

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies
de l'Information et de la Communication*
Spécialité : *Signal, Image and Vision*

Par

Diego DI CARLO

Echo-aware signal processing for audio scene analysis

The Call of Echo

Thèse présentée et soutenue à Rennes, le 04 December 2020

Unité de recherche : IRISA / INRIA

Thèse N° : 88666

Rapporteurs avant soutenance :

Laurent GIRIN Professeur GIPSA-Lab, Grenoble-INP
Simon DOCLO Full professor Carl von Ossietzky Universität, Oldenburg

Composition du Jury :

Président :	Laurent GIRIN	Professeur	GIPSA-Lab, Grenoble-INP
Examinateurs :	Simon DOCLO	Full professor	Carl von Ossietzky Universität, Oldenburg
	Renaud SEGUIER	Professeur	CentraleSupélec, Cesson-Sévigné
	Fabio ANTONACCI	Assistant professor	Politecnico di Milano
Dir. de thèse :	Nancy BERTIN	Chargée de recherche	IRISA, Rennes
Co-dir. de thèse :	Antoine DELEFORGE	Chargée de recherche	Inria Grand Est, Nancy

Abstract

Audio scene analysis aims at retrieving useful information from microphone recordings. Examples of these problems are sound source separation and sound source localization, where we are interested in estimating a speaker's content and position. As humans, we perform these tasks without effort. However, for computers and robots, they are still open challenges. One of the main limitations is that most of the available technologies solve audio scene analysis problems considering either sound semantic or spatial information.

The central theme of this theses is acoustic echoes: the sound propagation elements bridging between semantic and spatial information of sound sources. In particular, as repetitions a source sound, their semantic contribution can be integrated to enhance the sound source. As originated by the interaction with the environment, their path can be backtracked and used to estimate the audio scene's geometry. Based on these observations, recent echo-aware audio signal processing methods have been proposed. However, two are the main questions that arise: how we estimated acoustic echoes, and how we use their knowledge?

This thesis work aims at improving the current state-of-the-art for indoor audio signal processing along these two axes. It also provides new methodologies and data to process acoustic echoes and surpass current approaches' limits. To this end, we present two approaches: A novel approach based on the continuous dictionary framework which does not rely on parameter tuning or peak picking techniques; A deep learning model estimating the time difference of arrival of the first prominent echoes using physically-motivated regularizers. Furthermore, we present a fully annotated dataset specifically designed acoustic echo retrieval, echo-aware application for validating future echo-aware research.

The second part of this thesis regards extending existing methods for audio scene analysis in their echo-aware form. The Multichannel NMF framework for audio source separation, the SRP-PHAT localization methods, and the MVDR beamformer for speech enhancement are re-proposed in their echo-aware version. These applications show how a simple echo model can lead to a boost in performance.

Finally, we want to underline the difficulty related to the tasks of estimating and exploiting acoustic echoes to improve indoor audio processing. Therefore, this thesis consists only of a first attempting work that lays analytical foundations on how to model such problems, and it can serve as a starting point for new exciting directions.

- ▶ KEYWORDS: Acoustic echoes, acoustic echo retrieval, room impulse response estimation; audio scene analysis, room acoustics; audio source separation, room geometry estimation, spatial filtering, sound source localization; deep learning, continuos dictionary.

Résumé en français

L'analyse de la scène audio vise à récupérer des informations utiles à partir d'enregistrements de microphones. La séparation et la localisation de la source sonore sont des exemples de ces problèmes, pour lesquels nous nous intéressons à l'estimation du contenu et de la position d'un locuteur. En tant qu'humains, nous effectuons ces tâches sans effort. Cependant, pour les ordinateurs et les robots, ces tâches restent des défis à relever. L'une des principales limites est que la plupart des technologies disponibles résolvent les problèmes d'analyse des scènes sonores en tenant compte soit de la sémantique du son, soit des informations spatiales.

Le thème central de ces thèses est celui des échos acoustiques : les éléments de propagation du son faisant le pont entre les informations sémantiques et spatiales des sources sonores. En particulier, comme les répétitions d'une source sonore, leur contribution sémantique peut être intégrée pour améliorer la source sonore. Comme ils sont issus de l'interaction avec l'environnement, leur cheminement peut être retracé et utilisé pour estimer la géométrie de la scène sonore. Sur la base de ces observations, des méthodes récentes de traitement des signaux audio tenant compte de l'écho ont été proposées. Cependant, deux questions principales se posent : comment estimer les échos acoustiques et comment utiliser leurs connaissances ?

Ce travail de thèse vise à améliorer l'état actuel de la technique pour le traitement des signaux audio à l'intérieur des bâtiments selon ces deux axes. Il fournit également de nouvelles méthodologies et données pour traiter les échos acoustiques et dépasser les limites des approches actuelles. À cette fin, nous présentons deux approches : Une nouvelle approche basée sur le cadre du dictionnaire continu qui ne repose pas sur des techniques de réglage de paramètres ou de pic de crête; Un modèle d'apprentissage approfondi estimant le décalage temporel de l'arrivée des premiers échos importants à l'aide de régularisateurs physiquement motivés. En outre, nous présentons un ensemble de données entièrement annotées, spécialement conçu pour la récupération d'échos acoustiques, une application prenant en compte l'écho pour valider les futures recherches sur l'écho.

La deuxième partie de cette thèse concerne l'extension des méthodes existantes d'analyse de scènes audio dans leur forme adaptée à l'écho. Le cadre NMF multicanal pour la séparation des sources audio, les méthodes de localisation SRP-PHAT et le formateur de faisceau MVDR pour l'amélioration de la parole sont proposés dans leur version "echo-aware". Ces applications montrent comment un simple modèle d'écho peut conduire à une augmentation des performances.

Enfin, nous voulons souligner la difficulté liée aux tâches d'estimation et d'exploitation des échos acoustiques pour améliorer le traitement audio en intérieur. Cette thèse ne constitue donc qu'une première tentative de travail qui pose des bases analytiques sur la façon de modéliser de tels problèmes, et elle peut servir de point de départ à de nouvelles orientations passionnantes.

- ▶ MOTS-CLÉS : Echos acoustiques, récupération des échos acoustiques, estimation de la réponse impulsionnelle de la pièce ; analyse de la scène sonore, acoustique de la pièce ; séparation des sources audio, estimation de la géométrie de la pièce, filtrage spatial, localisation de la source sonore ; apprentissage approfondi, dictionnaire des continuos.

Résumé étendu en français

- ▶ CE RÉSUMÉ présente en français un aperçu des travaux abordés dans cette thèse. Le thème de l'analyse de la scène audio couvre de nombreuses tâches différentes qui visent à récupérer des informations utiles à partir d'enregistrements microphoniques. Des exemples de ces problèmes sont la séparation des sources sonores et la localisation des sources sonores, où nous nous intéressons à l'estimation du contenu et de la position d'un orateur. En tant qu'humains, nous effectuons ces tâches sans effort : imaginez que quelqu'un nous appelle de l'autre côté de la pièce. Votre réaction typique serait probablement de tourner votre attention vers lui ou même d'aller vers lui. Cependant, pour les ordinateurs et les robots, l'utilisation de techniques de traitement des signaux audio reste un défi à relever.

Les sons transmettent des informations sémantiques (ce que vos amis ont dit), temporelles et spatiales (quand il l'a dit, où il l'a dit). Nous pouvons modéliser ces contributions à l'aide de signaux décrivant le contenu du son et la réponse impulsionale de la pièce, en tenant compte de sa propagation dans l'espace. Certaines méthodes de traitement audio se concentrent sur le premier, ignorant ou décrivant grossièrement le second en raison de la difficulté de l'estimer. Les réponses impulsionales de la pièce intègrent tous les éléments de la propagation du son, tels que les échos, la réflexion diffuse et la réverbération.

Le thème central de ces thèses est l'écho acoustique. Ces éléments de propagation du son créent un pont entre les informations sémantiques et spatiales des sources sonores. Comme ce sont des répétitions et des copies du son source, nous pouvons peser davantage le son cible en intégrant leur contribution que les autres sources de bruit. Comme ces réflexions sont issues de l'interaction du son source avec l'environnement, grâce à leur temps d'arrivée, nous pouvons remonter leur parcours et ainsi reconstruire la géométrie de la scène sonore. Sur la base de ces observations, les méthodes de traitement du signal audio ont commencé à prendre en compte ces éléments de propagation du son pour résoudre le problème de l'analyse de la scène audio. Deux sont les principales questions qui se posent : comment nous avons estimé les échos acoustiques, et comment nous utilisons leurs connaissances ?

Ce travail de thèse vise à améliorer l'état actuel de la technique pour le traitement des signaux audio à l'intérieur des bâtiments selon ces deux axes. En particulier, il fournit de nouvelles méthodologies et données pour traiter les échos acoustiques et dépasser les limites des approches actuelles. Deuxièmement, il prolonge les méthodes classiques d'analyse de scènes audio dans leur forme adaptée à l'écho. Ces deux revendications sont développées dans les deux parties principales de la thèse, qui suivent après une introduction, comme résumé ci-dessous.

► PARTIE I, L'ACOUSTIQUE DES SALLES RENCONTRE LE TRAITEMENT NUMÉRIQUE DES SIGNAUX

Tout d'abord, nous donnons quelques définitions préliminaires du rôle du traitement du signal audio et énumérons quelques problèmes fondamentaux qui seront abordés tout au long de la thèse, à savoir la récupération de l'écho acoustique, la séparation des sources audio, la localisation des sources sonores, l'estimation de la géométrie de la pièce.

- Le chapitre 2 construira un premier pont important : de l'acoustique au traitement des signaux audio. Il définit d'abord le son, sa propagation dans l'environnement et l'écho de ses origines.
- Dans 3, nous passons de la physique au traitement numérique du signal où les échos sont modélisés comme des éléments de filtres, appelés Réponses Impulsionnelles de la Chambre (RIR), opérant sur le signal source. Comme le traitement dans le domaine temporel natif est compliqué, nous présentons la représentation de Fourier, qui facilite à la fois l'exposition des méthodes et la mise en œuvre des algorithmes.

Ce chapitre clôture la première partie introductive.

► PARTIE II - ACUSTIC ECHO ESTIMATION

Dans cette deuxième partie de la thèse, nous nous intéressons à l'estimation des premiers échos acoustiques à partir d'enregistrements microphoniques. Basée sur les modèles et la définition décrits dans la première partie, cette partie comprend d'abord un aperçu général des méthodes de récupération des échos, suivi de la présentation de deux travaux publiés lors de conférences internationales et d'un ensemble de données sur le point d'être publié.

- Tout d'abord, dans le chapitre 4, nous fournissons au lecteur des connaissances sur l'état de l'art de la récupération des échos acoustiques, à savoir comment estimer les propriétés des échos acoustiques. Après avoir présenté le problème, nous passons en revue la littérature selon la taxonomie typique utilisée dans le traitement du signal. Afin de fournir un aperçu complet de la récupération des échos acoustiques, certains ensembles de données et mesures d'évaluation récurrents dans la littérature et utilisés dans le chapitre suivant sont présentés. Les trois chapitres suivants présentent trois travaux que nous avons menés sur l'estimation de l'écho acoustique.
- Le chapitre 5 présente une nouvelle approche pour estimer les échos d'un enregistrement stéréophonique d'une source sonore inconnue telle que la parole. Contrairement aux méthodes existantes, elle s'appuie sur le récent cadre du dictionnaire continu et ne repose pas sur des techniques de réglage des paramètres ou de picorage. La précision et la robustesse de la méthode sont évaluées sur des configurations simulées difficiles avec des niveaux de bruit et de réverbération variables et sont comparées à deux méthodes de pointe. L'évaluation expérimentale sur des données synthétiques montre que des taux de récupération comparables ou légèrement inférieurs sont observés pour la récupération de sept échos ou plus. En revanche, de meilleurs résultats sont obtenus

pour un nombre d'échos inférieur, et la nature hors réseau de l'approche donne généralement des erreurs d'estimation plus faibles. Néanmoins, cela est prometteur puisque l'avantage pratique de connaître le moment où quelques échos par canal sont récupérés sera démontré dans la dernière partie de la thèse.

- Dans le chapitre 6, nous déployons des techniques d'apprentissage approfondi pour estimer les propriétés des échos acoustiques. À notre connaissance, il s'agit d'un des premiers exemples dans ces directions. La méthode proposée présente des points communs avec les techniques d'apprentissage approfondi déjà appliquées dans la localisation de sources sonores. Nous présenterons trois architectures différentes qui abordent le problème de l'estimation des échos acoustiques avec un ordre de complexité croissant : l'estimation du temps d'arrivée de la voie directe et des premiers échos proéminents; l'exécution de cette estimation de manière plus robuste; et enfin, l'extension à un nombre croissant d'échos.
- Enfin, pour conclure cette deuxième partie, dans 7, nous décrivons un ensemble de données que nous avons recueillies, spécifiquement conçu pour l'estimation de l'écho acoustique. Cet ensemble de données comprend des mesures de la réponse impulsionale multicanaux de la pièce (RIR), y compris des annotations des premiers échos et des positions 3D des microphones et des sources réelles et d'images sous différentes configurations de murs dans une pièce cubique. Ces données fournissent un nouvel outil pour l'évaluation comparative des méthodes récentes de traitement des signaux audio *echo-aware* et des utilitaires logiciels permettant d'accéder, de manipuler et de visualiser facilement les données.

► PARTIE III - ECHO-AWARE APPLICATION

La troisième et dernière partie de la thèse concerne les applications de traitement audio où la connaissance des premiers échos peut améliorer les performances par rapport aux méthodes standard. Pour l'occasion, nous supposons que les propriétés des échos sont disponibles *a priori* et nous construisons nos connaissances préalables. La structure de cette partie suit le format de la précédente.

