot \sum_{q=1}^{N_q} \cos\left[\underbrace{\angle X_{a_q}^{(t)}(k) - \angle X_{b_q}^{(t)}(k)}_{u_q^{(t)}(k)} + \underbrace{\pi\frac{k-1}{N_\mathrm{F}}\left(\mathrm{TOF}\left(\boldsymbol{\ell},\boldsymbol{\ell}_{a_q}\right) - \mathrm{TOF}\left(\boldsymbol{\ell},\boldsymbol{\ell}_{b_q}\right)\right)}_{-u_q^{\mathrm{th}}(k,\boldsymbol{\ell})}\right]\\ &= N_m + 2 \cdot \sum_{q=1}^{N_q} \cos\left[u_q^{(t)}(k) - u_q^{\mathrm{th}}(k,\boldsymbol{\ell})\right]\end{aligned}$$

Using the $\cos u = 1 - 2 \cdot \sin^2\left(\frac{u}{2}\right)$ equality, and (5.8), we obtain:

$$\begin{aligned}\mathrm{p}_{\text{SRP-PHAT}}\left(k, \boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right) &= N_m + 2 \cdot N_q - 4 \cdot \sum_{q=1}^{N_q} \sin^2\left(\frac{u_q^{(t)}(k) - u_q^{\mathrm{th}}(k,\boldsymbol{\ell})}{2}\right)\\ &= N_m + 2 \cdot N_q - 4 \cdot N_q \cdot d^2\left(\mathbf{u}^{(t)}(k), \mathbf{u}^{\mathbf{th}}(k,\boldsymbol{\ell})\right) &\text{(B.13)}\end{aligned}$$

Summing over the strictly positive discrete frequencies $k \in \{2, \cdots, N_{\mathrm{F}}+1\}$, we obtain:

$$\mathrm{P}_{\text{SRP-PHAT}}\left(\boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right) = N_{\mathrm{F}} \cdot (N_m + 2 \cdot N_q) - 4 \cdot N_q \cdot \sum_{k=2}^{N_{\mathrm{F}}+1} d^2\left(\mathbf{u}^{(t)}(k), \mathbf{u}^{\mathbf{th}}(k, \boldsymbol{\ell})\right) \tag{B.14}$$

Let us now consider the case where all strictly positive discrete frequencies are used to define the cost function Δ in (5.42): $\Upsilon = \{2, \cdots, N_{\mathrm{F}}+1\}$. We can then write:

$$\mathrm{P}_{\text{SRP-PHAT}}\left(\boldsymbol{\ell}, \mathbf{X}_1^{(t)}, \cdots, \mathbf{X}_{N_m}^{(t)}\right) = N_{\mathrm{F}} \cdot (N_m + 2 \cdot N_q) - 4 \cdot N_q \cdot N_{\mathrm{F}} \cdot \Delta\left(\left\{\mathbf{u}^{(t)}(k)\right\}, \Upsilon, \boldsymbol{\ell}\right) \tag{B.15}$$

In this case, there is a linear relationship between $\mathrm{P}_{\text{SRP-PHAT}}$ and the cost function Δ. In particular, the location $\hat{\boldsymbol{\ell}}$ that minimizes the cost function Δ also maximizes $\mathrm{P}_{\text{SRP-PHAT}}$. $\square$

Appendix C

Sector-Based Activeness Models and their EM Derivations

This appendix describes the two models used in Section 5.3.2 to model the activeness values $\zeta_{\check{s},t}$, where $\check{s}$ is the sector index and t the time frame center (expressed in sampling periods). Section C.1 describes the 1-dimensional model, where all activeness values are stacked in one dimension, irrespective of $\check{s}$ or t. Section C.2 describes the multidimensional model, where the activeness values of all sectors in the same time frame $\{\zeta_{1,t}, \cdots, \zeta_{\check{s},t}, \cdots, \zeta_{N_{\check{s}},t}\}$ are modelled jointly.

In each section, a description of the model is provided, followed by the complete Expectation-Maximization (EM) derivation (Dempster et al., 1977). For both models, implementation details are also provided, including data reduction and automatic initialization. The latter makes the EM fitting process an adaptive and fully deterministic process. Section C.1 is somewhat detailed. Some of the results in Section C.1 are reused in Section C.2, which is therefore less detailed.

We recall the notation $\{\zeta_{\check{s},t}\}$, which means the whole set of activeness values on which we want to fit a model (for example one recording):

$$\{\zeta_{\check{s},t}\} \stackrel{\text{def}}{=} \{\zeta_{\check{s},t} \mid (\check{s}, t) \in \{1, \cdots, N_{\check{s}}\} \times \{t_1, \cdots, t_{N_t}\}\} \tag{C.1}$$

where $N_{\check{s}}$ is the number of sectors in space, and N_t is the number of time frames.

All probabilities and likelihood are conditioned by a model $\mathcal{M}$ and its parameter values $\mathbf{\Lambda}(\mathcal{M})$. In this appendix, we abbreviate $p(\zeta_{\check{s},t} \mid \mathcal{M}, \mathbf{\Lambda}(\mathcal{M}))$ with $p(\zeta_{\check{s},t}, \mathbf{\Lambda})$ or $p(\zeta_{\check{s},t})$, whenever possible.

Moreover, we recall that $p(\zeta_{\check{s},t})$ is an abbreviation for $p\left(\underline{\zeta} = \zeta_{\check{s},t}\right)$, as defined in (2.17). On the other hand, any equation written using random variables without realization means that the equation is valid for *any* realization of the set of random variables. For example:

$$p(\underline{a}) \quad = \quad p(\underline{b}) \tag{C.2}$$

is strictly equivalent to:

$$\forall a, b \qquad p(\underline{a} = a) \quad = \quad p(\underline{b} = b) \tag{C.3}$$

As for probability distributions, the reader is referred to Section 2.4 for the definitions of the Dirac pdf $\delta_0(\xi)$, the Gamma pdf $\mathcal{G}_{\alpha,\beta}(\xi)$ and the Rice pdf $\mathcal{R}_{\sigma,V}(\xi)$ (Rice, 1944, 1945).

C.1 1-dimensional Model

This section describes the 1-dimensional probabilistic model used to model the distribution of $\{\zeta_{\check{s},t}\}$ in Section 5.3.2, and derives the EM algorithm for it. The set of all observed activeness values is stacked onto one dimension, irrespective of $\check{s}$ or t. The corresponding graphical model is shown in Figure C.3. **The space and time indices $(\check{s}, t)$ are irrelevant for the random variables $\underline{B}$ and $\underline{\zeta}$, because the whole data set $\{\zeta_{\check{s},t}\}$ was staked onto one dimension. This is an i.i.d. assumption across both space and time.** The proposed model is thus a 2-component mixture: $p\left(\underline{\zeta}\right) = w_0 \cdot f_0(\underline{\zeta}) + w_1 \cdot f_1(\underline{\zeta})$ where:

$$f_0\left(\underline{\zeta}\right) \overset{\text{def}}{=} p\left(\underline{\zeta} \mid \underline{B} = 0\right) \tag{C.4}$$

$$f_1\left(\underline{\zeta}\right) \overset{\text{def}}{=} p\left(\underline{\zeta} \mid \underline{B} = 1\right) \tag{C.5}$$

$$w_0 \overset{\text{def}}{=} P(\underline{B} = 0) \tag{C.6}$$

$$w_1 \overset{\text{def}}{=} P(\underline{B} = 1) \tag{C.7}$$

The priors w_0 and w_1 are, by definitions, values in $[0\ 1]$.

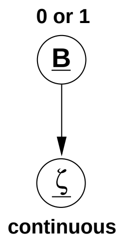


Figure C.1. Graphical model for the 1-dimensional model. The r.v. $\underline{B} \in \{0, 1\}$ is the sector state (inactive or active). The r.v. $\underline{\zeta} \geq 0$ is the sector activeness.

C.1.1 Description

As defined in (5.22), the activeness feature $\zeta_{\check{s},t} \in \{0, 1, \cdots, N_{\mathrm{F}}\}$ is the number of discrete frequencies where acoustic sources in sector $\mathbb{S}_{\check{s}}$ are dominant over the sources in other sectors, at time frame t. In the case that there is a speech source in a sector of space, the corresponding value $\zeta_{\check{s},t}$ will be large because speech is wideband.

Inactivity: Dirac + Rice

In the case that there is no active coherent source at all in a particular discrete frequency k, the frequency domain signals $(X_1(k), \cdots, X_{N_m}(k))$ are uncorrelated, and the choice (5.20) of the dominant sector is random, with equal probability $1/N_{\check{s}}$ for each sector. Consequently, in the case that a sector $\mathbb{S}_{\check{s}}$ is completely inactive at time t, the number $\zeta_{\check{s},t}$ of discrete frequencies attributed to this sector is a sum of realizations of such uniform random processes. It is therefore expected that $\underline{\zeta}$ follows a binomial pdf.

However, in real cases, even for a sector $\mathbb{S}_{\check{s}}$ that does not contain any active coherent source, $\zeta_{\check{s},t}$ will not only result from purely random decisions, but it will also capture acoustic activity due to background noise (e.g. computer fan) and reverberations. We found visually, on some inactive sectors in a separate set of development data, that the Gamma pdf has a better fit than the binomial pdf. The Gamma pdf can therefore be used to model inactive sectors.

Moreover, the parameters of the Gamma pdf used to model inactive sectors vary greatly between conditions, which can be roughly divided into two cases: whether at least one sector is active in a given frame or not. Therefore, ideally, two different pdfs should be used. This is the issue addressed by the multidimensional model in Section C.2.

In the case of a 1-dimensional model, we need to model a mixture of all inactive sectors (whether the frame is active or not) with a single pdf. We found visually that the Rice pdf provides a better fit than the Gamma.

Finally, since some (rare) values of $\zeta_{\check{s},t}$ are zeroes, the inactive data is modelled by a mixture of a Dirac pdf at zero and a Rice pdf:

$$f_0(\underline{\zeta}) \quad = \quad w_0^D \cdot \delta_0(\underline{\zeta}) + w_0^R \cdot \mathcal{R}_{\sigma_0, V_0}(\underline{\zeta}) \tag{C.8}$$

where w_0^D and w_0^R are the priors of each pdf (values in $[0\ 1]$).

Activity: Shifted Rice

The distribution of $\zeta_{\check{s},t}$ for acoustic activity (especially speech) follows a distribution that is quite complex, varying over time and not known *a priori*. We therefore chose to use the Rice pdf for activity, because it is a flexible way to model a distribution of positive values. The shape of the Rice can vary from a pointy, Dirac-like pdf to a Gaussian-like pdf or a Rayleigh pdf.

Furthermore, it is reasonable to assume that in a small range of values around the background noise level, $\zeta_{\check{s},t}$ does not give any information to discriminate between activity and inactivity. Hence, we only model values above the mean square value $\sqrt{2 \cdot \sigma_0^2 + V_0^2}$ (Greenstein et al., 1999) of the silence pdf $\mathcal{R}_{\sigma_0, V_0}$. Hence the "Shifted Rice" pdf for active sectors:

$$f_1(\underline{\zeta}) \quad = \quad \mathcal{R}_{\sigma_1, V_1}\left(\underline{\zeta} - \sqrt{2 \cdot \sigma_0^2 + V_0^2}\right) \tag{C.9}$$

Mixture of Inactivity and Activity

The likelihood $p\left(\underline{\zeta} \mid \underline{B}\right)$ is expressed as:

$$p\left(\underline{\zeta} \mid \underline{B}\right) \quad \stackrel{\text{def}}{=} \quad \delta_{Kr}(\underline{B} - 0) \cdot f_0(\underline{\zeta}) + \delta_{Kr}(\underline{B} - 1) \cdot f_1(\underline{\zeta}) \tag{C.10}$$

where $\delta_{Kr}(\underline{\zeta}) = \mathbf{1}_{\underline{\zeta}=0}$ is the Kronecker function, (not to be confused with the zero-centered Dirac pdf δ_0). $\mathbf{1}_{\text{proposition}}$ is the indicator function: $\mathbf{1}_{\text{proposition}} = 1$ if proposition is true, 0 otherwise.