- Un chapitre d'introduction (chapitre 8) rassemble les définitions standard et présente les approches actuelles de pointe pour le traitement audio en intérieur sous une même enseigne. Nous considérons trois problèmes fondamentaux : la séparation des sources audio, la localisation des sources sonores, le filtrage spatial et l'estimation de la géométrie de la pièce. Ces problèmes sont présentés tour à tour avec la revue de la littérature correspondante, en mettant en évidence les défis actuels. Ces problèmes particuliers seront les protagonistes des trois chapitres suivants, présentés sous leur forme échos.
- En chapitre 9, les échos sont utilisés pour améliorer les performances des méthodes classiques de séparation des sources audio et sont le

résultat d'une collaboration avec d'autres collègues, publiée lors d'une conférence internationale. Nous proposons notamment une interprétation physique des échos, à savoir des microphones d'image, qui permet de mieux comprendre comment les algorithmes tirent parti de leurs connaissances. Notre étude porte sur deux variantes du cadre de séparation des sources par factorisation matricielle non négative multicanaux : l'une qui utilise uniquement les amplitudes des fonctions de transfert et l'autre qui utilise les phases. Les résultats montrent que l'approche proposée bat sa variante vanille en n'utilisant que quelques échos et que les échos permettent la séparation là où elle était jugée inabordable.

- Le chapitre 10 aborde le problème de la localisation des sources audio dans le contexte de forts échos acoustiques. En utilisant le modèle de microphones image présenté dans le chapitre précédent, nous montrons que ces contributions parasites peuvent être utilisées pour modifier la manière classique dont la localisation de la source est effectuée. En particulier, nous montrons que dans un scénario simple impliquant deux microphones proches d'une surface réfléchissante et d'une source, l'approche proposée est capable d'estimer les angles d'azimut et d'élévation, tâche impossible en supposant une propagation idéale, comme le font les approches classiques. Ces résultats ont été fusionnés dans une publication, publiée lors d'une conférence internationale. En outre, l'étude est ensuite étendue aux réseaux de microphones à capteurs multiples et aux données du monde réel, fournies par une collaboration avec l'équipe de recherche de Honda.
 - Le chapitre 11 présente deux applications sensibles à l'écho qui peuvent bénéficier de l'ensemble de données dEchorate, présenté dans 7. Nous illustrons l'utilisation de ces données en considérant deux problèmes possibles d'analyse de scènes audio : le filtrage spatial conscient de l'écho et l'estimation de la géométrie de la pièce. Afin de valider les données et de montrer leur potentiel, des algorithmes de pointe bien connus sont utilisés. Par conséquent, pour chacune des applications, les méthodes envisagées sont contextualisées et résumées. Les résultats numériques confirmant la valeur de cet ensemble de données pour la communauté du traitement des signaux audio. L'ensemble de données et ces méthodes seront rendus publics afin que les contributeurs externes soient invités à les utiliser pour développer des méthodes de traitement audio plus robustes.
- LA DERNIÈRE PARTIE?? comprend le dernier chapitre (Chapitre ??), qui récapitule les principaux résultats présentés dans ce manuscrit et les perspectives liées à ce travail. Parmi ceux-ci, nous montrons comment peu d'échos acoustiques peuvent être estimés à partir de la seule observation d'enregistrements microphoniques comportant de la parole réverbérante en utilisant l'un ou l'autre modèle dérivé de la physique de la propagation du son et des modèles d'apprentissage profond formés sur des simulateurs acoustiques. De plus, nous démontrons les avantages d'inclure la connaissance des échos acoustiques dans les méthodes de traitement du son. Pour l'aspect lié à l'évaluation des méthodes tenant

compte des échos dans un scénario réel, nous préconisons que les ensembles de données de référence disponibles gratuitement manquent actuellement dans la littérature. Par conséquent, dans l'esprit de la recherche ouverte, nous construisons un nouvel ensemble de données qui sera bientôt publié. Ces données sont accompagnées d'annotations précises et d'outils algorithmiques pour une recherche consciente de l'écho, couvrant une grande partie des applications pour l'analyse des scènes audio.

Enfin, nous voulons souligner la difficulté liée à la tâche d'estimation et d'exploitation des échos acoustiques pour améliorer le traitement audio à l'intérieur. Cette thèse ne consiste donc qu'en une première tentative de travail qui pose des bases analytiques sur la façon de modéliser de tels problèmes. Comme toutes les premières investigations, beaucoup de choses peuvent être améliorées, et nous espérons qu'elle pourra servir de point de départ à de nouvelles recherches intéressantes et stimulantes.

Acknowledgements

Contents

ABSTRACT	iii
RÉSUMÉ EN FRANÇAIS	v
RÉSUMÉ ÉTENDU FRANÇAIS	vii
ACKNOWLEDGEMENTS	xii
CONTENTS	xv
NOTATIONS	xix
1 PROLOGUE	1
1.1 Audio Signal Processing	2
1.2 Audio Scene Analysis	3
1.3 Echo-aware approach	4
1.4 Thesis Outline and Main Contributions	5
1.5 List of Contribution	9
1.6 Don't Panic!	10
I ROOM ACOUSTIC MEETS SIGNAL PROCESSING	11
2 ELEMENTS OF ROOM ACOUSTICS	13
2.1 Sound wave propagation	13
2.2 Acoustic reflections	16
2.3 Room acoustics and room impulse response	19
2.4 Perception and some acoustic parameters	27
3 ELEMENTS OF AUDIO SIGNAL PROCESSING	29
3.1 Signal model in the time domain	29
3.2 Signal model in the spectral domain	33
3.3 Other (room) impulse response spectral models	40
II ACOUSTIC ECHO RETRIEVAL	44
4 ACOUSTIC ECHO RETRIEVAL	47
4.1 Problem Formulation	47
4.2 Taxonomy of Acoustic Echo Retrieval methods	48
4.3 Literature Review	49
4.4 Data and Evaluation	57
4.5 Proposed Approaches	60
5 blaster: KNOWLEDGE-DRIVEN ACOUSTIC ECHO RETRIEVAL	62
5.1 Introduction	62
5.2 Signal model	63
5.3 Background on On-grid Blind Channel Estimation	64
5.4 Proposed Approach	66
5.5 Experiments	70
5.6 Conclusion	73

6 lantern: DATA-DRIVEN ACOUSTIC ECHO RETRIEVAL	74
6.1 Introduction	74
6.2 Deep Learning Model	76
6.3 Learning the first echo	78
6.4 Robustly learning the first echo	81
6.5 Conclusion	84
7 dechorate: DATASETS FOR ACOUSTIC ECHO ESTIMATION	86
7.1 Introduction	86
7.2 Database realization	87
7.3 Dataset annotation	89
7.4 The dEchorate package	92
III ECHO-AWARE APPLICATION	94
8 AUDIO SCENE ANALYSIS MEETS SIGNAL PROCESSING	97
8.1 Audio Scene Analysis Problems	97
8.2 Overview on Multichannel Sound Source Separation	99
8.3 Overview on Spatial filtering	102
8.4 Overview on Sound Source Localization	104
8.5 Conclusion	106
9 separake: ECHO-AWARE SOUND SOURCE SEPARATION	107
9.1 Literature review in Echo-aware Audio Source Separation	107
9.2 Modeling	110
9.3 Source Separation by NMF	111
9.4 Echo-aware Source Separation	113
9.5 Numerical Experiments	113
9.6 Conclusion	117
10 mirage: ECHO-AWARE SOUND SOURCE LOCALIZATION	118
10.1 Literature review in Echo-aware Sound Source Localization	119
10.2 Proposed Approach	119
10.3 Background in microphone array SSL	120
10.4 Microphone Array Augmentation with Echoes	123
10.5 Experimental Results	124
10.6 Conclusion	130
11 ECHO-AWARE APPLICATIONS OF dechorate	132
11.1 Echo-aware Spatial Filtering	132
11.2 Room Geometry Estimation	140
11.3 Conclusions and Perspectives	142
IV EPILOGUE	145
12 ECHO-AWARE REFLECTIVE REFLECTION	147
12.1 Looking Back	147
12.2 Looking Ahead	148
SLIDING FRANK-WOLFE ALGORITHM & NON-NEGATIVE BLASSO	152
APPENDICES	152

DERIVATION OF THE ACOUSTIC WAVE	155
RIR AND RT60 MEASUREMENTS	158
RIR estimation	158
RT60 estimation	158
BIBLIOGRAPHY	159
BIBLIOGRAPHY	159

MIRAGE	Microphone Augmentation with Echoes	8
SEPARAKE	Sound Separation by Raking Echoes	
BLASTER	Blind and Sparse Technique for Echo Retrieval	7
LANTERN	LeArNing echo regression from room TransfEr functioNs ...	80
DECHORATE	Dataset dechorated by echoes	58
CASA	Computational Auditory Scene Analysis	42
SOTA	State of the Art	21
GA	Geometrical (room) acoustics	18
FEM	Finite Element Method	21
BEM	Boundary Element Method	21
FDTD	Finite-Difference-Time-Domain	21
DWM	Digital Waveguide Mesh	21
ISM	Image Source Method	20
TOA	Time of Arrival	26
RIR	Room Impulse Response	5
RTF	Room Transfer Function	33
FIR	Finite Impulse Response	54
ATF	Acoustic Transfer Function	20
AIR	Acoustic Impulse Response	20
TF	Time-Frequency	23
RT₆₀	Reverberation Time	75
SE	Speech Enhancement	56
SSL	Sound Source Localization	7
RooGE	Room Geometry Estimation	47
AER	Acoustic Echo Retrieval	5
FT	Fourier Transform	33
DFT	Discrete Fourier Transform	34
DTFT	Discrete-Time Fourier Transform	34
STFT	Short Time Fourier Transform	6
FFT	Fast Fourier Transform	39
ReIR	Relative Impulse Response	41
ReTF	Relative Transfer Function	41
ILD	Interchannel Level Difference	42
IPD	Interchannel Phase Difference	42
TDOA	Time Difference of Arrival	42
TDOAs	Time Differences of Arrival	
AWGN	Additive White Gaussian Noise	32

Glossary:

- A list of terms in a particular domain of knowledge with their definitions.
- From Latin *glossarium* “collection of glosses”, diminutive of *glossa* “obsolete or foreign word”.

DRR	Direct to Reverberant Ratio	27
AER	Acoustic Echo Retrieval	5
MLS	Minimum Length Sequence	49
ESS	Exponential Sine Sweep	50
ML	Maximum Likelihood	51
MUSIC	Multiple Signal Classification	51
ESPRIT	Estimation of Signal Parameters via Rational Invariance Techniques 51	
SSL	Sound Source Localization	7
RooGE	Room Geometry Estimation	47
JADE	Joint Angle and Delay Estimation	53
DOA	Direction of Arrival	53
DOAs	Directions of Arrival	54
SIMO	Single Input Multiple Output	54
BCE	Blind Channel Estimation	54
BSI	Blind System Identification	54
EM	Expectation Maximization	54
MULAN	Multichannel Annihilation	57
FRI	Finite Rate of Innovation	57
ASR	Finite Rate of Innovation	57
RMSE	Root Mean Square Error	59
NPM	Normalized Projection Misalignment	59
NMF	Nonnegative Matrix Factorization	54
CD	Continous Dictionary	7
LASSO	Least Absolute Shrinkage and Selection Operator	64
BLASSO	Beurling-LASSO	68
BSN	Blind Sparse Nonnegative Channel Identification	66
MDS	Multi-Dimensional Scaling	90
MU	Multiplicative Updates	111
CASA	Computational Auditory Scene Analysis	42
WSJ	Wall Street Journal	89
MIMO	Multiple Input Multiple Output	108
SDR	Signal to Distortion Ratio	116
SIR	Signal to Interference Ratio	116
DS	Delay-and-Sum	102
MVDR	Minimum-Variance-Distortionless-Response	102
LCMV	Linearly-Constrained-Minimum-Variance	102
SNR	Signal-to-Noise-Ratio	56

SINR	Signal-to-Interference-plus-Noise-Ratio	137
SNRR	Signal-to-Noise-plus-Reverberation-Ratio	140
MaxSINR	Maximum SINR	102
MaxSNR	Maximum SNR	102
PSD	Power Spectral Density	134
ReETF	Relative Early Transfer Function	138
SRMR	Speech-to-Reverberation-energy-Modulation Ratio	139
PESQ	Perceptual Evaluation of Speech Quality	139
GEVD	Generalized Eigenvalue Decomposition	138
NN	Neural Network	106
DNN	Deep Neural Network	74
GMM	Gaussian Mixture Model	75
GLLM	Gaussian Locally-Linear Mapping	75
CC	Cross Correlation	121
GCC	Generalized Cross Correlation	105
PHAT	Phase Transform	
GCC-PHAT	Generalized Cross Correlation with Phase Transform	81
SRP-PHAT	Steered Response Power with Phase Transform	79
SLAM	Source Localization and Mapping	119
iTDOA	Image TDOA	80
TDOE	Time Difference of Echo	80
AOA	Angle of Arrival	121
MLP	Multi-layer Perceptron	76
nRMSE	normalized Root Mean Squared Error	80
CNN	Convolutional Neural Network	75
RNN	Recurrent Neural Network	76
MDN	Mixture Density Network	82
STD	Standard Deviation	126
ReLU	Rectified Linear Unit	76
MSE	Mean Squared Error	80

Notations

LINEAR ALGEBRA

x, X	scalars
\mathbf{x}, \mathbf{x}	vectors
x_i	i -th entry of \mathbf{x}
$\mathbf{0}_I$	$I \times 1$ vector of zeros
\mathbf{x}^\top	transpose of the vector \mathbf{x}
\mathbf{x}^H	conjugate-transpose (hermitian) of the vector \mathbf{x}
$\text{Re}[x]$	real part scalar (vector) x (\mathbf{x})
$\text{Im}[x]$	imaginary part scalar (vector) x (\mathbf{x})
i	imaginary unit
\mathbb{N}	set of natural numbers
\mathbb{R}	set of real numbers
\mathbb{R}_+	set of real positive numbers
\mathbb{C}	set of complex number