The mixture model is then written:

$$\begin{aligned} p\left(\underline{\zeta}\right) &= P\left(\underline{B}=0\right)\cdot p\left(\underline{\zeta} \mid \underline{B}=0\right) + P\left(\underline{B}=1\right)\cdot p\left(\underline{\zeta} \mid \underline{B}=1\right) && \text{(C.11)} \\ &= w_0 \cdot f_0(\underline{\zeta}) + w_1 \cdot f_1(\underline{\zeta}) && \text{(C.12)} \end{aligned}$$

The complete list of parameters for the 1-dimensional model is:

$$\mathbf{\Lambda}_{\mathrm{1D}} \stackrel{\mathrm{def}}{=} \left(w_0, w_1, w_0^D, w_0^R, \sigma_0, V_0, \sigma_1, V_1\right) \tag{C.13}$$

C.1.2 EM Derivation

General

In the E-step, the posteriors are computed using the Bayes rule, for example, for a given set of parameter values $\mathbf{\Lambda}_{\mathrm{1D}}$ and a given observed activeness value $\zeta_{\breve{s},t}$:

$$\begin{aligned} P_{\breve{s},t}^{(0)}\left(\mathbf{\Lambda}_{\mathrm{1D}}\right) &\stackrel{\mathrm{def}}{=} p\left(\underline{B}=0 \mid \zeta_{\breve{s},t}, \mathbf{\Lambda}_{\mathrm{1D}}\right) = \frac{w_0 \cdot f_0(\zeta_{\breve{s},t})}{w_0 \cdot f_0(\zeta_{\breve{s},t}) + w_1 \cdot f_1(\zeta_{\breve{s},t})} && \text{(C.14)} \\ P_{\breve{s},t}^{(1)}\left(\mathbf{\Lambda}_{\mathrm{1D}}\right) &\stackrel{\mathrm{def}}{=} p\left(\underline{B}=1 \mid \zeta_{\breve{s},t}, \mathbf{\Lambda}_{\mathrm{1D}}\right) = 1 - P_{\breve{s},t}^{(0)}\left(\mathbf{\Lambda}_{\mathrm{1D}}\right) && \text{(C.15)} \end{aligned}$$

Let us assume that we have parameter values $\mathbf{\Lambda}_{\mathrm{1D}}$, in the M-step we look for new values $\hat{\mathbf{\Lambda}}_{\mathrm{1D}}$ that will increase the likelihood of the observed data $\{\zeta_{\breve{s},t}\}$. Let us write the KL divergence between the two pdfs[1] associated with current parameters $\mathbf{\Lambda}_{\mathrm{1D}}$ and new parameters $\hat{\mathbf{\Lambda}}_{\mathrm{1D}}$, conditionally to $\underline{\zeta}$:

$$\begin{aligned} & KL\left[P\left(\underline{B} \mid \underline{\zeta}, \mathbf{\Lambda}_{\mathrm{1D}}\right), \hat{P}\left(\underline{B} \mid \underline{\zeta}, \hat{\mathbf{\Lambda}}_{\mathrm{1D}}\right)\right] \stackrel{\mathrm{def}}{=} \\ & \left\langle \log P\left(\underline{B} \mid \underline{\zeta}, \mathbf{\Lambda}_{\mathrm{1D}}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}, \mathbf{\Lambda}_{\mathrm{1D}}\right)} - \left\langle \log \hat{P}\left(\underline{B} \mid \underline{\zeta}, \hat{\mathbf{\Lambda}}_{\mathrm{1D}}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}, \mathbf{\Lambda}_{\mathrm{1D}}\right)} \end{aligned} \tag{C.16}$$

where letters P and $\hat{P}$ represent the same function: they are only used to clarify where current parameters $\mathbf{\Lambda}_{\mathrm{1D}}$ or new parameters $\hat{\mathbf{\Lambda}}_{\mathrm{1D}}$ are used. In the following, $\mathbf{\Lambda}_{\mathrm{1D}}$ and $\hat{\mathbf{\Lambda}}_{\mathrm{1D}}$ are thus omitted whenever possible. The mean $\langle\cdot\rangle_{.}$ is calculated over all possible values of $\underline{B}$ (0 or 1). For example:

$$\left\langle \log \hat{P}\left(\underline{B} \mid \underline{\zeta}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} = \sum_{B=0}^{1} P\left(\underline{B}=B \mid \underline{\zeta}\right) \cdot \log \hat{P}\left(\underline{B}=B \mid \underline{\zeta}\right) \tag{C.17}$$

[1] In this particular case the conditional pdf of $\underline{B}$ is written as a posterior, because $\underline{B}$ is a discrete r.v.

The KL divergence is always positive, therefore (C.16) can be rewritten (omitting $\mathbf{\Lambda}_{1\mathrm{D}}$ and $\hat{\mathbf{\Lambda}}_{1\mathrm{D}}$):

$$\left\langle \log P\left(\underline{B} \mid \underline{\zeta}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} - \left\langle \log \hat{P}\left(\underline{B} \mid \underline{\zeta}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} \geq 0 \tag{C.18}$$

Using Bayes rule to decompose the second term we obtain:

$$\log \hat{p}\left(\underline{\zeta}\right) \geq -\left\langle \log P\left(\underline{B} \mid \underline{\zeta}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} + \left\langle \log \hat{p}\left(\underline{\zeta}, \underline{B}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} \tag{C.19}$$

The first term in the RHS $\langle \log P \rangle_P$ does not depend on $\hat{\mathbf{\Lambda}}_{1\mathrm{D}}$, therefore, one way to increase the likelihood $\log \hat{p}\left(\underline{\zeta}\right)$ is to find $\hat{\mathbf{\Lambda}}_{1\mathrm{D}}$ that maximizes the second term on the RHS $\left\langle \log \hat{P} \right\rangle_P$. The latter can be decomposed into:

$$\left\langle \log \hat{p}\left(\underline{\zeta}, \underline{B}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} = \left\langle \log \hat{p}\left(\underline{\zeta} \mid \underline{B}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} + \left\langle \log \hat{P}\left(\underline{B}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta}\right)} \tag{C.20}$$

In the M-step our purpose is to find the parameter values $\hat{\mathbf{\Lambda}}_{1\mathrm{D}}$ that maximize the likelihood of the observed data $\{\zeta_{\check{s},t}\}$:

$$\sum_{\check{s},t} \log \hat{p}\left(\underline{\zeta} = \zeta_{\check{s},t}\right) \tag{C.21}$$

which, using (C.19) and (C.20), can be done by maximizing $\Xi_1 + \Xi_2$, where:

$$\Xi_1 \stackrel{\text{def}}{=} \sum_{\check{s},t} \left\langle \log \hat{p}\left(\underline{\zeta} = \zeta_{\check{s},t} \mid \underline{B}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta} = \zeta_{\check{s},t}\right)} \tag{C.22}$$

$$\Xi_2 \stackrel{\text{def}}{=} \sum_{\check{s},t} \left\langle \log \hat{P}\left(\underline{B}\right)\right\rangle_{P\left(\underline{B} \mid \underline{\zeta} = \zeta_{\check{s},t}\right)} \tag{C.23}$$

Specific

Let us express both terms Ξ_1 and Ξ_2 as a function of the new parameters $\hat{\mathbf{\Lambda}}_{1\mathrm{D}} = \left(\hat{w}_0, \hat{w}_1, \hat{w}_0^D, \hat{w}_0^R, \hat{\sigma}_0, \hat{V}_0, \hat{\sigma}_1, \hat{V}_1\right)$.

$$\Xi_1 = \sum_{\check{s},t} \sum_{B=0}^{1} P_{\check{s},t}^{(B)} \cdot \log \hat{p}\left(\underline{\zeta} = \zeta_{\check{s},t} \mid \underline{B} = B, \hat{\mathbf{\Lambda}}_{1\mathrm{D}}\right) \tag{C.24}$$

From (C.10) we obtain:

$$\Xi_1 = \sum_{\check{s},t} \sum_{B=0}^{1} P_{\check{s},t}^{(B)} \cdot \log f_B\left(\zeta_{\check{s},t}, \hat{\mathbf{\Lambda}}_{1\mathrm{D}}\right) \tag{C.25}$$

From (C.8) and (C.9) we obtain:

$$\begin{aligned}\Xi_1 &= \sum_{\breve{s},t} P^{(0)}_{\breve{s},t} \cdot \log\left(\hat{w}^D_0 \cdot \delta_0(\zeta_{\breve{s},t}) + \hat{w}^R_0 \cdot \mathcal{R}_{\hat{\sigma}_0,\hat{V}_0}(\zeta_{\breve{s},t})\right) &\text{(C.26)}\\ &+ \sum_{\breve{s},t} P^{(1)}_{\breve{s},t} \cdot \log\left(\mathcal{R}_{\hat{\sigma}_1,\hat{V}_1}\left(\zeta_{\breve{s},t} - \sqrt{2\cdot\hat{\sigma}_0^2 + \hat{V}_0^2}\right)\right) &\text{(C.27)}\end{aligned}$$

$$\begin{aligned}\Xi_1 &= \sum_{\substack{\breve{s},t\\ \zeta_{\breve{s},t}=0}} P^{(0)}_{\breve{s},t} \cdot \log \hat{w}^D_0 \quad + \sum_{\substack{\breve{s},t\\ \zeta_{\breve{s},t}>0}} P^{(0)}_{\breve{s},t} \cdot \log \hat{w}^R_0 &\text{(C.28)}\\ &+ \sum_{\substack{\breve{s},t\\ \zeta_{\breve{s},t}=0}} P^{(0)}_{\breve{s},t} \cdot \log \delta_0(\zeta_{\breve{s},t}) &\text{(C.29)}\\ &+ \sum_{\substack{\breve{s},t\\ \zeta_{\breve{s},t}>0}} P^{(0)}_{\breve{s},t} \cdot \log \mathcal{R}_{\hat{\sigma}_0,\hat{V}_0}(\zeta_{\breve{s},t}) \quad + \sum_{\substack{\breve{s},t\\ \zeta_{\breve{s},t}>\sqrt{2\cdot\hat{\sigma}_0^2+\hat{V}_0^2}}} P^{(1)}_{\breve{s},t} \cdot \log \mathcal{R}_{\hat{\sigma}_1,\hat{V}_1}\left(\zeta_{\breve{s},t} - \sqrt{2\cdot\hat{\sigma}_0^2 + \hat{V}_0^2}\right) &\text{(C.30)}\end{aligned}$$

The term in $\log \delta_0(\cdot)$ is not finite, but does not involve any parameter in $\hat{\mathbf{\Lambda}}_{\mathrm{1D}}$. In the M-step, we therefore maximize the "partial likelihood", which is the sum of all other (finite) terms.

From (C.6) and (C.7):

$$\begin{aligned}\Xi_2 &= \sum_{\breve{s},t}\sum_{B=0}^{1} P^{(B)}_{\breve{s},t} \cdot \log \hat{P}\left(\underline{B} = B\right) &\text{(C.31)}\\ \Xi_2 &= \sum_{\breve{s},t} P^{(0)}_{\breve{s},t} \cdot \log \hat{w}_0 + \sum_{\breve{s},t} P^{(1)}_{\breve{s},t} \cdot \log \hat{w}_1 &\text{(C.32)}\end{aligned}$$

Our goal is to find $\hat{\mathbf{\Lambda}}_{\mathrm{1D}}$ that maximizes $\Xi_1 + \Xi_2$. From (C.28), (C.30) and (C.32), we can see that:

- The priors $\hat{w}_0$ and $\hat{w}_1$ only appear in Ξ_2 (C.32).
- The weights $\hat{w}^D_0$ and $\hat{w}^R_0$ of the silence mixture only appear in Ξ_1, in (C.28).
- The remaining parameters $\left(\hat{\sigma}_0, \hat{V}_0, \hat{\sigma}_1, \hat{V}_1\right)$ are tied in a non-linear fashion through (C.30). Finding their value can be done through joint, numerical optimization (e.g. simplex search, as in `fminsearch` in Matlab) of the corresponding sum (C.30).

Since $\hat{w}_1 = 1 - \hat{w}_0$, we have:

$$\frac{\partial \Xi_2}{\partial \hat{w}_0} = \sum_{\breve{s},t}\left(\frac{1}{\hat{w}_0} \cdot P^{(0)}_{\breve{s},t} - \frac{1}{1-\hat{w}_0} \cdot P^{(1)}_{\breve{s},t}\right) \quad \text{(C.33)}$$

The new parameter $\hat{w}_0$ does not appear in Ξ_1, it is therefore the maximum of Ξ_2 with respect to $\hat{w}_0$, which necessitates $\frac{\partial \Xi_2}{\partial \hat{w}_0} = 0$, therefore:

$$\hat{w}_0 \quad = \quad \frac{\sum_{\check{s},t} P^{(0)}_{\check{s},t}}{\sum_{B=0}^{1} \sum_{\check{s},t} P^{(B)}_{\check{s},t}} \quad = \quad \frac{1}{N_{\check{s}} \cdot N_t} \cdot \sum_{\check{s},t} P^{(0)}_{\check{s},t} \qquad \text{and} \quad \hat{w}_1 = 1 - \hat{w}_0 \tag{C.34}$$

This is the update of the priors of inactivity and activity.

Similarly, $\hat{w}^R_0 = 1 - \hat{w}^D_0$, and the new parameter $\hat{w}^D_0$ only appears in (C.28), it is therefore the maximum of (C.28) with respect to $\hat{w}^D_0$, which yields:

$$\hat{w}^D_0 \quad = \quad \frac{\sum\limits_{\substack{\check{s},t \\ \zeta_{\check{s},t}=0}} P^{(0)}_{\check{s},t}}{\sum\limits_{\substack{\check{s},t \\ \zeta_{\check{s},t}=0}} P^{(0)}_{\check{s},t} + \sum\limits_{\substack{\check{s},t \\ \zeta_{\check{s},t}>0}} P^{(0)}_{\check{s},t}} \qquad \text{and} \quad \hat{w}^R_0 = 1 - \hat{w}^D_0 \tag{C.35}$$

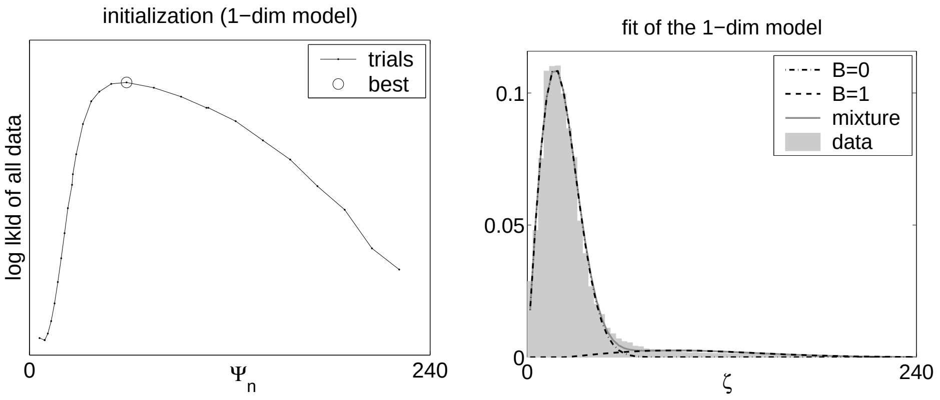
This is the update of the "inactivity" mixture weights.

Implementation Details

EM Implementation: in practice, we observed that the possibly large amount of data $\{\zeta_{\check{s},t}\}$ can be conveniently reduced to a very small number of samples (e.g. 100) with approximately the same pdf. This is done by ordering the samples (from min to max) and picking 100 samples at regular intervals along the ordered list. This way, the cost of each EM iteration is drastically reduced, and is independent of the amount of data (e.g. 20 minutes of recording are reduced to 100 samples).

An additional speedup can be obtained by replacing, in the M-step, the numerical optimization of $\left(\hat{\sigma}_0, \hat{V}_0, \hat{\sigma}_1, \hat{V}_1\right)$ with a moment-based update similar to the initialization described below. This way, the numerical optimization, which is itself a many-step process, is replaced with a direct, 1-step analytical update. Although this is an approximation, we observed in practice that after convergence of EM, the model parameters are almost the same as with the numerical optimization.

All results reported in the article for the 1-dimensional models use both simplifications. An example of data distribution fitted with the 1-dimensional model is depicted in Figure C.2b.



(a) (b)

Figure C.2. Fit of the 2-component mixture model described in Section C.1: (a) automatic initialization, (b) final model $p(\underline{\zeta}=\zeta)$ after convergence of EM.

Automatic Initialization: Using a given threshold value Ψ, the data $\{\zeta_{\check{s},t}\}$ is splitted between $N_{\text{low}}(\Psi)$ low values and $N_{\text{high}}(\Psi)$ high values. The non-zero low values are used to initialize the "inactivity" Rice pdf using the analytical moment-based approximation from (Greenstein et al., 1999). This is done by first computing the mean G_a and standard deviation G_v:

$$G_a = \frac{1}{N_{\text{low}}(\Psi)} \sum_{\substack{\check{s},t \\ 0<\zeta_{\check{s},t}<\Psi}} \zeta_{\check{s},t} \tag{C.36}$$

$$G_v = \left(\frac{1}{N_{\text{low}}(\Psi)} \sum_{\substack{\check{s},t \\ 0<\zeta_{\check{s},t}<\Psi}} (\zeta_{\check{s},t} - G_a)^2 \right)^{\frac{1}{2}} \tag{C.37}$$

and the parameters of the "inactivity" Rice pdf are initialized as follows:

$$V_0^{(\text{init})} \leftarrow \left[\max\left(0,\ G_a^2 - G_v^2\right)\right]^{\frac{1}{4}} \tag{C.38}$$

$$\sigma_0^{(\text{init})} \leftarrow \left[\frac{1}{2}\cdot\max\left(0,\ G_a - \left(V_0^{(\text{init})}\right)^2\right)\right]^{\frac{1}{2}} \tag{C.39}$$

The "activity" Shifted Rice pdf is initialized similarly, using data above:

$$\max\left(\Psi,\ \sqrt{2\cdot\left(\sigma_0^{(\text{init})}\right)^2 + \left(V_0^{(\text{init})}\right)^2}\right) \tag{C.40}$$

The mixture weights w_0^D and w_0^R are initialized by counting the number of zero samples.

The priors w_0 and w_1 are initialized as follows:

$$w_0^{(\text{init})} \leftarrow \max\left(0.1,\ \min\left(0.9,\ \frac{N_{\text{low}}(\Psi)}{N_{\text{low}}(\Psi) + N_{\text{high}}(\Psi)}\right)\right) \quad \text{(C.41)}$$

$$w_1^{(\text{init})} \leftarrow 1 - w_0^{(\text{init})} \quad \text{(C.42)}$$

where the restriction to the $[0.1, 0.9]$ interval avoids a "wrong" local maxima such as $w_0^{(init)} = 0$.

In order to have a fully automatic initialization process, a series of thresholds $\{\Psi_1, \cdots, \Psi_{N_\Psi}\}$ are derived from the data $\{\zeta_{\check{s},t}\}$ itself (e.g. $N_\Psi = 30$: 15 equal-interval percentiles and 15 equal intervals between minimum and maximum). For each threshold Ψ_n, the moment-based initialization is done as explained above and the likelihood of the whole data is computed. The initialization yielding the maximum likelihood is selected, as depicted in Figure C.2a. This way, we avoid starting from a "wrong" local maxima (e.g. the "inactive" component capturing all data and the "active" component capturing none, or vice-versa). Moreover, this automatic initialization is deterministic, therefore the whole EM fitting process is also deterministic. This means that for a given set of data $\{\zeta_{\check{s},t}\}$, the EM fitting process always yields the same values for model parameters $\mathbf{\Lambda}_{\text{1D}}$.

C.2 Multidimensional Model

This section describes a multidimensional model that models activeness for all sectors $(\zeta_{1,t}, \cdots, \zeta_{\check{s},t})$ jointly, at any given time frame t. It is used in Section 5.3.2. In many places, reasoning exposed in details in the 1-dimensional case (Section C.1) is reused here in a brief form.

C.2.1 Description

A property of the SAM-SPARSE-MEAN activeness is (5.23): for a given time frame t, the activeness values of all sectors sum to a constant: $\sum_{\check{s}} \zeta_{\check{s},t} = N_{\text{F}}$. This knowledge is enough to expect two cases:

- At a given time frame t, there is no activity. Then, as explained in C.1.1, the N_{F} discrete frequencies of the frequency spectrum are attributed to the various sectors in a uniformly random fashion. It is therefore expected that $\zeta_{\check{s},t}$ will be $N_{\text{F}}/N_{\check{s}}$ in average. As mentioned in C.1.1, the Gamma pdf $\mathcal{G}_{\gamma,\beta}$ fits well (visual trials on real data). In this case, we could expect the first moment of the Gamma pdf to be equal to the average activeness value: $\gamma \cdot \beta = \frac{N_{\text{F}}}{N_{\check{s}}}$.

- At a given time frame t, at least one sector contains at least one active wideband source (e.g. speech source). In such a case, the $\zeta_{\check{s},t}$ values corresponding to active sector(s) will be larger than the average $N_{\mathrm{F}}/N_{\check{s}}$, thus leaving less discrete frequencies to be randomly attributed to inactive sectors. We propose to model this case with a (Gamma + Shifted Rice) mixture, similarly to Section C.1.1. For the Gamma pdf, we could expect that $\gamma \cdot \beta < \frac{N_{\mathrm{F}}}{N_{\check{s}}}$.

At this point we need to incorporate these two frame-level cases into the model. Thus, we define the sector state binary random variables $\left(\underline{B}_1, \cdots, \underline{B}_{\check{s}}, \cdots \underline{B}_{N_{\check{s}}}\right)$, and the frame state random variable $\underline{A} \in \{0, 1\}$ which indicates whether or not at least one sector is active in a given time frame:

$$\underline{A} \stackrel{\mathrm{def}}{=} \max_{1 \leq \check{s} \leq N_{\check{s}}} \underline{B}_{\check{s}} \tag{C.43}$$

and a realization A_t of $\underline{A}$ is defined by:

$$A_t \stackrel{\mathrm{def}}{=} \max_{1 \leq \check{s} \leq N_{\check{s}}} B_{\check{s},t} \tag{C.44}$$

Next, we define the sector activeness random variables $\left(\underline{\zeta}_1, \cdots, \underline{\zeta}_{\check{s}}, \cdots, \underline{\zeta}_{N_{\check{s}}}\right)$ associated with the activeness values of each sector, for any time frame. Let us define the joint random variables:

$$\underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} \stackrel{\mathrm{def}}{=} \left[\underline{\zeta}_1, \cdots, \underline{\zeta}_{N_{\check{s}}}\right]^{\mathrm{T}} \quad \text{and} \quad \underline{\mathbf{B}}_{1:N_{\check{s}}} \stackrel{\mathrm{def}}{=} \left[\underline{B}_1, \cdots, \underline{B}_{N_{\check{s}}}\right]^{\mathrm{T}} \tag{C.45}$$

We assume that the knowledge of the frame state $\underline{A}$ (the two cases mentioned above) is sufficient to determine the pdf of $\underline{\zeta}_{\check{s}}$, and that further interdependences between the activeness values of the various sectors need not to be modelled. This amounts to the following assumption:

$$\boxed{p\left(\underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}},\ \underline{\mathbf{B}}_{1:N_{\check{s}}},\ \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) = \frac{1}{Z} \cdot \prod_{\check{s}=1}^{N_{\check{s}}} p\left(\underline{\zeta}_{\check{s}},\ \underline{B}_{\check{s}},\ \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right)} \tag{C.46}$$

where the normalization factor is $Z = \left[P\left(\underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right)\right]^{N_{\check{s}}-1}$. A graphical model is shown in Figure C.3. The independence between sectors is somewhat realistic, because the sparsity assumption (5.20) eliminates a lot of the spatial leakage between neighboring sectors – as compared to delay-sum beamforming, for example. $\boldsymbol{\Lambda}_{\mathrm{ND}}$ denotes a set of parameter values of the multidimensional model, as formally defined further below, after the model is fully described.

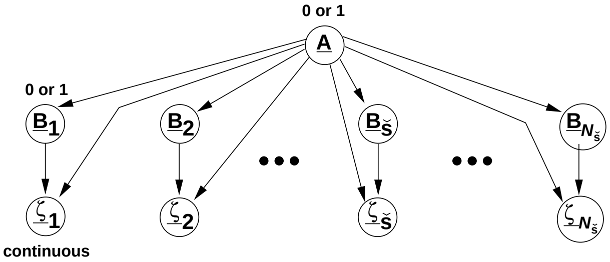


Figure C.3. Graphical model for the independence assumption (C.46) used in the multidimensional model. The r.v. $\underline{A}$ is the frame state (inactive or active) and the r.v. $\underline{B}_{\breve{s}}$ is the state (inactive or active) of a given sector $\mathbb{S}_{\breve{s}}$. The r.v. $\underline{\zeta}_{\breve{s}} \geq 0$ is the activeness of sector $\mathbb{S}_{\breve{s}}$. On an active frame ($\underline{A}= 1$) at least one sector is active ($\exists \breve{s} \quad \underline{B}_{\breve{s}} = 1$).

From (C.46), the following results can be shown without any additional assumption:

$$p\left(\underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} \,\middle|\, \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) = \prod_{\breve{s}=1}^{N_{\breve{s}}} p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{C.47}$$

$$P\left(\underline{B}_{\breve{s}} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\ \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) = P\left(\underline{B}_{\breve{s}} \mid \underline{\zeta}_{\breve{s}},\ \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{C.48}$$

$$P\left(\underline{\mathbf{B}}_{1:N_{\breve{s}}} \mid \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) = \prod_{\breve{s}=1}^{N_{\breve{s}}} P\left(\underline{B}_{\breve{s}} \mid \underline{A},\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{C.49}$$

(C.47), (C.48) and (C.49) are used further below, in the EM derivation (Section C.2.2).

Inactive sector: Dirac + Gamma

Similarly to C.1.1, we model an inactive sector $\underline{B}_{\breve{s}} = 0$ with a (Dirac + Gamma) mixture. Two such mixtures g_{00} and g_{01} are defined, depending on the state of the frame: inactive frame $\underline{A} = 0$ or active frame $\underline{A} = 1$:

$$g_{00}\left(\underline{\zeta}_{\breve{s}}\right) \stackrel{\text{def}}{=} p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}} = 0,\ \underline{A} = 0,\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{C.50}$$

$$\stackrel{\text{def}}{=} v_{00}^{D} \cdot \delta_0(\underline{\zeta}_{\breve{s}}) \quad + \quad v_{00}^{G} \cdot \mathcal{G}_{\gamma_{00},\beta_{00}}\left(\underline{\zeta}_{\breve{s}}\right) \tag{C.51}$$

$$g_{01}\left(\underline{\zeta}_{\breve{s}}\right) \stackrel{\text{def}}{=} p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}} = 0,\ \underline{A} = 1,\ \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{C.52}$$

$$\stackrel{\text{def}}{=} v_{01}^{D} \cdot \delta_0(\underline{\zeta}_{\breve{s}}) \quad + \quad v_{01}^{G} \cdot \mathcal{G}_{\gamma_{01},\beta_{01}}\left(\underline{\zeta}_{\breve{s}}\right) \tag{C.53}$$

We further constrain $\gamma_{00} > 1$ and $\gamma_{01} > 1$ so that $\mathcal{G}_{\gamma_{00},\beta_{00}}(0) = 0$ and $\mathcal{G}_{\gamma_{01},\beta_{01}}(0) = 0$. This way, zero values and strictly positive values are separately modelled by the Dirac and Gamma functions.