COMMON INDEXING

i	microphone or channel index in $\{0, \dots, I - 1\}$
j	source index in $\{0, \dots, J - 1\}$
r	reflection (echo) in $\{0, \dots, R - 1\}$
t	continuous sample index
n	discrete sample index in $0, \dots, N - 1\}$
f	continuous frequency index
k	discrete frequency index in $\{0, \dots, K - 1\}$
l	discrete time-frame index $\{0, \dots, L - 1\}$
τ	tap index in $\{0, \dots, T - 1\}$

GEOMETRY

$\underline{\mathbf{x}}_i$	3D location of microphone i recording $x_i(t)$
$\underline{\mathbf{x}}_i$	3D position of the microphone i recording $x_i(t)$
$\underline{\mathbf{s}}_j$	3D position of the source j emitting $s_j(t)$
$d_{ii'}$	distance between microphone i and i'
q_{ij}	distance between microphone i and source j
$\underline{\mathbf{s}}_j$	3D location of (target) point source j emitting $s_j(t)$
$\underline{\mathbf{q}}_j$	3D location of (interfering) point source j emitting $q_j(t)$
r_j	distance of source j wrt to the array origin
θ_j	azimuth of source j wrt to the array origin
φ_j	elevation of source j wrt to the array origin

SIGNALS

x_i	input signal recorded at microphone i
\mathbf{x}	$I \times 1$ multichannel input signal, i.e. $\mathbf{x} = [x_0, \dots, x_{I-1}]$
\mathbf{X}	matrix of multichannel input signals
s_j	(target) point source signal j
q_j	(interfering) point source signal j
c_{ij}	spatial image source j as recorded at microphone i
a_{ij}	acoustic impulse response from source j to microphone i
h_{ij}	generic filter from source j to microphone i
n_i	(white or distortion) noise signal at microphones i
u_i	generic interfering and distortion noise signal at microphone i
ε_i	generic noise signal due to mis- or under-modeling i

ACOUSTIC

α_r	attenuation coefficient at reflection r
β_r	reflection coefficient at reflection r
τ_r	time location of the reflection r
c_{air}	speed of sound in air
T	temperature
H	relative humidity
p	sound pressure
h_{ij}	Room Impulse Response between source j to microphone i

MATHEMATICAL OPERATION

- ★ cross-correlation
- ⊗ generalized cross-correlation
- * convolution

EXAMPLES

Acoustic Impulse Response for single source scenario:

$$a_i(t) = \sum_{r=0}^{R_i} \frac{\alpha_{ir}}{4\pi c_{\text{air}} \tau_{ir}} \delta(t - \tau_{ir}) \quad (1)$$

Acoustic Transfer Function for single source scenario:

$$a_i(f) = \sum_{r=0}^{R_i} \frac{\alpha_{ir}}{4\pi c_{\text{air}} \tau_{ir}} e^{-j2\pi f \tau_{ir}} \quad (2)$$

Time of Arrival between source and microphone

$$\tau_{ij} = \frac{\|\underline{\mathbf{x}}_i - \underline{\mathbf{s}}_j\|}{c_{\text{air}}} \quad (3)$$

1

Prologue

- IN A NUTSHELL, this Ph. D. thesis is about acoustic ECHOES.
We live immersed in a complex acoustical world, where every concrete thing can sound, resound, and echo. For humans, it is difficult to internalize what is sound, its constituents, and its generation. It is processed by our auditory systems and brain so efficiently that our attention is detached from the physical laws governing it. Therefore, when listening to something, we typically focus directly on its *semantic content*. Evolution lead us to conduct this process without any efforts, despite the presence of a high level of background noise, for instance during a concert. This outstanding capability is not limited to humans and is common to many the creatures we are sharing the physical world with.

Nonetheless, we process many other information of the complex *acoustic scenes* we are immersed into. In addition to the semantic content, a sound also conveys *temporal* and *spatial* information. For instance, the ticking of a metronome or clock provides units of time and when hearing someone shouting, we unconsciously know where to turn our attention. However, this information is determined by how sound *propagates* in space and not by the source itself.

Before reaching the ears, sound propagates in all directions and a portion of its energy arrives at us directly, others indirectly after being reflected around. This process leads to the creation of *echoes* and *reverberation*. Typical examples are the echoes produced by large rocky mountains or walls in monumental buildings, such as the Panthéon in Rome or the Pont de Neuilly in Paris.¹ In common language, echoes refer to distinctive reflected sounds which can be heard, thus, characterized by a specific *time of arrival* and *attenuation*. In smaller environments, echoes are still present but are typically less perceived as they arrive more quickly and densely. What is perceived here is the so-called reverberation, for which large empty rooms or churches are great examples.

Some animals evolved to “see” thanks to echoes. Two of the most striking examples are bats and whales which use them for navigation and hunting. By emitting sound patterns and listening to their reflections returned from the environment, these animals scan the surrounding space, identifying and locating objects. Here, the echoes are voluntarily produced and this is referred to as *active echo-location* or (bio) sonar. In contrast, in *passive* echo-location,

ECHOES
ECHOES

“Echoes” shows the direction that we’re moving in.”

—David Gilmour,
about the making of “The Dark Side Of The Moon”

Resources:

-  Testing The World's Longest Echo
-  SKUNK BEAR : What Does Sound Look Like?
-  ARTE : La Magie Du Son
-  Daniel Kish: How I use sonar to navigate the world

For his experiments, Galileo Galilei was measuring time using the sound of a metronome.

¹ “Écho. Citer ceux du Panthéon et du pont de Neuilly.” — Gustave Flaubert, *Dictionnaire des idées reçues*.

external, unknown sound sources are used instead. “Locating it” means estimating its delay with respect to the direct sound. These delays are then processed as distances in the brain, in the same way our grandparents taught us to localize a storm by counting the time between a lightning and its thunder. That is how bats and whales find preys, see obstacles, and orientate themselves in dark caves or deep seas. However, the term “echo-location” here could be misleading as it may solely refer to the only problem of locating objects. As we will discuss later, the application of echoes goes beyond simple localization. Therefore, in this thesis, we will prefer the terminology of *echo estimation*.

Remarkable examples of passive echo estimation in nature are not very well known. Sand scorpions use the propagation of vibrations in the sand to follow the movement of other insects in the dark night. By using their 8 legs as a radar, they perform passive (seismic) echo-location with inevitable consequences for the prey. This technique is common to spiders who sense reverberation in their complex web². They are not only able to localize the preys fast, but also to identify them, and disambiguate them from simple objects moved by the wind or malicious visitors. In this case, instead of emitting sounds, evolution taught them to use complex structures (for scorpions their legs, for spiders webs) in order to feed and survive.

Echoes do not only serve for computing distances or localizing preys. For instance, they make speech more intelligible, improve our sense of orientation and balancing[Huggett 1953, Chapter 5.4], and provide music with “dimensionality”[Sacks 2014]. This phenomenon is a branch of study in *room acoustics*, *psychoacoustics* and *sound design*. In particular, the former study acoustic echoes for designing theatres, auditoriums and meeting rooms, , with the aim of a good listening quality.

The problems addressed in this thesis are indicated in the thesis title: *Echo-aware signal processing for audio scene analysis*. There are three parts in the sentence that deserve an explanation: *echo-aware*, *signal processing* and *audio scene analysis*. In order, we will first elaborate the last two as they contextualize this thesis, and after, we will explain why and how echoes help.

1.1 AUDIO SIGNAL PROCESSING

Signal processing is the process of analyzing and modifying a *signals*, which are mathematical representations of quantities carrying information about a phenomenon. In *audio signal processing*, this phenomenon is sound. Typical signals are speech or music, and various mathematical and algorithmic techniques have been developed to process them. There are multiple reasons to do this, such as producing new signals with higher quality or and retrieving high-level information that the signal carry. To this end, complex systems are built which can be represented as a collection of simpler subsystems, with well-defined tasks, interacting with each other. In (audio) signal processing, these subsystems roughly fall into four categories: *representation*, *enhance-*

This technique is developed instinctively by some blind people as well. By tapping their canes or clicking their tongues, they are able for instance to avoid obstacles when walking. In the 18th century, French philosopher Denis Diderot recorded the this incredible ability, which was labeled as “echo-location” only 300 years later by Donald Griffin. [Kolarik et al. 2014].

² According to some recent studies, spiders appear to offload cognitive tasks to their webs. The web may then act then as a complex system processing and filtering the information, which is then returned to their owner. [Sokol 2017]

Audio is a used as a technical term, referring to sound coming from a recording and transmission. Acoustic, instead, refer to the physical aspect of sound.

ment, estimation, and adaptive processing. Many related problems can be then decomposed into blocks which include one or more of the following steps.

- ▶ **REPRESENTATION.** The signals can be represented in many different ways, so that the *information* they contain becomes more easily accessible for specific tasks. It is generally implemented through *change of domain* or *feature extraction*. In audio, the most widely used representation is the Fourier basis, which changes the signal domain from time to frequencies.
- ▶ **ENHANCEMENT.** Observed signals may be affected by distortions, which corrupts and hides the relevant information. Examples of these are measurement and quantization noise, noisy sources, reverberation, clipping, etc. Therefore, signal enhancement, namely, reducing any form of distortion, is typically a necessary step. Practical examples of enhancement are removing background noise from a mobile phone recording, isolating instrument tracks from in a song, etc.
- ▶ **ESTIMATION.** Often we wish to estimate some key properties of the target signal, which may be used as inputs to a different algorithm. For instance, we may be interested in estimating a speaker's position in a recording, the time of arrival of an echo, or the pitch of a sound.
- ▶ **ADAPTIVE PROCESSING.** It deals with algorithms that are able to adapt themselves to the data. Typically process the data on the fly and are controlled by variable parameters resulting from previous estimation blocks. They usually rely optimization of an objective function designed to meet specific requirements. Examples of these algorithms are present in noise-canceling headphones or echo cancellation modules implemented in video conference call systems.

1.2 AUDIO SCENE ANALYSIS

Pay attention to what are you listening now: there might be music, someone talking to you, footsteps echoing in the other room, background noise due to cars, a heating system, maybe rain or wind, the sound of your movements, and many others. Everything you hear now as well as its location in space is what is called the *audio scene*³. Therefore, *audio scene analysis* is trivially the analysis of it. More specifically, the extraction and organization of all the information contained in the sound associated with an audio scene.

In audio signal processing, this process involves using algorithmic and mathematical tools to retrieve and organize such information. After recording the audio scene with microphones, complex systems, as described above, are used to access the information. Accessing different types of information at different levels of complexity leads to the definition of different *problems*. These problems focus on well-defined tasks and some are referred to with established names. [Table 8.1](#) lists some selected audio scene analysis problems that will be considered later in this thesis.

³ The correct terminology for it is *auditory scene*, which relates to human perception. Psychologist Albert Bregman in [Bregman 1990] coined it. However, we will use this terminology since we extend this concept to audio signal processing, and as it is commonly accepted in the literature.

From the ancient greek, analysis means dismantling into constituent elements. It allows then to reach information otherwise obfuscated by the big picture. It is opposed to synthesis, which instead combines parts into a whole.

Problems	<i>From the recordings, can we...</i>
Audio Source Separation	... estimate the audio signal of sound sources?
Audio Source Enhancement	... estimate the audio signal of a target sound source?
Sound Source Localization	... estimate the positions of sounds sources?
Microphone Calibration	... positions and properties of the microphones??
Room Geometry Estimation	... estimate the shape of the room?
Acoustic Echo Estimation	... estimate the echoes' properties?
Acoustic measurement	... estimate physical properties related to sound propagation?
Source Identification	... estimated the type of source signal?
Speech Diarization	... who is speaking and when ?
Source Counting	... count the number of speaker s?
Automatic Speech Recordings	... the content of the speech ?

TABLE 1.1: List of selected audio scene analysis problems. The one above the line are considered in this thesis.

Without getting too philosophical, it is possible to re-cast these problems to some (simple) human interrogations:

- *What?* Answered by Audio Source Separation and Enhancement, Automatic Speech Recognition, and Source Identification, operating on the source signals' semantic content;
- *Where?* Answered by Sound Source Localization, Microphone Calibration, and Room Geometry Estimation, by elaborating the spatial information of the sound propagation;
- *When?* Answered by Speech Diarization, or leveraging the sound temporal information;

Our brain and the auditory system can instantly and often effortlessly solve these problems, such that they may seem like trivial tasks. However, they hide many difficult challenges when it comes to designing efficient and robust algorithms. Moreover, most of these problems may exhibit strong interconnections, and the solution of one of them depends on the solution of others. For instance, knowing when someone is speaking and its location in the room, sound source separation can be achieved more easily. This should not be surprising since it bears a strong parallelism with our everyday experience.

Similarly, echoes may help audio signal processing

1.3 ECHO-AWARE APPROACH

As proven by natural behaviors, acoustic echoes are important for humans and animals to analyze the audio scene: As repetitions of a sound, they convey information about that sound. As characterized by temporal instants and attenuation related to distances, they convey spatial information about the audio scene. As modified by the frequency description of the object that generates them, they convey acoustic information about it.

This observation motivated many researchers to include echoes in signal processing applications, not only limited to audio⁴. However, it was not always the case. Many audio scene analysis methods make limiting assumptions on of the sound propagation to derive efficient algorithms. One of the most common

“Everything is connected”
—Douglas Adams, *Dirk Gently’s Holistic Detective Agency*

⁴ The idea of integrating reflection in models is also studied in other fields of engineering. In telecommunication and networking, for instance, where these phenomena are referred to as *multipath propagation*.

ones is the so-called *anechoic* or *free-field* scenario, assuming neither echoes nor reverberation is present in the audio scene. Even if this assumption can be seen as reasonable in some scenarios, it is easy to understand its underlying limitations when applied to real-world recordings. Furthermore, in some cases, echoes are considered a source of noise and interference and then modeled as something to cancel out.

Instead, some researchers proposed to explicitly include acoustic echoes in leading to what we will refer to as *echo-aware methods*. Some of the earliest examples in this direction are the works of Flanagan et al [Flanagan et al. 1993; Jan et al. 1995; Jan and Flanagan 1996] for in source enhancement. However, only recently, these methods have regained interest for audio processing as manifested by the European project SCENIC [Annibale et al. 2011] and the UK research S³A project. In some recent studies, echoes are used to boost performances of typical audio scene analysis problems, e.g., speech enhancement [Dokmanić et al. 2015; Kowalczyk 2019], sound source localization [Ribeiro et al. 2010a], and source separation [Scheibler et al. 2018c; Leglaive et al. 2016; Remaggi et al. 2019], and room geometry estimation from sound [Remaggi et al. 2016; Dokmanić et al. 2013; Crocco et al. 2017].

All these methods show the importance and the benefits of modeling acoustic reflection. However, prior to all them is the Acoustic Echo Retrieval (**AER**) problem. This step, which is most often given for granted in the above applications, is extremely challenging, as shown throughout this entire thesis work.

1.4 THESIS OUTLINE AND MAIN CONTRIBUTIONS

The goal of this thesis is to improve the current state-of-the-art for indoor audio signal processing along two axes:

“Sometimes a scream is better than a thesis.”
—Ralph Waldo Emerson

1. Provide new methodologies and data to process and estimate acoustic echoes and surpass the limits of current approaches.
2. Extend previous classical methods for audio scene analysis by incorporating the knowledge of echo properties in sound propagation models.

These two goals are elaborated in the two main parts of the thesis, which follow after an introductory one, as summarized below. However the parts are largely interconnected, as shown in [Figure 1.1](#).

► PART I ROOM ACOUSTIC MEETS SIGNAL PROCESSING

Chapter 2 This chapter builds a first important bridge: from acoustics to audio signal processing. It first defines sound and how it propagates in the environment § 2.1, presenting the fundamental concepts of this thesis: echoes in § 2.2 and Room Impulse Responses (RIRs) in § 2.3. By assuming some approximations, RIRs will be described by parts in relation to methods to compute them. Finally, in § 2.4, how the human auditory system perceives reverberation will be reported.

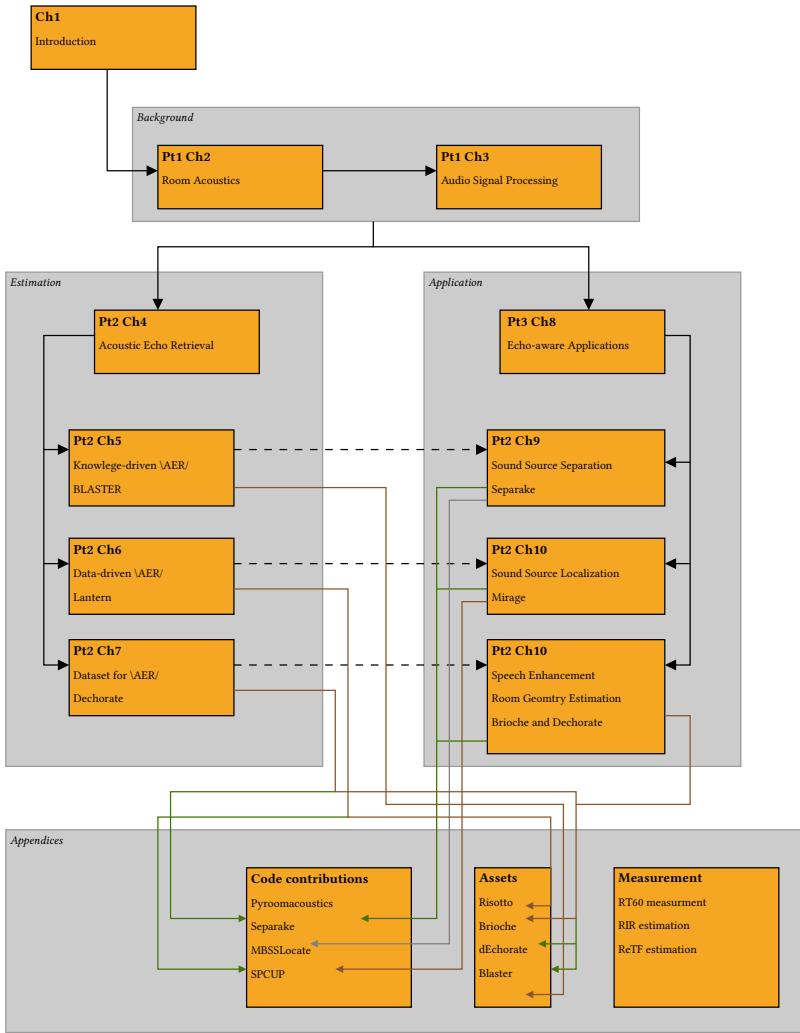


FIGURE 1.1: Schematic organization of the thesis, dependencies between chapters linked to author contributions.

Chapter 3 Let us now move from physics to digital signal processing. At first, in § 3.1, this chapter formalizes fundamental concepts of audio signal processing such as signals, mixtures, filters and noise in the time domain. In § 3.2, we will present a fundamental signal representation that we will use throughout the entire thesis: the Short Time Fourier Transform (**STFT**). Finally, in § 3.3, some essential models of the **RIR** are described.

► PART II ACOUSTIC ECHOES ESTIMATION

This part focuses on how to estimate echoes from the only observation of microphone recordings.

Chapter 4 This chapter aims to provide the reader with knowledge in the state-of-the-art of Acoustic Echo Retrieval (**AER**). After presenting the **AER** problem in § 4.1, the chapter is divided into three main sections: § 4.2 defines the categories of methods according to which the literature can be clustered and analyzed in details later in § 4.3. Finally, in § 4.4 some datasets and evaluation metrics for **AER** are presented.

Chapter 6 In this chapter, we will use virtually supervised deep learning models to learn the mapping from microphone recordings to the echoes' timings. After presenting a quick overview of deep learning techniques (6.2), we will present a first simple model to estimate the echo coming from a close surface (6.3). This case is motivated by an application in Sound Source Localization (**SSL**), which will be discussed in detail in Chapter 10. Finally, we will present a possible way to achieve a more robust estimation (??) and discuss a possible way to scale this approach to multiple echoes (??).

Chapter 5 This chapter proposes a novel approach for *off-grid AER* from a stereophonic recording of an unknown sound source such as speech. In order to address some limitation of existing methods, we propose a new approach, named Blind and Sparse Technique for Echo Retrieval (**BLASTER**). It builds on the recent framework of Continous Dictionary (**CD**), and it does not rely on parameter tuning nor peak picking techniques by working directly in the parameter space of interest. The method's accuracy and robustness are assessed on challenging simulated setups with varying noise and reverberation levels and are compared to two state-of-the-art methods. While comparable or slightly worse recovery rates are observed for recovering seven echoes or more, better results are obtained for fewer echoes, and the off-grid nature of the approach yields generally smaller estimation errors.

Chapter 7 This chapter presents dEchorate: a new database of measured multichannel room impulse response (RIRs) including annotations of early echoes and 3D positions of microphones, real and image sources under different wall configurations in a cuboid room. These data provide a tool for benchmarking recent methods in *echo-aware* speech enhancement, room geometry estimation, RIR estimation, acoustic echo retrieval, microphone calibration, echo labeling, and reflectors estimation. The database is accompanied by software utilities to easily access, manipulate, and visualize the data and baseline methods for echo-related tasks.

► PART III ECHO-AWARE AUDIO SCENE ANALYSIS

In this part, we present some audio scene analysis problems that will be later discussed in their echo-aware extension.

Chapter 8 In this chapter, we will present a selection of algorithms and methodologies which we identified as potential beneficiaries of echo-aware additions. At first, in section § 8.1, we present a typical scenario that highlights some cardinal problems. Then, in section ??, state-of-the-art approaches to address these problems are listed and commented, highlighting the relationship with some acoustic propagation models. The content presented here serves as a basis for a deeper investigation conducted in each of the following chapters.

Chapter 9 In this chapter, echoes are used for improving performance of classical Audio Source Separation methods. First, we describe existing

methods that either ignore the acoustic propagation or attempt to estimate it fully. Instead, this work investigates whether sound separation can benefit from the knowledge of early acoustic echoes. These echoes are derived from the known locations of a few *image microphones*. The improvements are shown for two variants of a method based on non-negative matrix factorization: one that uses only magnitudes of the transfer functions and one that uses the phases as well. The experimental part shows that the proposed approach beats its vanilla variant by using only a few echoes and that with magnitude information only, echoes enable separation where it was previously impossible.

Chapter 10 This chapter addresses the problem of audio source localization in the context of strong acoustic echoes. Classic [SSL](#) methods are deceived by strong acoustic echoes: rather than estimating the location of a true source, they might be fooled by its strong correlated echos. Instead, we show early-echo characteristics can, in fact, benefit [SSL](#). To this end, we introduce the concept of microphone array augmentation with echoes Microphone Augmentation with Echoes ([MIRAGE](#)), using the model of image microphones presented in the previous chapter. In particular, we show that in a simple scenario involving two microphones close to a reflective surface and one source, the proposed approach can estimate both azimuthal and elevation angles, an impossible task assuming an ideal propagation, as classical approaches do. Later, the proposed approach is extended to multichannel recording and tested on real data scenario.

Chapter 11 This chapter presents two echo-aware applications that can benefit from the dataset [dEchorate](#). In particular, we exemplify the utilization of these data considering two possible use-cases: echo-aware speech enhancement ([§ 11.1](#)) and room geometry estimation ([§ 11.2](#)). This investigation is conducted using state-of-the-art algorithms described and contextualized in the corresponding sections. In the final section ([§ 11.3](#)), the main results are summarized, and future perspectives are presented.

- ▶ FINALLY, the dissertation concludes with [Chapter 12](#), which summarizes the contributions and raises several additional research questions.

1.5 LIST OF CONTRIBUTION

This dissertation draws heavily on the work and writing of the following papers, written jointly with several collaborators:

- **Di Carlo, Diego**, Pinchas Tandeitnik, Sharon Gannot, Antoine Deleforge, and Nancy Bertin (2021). “dEchorate: a calibrated Room Impulse Response database for acoustic echo retrieval”. In: *Work in progress*
- **Di Carlo, Diego**, Clement Elvira, Antoine Deleforge, Nancy Bertin, and Rémi Gribonval (2020). “Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 156–160
- **Di Carlo, Diego**, Antoine Deleforge, and Nancy Bertin (2019). “Mirage: 2D source localization using microphone pair augmentation with echoes”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 775–779
- Deleforge, Antoine, **Di Carlo, Diego**, Martin Strauss, Romain Serizel, and Lucio Marcenaro (2019). “Audio-Based Search and Rescue With a Drone: Highlights From the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]”. In: *IEEE Signal Processing Magazine* 36.5, pp. 138–144
- Lebarbenchon, Romain, Ewen Camberlein, **Di Carlo, Diego**, Clément Gaultier, Antoine Deleforge, and Nancy Bertin (2018a). “Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge”. In: *Proc. of LOCATA Challenge Workshop-a satellite event of IWAENC*
- Scheibler, Robin, **Di Carlo, Diego**, Antoine Deleforge, and Ivan Dokmanić (2018d). “Separake: Source separation with a little help from echoes”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901

1.6 DON'T PANIC!

The reader will have already noticed that a large margin is left free on each manuscript page. We will use it to insert personal comments, historical notes, additional insights, and figures and tables to complete each subject. This graphic template is inspired by the work of Tufte and Graves-Morris [Tufte and Graves-Morris 1983]⁵. We emphasize the presence of clickable links by the ☰ logo. Code libraries are written in typewriter font, e. g. dEchorate.

⁵ The colophon of the thesis reports more information on the template.

1.6.1 Quick vademe^{cum}

For the readers:

- orange is used for clickable internal reference, such as for sections § 1.1 and acronyms FFT;
- grey and ☰ is used for clickable external link, such as my website ☰;
- reference sidenotes on the margin are used as footnotes, providing additional insights;
- italic sidenotes and figures without proper reference numbers on the margin are meant to provide optional information and can be read in a second moment;
- ▶ should capture the reader attention towards the important points;
- ? indicates a question-and-answer paragraph;
- ⇛ indicates the presence of definitions by dichotomy;
- * denoted the conclusive paragraph of a chapter;

1.6.2 The golden ratio of the thesis

This thesis has been written the following personal stylistic rules:

- after the colon punctuation mark, “:”, small caps. In case of list, items are delimited by semicolon punctuation mark “;”;
- sidenotes number referring to a word superscripts the word itself, sidenotes referring to a whole sentence superscripts the corresponding full stop;
- N'ilazo color palette: black, grey and orange;
- at most three levels of sub-headings: section, subsection, and Tufte's *new-thought* [Tufte and Graves-Morris 1983] to capture attention;
- the usage of dichotomies is privileged when presenting concepts and definitions, thus they are emphasized;
- each section should be introduced briefly at the end of the previous one to promote reading flow;
- no indentation, but well-separated text blocks;

Part I

ROOM ACOUSTIC MEETS SIGNAL PROCESSING

2 ELEMENTS OF ROOM ACOUSTICS

2.1	Sound wave propagation	13
2.1.1	The acoustic wave equation	14
2.1.2	... and its Green solution	15
2.2	Acoustic reflections	16
2.2.1	Large smooth surfaces, absorption and echoes	18
2.2.2	Diffusion, scattering and diffraction of sound	19
2.3	Room acoustics and room impulse response	19
2.3.1	The room impulse response	20
2.3.2	Simulating room acoustics	21
2.3.3	The method of images and the image source model	24
2.4	Perception and some acoustic parameters	27
2.4.1	The perception of the RIR's elements	27
2.4.2	Mixing time	28
2.4.3	Reverberation time	28
2.4.4	Direct-to-Reverberant ratio and the critical distance	28

3 ELEMENTS OF AUDIO SIGNAL PROCESSING

3.1	Signal model in the time domain	29
3.1.1	The mixing process	30
3.1.2	Noise, interferer and errors	32
3.2	Signal model in the spectral domain	33
3.2.1	Discrete time and frequency domains	34
3.2.2	The DFT as approximation of the FT	35
3.2.3	Signal models in the discrete Fourier domain	36
3.2.4	Time-Frequency domain representation	38
3.2.5	The final model	40
3.3	Other (room) impulse response spectral models	40
3.3.1	Steering vector model	41
3.3.2	Relative transfer function and interchannel models	41

2

Elements of Room Acoustics

- ▶ **SYNOPSIS** This chapter builds a first important bridge: from acoustics to audio signal processing. It first defines sound and how it propagates in the environment § 2.1, presenting the fundamental concepts of this thesis: echoes in § 2.2 and Room Impulse Responses (RIRs) in § 2.3. By assuming some approximations, RIRs will be described by parts in relation to methods to compute them. Finally, in § 2.4, how the human auditory system perceives reverberation will be reported.
The material on waves and acoustic reflection is digested from classic texts on room acoustics [Kuttruff 2016a; Pierce 2019] and on partial differential equations [Duffy 2015].