Active sector: Shifted Rice

Similarly to Section C.1.1, on an active frame $\underline{A} = 1$, we model an active sector with a shifted Rice pdf, where the shift is equal to the first moment $\gamma_{01} \cdot \beta_{01}$ of the Gamma pdf $\mathcal{G}_{01}$:

$$p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}} = 1,\ \underline{A} = 1,\ \mathbf{\Lambda}_{\mathrm{ND}}\right) \stackrel{\text{def}}{=} g_{11}\left(\underline{\zeta}_{\breve{s}}\right) \quad \text{(C.54)}$$

$$\stackrel{\text{def}}{=} \mathcal{R}_{\sigma_{11},V_{11}}\left(\underline{\zeta}_{\breve{s}} - \gamma_{01} \cdot \beta_{01}\right) \quad \text{(C.55)}$$

Complete model

Let v_0 and v_1 denote the frame-level priors ($v_0 + v_1 = 1$):

$$v_0 \stackrel{\text{def}}{=} P\left(\underline{A} = 0 \mid \mathbf{\Lambda}_{\mathrm{ND}}\right) \quad \text{(C.56)}$$

$$v_1 \stackrel{\text{def}}{=} P\left(\underline{A} = 1 \mid \mathbf{\Lambda}_{\mathrm{ND}}\right) \quad \text{(C.57)}$$

Let v_{01} and v_{11} denote the sector-level conditional priors ($v_{01} + v_{11} = 1$), given that a frame is active:

$$v_{01} \stackrel{\text{def}}{=} P\left(\underline{B}_{\breve{s}} = 0 \mid \underline{A} = 1,\ \mathbf{\Lambda}_{\mathrm{ND}}\right) \quad \text{(C.58)}$$

$$v_{11} \stackrel{\text{def}}{=} P\left(\underline{B}_{\breve{s}} = 1 \mid \underline{A} = 1,\ \mathbf{\Lambda}_{\mathrm{ND}}\right) \quad \text{(C.59)}$$

The set of parameters of the multidimensional model is:

$$\mathbf{\Lambda}_{\mathrm{ND}} \stackrel{\text{def}}{=} \left(v_0,\ v_1,\ v_{01},\ v_{11},\ v_{00}^D,\ v_{00}^G,\ \gamma_{00},\ \beta_{00},\ v_{01}^D,\ v_{01}^G,\ \gamma_{01},\ \beta_{01},\ \sigma_{11},\ V_{11}\right) \quad \text{(C.60)}$$

From (C.47), (C.51), (C.53), (C.55), (C.56), (C.57), (C.58) and (C.59), the complete model can be written as follows.

The priors of the frame state:

$$P\left(\underline{A}\right) = v_0^{\delta_{Kr}(\underline{A}-0)} \cdot v_1^{\delta_{Kr}(\underline{A}-1)} \quad \text{(C.61)}$$

where δ_{Kr} is the Kronecker function (not to be confused with the Dirac pdf δ_0).

The conditional priors of the sector state, for one sector $\mathbb{S}_{\breve{s}}$:

$$P\left(\underline{B}_{\breve{s}} \mid \underline{A}\right) = \delta_{Kr}\left(\underline{B}_{\breve{s}} - 0\right) \cdot \delta_{Kr}\left(\underline{A} - 0\right) + v_{01}^{\delta_{Kr}\left(\underline{B}_{\breve{s}}-0\right)} \cdot v_{11}^{\delta_{Kr}\left(\underline{B}_{\breve{s}}-1\right)} \cdot \delta_{Kr}\left(\underline{A} - 1\right) \quad \text{(C.62)}$$

The likelihood of the data $\underline{\zeta}_{\breve{s}}$ for *one* sector $\mathbb{S}_{\breve{s}}$, given the frame and sector states:

$$\begin{aligned} p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}},\, \underline{A}\right) &= \delta_{Kr}\left(\underline{B}_{\breve{s}} - 0\right) \cdot \delta_{Kr}\left(\underline{A} - 0\right) \cdot \left[v_{00}^{D} \cdot \delta_0\left(\underline{\zeta}_{\breve{s}}\right) + v_{00}^{G} \cdot \mathcal{G}_{\gamma_{00},\beta_{00}}\left(\underline{\zeta}_{\breve{s}}\right)\right] \\ &+ \delta_{Kr}\left(\underline{B}_{\breve{s}} - 0\right) \cdot \delta_{Kr}\left(\underline{A} - 1\right) \cdot \left[v_{01}^{D} \cdot \delta_0\left(\underline{\zeta}_{\breve{s}}\right) + v_{01}^{G} \cdot \mathcal{G}_{\gamma_{01},\beta_{01}}\left(\underline{\zeta}_{\breve{s}}\right)\right] \\ &+ \delta_{Kr}\left(\underline{B}_{\breve{s}} - 1\right) \cdot \delta_{Kr}\left(\underline{A} - 1\right) \cdot \left[\mathcal{R}_{\sigma_{11},V_{11}}\left(\underline{\zeta}_{\breve{s}} - \gamma_{01} \cdot \beta_{01}\right)\right] \end{aligned} \tag{C.63}$$

C.2.2 EM Derivation

E step

In the E-step, we need to estimate the posteriors of the $(\underline{A} = 1)$ and $(\underline{B}_{\breve{s}} = 1,\ \underline{A} = 1)$ events. The posteriors of the other events $(\underline{A} = 0)$ and $(\underline{B}_{\breve{s}} = 0,\ \underline{A} = 1)$ are their respective 1-complements.

The first posterior is directly obtained from the Bayes rule:

$$P\left(\underline{A} = 1 \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) = \frac{p\left(\underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} \,\middle|\, \underline{A} = 1,\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \cdot v_1}{p\left(\underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} \,\middle|\, \underline{A} = 0,\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \cdot v_0 + p\left(\underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} \,\middle|\, \underline{A} = 1,\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \cdot v_1} \tag{C.64}$$

and each of the 3 terms $p\left(\underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} \,\middle|\, \underline{A},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right)$ expands as a product of individual likelihoods, given by (C.47). Note that in the case of a zero value $\underline{\zeta}_{\breve{s}} = 0$, an (indefinite) Dirac term will appear in all 3 terms, hence it simplifies out and only the corresponding (finite) Dirac weights remain.

The second posterior can be obtained from the following decomposition:

$$P\left(\underline{B}_{\breve{s}},\, \underline{A} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) = P\left(\underline{B}_{\breve{s}} \mid \underline{A},\, \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \cdot P\left(\underline{A} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{C.65}$$

which, using (C.48), becomes:

$$P\left(\underline{B}_{\breve{s}},\, \underline{A} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) = P\left(\underline{B}_{\breve{s}} \mid \underline{A},\, \underline{\zeta}_{\breve{s}},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \cdot P\left(\underline{A} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \boldsymbol{\Lambda}_{\mathrm{ND}}\right) \tag{C.66}$$

The last term of the RHS of (C.66) is given by (C.64), and the first term develops into:

$$P\left(\underline{B}_{\breve{s}} \mid \underline{\zeta}_{\breve{s}},\, \underline{A},\, \mathbf{\Lambda}_{\mathrm{ND}}\right) = \frac{p\left(\underline{\zeta}_{\breve{s}},\, \underline{B}_{\breve{s}},\, \underline{A} \,\middle|\, \mathbf{\Lambda}_{\mathrm{ND}}\right)}{p\left(\underline{\zeta}_{\breve{s}},\, \underline{A} \,\middle|\, \mathbf{\Lambda}_{\mathrm{ND}}\right)} \tag{C.67}$$

$$= \frac{p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}},\, \underline{A},\, \mathbf{\Lambda}_{\mathrm{ND}}\right) \cdot P\left(\underline{B}_{\breve{s}} \mid \underline{A},\, \mathbf{\Lambda}_{\mathrm{ND}}\right) \cdot P\left(\underline{A} \mid \mathbf{\Lambda}_{\mathrm{ND}}\right)}{p\left(\underline{\zeta}_{\breve{s}},\, \underline{A} \,\middle|\, \mathbf{\Lambda}_{\mathrm{ND}}\right)} \tag{C.68}$$

$$= \frac{p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}},\, \underline{A},\, \mathbf{\Lambda}_{\mathrm{ND}}\right) \cdot P\left(\underline{B}_{\breve{s}} \mid \underline{A},\, \mathbf{\Lambda}_{\mathrm{ND}}\right)}{p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{A},\, \mathbf{\Lambda}_{\mathrm{ND}}\right)} \tag{C.69}$$

This decomposition is valid for both $(\underline{B}_{\breve{s}} = 0,\ \underline{A} = 1)$ and $(\underline{B}_{\breve{s}} = 1,\ \underline{A} = 1)$ events, hence:

$$P\left(\underline{B}_{\breve{s}} = 1 \mid \underline{\zeta}_{\breve{s}},\, \underline{A} = 1,\, \mathbf{\Lambda}_{\mathrm{ND}}\right) \tag{C.70}$$

$$= \frac{p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}} = 1,\, \underline{A} = 1,\, \mathbf{\Lambda}_{\mathrm{ND}}\right) \cdot P\left(\underline{B}_{\breve{s}} = 1 \mid \underline{A} = 1,\, \mathbf{\Lambda}_{\mathrm{ND}}\right)}{\sum_{B=0}^{1} p\left(\underline{\zeta}_{\breve{s}} \,\middle|\, \underline{B}_{\breve{s}} = B,\, \underline{A} = 1,\, \mathbf{\Lambda}_{\mathrm{ND}}\right) \cdot P\left(\underline{B}_{\breve{s}} = B \mid \underline{A} = 1,\, \mathbf{\Lambda}_{\mathrm{ND}}\right)} \tag{C.71}$$

$$= \frac{g_{11}\left(\underline{\zeta}_{\breve{s}}\right) \cdot v_{11}}{g_{01}\left(\underline{\zeta}_{\breve{s}}\right) \cdot v_{01} + g_{11}\left(\underline{\zeta}_{\breve{s}}\right) \cdot v_{11}} \tag{C.72}$$

M step

To derive the M-step, the KL divergence

$$KL\left[P\left(\underline{B}_{\breve{s}},\, \underline{A} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \mathbf{\Lambda}_{\mathrm{ND}}\right),\, \hat{P}\left(\underline{B}_{\breve{s}},\, \underline{A} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}},\, \hat{\mathbf{\Lambda}}_{\mathrm{ND}}\right)\right] \tag{C.73}$$

can be written similarly to (C.16), where $\mathbf{\Lambda}_{\mathrm{ND}}$ and $\hat{\mathbf{\Lambda}}_{\mathrm{ND}}$ are the current parameters and new parameters, respectively. As in Section C.1.2, from now on we omit $\mathbf{\Lambda}_{\mathrm{ND}}$ whenever possible, using the P and $\hat{P}$ notations to distinguish between current parameter values $\mathbf{\Lambda}_{\mathrm{ND}}$ and new parameter values $\hat{\mathbf{\Lambda}}_{\mathrm{ND}}$. Similarly to Section C.1.2, a lower bound on the likelihood of the observed data can be found using the fact that the KL divergence is always positive:

$$\sum_t \log \hat{p}\left(\underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} = \boldsymbol{\zeta}_t\right) \geq \sum_t \left\langle \log \hat{p}\left(\underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} = \boldsymbol{\zeta}_t,\, \underline{\mathbf{B}}_{1:N_{\breve{s}}},\, \underline{A}\right)\right\rangle_{P\left(\underline{\mathbf{B}}_{1:N_{\breve{s}}},\, \underline{A} \mid \underline{\boldsymbol{\zeta}}_{1:N_{\breve{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.74}$$

where $\boldsymbol{\zeta}_t \stackrel{\mathrm{def}}{=} \left[\zeta_{1,t}, \cdots, \zeta_{N_{\breve{s}},t}\right]^{\mathrm{T}}$ is the vector of activeness values for all sectors, in the time frame t.

From the decomposition:

$$\hat{p}\left(\underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}},\, \underline{\mathbf{B}}_{1:N_{\check{s}}},\, \underline{A}\right) \;=\; \hat{p}\left(\underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} \;\middle|\; \underline{\mathbf{B}}_{1:N_{\check{s}}},\, \underline{A}\right) \cdot \hat{P}\left(\underline{\mathbf{B}}_{1:N_{\check{s}}} \mid \underline{A}\right) \cdot \hat{P}\left(\underline{A}\right), \tag{C.75}$$

the RHS of (C.74) can be decomposed into a sum of 3 terms $\Xi_3 + \Xi_4 + \Xi_5$, where:

$$\Xi_3 \stackrel{\text{def}}{=} \sum_t \left\langle \log \hat{p}\left(\underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t \;\middle|\; \underline{\mathbf{B}}_{1:N_{\check{s}}},\, \underline{A}\right)\right\rangle_{P\left(\underline{\mathbf{B}}_{1:N_{\check{s}}},\, \underline{A} \,\mid\, \underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.76}$$

$$\Xi_4 \stackrel{\text{def}}{=} \sum_t \left\langle \log \hat{P}\left(\underline{\mathbf{B}}_{1:N_{\check{s}}} \mid \underline{A}\right)\right\rangle_{P\left(\underline{\mathbf{B}}_{1:N_{\check{s}}},\, \underline{A} \,\mid\, \underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.77}$$

$$\Xi_5 \stackrel{\text{def}}{=} \sum_t \left\langle \log \hat{P}\left(\underline{A}\right)\right\rangle_{P\left(\underline{\mathbf{B}}_{1:N_{\check{s}}},\, \underline{A} \,\mid\, \underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.78}$$