2.1 SOUND WAVE PROPAGATION

According to common dictionaries and encyclopedias,

sound is the sensation caused by the vibration of air and perceived by the ear.

This definition highlights two aspects of sound: a physical one, characterized by the air particles vibration; and a perceptual one, involving the auditory system. Focusing on the former phenomenon, when vibrating objects excites air, surrounding air molecules starts oscillating, producing zones with different air densities leading to a compressions-rarefactions phenomenon. Such a vibration of molecules takes place in the direction of the excitement, with the next layer of molecules excited by the previous one. Pushing layer by layer forward, a *longitudinal mechanical wave*⁶ is generated. Notice that therefore sound needs a medium to travel: it cannot travel through a vacuum and no sound is present in outer space.

Thus, sound propagates through a medium, which can be solid, liquid or gaseous. The propagation happens at a certain speed which depends on the physical properties of the medium, such as its density. The medium assumed throughout the entire thesis is air, although extensions of the developed methods to other media could be envisioned. Under the fair assumption of air being homogeneous and steady, the speed of sound can be approximated as follows:

$$c_{\text{air}} = 331.4 + 0.6T + 0.0124H \quad [\text{m/s}], \quad (2.1)$$

where T is the air temperature [$^{\circ}\text{C}$] and H is the relative air humidity [%].

“*Sound, a certain movement of air.*”
—Aristotele, De Anima II.8 420b12



Imagine a calm pond. The surface is flat and smooth. Drop a rock into it. *Kerploop!* The surface is now disturbed. The disturbances propagate, as waves. The medium here is the water surface.

⁶As opposed to mechanical vibrations in a string or (drum) membrane, acoustic vibrations are *longitudinal* rather than *transversal*, i.e. the air particles are displaced in the same direction of the wave propagation.

The air pressure variations at one point in space can be represented by a *waveform*, which is a graphical representation of a sound, as shown in Figure 2.2.

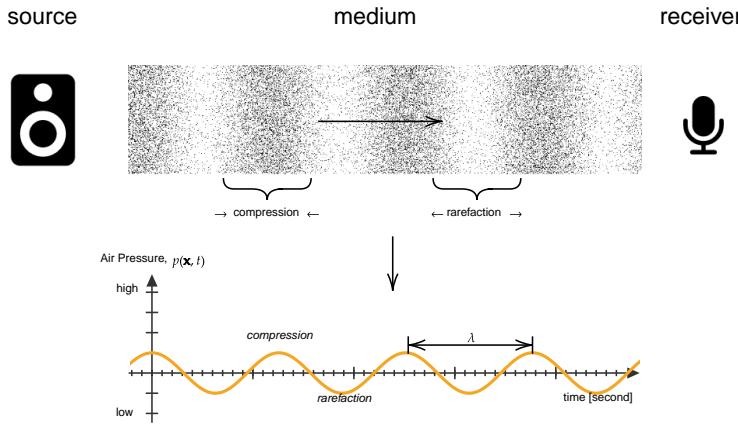


FIGURE 2.2: Illustration of the molecules under sound compression and rarefaction due to longitudinal sound wave and its waveform representation.

We can think of this process in the light of the classic *source-medium-receiver* model of communication theory: the *source* is anything that emits waves⁷, the *medium* carries the waves from one point to another, and the *receiver* receives them.

2.1.1 The acoustic wave equation

The acoustic wave equation is a second-order partial differential equation⁸ which describes the evolution of acoustic pressure p as a function of the position \mathbf{x} and time t

$$\nabla^2 p(\mathbf{x}, t) - \frac{1}{c^2} \frac{\partial^2 p(\mathbf{x}, t)}{\partial t^2} = 0, \quad (2.2)$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ stands for the 3-dimensional *Laplacian* operator. The constant c is the sound velocity in the medium and has dimension $[\frac{\text{m}}{\text{s}}]$. Boundary condition and free field definition here.

The wave equation is linear, which implies the followings:

- the pressure field at any time is the sum of the pressure fields resulting from each source at that time;
- the pressure field emitted at a given position propagates over space and time according to a linear operation.

Assuming the propagation of the wave in a homogeneous medium, one can obtain the equation above by combining three fundamental physical laws:

- the *conservation of momentum*,
- the *conservation of mass*, and

⁷example of sources are vibrating solids (e.g. loudspeakers membrane), rapid compression or expansion (e.g. explosions or implosions) or air vortices (e.g. flute and whistles).

⁸In 1746, d'Alembert discovered the one-dimensional wave equation for music strings, and within ten years Euler discovered the three-dimensional wave equation for fluids.

- the *polytropic process relation*, meaning that the medium is an ideal gas undergoing a reversible adiabatic process.

However, media are not uniform and feature inhomogeneities of two types: scalar inhomogeneities, e.g. due to temperature variation, and vector inhomogeneities, e.g. due to presence of fans or air conditioning. Although these affect the underlying assumption of the model, the effects are small in typical application of speech and audio signal processing. Therefore they are commonly ignored.

► THE HELMHOLTZ'S EQUATION

The equation 2.2 is expressed in the space-time domain (\underline{x}, t) . By applying the temporal Fourier transform, we obtain the *Helmholtz equation*:

$$\nabla^2 P(\underline{x}, f) + k^2 P(\underline{x}, f) = 0, \quad (2.3)$$

where $k = \frac{2\pi f}{c}$ is known as *wave number* and relates the frequency f to the propagation velocity c .

Both the wave (2.2) and the Helmholtz's equation (2.3) are source-independent, namely no source is present in the medium. Therefore they are said to be *homogeneous* as the right-hand term is zero. Normally the sound field is a complex field generated by acoustic sources. As consequence, the two equations become inhomogeneous as some non-zero terms needs to be added to the right-hand sides.

In the presence of a sound source producing waves with source function $s(t, \underline{x})$, the wave equation can be written

$$\nabla^2 p(\underline{x}, t) - \frac{1}{c^2} \frac{\partial^2 p(\underline{x}, t)}{\partial t^2} = s(t, \underline{x}). \quad (2.4)$$

Thus, the corresponding Helmholtz's equation writes

$$\nabla^2 P(\underline{x}, f) - k^2 P(\underline{x}, f) = S(\underline{x}, f). \quad (2.5)$$

For instance one can assume an infinitesimally small sphere locate at \underline{s} emitting a sound at frequency f , i.e. $S(\underline{x}) = \delta(\underline{x} - \underline{s})$. At the receiver position $\underline{x} \neq \underline{s}$, the Helmholtz's equation writes

$$\nabla^2 H(f, \underline{x} | \underline{s}) - k^2 H(f, \underline{x} | \underline{s}) = \delta(\underline{x} - \underline{s}), \quad (2.6)$$

The function $H(f, \underline{x} | \underline{s})$ satisfying Eq. (2.6) is called the *Green's function* and is associated to Eq. (2.3), for which it is also a solution.

2.1.2 ... and its Green solution

Green's Functions are mathematical tools for solving linear differential equations with specified initial- and boundary- conditions [Duffy 2015]. They have been used to solve many fundamental equations, among which Eqs. (2.2) and (2.3) for both free and bounded propagation. They can be seen as a concept analogous to *impulse responses*⁹ in signal processing. Under this light, the physic so-far can be rewritten using the vocabulary of the communication theory, namely *input, filter and output*.

By 1950 Green's functions for Helmholtz's equation were used to find the wave motions due to flow over a mountain and in acoustics. Green's functions for the wave equation lies with Gustav Robert Kirchhoff (1824–1887), who used it during his study of the three-dimensional wave equation. He used this solution to derive his famous *Kirchhoff's theorem* [Duffy 2015].

⁹Impulse responses in time domain, transfer functions in the frequency domain.

According to Green's method, the equations above can be solved in the frequency domain for arbitrary source as follows:

$$P(f, \underline{x}) = \iiint_{\mathcal{V}_{\tilde{s}}} H(f, \underline{x} | \underline{s}) S(f, \underline{s}) d\underline{s}, \quad (2.7)$$

where $\mathcal{V}_{\tilde{s}}$ denotes the source volume, and $d\underline{s} = dx_{\tilde{s}} dy_{\tilde{s}} dz_{\tilde{s}}$ the differential volume element at position \underline{s} . If one ignores the space integral, one can see the close relation with a transfer function. Finally, the requested sound pressure $p(\underline{x}, t)$ can be computed by taking the frequency-directional inverse Fourier transform of Eq. (2.7).

It can be shown [Kuttruff 2016a] that the Green's function for Eqs. (2.3) and (2.6) writes

$$H(f, \underline{x} | \underline{s}) = \frac{1}{4\pi \|\underline{x} - \underline{s}\|} e^{-\frac{i2\pi f \|\underline{x} - \underline{s}\|}{c}} \quad (2.8)$$

where $\|\cdot\|$ denotes the Euclidean norm. By applying the inverse Fourier transform to the result above, we can write the time-domain Green's function as

$$h(t, \underline{x} | \underline{s}) = \frac{1}{4\pi \|\underline{x} - \underline{s}\|} \delta\left(t - \frac{\|\underline{x} - \underline{s}\|}{c}\right) \quad (2.9)$$

where $\delta(\cdot)$ is the time-directional Dirac delta function.

As consequence, the *free field*, that is open air without any obstacle, the sound propagation incurs a delay q/c and an attention $1/(4\pi q)$ as function of the distance $q = \|\underline{x} - \underline{s}\|$ from the source to the microphone.

According to Eq. (2.9), the sound propagates away from a point source with a spherical pattern. When the receiver is far enough from the source, the curvature of the *wavefront* may be ignored. The waves can be approximated as *plane waves* orthogonal to the propagation direction. This scenario depicted in Figure 2.3 is known as *far-field*. In contrast, when the distance between the source and the receiver is small, the scenario is called *near field*. These distances can be quantified exactly based on the sound wavelength,

$$\lambda = \frac{2\pi}{k} = \frac{c}{f} \quad [\text{m}], \quad (2.10)$$

where f is the frequency of the sound wave. As depicted in Figure 2.2, λ measures the spatial distance between two points around which the medium has the same value of pressure. In practical acoustic applications, the boundary between near- and far-field can be empirically set at 2λ . For instance, for a sound source emitting a pure sinusoidal tone at 1 kHz, the receiver sensor can be considered in far-field if placed ~ 69 cm away from the source.

Eqs. (2.8) and (2.9) are respectively the free-field transfer function and the impulse response.

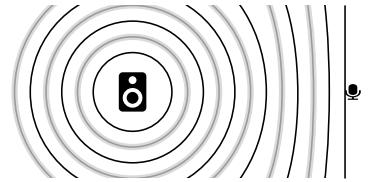


FIGURE 2.3: Visualization of the sound propagation. Since the sensor (i.e. a microphone) is drawn in the far field, the incoming waves can be approximated as plane waves.

2.2 ACOUSTIC REFLECTIONS

The equations derived so far assumed unbounded medium, i. e. free space: a rare scenario in everyday applications. Real mediums are typically bounded, at least partially. For instance in a room, the air (propagation medium) is bounded by walls, ceiling, and floor. When sound travels outdoor, the ground acts as a boundary for one of the propagation directions. Therefore, the sound wave does not just stop when it reaches the end of the medium or when it

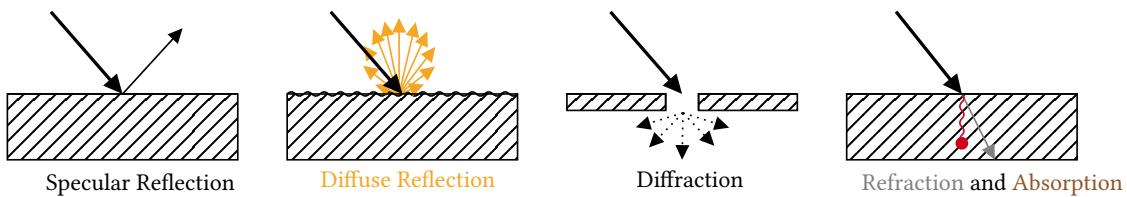


FIGURE 2.4: Different types of sound interact with a surface.

encounters an obstacle in its path. Rather, a sound wave will undergo certain behaviors depending on the obstacles' acoustics and geometrical properties, including

- *reflection* off the obstacle,
- *diffraction* around the obstacle, and
- *transmission* into the obstacle, causing
 - *refraction* through it, and
 - *dissipation* of the energy.

Reflections typically arise when a sound wave hits a large surface, like a room wall. When the sound meets a wall edge or a slit, the wave diffracts, namely it bends around the corners of an obstacle. The point of diffraction effectively becomes a secondary source which may interact with the first one. The part of energy transmitted to the object may be absorbed and refracted. Objects are characterized by a proper acoustic resistance, called *acoustic impedance* (see § 2.2.1), which describes their acoustic inertia as well as the energy dissipation. The remaining contribution may continue to propagate resulting in the refraction phenomenon.

When sound reflects on an solid surface, two types of acoustic reflections can occur: part of the sound energy

- is reflected *specularly*, i. e., the angle of incidence equals the angle of reflection; and
- is reflected *diffusely* - or *scattered*, i. e., scatter in every direction).

All the phenomena occur with different proportions depending on the acoustics and geometrical properties of surfaces and the frequency content of the wave. In acoustics, it is common to define the *operating points* and different *regimes*, e. g. for instance near- vs. far-field, according the wavelength λ . Therefore, we can identify the following three responses of objects (irregularities) of size d to a plane-wave, as depicted in Figure 2.6

- $\lambda \gg d$, the irregularities are negligible and the sound wave reflection is of specular type;
- $\lambda \approx d$, the irregularities break the sound wave which is reflected towards every direction;
- $\lambda \ll d$, each irregularities is a surface reflecting specularly the sound waves.

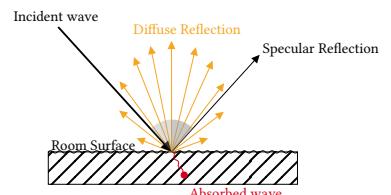


FIGURE 2.5: Specular and diffuse reflection.

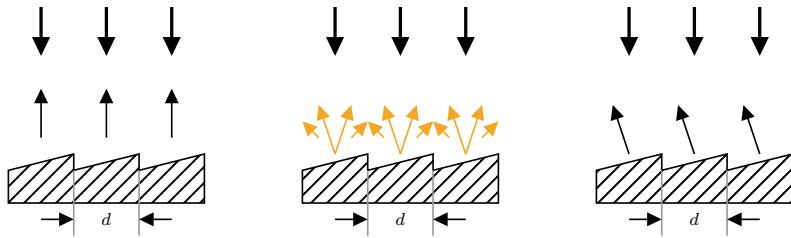


FIGURE 2.6: A reflector having irregularities on its surface with width d much smaller than the sound wavelength λ . Image courtesy of [Kuttruff 2016a].

This presented behavior can be described with the wave equation by imposing adequate boundary conditions. A simplified yet effective approach - just as in optics - is to model incoming sound waves as *acoustic rays* [Davis and Fleming 1926; Krokstad et al. 1968]. A ray has well-defined direction and velocity of propagation perpendicular to the wavefront, and conveys a total wave energy which remains constant. This simplified description undergoes with the name of Geometrical (room) acoustics (**GA**) [Savioja and Svensson 2015], and share many fundamentals with geometrical optics. This model will be convenient to describe and visualize the reflection behavior hereafter.

2.2.1 Large smooth surfaces, absorption and echoes

Specular reflections are generated by surfaces which can be modelled as infinite, flat, smooth and rigid. As mentioned above, this assumption is valid as long as the surface has dimension much larger than the sound wavelength. Here the acoustic ray is reflected according to the *law of reflection*, stating that (i) the reflected ray remains in the plane identified by the incident ray and the normal to the surface, and (ii) the angles of the incident and reflected rays with the normal are equal.

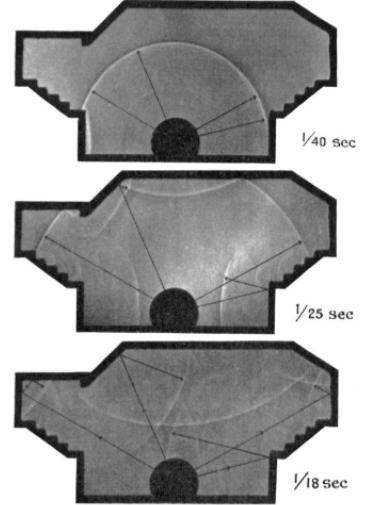
If the surface S is not perfectly rigid or impenetrable, its behavior is described by the *acoustic impedance*, $Z_S(f) \in \mathbb{C}$. Analytically, it is defined as a relation between sound pressure and particle velocity at the boundary. It consists of a real and imaginary part, called respectively acoustic *resistance* and *reactance*. The former can be seen as the part of the energy which is lost, and the latter as the part which is stored.

- ▶ THE REFLECTION COEFFICIENT β can be derived from the acoustic impedance for plane waves, i. e. under assuming a far-field regime between source, receiver and surface. It measures the portion of energy absorbed by the surface and the incident acoustic wave. Analytically, it is defined as [Kuttruff 2016a; Pierce 2019]

$$\beta(f, \theta) = \frac{Z_S(f) \cos \theta - \xi_{\text{air}}(f)}{Z_S(f) \cos \theta + \xi_{\text{air}}(f)}, \quad (2.11)$$

where $Z_S(f)$ are the frequency-dependent impedance of the surface, $\xi_{\text{air}}(f)$ is a parameter of the propagation medium (air) sometimes called as *intrinsic impedance*, and θ is the angle of incidence.

- ▶ THE ABSORPTION COEFFICIENT is typically used instead in the context of **GA** and audio signal processing. It comes from the following approximations [Savioja and Svensson 2015]: (i) the energy or intensity of the plane wave¹⁰ is considered instead of the acoustic pressure; (ii) dependency on the angle of incidence is relaxed in favor of the averaged quantities; (iii) local dependency on fre-



Schlieren photographs showing successive stages in the progress of a sound pulse in a section of a Debating Chamber. Image courtesy of [Davis and Fleming 1926]. Sabine used Schlieren photography to investigate sound propagation paths in the early 1900s. Their impressive visualizations show wavefronts that are augmented with rays that are perpendicular to the wavefronts

¹⁰Since it is the square magnitude of the acoustic pressure, the phase information is lost.

quencies is relaxed in favor of a frequency-independent scalar or at most a description per octave-band. These assumptions are motivated by the difficulty of measuring the acoustic impedance and the possibility to compute an equivalent coefficient a posteriori.

Therefore, it is customary to use the absorption coefficient, defined as

$$\alpha(f) = 1 - |\bar{\beta}(f)|^2, \quad (2.12)$$

where $\bar{\beta}$ is the reflection coefficient averaged over the angles θ .

- ▶ ECHOES ARE SPECULAR REFLECTIONS which distinguish themselves in terms of gain or timing. Originally this term is used to refer to sound reflections which are subjectively noticeable as a separated repetition of the original sound signal. These can be heard consciously in outdoor, for instance in mountains. However, they are less noticeable to the listener in close rooms. In § 2.3.1 a proper definition of echoes will be given with respect to the temporal distribution of the acoustic reflections.

The word echo derives from the Greek *echos*, literally “sound”. In the folk story of Greek, Echo is a mountain nymph whose ability to speak was cursed: she only able to repeat the last words anyone spoke to her.

2.2.2 Diffusion, scattering and diffraction of sound

Real-world surfaces are not ideally flat and smooth; they are rough and uneven. Examples of such surfaces are coffered ceilings, faceted walls, raw brick walls as well as the entire audience area of a concert hall. When such irregularities are in the same order as the sound wavelength, *diffuse reflections* is observed.

In the context of GA, the acoustic ray associated to a plane-wave can be thought of as a bundle of rays traveling in parallel. When it strikes such a surface, each individual rays are bounced off irregularly, creating *scattering*: a number of new rays are created, uniformly distributed in the original half-space. The energy carried by each of the outgoing ray is angle dependent and it is well modeled through the *Lambert's cosine law*, originally used to describe optical diffuse reflection.

The total amount of energy of this reflection may be computed a-priori knowing the *scattering coefficient* of the surface material. Alternatively, it can be derived a-posteriori with the *diffusion coefficient*, namely the ratio between the specularly reflected energy over the total reflected energy.

Diffraction waves occur when the sound confronts the edge of a finite surface, for instance around corners or through door openings. This effect is shown in Figure 2.8 At first the sound wave propagates spherically from the source. Once it reaches the reflector's apertures, the wave is diffracted, i. e. bended, behind it. It is interesting to note that the diffraction waves produced by the semi-infinite reflector edge allow the area that is “behind” the reflector to be reached by the propagating sound. This physical effect is exploited naturally by the human auditory system to localize sound sources.

2.3 ROOM ACOUSTICS AND ROOM IMPULSE RESPONSE

Room acoustics studies acoustic waves propagating in enclosed volume delimited by surfaces (walls, floors, etc.), with which an incident wave interacts as described in § 2.2. In this context,

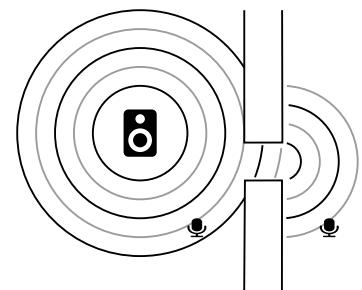


FIGURE 2.8: Schematic representation of sound diffraction. This effect allows to hear “behind walls”.

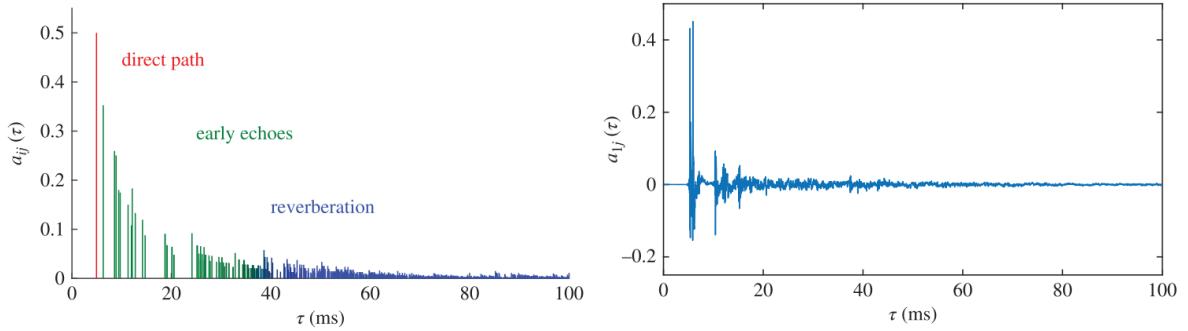


FIGURE 2.9: Schematic illustration of the shape of an RIR and the first 100 ms of a measured one. Image taken from [Vincent et al. 2018]

a room is a physical enclosure containing a propagation medium with boundaries limiting the sound propagation.

Mathematically, the sound propagation is described by the wave equation (2.2).

By solving it, the Acoustic Impulse Response (AIR)¹¹ from a source to a microphone can be obtained. In the context of room acoustics, it is commonly referred to as the Room Impulse Response (RIR), usually stressing the geometric relation between reflections and the geometry of the scene. In this thesis the two terms will be used indistinctly.

¹¹The Acoustic Transfer Function (ATF) is the Fourier transform of the AIR

2.3.1 The room impulse response

Room Impulse Response (RIR) is where physical room acoustics and indoor audio signal processing meet and from now on, we will adopt a signal processing perspective. Therefore,

a RIR as a causal time-domain filter that accounts for the whole indoor sound propagation from a source to a receiver.

Figure 2.9 provides a schematic illustration of the shape of a RIR compared to a measured one. The RIRs usually exhibit common structures. Based on the consideration of § 2.2, they are commonly divided into three partially overlapping components:

$$h(t) = h^d(t) + h^e(t) + h^l(t), \quad (2.13)$$

where

- *the direct path* $h^d(t)$ is the line-of-sight contribution of the sound wave. This term coincides with the “pure delay” modeled by the free-field propagation model (2.9).
- *the acoustics echoes or early reflections* are included in $h^e(t)$ comprising few disjoint reflections coming typically from room surfaces. They are usually characterized by *sparsity* in the time domain and amplitude prominence greater than the later reflections. These first reflections are typically specular and are well modeled in general by the Image Source Method (ISM) explained in § 2.3.3.

- the late reverberation, or simply reverberation, $h^l(t)$ collects many reflections occurring simultaneously. This part is characterized by a diffuse sound field with exponentially decreasing energy.

These three components are not only “visible” when plotting the RIR against time, but are characterized by different perceptual features, as explained § 2.4.

To conclude, let \tilde{s} and \tilde{h} be the continuos-time source signal and RIR, respectively. Then, assuming that the RIR is time-invariant, the received sound writes

$$\tilde{x}(t) = (\tilde{h} \star \tilde{s})(t) \stackrel{\text{def}}{=} \int \tilde{h}(u) \tilde{s}(t-u) du, \quad (2.14)$$

where the symbol \star is the continuos-time convolution operator.

Apart for certain simple scenarios, computing RIRs in closed forms is a cumbersome task. Therefore numerical solvers or approximate models are used instead.

2.3.2 Simulating room acoustics

Most available simulators falls in three main categories:

- *Wave-based simulators* aims at solving the wave equation numerically;
- *Geometric simulators* make some simplifying assumption about the wave propagation. They typically ignore the wave physic, instead they adopt much lighter models such as *rays* or *particles*;
- *Hybrid simulators* combining both approaches.

The documentation of the Wayverb¹² acoustic simulator offers a complete overview of the State of the Art (SOTA) in acoustic simulator methods [Thomas 2017].

- WAVE-BASED METHODS are iterative methods that divide the 3D bounded enclosure into a grid of interconnected nodes¹². For instance, the Finite Element Method (FEM) divides the space into small volume elements smaller than the sound wavelengths, while the Boundary Element Method (BEM) divides only the boundaries of the space into surface elements. These nodes interact with each other according to the math of the wave equation. Unfortunately, simulating higher frequencies requires denser interconnection, so the computational complexity increases. The Finite-Difference-Time-Domain (FDTD) method replaces the derivatives with their discrete approximation, i. e. finite differences. The space is divided into a regular grid, where the changes of a quantity (air pressure or velocity) is computed over time at each grid point. Digital Waveguide Mesh (DWM) methods are a subclass of FDTD often used in acoustics problem.