The aim of the M-step is to find new parameter values $\hat{\boldsymbol{\Lambda}}_{\text{ND}}$ that will maximize the sum $\Xi_3 + \Xi_4 + \Xi_4$, in order to increase the overall likelihood of the observed data $\{\zeta_{\check{s},t}\}$. From (C.47) we obtain:

$$\Xi_3 = \sum_{\check{s},t} \left\langle \log \hat{p}\left(\underline{\zeta}_{\check{s}} = \zeta_{\check{s},t} \;\middle|\; \underline{B}_{\check{s}},\, \underline{A}\right)\right\rangle_{P\left(\underline{\mathbf{B}}_{1:N_{\check{s}}},\, \underline{A} \,\mid\, \underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.79}$$

which can be shown to be equal to:

$$\Xi_3 = \sum_{\check{s},t} \left\langle \log \hat{p}\left(\underline{\zeta}_{\check{s}} = \zeta_{\check{s},t} \;\middle|\; \underline{B}_{\check{s}},\, \underline{A}\right)\right\rangle_{P\left(\underline{B}_{\check{s}},\, \underline{A} \,\mid\, \underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.80}$$

As for Ξ_4, from (C.49) we obtain:

$$\Xi_4 = \sum_{\check{s},t} \left\langle \log \hat{P}\left(\underline{B}_{\check{s}} \mid \underline{A}\right)\right\rangle_{P\left(\underline{B}_{\check{s}},\, \underline{A} \,\mid\, \underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.81}$$

As for Ξ_5, it is directly equal to:

$$\Xi_5 = \sum_t \left\langle \log \hat{P}\left(\underline{A}\right)\right\rangle_{P\left(\underline{A} \,\mid\, \underline{\boldsymbol{\zeta}}_{1:N_{\check{s}}} = \boldsymbol{\zeta}_t\right)} \tag{C.82}$$

Considering the definition of the model (C.61), (C.62) and (C.63), it can be shown that:

- The priors v_0 and v_1 only appear in Ξ_5, under a form similar to w_0 and w_1 in Ξ_2 (Section C.1.2).
- The conditional priors v_{01} and v_{11} only appear in Ξ_4, similarly to w_0 and w_1 in Ξ_1 (Section C.1.2).
- The weights $\left(v_{00}^D,\, v_{00}^G\right)$, $\left(v_{01}^D,\, v_{01}^G\right)$ only appear in Ξ_3, similarly to $\left(w_0^D,\, w_0^R\right)$ in Ξ_1 (Section C.1.2).

- The parameters $\gamma_{00},\ \beta_{00}$ appear only in Ξ_3, not tied to any other parameter.
- The parameters $\gamma_{01},\ \beta_{01},\ \sigma_{11},\ V_{11}$ appear only in Ξ_3 and are tied together in a non-linear fashion, similarly to the Rice and the Shifted Rice in Section C.1.2.

Therefore, a reasoning similar to Section C.1.2 leads to the following update equations.

For the frame-level priors:

$$\hat{v}_0 \quad = \quad \frac{\sum\limits_{\check{s},t} P_{\check{s},t}^{(0)}}{\sum\limits_{\check{s},t} P_{\check{s},t}^{(0)} + \sum\limits_{\check{s},t} P_{\check{s},t}^{(1)}} \quad = \quad \frac{1}{N_{\check{s}} \cdot N_t} \cdot \sum_{\check{s},t} P_{\check{s},t}^{(0)} \qquad \text{and} \quad \hat{v}_1 = 1 - \hat{v}_0 \tag{C.83}$$

For the sector-level conditional priors:

$$\hat{v}_{01} \quad = \quad \frac{\sum\limits_{\check{s},t} P_{\check{s},t}^{(01)}}{\sum\limits_{\check{s},t} P_{\check{s},t}^{(01)} + \sum\limits_{\check{s},t} P_{\check{s},t}^{(11)}} \qquad \text{and} \quad \hat{v}_{11} = 1 - \hat{v}_{01} \tag{C.84}$$

For the “inactive frame, inactive sector” (Dirac + Gamma) mixture weights:

$$\hat{v}_{00}^{D} \quad = \quad \frac{\sum\limits_{\substack{\check{s},t \\ \zeta_{\check{s},t}=0}} P_{\check{s},t}^{(00)}}{\sum\limits_{\substack{\check{s},t \\ \zeta_{\check{s},t}=0}} P_{\check{s},t}^{(00)} + \sum\limits_{\substack{\check{s},t \\ \zeta_{\check{s},t}>0}} P_{\check{s},t}^{(00)}} \qquad \text{and} \quad \hat{v}_{00}^{G} = 1 - \hat{v}_{00}^{D} \tag{C.85}$$

Replacing “00” with “01” in (C.85) gives the update equations for $\hat{v}_{01}^{D}$ and $\hat{v}_{01}^{G}$.

Parameters γ_{00} and β_{00} are updated within the $\{\gamma_{00} > 1,\ \beta_{00} > 0\}$ space through numerical optimization, by maximizing the following sum (e.g. using the simplex method):

$$\sum_{\substack{\check{s},t \\ \zeta_{\check{s},t}>0}} P_{\check{s},t}^{(00)} \cdot \log \mathcal{G}_{\gamma_{00},\beta_{00}}\left(\zeta_{\check{s},t}\right) \tag{C.86}$$

Parameters $\gamma_{01},\ \beta_{01},\ \sigma_{11}$ and V_{11} are updated within the $\{\gamma_{01} > 1,\ \beta_{01} > 0,\ \sigma_{11} > 0,\ V_{11} \geq 0\}$ space, through numerical optimization, by maximizing the following sum:

$$\sum_{\substack{\check{s},t \\ \zeta_{\check{s},t}>0}} P_{\check{s},t}^{(01)} \cdot \log \mathcal{G}_{\gamma_{01},\beta_{01}}\left(\zeta_{\check{s},t}\right) \quad + \quad \sum_{\substack{\check{s},t \\ \zeta_{\check{s},t}>\gamma_{01}\beta_{01}}} P_{\check{s},t}^{(11)} \cdot \log \mathcal{R}_{\sigma_{11},V_{11}}\left(\zeta_{\check{s},t} - \gamma_{01} \cdot \beta_{01}\right) \tag{C.87}$$

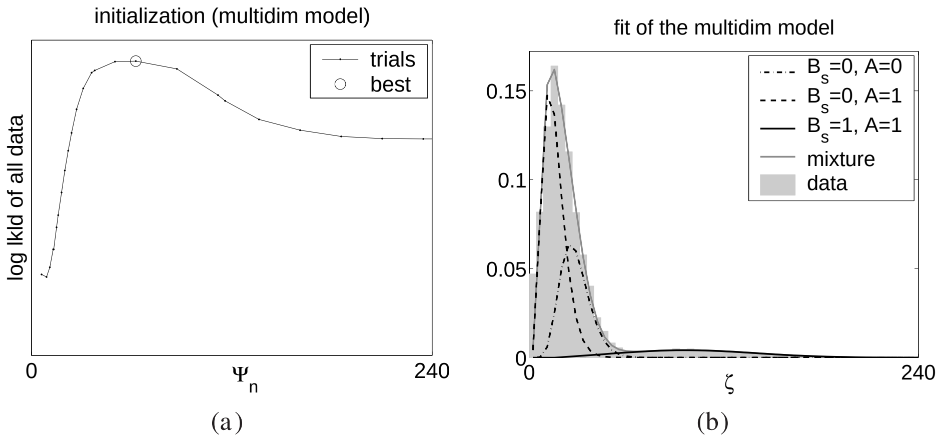


Figure C.4. Fit of the multidimensional model described in Section C.2: (a) automatic initialization, (b) final pdfs $\mathcal{G}_{00}, \mathcal{G}_{01}, \mathcal{R}_{11}$ and the mixture, after convergence of EM.

Implementation Details

EM Implementation: similarly to Section C.1.2, the possibly large data $\{\zeta_{\check{s},t}\}$ (e.g. 70000 frames for 20 minutes) is reduced to a fixed, small number of frames (e.g. 1000), by first ordering the frames $(\boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_t, \cdots, \boldsymbol{\zeta}_T)$ by their maximum value $\max_{\check{s}} (\zeta_{\check{s},t})$, and second picking 1000 frames at equal intervals along the ordered list. This way the computational cost of each EM iteration is drastically reduced, and the cost is independent of the size of the data.

As for the M-step, contrary to the 1-dimensional case, we found in practice that a moment-based approximation produces quite different results after the convergence of EM, than the numerical optimization of the *exact* sums (C.86) and (C.87). Therefore, the above-described data reduction and the numerical optimization are used for all multidimensional results reported in this thesis.

Initialization: an approach similar to Section C.1.2 is used. *For the initialization only*, all data is stacked into 1-dimension, and considered as a mixture of a Gamma with parameters $\gamma^{(\mathrm{init})}$, $\beta^{(\mathrm{init})}$ and a Shifted Rice with shift $\gamma^{(\mathrm{init})} \cdot \beta^{(\mathrm{init})}$. As in Section C.1.2, moment-based methods are used to initialize the Gamma, then the Shifted Rice, within an automatic, multiple initialization approach. Finally both $\mathcal{G}_{00}$ and $\mathcal{G}_{01}$ are initialized with the same parameters $\gamma^{(\mathrm{init})}$, $\beta^{(\mathrm{init})}$. As in the 1-dimensional case, this automatic initialization makes the whole EM fitting process fully deterministic. Figures C.4a and C.4b respectively depict an example of automatic initialization, and the final pdfs after convergence of EM.

Appendix D

Comparison of Detection Features for Localization

This appendix presents an experimental comparison of four different features for speech detection. This task differs from the usual speech/silence classification task, because the purpose is to determine, for each time frame, whether an active speaker can be *correctly localized* or not. The evaluation is conducted using two different azimuth localization methods, GCC-PHAT (Knapp and Carter, 1976) and SRP-PHAT (DiBiase, 2000). Ideally, a “good” feature for localization-oriented speech detection would exhibit smaller localization errors when the detection threshold is more conservative. In the following, we first describe the four detection features and their usage, then the two localization methods GCC-PHAT and SRP-PHAT. The single source recording `seq01` from the AV16.3 Corpus (Chapter 4) was used for evaluation, using the first circular 8-microphone array. In both GCC-PHAT and SRP-PHAT cases, we are not only evaluating each detection feature, but rather its complete integration within an automatic threshold selection system, as in Section 5.3.2. The goal is to determine whether a feature that allows for good classification of *all* time frames (speech and silence) also allows for good localization precision (azimuth error on speech frames only). The selected range of threshold values corresponds to target False Alarm Rate values FAR_T from 1e-14 % to 99.0%. As explained in Section 5.3.2, no training data is needed for this automatic threshold selection strategy. For each of the SNR, energy and SRP-PHAT features, we tried several *types* of models, and picked the one providing a decent fit to the data, as described below. For the SAM-SPARSE-MEAN feature we used the multidimensional approach introduced in Section 5.3.2.

SNR estimate for frame-level detection: The multimicrophone SNR estimate presented in (Chen and Ser, 2000) was implemented to evaluate the instantaneous SNR within each time frame (a non-negative value). The probabilistic model described in Section C.1 is fitted in an unsupervised manner on *all* SNR estimate values, where silence and speech are each modelled with a Rice pdf. Using the fitted model, the posterior probability of having speech is then estimated for each time frame. Based on all estimated posteriors, a threshold on the posterior probability is then selected, corresponding to a given target $\mathrm{FAR_T}$, as in (5.32) and (5.33). Finally, each time frame is classified as “silence” or “speech” by comparing the posterior probability of activity to the threshold.

Energy for frame-level detection: The same type of approach is used as for the SNR estimate. The feature is the log frame energy. Speech and silence are each modelled with a Gaussian pdf.

SRP-PHAT value for frame-level detection: The same type of approach is used as for the SNR estimate. Within each time frame, the multimicrophone location-dependent SRP-PHAT metric defined in (DiBiase, 2000) is maximized, by searching through locations in space. If negative, the obtained maximum value is replaced with zero, thus yielding a feature value between 0 and 1. Speech and silence are each modelled with a Rice pdf, as detailed in Section C.1.

SAM-SPARSE-MEAN for frame-level detection: We used the multidimensional sector-based detection-localization approach described in Section 5.3.2. The frame-level posterior probability of activity was extracted using (C.64), and the threshold was selected as in (5.32) and (5.33).

Localization methods: In order to evaluate the four detection features, we test them as a prior detection step for two different localization methods: GCC-PHAT (Knapp and Carter, 1976) and SRP-PHAT (DiBiase, 2000). GCC-PHAT is implemented using the two squares of microphones defined by the array, as in the FASTTDE implementation described in Section 5.4.5. Whenever none of the two squares produces solvable equations, or when they produce two resulting direction estimates differing by more than 90 degrees, the time frame is dropped. This happened in 13.4 % of the time frames. The SRP-PHAT method (DiBiase, 2000) uses all 8 microphones to find the location in space that maximizes the SRP-PHAT metric. It is always solvable, thus no frame is dropped.

Results are reported in Figures D.1 and D.2. “RMS” stands for Root Mean Square azimuth error. From Figures D.1a and D.2a, energy appears *not* to be an adequate measure for speech detection in the localization context, because increasing the energy threshold to high values does not permit to reach a RMS localization error smaller than 10 degrees. This confirms the study conducted in Section 3.1.3. The other three measures exhibit decent behaviour: in both GCC-PHAT and SRP-PHAT cases, increasing the detection threshold permits to reach a precision below 10 degree in terms of RMS error.

A more detailed analysis is presented in Figures D.1b and D.2b, where only results below 10 degree RMS error are considered. It appears clearly that only SAM-SPARSE-MEAN allows to reduce the RMS error, as the detection threshold is increased. A possible reason for this success is that a high SAM-SPARSE-MEAN value corresponds to a large bandwidth occupied by the speech source. This in turn directly impact on the precision of both GCC-PHAT and SRP-PHAT location estimates.