¹²i. e. mechanical unit with simple degrees of freedoms, like mass-spring system or one-sample-delay unit

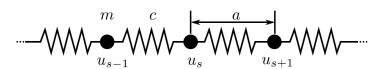


FIGURE 2.10: Example of a mass-spring linear mesh used to simulate a 1D transversal wave.

The main drawback of these methods is discretisation: less dense grids may simplify too much the simulation, while denser grids increase the computational load and other intrinsic limitations.¹³ Moreover, they require delicate definitions of the boundary condition at the physical level, like knowing complex impedances, which are rarely available in practice. On the other hand these methods inherently account for many effects such as occlusions, reflections, diffusion, diffractions and interferences. In particular, by simulating

¹³ For more details on this, the reader can refers to the *Courant–Friedrichs–Lowy conditions* and the documentation of the Wayverb¹².

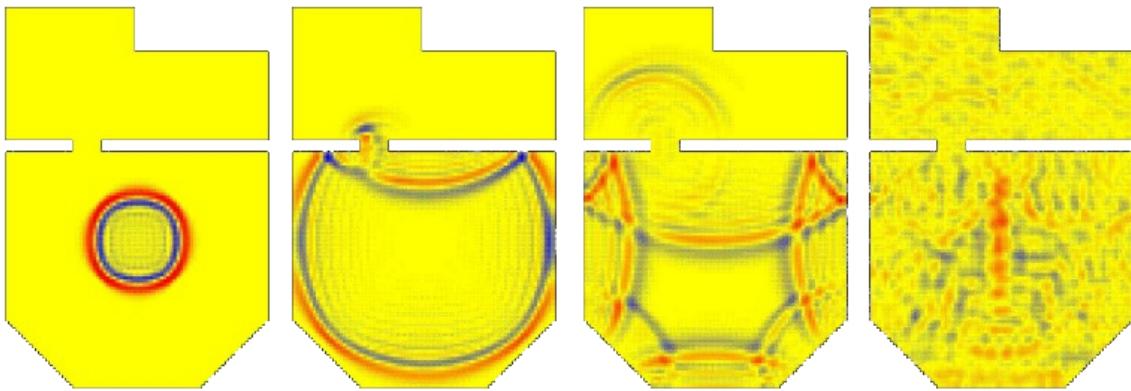


FIGURE 2.11: Simulation of Sound propagation at four consecutive timestamps using the **DWM** technique. A short, sharp, impulsive sound fired into the larger of two rooms causes a circular wavefront to spread out from the sound source. The the wave is reflected from the walls and part of it passes through a gap into the smaller room. In the larger room, interference effects are clearly visible; in the smaller room, the sound wave has spread out into an arc, demonstrating the effects of diffraction. A short while after the initial event, the sound energy has spread out in a much more random and complex fashion. Image taken form University of York's AudioLab website.

accurately low-frequency components of the **RIR**, they are able to well characterize the *room modes*¹⁴, namely, collections of resonances that exist in a room and characterize it.

As stated in [Välimäki et al. 2016], among the wave-based methods, the **DWMS** are usually preferred: they run directly in the time domain, requiring typically an easier implementation, and they exhibit a high level of parallelism.

- ▶ **GEOMETRIC METHODS** can be sub-grouped into *stochastic* and *deterministic* approaches. They typically compute the reflection path(s) between the source and the receivers, assuming that the wave behaves like a particle or a ray carrying the acoustic energy around the scene.

STOCHASTIC METHODS are approximate by nature. They are based on statistical modeling of the **RIRs** or Monte Carlo simulation methods. The former ones write statistical signal processing models based on prior knowledge, such as probability distribution of the **RIR** in regions of the time-frequency domain [Badeau 2019]. Rather than the detailed room geometry, these methods generally use high-level descriptors¹⁵ to synthesize **RIRs** and in some application are preferable.

The latter ones randomly and repeatedly subsample the problem space, e.g. tracing the path of random reflections, recording samples which fulfil some correctness criteria, and discarding the rest. By combining the results from multiple samples, the probability of an incorrect result is reduced, and the accuracy is increased. Typically the trade-off between quality and speed of these approaches is based on the number of samples and the quality of the prior knowledge modeled.

Ray-tracing [Kulowski 1985] is one the most common methods that fall in this category and is very popular in the field of computer graphics for light simulation. The basic idea is to collect “valid” paths of discrete rays traced around the room. Many technique have been proposed to reduce the computational load,

¹⁴ Room modes have the effect of amplifying and attenuating specific frequencies in the **RIR**, and produce much of the subjective sonic “colour” of a room. Their analysis and synthesis is of vital importance for evaluating acoustic of rooms, such as concert halls and recording studios or when producing musically pleasing reverbs.

For a detailed discussion about geometric acoustic methods, please refer to [Savioja and Svensson 2015].

¹⁵such as the amount of reverberation or source-to-receiver distance.

among which the *diffuse rain algorithm* [Schröder et al. 2007; Heinz 1993] is commonly used in many acoustic simulators. Each ray trajectory is reflected in a random direction every time it hits a wall and its energy is scaled according to the wall absorption. The process of tracing a ray is continued until the ray's energy falls below a predefined threshold. At each reflection time and for each frequency (bin or band), the ray's energy and angle of arrival are recorded in histograms, namely a *directional-time-frequency energy map* of the room's diffuse sound field for a given receiver location (see Figure 2.12). This map is then used as a prior distribution for drawing random sets of impulses which are used to form the RIR. While lacking a detailed description of early reflections and room modes, these methods are good to capture and simulate the statistical behavior of the diffuse sound field at a low computational cost.

DETERMINISTIC METHODS are good to simulate early reflections instead: they accurately trace the exact direction and the timing of the main reflections' paths. The most popular method is the Image Source Method (**ISM**), proposed by Allen and Berkley in [Allen and Berkley 1979]. Even if the basic idea is rather simple, the model is able to produce the exact solution to the wave equation for a 3D shoebox with rigid walls. It models only specular reflections, ignoring diffuse and diffracted components. It only approximates arbitrary enclosures and the late diffuse reflections.

The implementation reflects the sound source against all surfaces in the scene, resulting in a set of *image* sources. Then, each of these image sources is itself reflected against all surfaces and so on, up to a predefined order.

There are two main limitations of this method. First, in a shoebox the complexity of the algorithm is cubic in the order of reflections. Therefore when a high order is required, the algorithm becomes impractical. Second, it models only the specular reflection, neglecting the diffuse sound field. For these reasons, the image-source method is generally combined with a stochastic method to model the full impulse response.

- ▶ HYBRID METHODS combines the best of these two approaches. As discussed above, the image-source method is accurate for early reflections, but slow and not accurate for longer responses. The ray tracing method is by nature an approximation, but produces acceptable responses for diffuse fields. And in general geometric methods fail to properly model lower frequencies and room modes. The waveguide method models physical phenomena better than geometric methods, but is expensive at high frequencies. All these limitations correspond to three regions in the Time-Frequency (**TF**) representation of the RIR. As depicted in Figure 2.14,
 - in the time domain, a transition can be identified between the early vs. late reflection, corresponding to the validity of the deterministic vs. stochastic models; and
 - in the frequency domain, between geometric vs. wave-based modeling.

Providing these time- and frequency-domain crossover points [Badeau 2019] it is possible to combine image-source, ray-tracing and wave models. The time-domain cross-over point is called *transition time*, or *mixing time*. It

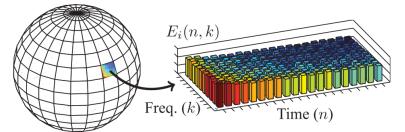


FIGURE 2.12: Directional-time-frequency energy map resulting from the diffuse rain algorithm [Schröder et al. 2007]. For each direction, that is receiver's spherical bin, a time-frequency histogram collects the energy of incoming rays. Image taken from [Schimmel et al. 2009]

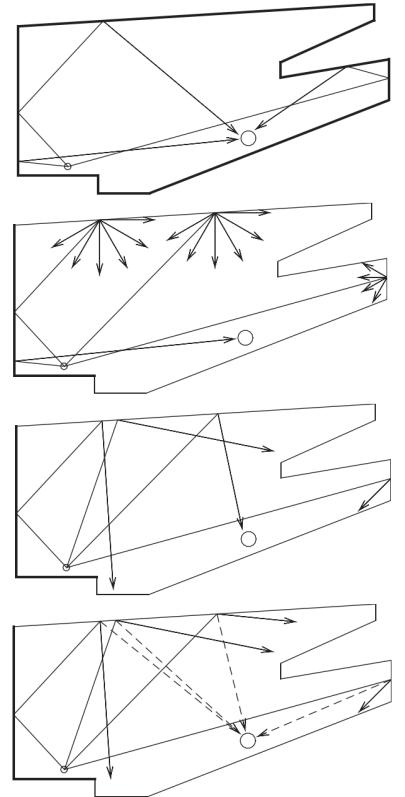


FIGURE 2.13: Visualization of ray-tracing method. From top to bottom: first the method will eventually find specular reflection; then diffuse reflections can be modeled either by splitting a ray into several new rays or a single random one. In the diffuse rain technique a shadow-ray is cast from each diffuse reflection point to the receiver to speed-up convergence of the simulation. Image courtesy of [Savioja and Svensson 2015]

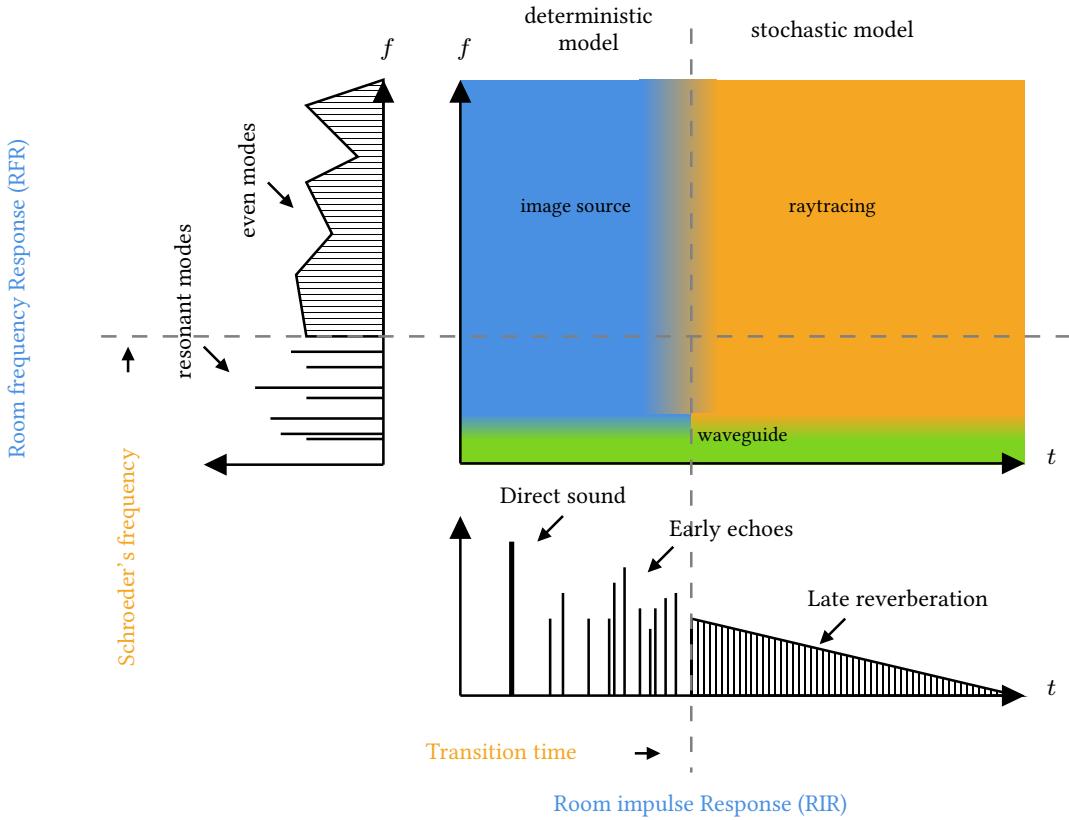


FIGURE 2.14: Time-Frequency regions of the RIR associated to the method that better simulate them. Image adapted from [Thomas 2017; Badeau 2019].

identifies the moment after which reflections are so frequent that they form a continuum. After it, the sound level decays exponentially over time, as it is partially absorbed by the room surfaces at every reflection. This point can be then used to set a cross-fade between the deterministic and the stochastic process. The crossover point in the frequency domain is called *Schroeder's frequency*. It splits the spectrum of the RIR into a region with a few isolated modes and one denser, called respectively the *resonant* and *even* behaviors. Therefore, this point defines the cross-fade between the geometrical and wave-based model.

Each simulator available has its own way to compute and implement this crossover points as well as mixing the results of the three methods.

2.3.3 The method of images and the image source model

The *Method of Images* is a mathematical tool for solving a certain class of differential equations subjected to boundary conditions. By assuming the presence of a “mirrored” source, certain boundary conditions are verified facilitating the solution of the original problem. This method is widely used in many fields of physics, and interestingly with specific applications to Green’s functions. Its application to acoustics was originally proposed by Allen and Berkley in [Allen and Berkley 1979] and it is known as the Image Source

Method (**ISM**). Now **ISM** is probably the most used technique for deterministic **RIR** simulation due to its conceptual simplicity and its flexibility.

The **ISM** is based on purely specular reflection and it assumes that the sound energy travels around a scene in “rays”. In the appendix of [Allen and Berkley 1979], the authors also proved that this method produces a solution the Helmholtz’s equation for cuboid enclosures with rigid boundaries.