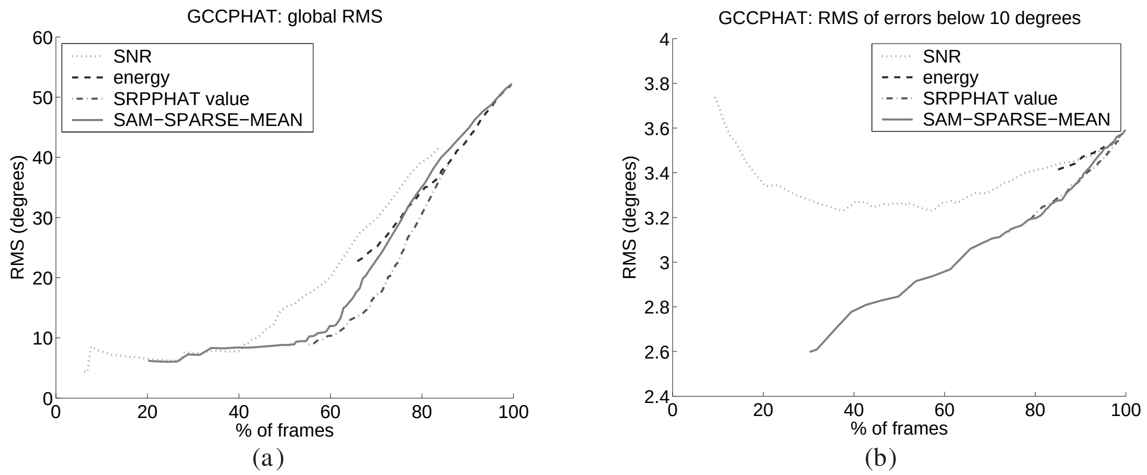


Figure D.1. GCC-PHAT localization: Variation of the RMS azimuth error (in degrees) when the detection threshold is varied. "% of frames" is the proportion of active frames that are above a given value of the detection threshold. In (b), only frames with a localization error below 10 degrees are considered.

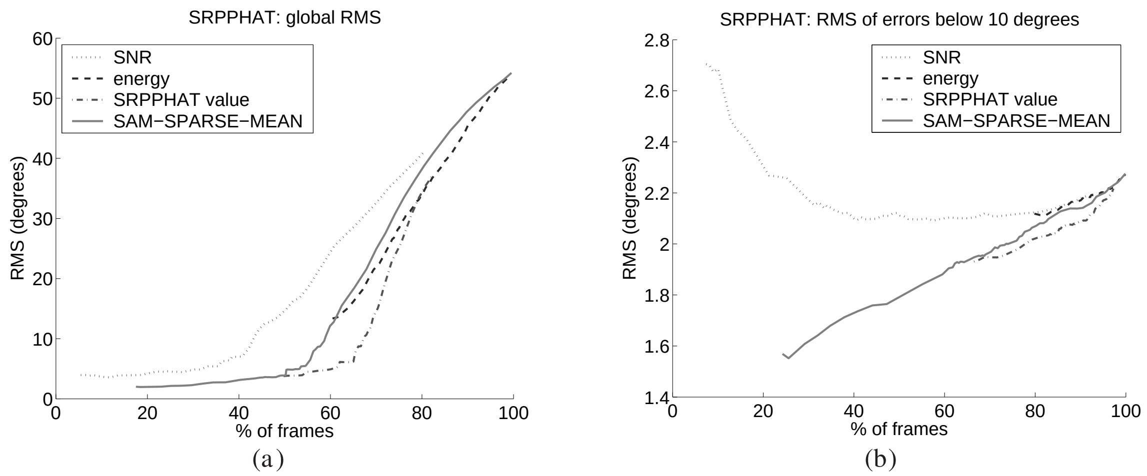


Figure D.2. SRP-PHAT localization: Variation of the RMS azimuth error (in degrees) when the detection threshold is varied. "% of frames" is the proportion of active frames that are above a given value of the detection threshold. In (b), only frames with a localization error below 10 degrees are considered.

Appendix E

Some Analytical Formulas for Single Gaussians

All formulas in this Appendix are valid in any dimensionality $D \in \mathbb{N} \setminus \{0\}$. Assume that N data samples in $\mathbb{R}^D$, denoted $\boldsymbol{\xi}_{1:N} \stackrel{\text{def}}{=} (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_N)$, are modeled with a single Gaussian of mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$. For the covariance matrix $\boldsymbol{\Sigma}$, we use the best unbiased estimate (normalization by $N-1$). The log likelihood of the N data samples, given the single Gaussian model $\mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, can be computed *without the data*, by using the analytical formula:

$$\log p\left(\boldsymbol{\xi}_{1:N} \mid \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\right) = -\frac{N \cdot D}{2} \cdot \log 2\pi - \frac{N}{2} \cdot \log |\boldsymbol{\Sigma}| - \frac{(N-1) \cdot D}{2}$$

Similarly, assume two data sets of N_1 and N_2 samples, each modelled with a single Gaussian, of respective parameters $(\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2,\ \boldsymbol{\Sigma}_2)$. The merge of the two data sets can be modelled with a single Gaussian, whose parameters $\left(\boldsymbol{\mu}_{1+2},\ \boldsymbol{\Sigma}_{1+2}\right)$ can be calculated *without the data*, using the analytical formulas:

$$\boldsymbol{\mu}_{1+2} = \frac{N_1 \cdot \boldsymbol{\mu}_1 + N_2 \cdot \boldsymbol{\mu}_2}{N_1 + N_2}$$

$$\boldsymbol{\Sigma}_{1+2} = \frac{(N_1 - 1) \cdot \boldsymbol{\Sigma}_1 + N_1 \cdot \boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_1^{\mathrm{T}} + (N_2 - 1) \cdot \boldsymbol{\Sigma}_2 + N_2 \cdot \boldsymbol{\mu}_2 \cdot \boldsymbol{\mu}_2^{\mathrm{T}} - (N_1 + N_2) \cdot \boldsymbol{\mu}_{1+2} \cdot \boldsymbol{\mu}_{1+2}^{T}}{N_1 + N_2 - 1}$$

Appendix F

Proof of the Rayleigh-Distributed Magnitude Spectrum

In this section we derive the Rayleigh magnitude-domain silence model of $|X^{(t)}(k)|$ (Section 8.2.1), from a white Gaussian assumption on the pre-emphasized signal $x(t) - 0.97 \cdot x(t-1)$.

First, let us recall a result from (Rice, 1944, 1945). Rice showed that given two zero-mean Gaussian, uncorrelated r.v.s $\underline{A}$ and $\underline{B}$ with same standard deviation σ, and $\underline{R} \stackrel{\text{def}}{=} |\underline{A} + j\underline{B}|$, the r.v. $\underline{R}$ has a Rayleigh pdf:

$$p(R) \quad = \quad \frac{R}{\sigma} \cdot e^{-\frac{R^2}{2\sigma^2}} \quad \text{for} \quad R > 0. \tag{F.1}$$

Let us now define $\underline{\mathbf{x}} \stackrel{\text{def}}{=} [\underline{x}(1), \cdots, \underline{x}(2N_{\mathrm{F}})]^{\mathrm{T}}$, a vector of $2N_{\mathrm{F}}$ uncorrelated[1] zero-mean Gaussian r.v.s $\underline{x}(1), \cdots, \underline{x}(2N_{\mathrm{F}})$ with same standard deviation σ. The DFT of $\underline{\mathbf{x}}$ is: $\underline{\mathbf{X}} \stackrel{\text{def}}{=} \mathbf{F} \cdot \underline{\mathbf{x}}$, where the matrix $\mathbf{F}$ is defined by (2.7).

For $k \in \{1, \cdots, 2N_{\mathrm{F}}\}$, let us define the two r.v.s $\underline{A}(k)$ and $\underline{B}(k)$:

$$\left\{ \begin{array}{lclcl} \underline{A}(k) & \stackrel{\text{def}}{=} & \Re(\underline{X}(k)) & = & \displaystyle\sum_{n=1}^{2N_{\mathrm{F}}} \underline{x}(n) \cdot \cos\left(-\pi \cdot (k-1) \cdot \frac{n-1}{N_{\mathrm{F}}}\right) \\ \underline{B}(k) & \stackrel{\text{def}}{=} & \Im(\underline{X}(k)) & = & \displaystyle\sum_{n=1}^{2N_{\mathrm{F}}} \underline{x}(n) \cdot \sin\left(-\pi \cdot (k-1) \cdot \frac{n-1}{N_{\mathrm{F}}}\right) \end{array} \right. \tag{F.2}$$

[1] Uncorrelation and independence are equivalent for Gaussian r.v.s.

For $k=1$ and $k=N_{\mathrm{F}}+1$ we obtain:

$$\begin{array}{lllll} \underline{A}(1) & = & \sum_{n=1}^{2N_{\mathrm{F}}} \underline{x}(n) & \text{and} & \underline{B}(1) & = & 0 \\ \underline{A}(N_{\mathrm{F}}+1) & = & \sum_{n=1}^{2N_{\mathrm{F}}} \underline{x}(n)\cdot(-1)^{n-1} & \text{and} & \underline{B}(N_{\mathrm{F}}+1) & = & 0 \end{array}$$

For $k \in \{2,\cdots,N_{\mathrm{F}},\ N_{\mathrm{F}}+2,\cdots,2N_{\mathrm{F}}\}$: the r.v. $\underline{A}(k)$ (resp. $\underline{B}(k)$) is a weighted sum of zero-mean, single Gaussian r.v.s, therefore (Delmas, 1993, p. 99) it is also a zero-mean, single Gaussian r.v. with variance:

$$\left\{ \begin{array}{l} \sigma^2_{\underline{A}(k)} = \sigma^2 \cdot \sum_{n=1}^{2N_{\mathrm{F}}} \cos^2\left(\pi\cdot(k-1)\cdot\frac{n-1}{N_{\mathrm{F}}}\right) \\ \sigma^2_{\underline{B}(k)} = \sigma^2 \cdot \sum_{n=1}^{2N_{\mathrm{F}}} \sin^2\left(\pi\cdot(k-1)\cdot\frac{n-1}{N_{\mathrm{F}}}\right) \end{array} \right. \tag{F.3}$$

Given that $\cos^2 u = \frac{1}{2}\cdot(1+\cos 2u)$ and $\sin^2 u = \frac{1}{2}\cdot(1-\cos 2u)$ we can write:

$$\left\{ \begin{array}{l} \sigma^2_{\underline{A}(k)} = \frac{\sigma^2}{2}\left(2N_{\mathrm{F}} + \sum_{n=1}^{2N_{\mathrm{F}}} \cos\left(2\pi\cdot(k-1)\cdot\frac{n-1}{N_{\mathrm{F}}}\right)\right) \\ \sigma^2_{\underline{B}(k)} = \frac{\sigma^2}{2}\left(2N_{\mathrm{F}} - \sum_{n=1}^{2N_{\mathrm{F}}} \cos\left(2\pi\cdot(k-1)\cdot\frac{n-1}{N_{\mathrm{F}}}\right)\right) \end{array} \right. \tag{F.4}$$

Let us now define $\alpha(k) \stackrel{\text{def}}{=} e^{j2\pi\frac{k-1}{N_{\mathrm{F}}}}$. For $k \in \{2,\cdots,N_{\mathrm{F}},\ N_{\mathrm{F}}+2,\cdots,2N_{\mathrm{F}}\}$ we have $\alpha(k) \neq 1$. We also have $[\alpha(k)]^{2N_{\mathrm{F}}} = 1$. Hence:

$$\sum_{n=1}^{2N_{\mathrm{F}}} e^{j2\pi(k-1)\frac{n-1}{N_{\mathrm{F}}}} = \sum_{n=0}^{2N_{\mathrm{F}}-1} [\alpha(k)]^n = \frac{1-[\alpha(k)]^{2N_{\mathrm{F}}}}{1-\alpha(k)} = 0 \tag{F.5}$$

From (F.4), and the real part of (F.5), we conclude that: $\sigma_{\underline{A}(k)} = \sigma_{\underline{B}(k)} = \sigma\sqrt{N_{\mathrm{F}}}$. Similarly, the cross-correlation $\sigma_{\underline{A}(k)\underline{B}(k)} \stackrel{\text{def}}{=} \mathrm{E}\{\underline{A}(\cdot)\,k\underline{B}(k)\}$ can be shown to be zero, using the imaginary part of (F.5) and the uncorrelation hypothesis on the r.v.s $\underline{x}(1),\cdots,\underline{x}(2N_{\mathrm{F}})$. To conclude, we have shown that the r.v.s $\underline{A}(k)$ and $\underline{B}(k)$ are zero-mean, uncorrelated single Gaussian r.v.s of same standard deviation $\sigma\sqrt{N_{\mathrm{F}}}$, therefore the result of Rice applies to $|\underline{X}(k)| = |\underline{A}(k) + j\underline{B}(k)|$:
For $k \in \{2,\cdots,N_{\mathrm{F}},\ N_{\mathrm{F}}+2,\cdots,2N_{\mathrm{F}}\}$, $|\underline{X}(k)|$ has a Rayleigh pdf of parameter $\sigma\sqrt{N_{\mathrm{F}}}$. $\square$

Bibliography

AAAI (2006). Smart rooms, smart houses and households appliances. American Association for Artificial Intelligence (AAAI), http://www.aaai.org/AITopics/html/rooms.html.

Ajmera, J. (2004). *Robust audio segmentation*. PhD thesis, EPFL/LIDIAP.

Ajmera, J., Lathoud, G., and McCowan, I. (2004). Segmenting and clustering speakers and their locations in meetings. In *Proceedings the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ajmera, J., McCowan, I., and Bourlard, H. (2002). Robust HMM-based speech/music segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ajmera, J. and Wooters, C. (2003). A robust speaker clustering algorithm. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Algazi, V., Duda, R., and Thompson, D. (2001). The CIPIC HRTF database. In *Proceedings the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-01)*.

Anguera, X., Wooters, C., Peskin, B., and Aguilo, M. (2005). Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In *Proceedings of the NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation (NIST-RT05s)*.

Baumgarte, F. and Faller, C. (2003). Binaural cue coding – part I: Psychoacoustic fundamentals and design principles. *IEEE Transactions on Speech and Audio Processing*, 11(6).

Benesty, J. (2000). Adaptative eigenvalue decomposition algorithm for passive acoustic source localization. *Journal of the Acoustical Society of America (JASA)*, 107(1):384–391.

Bengio, S., Mariéthoz, J., and Keller, M. (2005). The expected performance curve. In *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2).

Bouguet, J. Y. (2004). Camera Calibration Toolbox for Matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/`.

Bourgeois, J., Freudenberger, J., and Lathoud, G. (2005). Implicit control of noise canceller for speech enhancement. In *Proceedings of Interspeech*.

Brandstein, M. (1995). *A Framework for Speech Source Localization Using Sensor Arrays*. PhD thesis, Brown University.

Brandstein, M. and Ward, D., editors (2001). *Microphone Arrays*. Springer.

Buchner, H., Aichner, R., and Kellerman, W. (2004). TRINICON: A versatile framework for multichannel blind signal processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Buchner, H., Aichner, R., and Kellermann, W. (2005a). Relation between blind system identification and convolutive blind source separation. In *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Piscataway, NJ, USA.

Buchner, H., Aichner, R., Stenglein, J., Teutsch, H., and Kellermann, W. (2005b). Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, PA, USA.

Cerwin, S. (2004). Ears in the sky. *Technology Today*.

Cetin, O. and Shriberg, E. (2006). Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chen, B. and Loizou, P. (2005). Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chen, J., Benesty, J., and Huang, A. (2005). MIMO acoustic signal processing. Invited Talk at the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA).

Chen, J., Benesty, J., and Huang, Y. (2006). Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing, special issue on Advances in Multimicrophone Processing*.

Chen, J. and Ser, W. (2000). Speech detection using microphone array. *Electronic Letters*, 36(2).

Chen, S. and Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the Bayesian information criterion. Technical report, IBM T.J. Watson Research Center.

Chen, S. and Gopalkrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. *IBM Technical Journal*.

Chu, W. and Warnock, A. (2002). Detailed directivity of sound fields around human talkers. Technical Report IRC-RR-104, IRC-CNRC, Canada.

Claudio, E. D. and Parisi, R. (2001). Multi-source localization strategies. In Brandstein, M. and Ward, D., editors, *Microphone Arrays*, chapter 9, pages 181–201. Springer.

Cohen, J. (1989). Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America (JASA)*, 85(6).

Cole, R. A., Fanty, M., Noel, M., and Lander, T. (1994). Telephone speech corpus development at CSLU. In *Proceedings of the International Conference on Speech and Language Processing*.

Conti, S., de Rosny, J., Roux, P., and Demer, D. (2006). Characterization of scatterer motion in a reverberant medium. *Journal of the Acoustical Society of America (JASA)*, 119(2).

Crawford, M., Brown, G., Cooke, M., and Green, P. (1994). Design, collection and analysis of a multi-simultaneous- speaker corpus. In *Proceedings of The Institute of Acoustics*, volume 16, pages 183–190.

Delmas, J.-P. (1993). *Introduction aux Probabilités*. Ellipses Marketing.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

DiBiase, J. (2000). *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments*. PhD thesis, Brown University, Providence RI, USA.

DiBiase, J., Silverman, H., and Brandstein, M. (2001). Robust localization in reverberant rooms. In Brandstein, M. and Ward, D., editors, *Microphone Arrays*, chapter 8, pages 157–180. Springer.

Dielmann, A. and Renals, S. (2004). Multistream dynamic bayesian network for meeting segmentation. In *Proceedings of the workshop on Machine Learning for Multimodal Interfaces (MLMI)*.

Doron, M., Doron, E., and Weiss, A. (1993). Coherent wide-band processing for arbitrary array geometry. *IEEE Transactions on Signal Processing*, 41(1).

Duraiswami, R., Zotkin, D., and Davis, L. (2001). Active speech source localization by a dual coarse-to-fine search. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Dvorkind, T. and Gannot, S. (2005). Speaker localization using the Unscented Kalman Filter. In *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*.

Ellis, D. and Liu, J. (2004). Speaker turn segmentation based on between-channel differences. In *Proceedings the NIST Meeting Recognition Workshop*, Montreal.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

ETSI (2003a). Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced frond-end feature extraction algorithm; Compression algorithms. ETSI standard doc. ES 202 050 V1.1.3.

ETSI (2003b). Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Frond-end feature extraction algorithm; Compression algorithms. ETSI standard doc. ES 201 108 V1.1.3.

Faller, C. and Baumgarte, F. (2003). Binaural cue coding – part II: Schemes and applications. *IEEE Transactions on Speech and Audio Processing*, 11(6).

Friedlander, B. and Weiss, A. (1993). Direction finding for wide-band signals using an interpolated array. *IEEE Transactions on Signal Processing*, 41(4).

Fuchs, J. (2001). On the application of the global matched filter to doa estimation with uniform circular arrays. *IEEE Transactions on Signal Processing*, 49(4).

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J., and Gravier, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of Interspeech*.

Gannot, S., Benesty, J., Bitzer, J., Cohen, I., Doclo, S., Martin, R., and Nordholm, S. (2006). Advances in multimicrophone speech processing - editorial. *EURASIP Journal on Applied Signal Processing, special issue on Advances in Multimicrophone Processing*, 2006:1–3.

Gannot, S., Burshtein, D., and Weinstein, E. (2001). Beamforming methods using nonstationarity with application to speech processing. *IEEE Transactions on Signal Processing*, 49(8):1614–1626.

Gatica-Perez, D., Lathoud, G., McCowan, I., and Odobez, J.-M. (2003). A Mixed-State I-Particle Filter for Multi-Camera Speaker Tracking. In *Proceedings of the IEEE ICCV Workshop on Multimedia Technologies in E-Learning and Collaboration (ICCV-WOMTEC)*, Nice, France.

Gatica-Perez, D., Lathoud, G., Odobez, J.-M., and McCowan, I. (2006). Audio-visual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech and Language Processing*.

Gelbart, D. and Morgan, N. (2001). Evaluating long-term spectral subtraction for reverberant asr. In *Proceedings the IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gemello, R., Mana, F., and Mori, R. D. (2004). A modified Ephraim-Malah noise suppression rule for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian bayesian state estimation. In *IEEE Proceedings*, volume 140, pages 107–113.

Greenstein, L., Michelson, D., and Erceg, V. (1999). Moment-method estimation of the ricean K-factor. *IEEE Communications Letters*, 3(6).

Grenier, Y. (1994). Wideband source location through frequency-dependent modeling. *IEEE Transactions on Signal Processing*, 42(5).

Griffiths, L. and Jim, C. (1982). An Alternative Approach to Linearly Constrained Adaptive Beamforming. *IEEE Transactions on Antennas and Propagations*, 30(1):27–34.

Grinstead, C. M. and Snell, J. L. (1997). *Introduction to Probability*. American Mathematical Society.

Gustafsson, T., Rao, B., and Trivedi, M. (2003). Source localization in reverberant environment: Modeling and statistical analysis. *IEEE Transactions on Speech and Audio Processing*, 11(6):791–803.

Herbordt, W., Kellermann, W., and Nakamura, S. (2004). Joint optimization of LCMV beamforming and acoustic echo cancellation. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*.

Herbordt, W., Trini, T., and Kellermann, W. (2003). Robust spatial estimation of the Signal-to-Interference Ratio for non-stationary mixtures. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*.

Hirsch, H. and Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of the ISCA Tutorial and Research Workshop ASR2000 (ISCA-ITRW-ASR2000)*.

Hoshuyama, O. and Sugiyama, A. (1996). A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix using Constrained Adaptive Filters. In *Proceedings the IEEE International Conferance on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, GA.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Jorge, P., Marques, J., and Abrantes, A. (2004). Estimation of the bayesian network architecture for object tracking in video sequences. In *Proceedings of the International Conference on Image Processing (ICIP).*

Julier, S. and Uhlmann, J. (1997). A new extension of the Kalman filter to nonlinear systems. In *Proceedings of AeroSense: the 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*. Multi Sensor Fusion, Tracking and Resource Management II, SPIE.

Julier, S., Uhlmann, J., and Durrant-Whyte, H. (1995). A new approach for filtering nonlinear systems. In *Proceedings of the 1995 American Control Confernce*, pages 1628–1632.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers (ASME), Journal of basic engineering*, 82:35–45.

Kellermann, W. (1991). A self-steering digital microphone array. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Kinsler, Frey, and et al., C. (1999). *Fundamentals of Acoustics*. John Wiley and Sons Canada Ltd., 4th edition.

Knapp, C. and Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustic, Speech and Signal Processing*, ASSP-24(4):320–327.

Korzybski, A. (1994). *Science and Sanity*. Institute of General Semantics, Fort Worth, TX, USA, 5th edition.

Krim, H. and Viberg, M. (1996). Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine*, 13:67 – 94.

Larocque, J., Reilly, J., and Ng, W. (2002). Particle filters for tracking an unknown number of sources. *IEEE Transactions on Signal Processing*, 50(12).

Lathoud, G., Bourgeois, J., and Freudenberger, J. (2006a). Sector-based detection for hands-free speech enhancement in cars. *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Multimicrophone Speech Processing*.

Lathoud, G. and Magimai.-Doss, M. (2005). A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lathoud, G., Magimai.-Doss, M., and Bourlard, H. (2006b). Channel Normalization for Unsupervised Spectral Subtraction. IDIAP-RR 06-09.

Lathoud, G., Magimai.-Doss, M., and Mesot, B. (2005a). A Spectrogram Model for Enhanced Source Localization and Noise-Robust ASR. In *Proceedings of Interspeech*, Lisbon, Portugal.

Lathoud, G., Magimai.-Doss, M., Mesot, B., and Bourlard, H. (2005b). Unsupervised Spectral Subtraction for Noise-Robust ASR. In *Proceedings the IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Lathoud, G. and McCowan, I. (2003). Location based speaker segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong.

Lathoud, G., McCowan, I., and Odobez, J. (2004). Unsupervised location-based segmentation of multi-party speech. In *Proceedings the NIST Meeting Recognition Workshop*.

Lathoud, G., McCowan, I. A., and Moore, D. C. (2003). Segmenting multiple concurrent speakers using microphone arrays. In *Proceedings of Eurospeech*, Geneva, Switzerland.

Lathoud, G., Odobez, J.-M., and Gatica-Perez, D. (2005c). AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking. In *Proceedings the workshop on Machine Learning for Multimodal Interfaces (MLMI)*.

LaViola, J. (2003). A comparison of Unscented and Extended Kalman Filtering for estimating quaternion motion. In *Proceedings the 2003 American Control Conference*, pages 2435–2440. IEEE Press.

Lehmann, E. (2004). *Particle Filtering Methods for Acoustic Source Localisation and Tracking*. PhD thesis, Australian National University.

Lehmann, E. (2005). Importance sampling particle filter for robust acoustic source localisation and tracking in reverberant environments. In *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Piscataway, NJ, USA.

Leonard, R. (2004). A database for speaker-independent digit recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Levitan, E. and Merhav, N. (2002). A competitive Neyman-Pearson approach to universal hypothesis testing with applications. *IEEE Transactions on Information Theory*, 48(8):2215–2229.

Li, S. (1995). *Markov Random Field Modeling in Computer Vision*. Springer Verlag.

Lu, L. and Zhang, H. J. (2002). Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7).

Mader, A., Puder, H., and Schmidt, G. (2000). Step-Size Control for Acousic Echo Cancellation Filters - An Overview. *Signal Processing*, 80(9):1697–1719.

Martin, R. and Breithaupt, C. (2003). Speech enhancement in the dft domain using Laplacian speech priors. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*.

Matsumoto, M. and Hashimoto, S. (2005). Multiple signal classification by aggregated microphones. *IEICE Transactions Fundamentals*, E88-A(7).

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):305–317.

Moller, M. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6:525–533.

Moon, T. K. and Stirling, W. C. (2000). *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall.

Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing, fourth edition*. Academic Press.

Moore, D. (2002). The IDIAP Smart Meeting Room. IDIAP-COM 02-07, IDIAP.

Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., and Stolcke, A. (2001). The meeting project at ICSI. In *Proc. Human Language Technology Conference*.

NIST (2003). The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan. Technical report.

Nix, J. and Hohmann, V. (2006). Sound source localization in real sound fields based on empirical statistics of interaural parameters. *Journal of the Acoustical Society of America (JASA)*, (1):463–479.

Odobez, J.-M., Gatica-Perez, D., and Ba, S. (2006). Embedding motion in model-based stochastic tracking. *IEEE Transactions on Image Processing*, 15(11).

Oppenheim, A., Schafer, R., and Buck, J. (1999). *Discrete-Time Signal Processing*. Prentice Hall, 2nd edition.

Patterson, E., Gurbuz, S., Tufekci, Z., and Gowdy, J. (2002). Moving talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus. *Eurasip Journal on Applied Signal Processing*, 11:1189–1201.

Perez, P., Hue, C., Vermaak, J., and Gangnet, M. (2002). Color-based probabilistic tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Pham, T. and Fong, M. (1997). Real-time implementation of MUSIC for wideband acoustic detection and tracking. In *Proceedings SPIE AeroSence '97: Automatic Target Recognition VII*.

Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, NJ.

Rauch, H. E., Tung, G., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *Journal of American Institute of Aeronautics and Astronautics*, 3(8):1445Â–1450.

Rice, S. (1944). Mathematical analysis of random noise. *Bell System Technical Journal*, 23:282–332.

Rice, S. (1945). Mathematical analysis of random noise (conclusion). *Bell System Technical Journal*, 24:46–156.

Roweis, S. (2003). Factorial Models and Refiltering for Speech Separation and Denoising. In *Proceedings of Eurospeech*.

Rowland, T. (2002). Generalized function. From MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. `http://mathworld.wolfram.com/GeneralizedFunction.html`.

Roy, R. and Kailath, K. (1989). ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(7):984–995.

Rui, Y., Florencio, D., Lam, W., and Su, J. (2005). Sound source localization for circular arrays of directional microphones. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Schmidt, R. (1986). Multiple Emitter Location and Signal Parameter Estimation. *IEEE Transactions on Antennas and Propagation*, AP-34:276–280.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Schwetz, I., Gruhler, G., and Obermayer, K. (2004). Correlation and stationarity of speech radiation. *IEEE Transactions on Speech and Audio Processing*, 12(5).

Sekiya, T. and Kobayashi, T. (2004). Speech enhancement based on multiple directivity patterns using a microphone array. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Shriberg, E., Stolcke, A., and Baron, D. (2001). Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation and disfluencies, and overlapping speech. In *Proceedings the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding (Prosody-2001)*.

Silverman, H. F., Yu, Y., Sachar, J. M., and III, W. R. P. (2005). Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing*, 13(4).

Smith, K., Ba, S., Odobez, J., and Gatica-Perez, D. (2005). Evaluating multi-object tracking. In *Proceedings of the workshop on Empirical Evaluation Methods in Computer Vision (EEMCV)*.

Smith, S. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 2nd edition.

Sony Corp. (2006). Sony AIBO^{TM}. `http://www.sony.net/Products/aibo/`.

Sorenson, H. (1985). *Kalman Filtering: Theory and Application*. IEEE Press.

Spriet, A. (2004). *Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids*. PhD thesis, Faculty of Engineering, K.U.Leuven, Leuven, Belgium.

Stauffer, C. and Grimson, W. (2000). Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8).

Stoica, P. and Mose, R. (1997). *Introduction to Spectral Analysis*. Prentice-Hall.

Su, G. and Morf, M. (1983). Signal subspace approach for multiple wide-band emitter location. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(12):1502–1522.

Sugiyama, M., Murakami, J., and Watanabe, H. (1993). Speech segmentation and clustering based on speaker features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 395–398.

Svoboda, T. (2003). Multi-Camera Self-Calibration. `http://cmp.felk.cvut.cz/~svoboda/SelfCal`.

Tewfik, A. and Hong, W. (1992). On the application of uniform linear array bearing estimation techniques to uniform circular arrays. *IEEE Transactions on Signal Processing*, 40(4).

Valente, F. (2006). Infinite models for speaker clustering. In *Proceedings of Interspeech*.

Van Compernolle, D. (1989). Noise adaptation in a hidden markov model speech recognition system. *Computer Speech and Language*, 13(1).

van Laarhoven, P. J. M. and Aarts, E. H. L. (1987). *Simulated Annealing: Theory and Practice*. Kluwer Academic Publishers, Dordrecht, Holland.

Vermaak, J. and Blake, A. (2001). Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Vermaak, J., Doucet, A., and Pérez, P. (2003). Maintaining multi-modality through mixture tracking. In *Proceedings ICCV*.

Wang, H. and Kaveh, M. (1985). Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(4).

Wang, L., Kitaoka, N., and Nakagawa, S. (2005). Robust distant speaker recognition based on position dependent cepstral mean normalization. In *Proceedings of Interspeech*.

Ward, D., Lehmann, E., and Williamson, R. (2003). Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions Speech and Audio Processing*, 11(6).

Ward, D. and Williamson, R. (2002). Particle filter beamforming for acoustic source localization in a reverberant environment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Weisstein, E. (2006a). Bell number. From Mathworld–A Wolfram Web Resource. `http://mathworld.wolfram.com/BellNumber.html`.

Weisstein, E. (2006b). Probability space. From Mathworld–A Wolfram Web Resource. `http://mathworld.wolfram.com/ProbabilitySpace.html`.

Welch, G. and Bishop, G. (2004). An introduction to the Kalman filter. TR 95-041, Department of Computer Science, University of North Carolina at Chapel Hill.

Widrow, B. and Stearns, S. (1985). *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.

Williams, G. and Ellis, D. (1999). Speech/music discrimination based on posterior probabilities. In *Proceedings of the European Conference on Speech Communications and Technology*.

Wrigley, S., Brown, G., Wan, V., and Renals, S. (2005). Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing*, 13(1):84–91.

Zotkin, D. and Duraiswami, R. (2004). Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Speech and Audio Process.*, 12(5).

Curriculum Vitae

Guillaume Lathoud
Rossettan 40
CH-1920 Martigny, Switzerland
+41 76 495 4750

30 year old
French
Military service accomplished
glathoud@yahoo.fr

STRONG POINTS

- "Out-of-the-box" thinking & adaptation to a new field.
- Very rigorous engineering practices (math, implementation & evaluation).
- Interaction & collaboration in a multidisciplinary environment.

EXPERIENCES AND SKILLS

2002-06 (4 years 1/2)	IDIAP Research Institute (EPFL/LIDIAP), Martigny, Switzerland. PhD student in speech and signal processing.
Task	Build an integrated system for detection, spatial localization and tracking of multiple persons speaking spontaneously, using arrays of microphones.
Achievements	• Designed & collected a test database which is now used worldwide: `http://mmm.idiap.ch/Lathoud/av16.3_v6` • Proposed robust theoretical & practical solutions (little or no training data), with an emphasis on non-technical end-users. • Successfuly tested the solutions on meeting room recordings, as well as in-car speech acquisition and automatic speech recognition in noise. • Publications: 19 conferences, 3 journals. • IDIAP Best Paper Award 2004. • IDIAP Best PhD Student Award 2006. • Reviewer for IEEE Transactions and Elsevier.
Skills developed	Research, implementation, evaluation, scientific publication & presentation.

1999-01 (2 years 1/2)	National Institute of Standards and Technology, Washington, DC, USA. Java development engineer in the Digital TV team (DASE).
Task	Implement the ATSC set-top-box standard software, and provide feedback during ATSC standardization meetings.
Achievements	• Developed and documented an API to access DTV multimedia content. • Participated to standardization comittees (with Sun, HP, Microsoft). • Provided technical support to users in the USA, in Ireland and South Korea.
Skills developed	Team software development.

EDUCATION

1996-99 (3 years)	Institut National des Télécommunications, Evry, France. Diplôme d'Ingénieur (equiv. to a M.Sc. in Telecomm. and Computer Sc.). • 6-month specialization in Parallel and Distributed Computing. • 1-month student project in geomatics.
1994-1996 (2 years)	Classes Préparatoires aux Grandes Ecoles, Lycée Champollion, Grenoble, France. • Highly intensive studies in math and physics ("Math Sup. & Spé.").
1994	High school diploma with honors ("mention bien").

TECHNICAL SKILLS

General	• Math & physics, digital signal processing, probabilistic modelling. • Microphone array processing, camera calibration, multi-target tracking. • Spontaneous speech & audio-visual processing, noise filtering.
Computer Science	• Design, optimization, implementation & testing of complex algorithms. • OS: UNIX/Linux, Windows. Languages: Java, C/C++, Matlab, Python, LaTeX.

LANGUAGES

French	Mother tongue.
English	Fluent, spoken and written: more than 2 years in the USA + 2 months in Vancouver, Canada (certified work & study program).
German	Enough to travel.
Spanish	Notions.

LEISURE

Sports	Alpinism, climbing, rowing competition in teams (1990-94).
Travels	Europe (Czech & Slovak Rep., Greece, Portugal), USA, Canada, South America.

Publications:

Keywords: Speech, detection, localization, segmentation, tracking, clustering, multimodal, events, summarization, human interaction, audio-visual processing.

Related projects: Segmentation & tracking: IM2 RTMAP, HOARSE, AMI and M4 WP2. Audio-visual corpus: AMI WP4.

Journal Publications:

- G. Lathoud, J. Bourgeois and J. Freudenberger. "Sector-Based Detection for Hands-Free Speech Enhancement in Cars". In the Eurasip special issue on *Advances in Multimicrophone Speech Processing*, 2006.
- D. Gatica-Perez and G. Lathoud and J.-M. Odobez and I. McCowan. "Audio-visual probabilistic tracking of multiple speakers in meetings". In *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.
- I.A. McCowan, D. Gatica-Perez, S. Bengio, and G. Lathoud. "Automatic Analysis of Multimodal Group Actions in Meetings". In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2005.

Conference Publications:

- G. Lathoud, M. Magimai.-Doss, J.M. Odobez and H. Bourlard. "Threshold Selection for Unsupervised Detection, with an Application to Microphone Arrays". In *Proceedings of ICASSP*, 2006.
- G. Lathoud, M. Magimai.-Doss, B. Mesot and H. Bourlard. "Unsupervised Spectral Substraction for Noise-Robust ASR". In *Proceedings of ASRU*, 2005.
- G. Lathoud, M. Magimai.-Doss and B. Mesot. "A Spectrogram Model for Enhanced Source Localization and Noise-Robust ASR". In *Proceedings of Interspeech*, 2005.
- J. Bourgeois, J. Freudenberger and G. Lathoud. "Implicit Control of Noise Canceller for Speech Enhancement". In *Proceedings of Interspeech*, 2005.
- D. Gatica-Perez, G. Lathoud, J.M. Odobez and I. McCowan. "Multimodal Multispeaker Probabilistic Tracking in Meetings". In *Proceedings of ICMI*, 2005.
- D. Gatica-Perez, J.M. Odobez, K. Smith, S. Ba and G. Lathoud. "Tracking People in Meetings with Particles". Paper invited to *WIAMIS'05, Montreux, Switzerland*, 2005.
- G. Lathoud, J. Bourgeois and J. Freudenberger. "Multichannel Speech Enhancement in Cars: Explicit vs. Implicit Adaptation Control". In *Proceedings of the HSCMA'05 Workshop*, 2005.
- G. Lathoud and M. Magimai.-Doss. "A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers". In *Proceedings of ICASSP'05*, 2005.
- G. Lathoud, J.M. Odobez and D. Gatica-Perez. "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking". In *Proceedings of the MLMI'04 Workshop*, 2005.
- D. Zhang, D. Gatica-Perez, S. Bengio, I.A. McCowan and G. Lathoud. "Multimodal Group Action Clustering in Meetings". In *Proceedings of the ACM-VSSN'04 Workshop*, 2004.
- G. Lathoud and I.A. McCowan. "A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays". In *Proceedings of the ISCA-SAPA'04 Workshop*, 2004.

- G. Lathoud, I.A. McCowan, and J.M. Odobez. "Unsupervised Location-Based Segmentation of Multi-Party Speech". In *Proceedings of the NIST-RT04 Workshop*, 2004.
- J. Ajmera, G. Lathoud and I.A. McCowan. "Clustering and Segmenting Speakers and their Locations in Meetings". In *Proceedings of ICASSP'04*, 2004.
- D. Zhang, D. Gatica-Perez, S. Bengio, I.A. McCowan and G. Lathoud. "Modeling Individual and Group Actions in Meetings: a Two-Layer HMM Framework". In *Proceedings of CVPR'04*, 2004.
- D. Gatica-Perez, G. Lathoud, I.A. McCowan and J.M. Odobez. "A Mixed-State I-Particle Filter for Multi-Camera Speaker Tracking". In *Proceedings of the ICCV-WOMTEC Workshop*, 2003.
- G. Lathoud, I.A. McCowan, and D.C. Moore. "Segmenting Multiple Concurrent Speakers Using Microphone Arrays". In *Proceedings of Eurospeech*, 2003.
- D. Gatica-Perez, G. Lathoud, I.A. McCowan, J.M. Odobez and D.C. Moore. "Audio-Visual Speaker Tracking with Importance Particle Filters". In *Proceedings of ICIP*, 2003.
- G. Lathoud and I.A. McCowan. "Location based speaker segmentation". In *Proceedings of ICASSP*, 2003.
- I.A. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D.C. Moore, P. Wellner and H. Bourlard. "Modeling human interactions in meetings". In *Proceedings of ICASSP*, 2003.

Research Reports:

- G. Lathoud, M. Magimai.-Doss and H. Bourlard. "Channel Normalization for Unsupervised Spectral Subtraction". *IDIAP RR 06-09*, 2006.
- G. Lathoud. "Further Applications of Sector-Based Detection and Short-Term Clustering". *IDIAP RR 06-26*, 2006.
- G. Lathoud. "Observations on Multi-Band Asynchrony in Distant Speech Recordings". *IDIAP RR 06-74*, 2006.