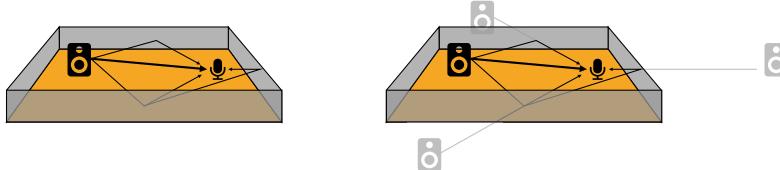


FIGURE 2.15: 3-D representation of the Image Source Method (**ISM**) and its propagation paths for selected echoes.

The image source defines the interaction of the propagating sound and the surface. It is based on the observation that when a ray is reflected, it spawns a secondary source “behind” the boundary surface.

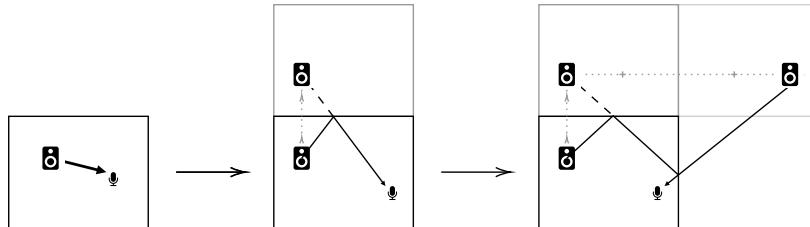


FIGURE 2.16: From left to right, path involving the direct path, one reflection obtained using first-order image, and two reflections obtained using two images. Image inspired from [Habets 2006].

As shown in Figure 2.16, this additional source is located on a line perpendicular to the wall, at the same distance from it as the original source, as if the original source had been “mirrored” in the surface. In this way, each wavefront that arrives to the receiver from each reflection off the walls corresponds to the direct path received from an equivalent (or image) source.

The **ISM** makes use of the following assumptions:

- sound source and receiver are points in a cuboid enclosure;
- purely specular reflection paths between a source and a receiver;
- this process is simplified by assuming that sound propagates only along straight lines or rays;
- and rays are perfectly reflected at boundaries.

Finally, the **RIR** is found by summing the contribution from each (image) source, delayed and attenuated appropriately depending on their distance from the receiver. Therefore, in the time domain, the **RIR** associated to the source at position \underline{s} and the receiver at \underline{x} reads

$$h_{\text{ISM}}(t, \underline{x} | \underline{s}) = \sum_{r=0}^R \frac{\bar{\alpha}_r}{4\pi \|\underline{x} - \underline{s}_r\|} \delta\left(t - \frac{\|\underline{x} - \underline{s}_r\|}{c}\right) \quad (2.15)$$

where \underline{s}_r is the r -th image of the source and $\bar{\alpha}_r$ is the total frequency-independent¹⁶

¹⁶Which is equivalent to consider perfectly rigid and reflective walls

damping coefficient related to the r -th image. Such coefficient accounts for all the dissipation effects encountered in the reflection path, e. g. absorption, air attenuation and scattering. In this thesis, we will refer to Eq. (2.15) as the time-domain *echo model*.

The reader should notice that the summation in the echo models of Eq. (2.15) induce an “order” among reflections indexed by r . Reflections are usually sorted for increasing Time of Arrival (TOA), $\tau_r = \|\underline{x} - \underline{s}_r\|/c$, or decreasing amplitudes, $\bar{\alpha}_r/(4\pi\|\underline{x} - \underline{s}_r\|)$. Alternatively, one can sort them according to their “image” generation, e. g. direct path, first-, second-order images etc. This would require an arbitrary order within the same generation, based typically on arbitrary wall sequence. Notice that the resulting sorted sequences can differ substantially as shown in ???. This translates into non-trivial definition of evaluation metrics for the task of estimating echoes.

In the original formulation of the ISM, $\bar{\alpha}_0 = 1$ is assumed for the direct propagation, while for the first order images, it coincides with the frequency-independent surface absorption coefficient of the surface. For the subsequent orders of images, the product of all the coefficients of the surfaces encountered in the reflection path is considered.

In order to easily incorporate frequency-dependent damping effects, the Fourier transform of Eq. (2.15) is considered instead, where each reflection term is appropriately scaled

$$H_{ISM}(f, \underline{x} | \underline{s}) = \sum_{r=0}^R \frac{\alpha_r(f)}{4\pi\|\underline{x} - \underline{s}_r\|} \exp\left(-i2\pi f \frac{\|\underline{x} - \underline{s}_r\|}{c}\right), \quad (2.16)$$

where now the r -th damping coefficient α_r is frequency dependent. In this thesis, we will refer to Eq. (2.16) as the frequency-domain echo model.

Notice that now the damping coefficients correspond to filters, requiring Eq. (2.15) to be written as sum of convolutions. This has a strong implication when modeling and estimating the RIRs as stream of Dirac impulses. Ideally they consist of scaled Diracs with well-defined time locations. The probability that two or more Diracs arrive at the same time is then very small. However, if we now assume that each reflection has a non-flat frequency response, filters are observed in the time domain. Such filters have arbitrary long time-domain description and now the probability that two or more overlap is much higher.

- ? CAN ECHOES BE LOUDER THAN THE DIRECT-PATH? Yes, in certain cases reflections maybe carry energy comparable or stronger than the direct contribution. This happens for instance when directional sources are directed towards reflectors. Typical scenarios are when a speaker is giving a talk while facing the slides projected on a wall giving the shoulders to the microphones. Moreover, when multiple reflections arrive within a very short time, due to the recording process (such as pre-filtering and sampling), their energy is aggregated together. This effect will be studied in detail in Chapter 5. Nevertheless, aggregating echoes is something that the human auditory system does to enhance speech intelligibility, as discussed in the next section.

2.4 PERCEPTION AND SOME ACOUSTIC PARAMETERS

So far we have analyzed reverberation from a purely mathematical point of view. However in many applications it is important to correlate physical measurements to subjective and perceptual qualities. This will be important in order to define evaluation scenarios later in this thesis.

2.4.1 The perception of the RIR's elements

It is commonly accepted that the RIR components defined in § 2.3.1 play rather separate roles in the perception of sound propagation.

- ▶ THE DIRECT PATH is the delayed and attenuated version of source signal itself. It coincides with the free-field sound propagation and, as we will see in ??, it reveals the direction of the source.
- ▶ THE EARLY REFLECTIONS AND ECHOES are reflections which are by nature highly correlated with the direct sound. They convey a sense of geometry which modifies the general perception of the sound:
 - *The precedence effect* occurs when two correlated sounds are perceived as a single auditory event [Wallach et al. 1973]. This happens usually when they reach the listener with a delay within 5 ms to 40 ms. However, the perceived spatial location carried by the first-arriving sound suppressing the perceived location of the lagging sound. This allows human to accurately localize the direction of the main source, even in presence of its strong reflections.
 - *The comb filter effect* indicates the change in timbre of the perceived sound, named *coloration*. This happens when multiple reflections arrive with periodic patterns and some constructive or destructive interferences arise. Such phenomena can be well modeled with a comb filter [Barron 1971].
 - *Apparent source width* is the audible impression of a spatially extended sound source [Griesinger 1997]. By the presence of early reflections, the perceived energy increases, providing the impression that a source is larger than its true size.
 - *Distance and depth perception* provides the listener with cues about the source 3D location. “Distance” refers to only the range and applies to outdoor scenarios, as opposed to “depth” which relates to the indoor case. A fundamental cue for distance perception is the Direct to Reverberant Ratio (DRR), i. e. the ratio between the direct path ratio and the remaining portion of the RIR (see § 2.4.4). Regarding the depth perception, early reflections are the main responsible. In the context of virtual reality, correctly modeling these quantities is essential in order to maintain a coherent depth impression [Kearney et al. 2012].
- ▶ THE LATE REVERBERATION is indicative of the size of the environment and the materials within [Välimäki et al. 2016]. It produces the so-called the *listener*

envelopment, i. e. the degree of immersion in the sound field [Griesinger 1997]. This portion of the RIR is mainly characterized by sound diffusion, which depends on the surfaces roughness.

2.4.2 Mixing time

Perceptually, the mixing time define the instant when the reverberation cannot be distinguished from that of any other position of the listener in the room. Analytically, it can be identified as the point dividing the early reflections from the late reverberation in a RIR. The work of [Lindau et al. 2012] shows the equivalence between the two definitions. Due to this, it is an important parameter also in the context of RIRs synthesis as it defines cross-over point for room acoustics simulator using hybrid methods [Savioja and Svensson 2015](see § 2.3.2).

2.4.3 Reverberation time

The *reverberation time* measures the time that takes the sound to “fade away” after the source has ceased to emit. In order to quantify it, acoustics and audio signal processing use the *Reverberation Time at 60 dB* (RT_{60}), the time after which the sound energy dropped by 60 dB. It depends on the size and absorption level of the room (including obstacles), but not on the specific positions of the source and the receiver.

Real measurements of RIRs are affected by noise. As a consequence, it is not always possible to consider a dynamic range of 60 dB, i. e. the energy gap between the direct path and the ground noise level. In this case, the RT_{60} value must be approximated with other methods.

By knowing the room geometry and the surfaces acoustics profiles, it is possible to use the empirical *Sabine’s formula*:

$$RT_{60} \approx 0.161 \frac{V_{TOT}}{\sum_l \alpha_l S_l} \quad [\text{s}], \quad (2.17)$$

where V_{TOT} is the total volume of the room [m^3] and α_l and S_l are the absorption coefficient and the area [m^2] of the l -th surface.

2.4.4 Direct-to-Reverberant ratio and the critical distance

The Direct to Reverberant Ratio (DRR) quantifies the power of direct against indirect sound [Zahorik 2002]. It varies with the size and the absorption of the room, but also with the distance between the source and the receiver according to the curves depicted in Figure 2.17. The distance beyond which the power of indirect sound becomes larger than that of direct sound is called the *critical distance*. These two quantities represent an important parameter to assert the robustness of audio signal processing methods, since they basically measure the validity of the free-field assumption.

- ★ THIS CHAPTER presented fundamental notions of room acoustics, and it defined echoes and RIRs, which are the main protagonist of this thesis. In the following chapter, we will study these quantities under the light of audio signal processing and, in particular, how we can process continuos-time acoustic quantities with computers.

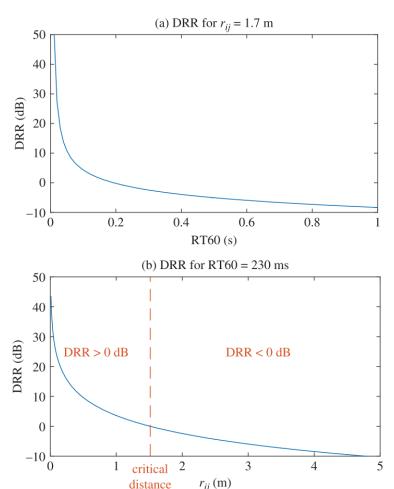


FIGURE 2.17: DRR as a function of the RT_{60} and the source distance r_{ij} based on Eyring’s formula (Gustafsson et al., 2003). These curves assume that there is no obstacle between the source and the microphone, so that the direct path exists. The room dimensions are the same as in Figure 3.1.

3

Elements of Audio Signal Processing

- ▶ **SYNOPSIS** Let us now move from physics to digital signal processing. At first, in § 3.1, this chapter formalizes fundamental concepts of audio signal processing such as signals, mixtures, filters and noise in the time domain. In § 3.2, we will present a fundamental signal representation that we will use throughout the entire thesis: the Short Time Fourier Transform (STFT). Finally, in § 3.3, some essential models of the RIR are described.
- Unless specified, the notation and definitions presented in this chapter for the audio signal model are excerpted from Vincent et al.'s book *Audio source separation and speech enhancement*. The material used for illustrating concepts of digital signal processing are taken from standard book on the topics.

3.1 SIGNAL MODEL IN THE TIME DOMAIN

In the previous chapter we formalized the physics that rule the sound propagation from the source to the microphone. A raw *audio signal* encodes the variation of pressure over time on the microphone membrane. Mathematically it is denoted as the function

$$\begin{aligned} \tilde{x} : \mathbb{R} &\rightarrow \mathbb{R} \\ t &\mapsto \tilde{x}(t), \end{aligned} \tag{3.1}$$

continuous both in time $t \in \mathbb{R}$ and amplitudes.

Today signals are typically processed, stored and analyzed by computers as *digital audio signal*. This corresponds to finite and discrete-time signal $x[n]$ obtained by periodically sampling the continuous-time signal \tilde{x} at rate F_s [Hz], truncate it to N samples. As common to most measurement models, we assume that the sampling process involves two steps: first, the impinging signal undergoes an ideal low-pass filter $\tilde{\phi}_{LP}$ with frequency support in $] -F_s/2, F_s/2]$; ¹⁷ then its time-support is regularly discretized, $t = n/F_s$ for $n \in \mathbb{Z}$. This is expressed by

$$x[n] = \left(\tilde{\phi}_{LP} \star \tilde{x} \right) \left(\frac{n}{F_s} \right) \in \mathbb{R}, \tag{3.2}$$

where \star is the continuous-time convolution operator. This will restrict the frequency support of signal to satisfy the *Nyquist–Shannon sampling theorem* and avoid aliasing effect.

Finally, at the end of the discretisation process, the $\tilde{x}(t)$ is represented as the

*"Everything in its right place
There are two colours in my head"*
—Radiohead *Everything In Its Right Place*

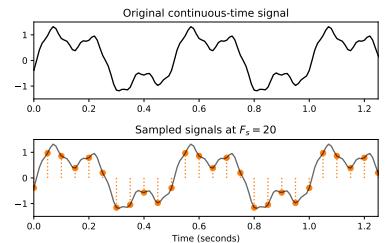


FIGURE 3.1: Continuous-time signal and its sampled version.

Strictly speaking, the digital representation of a continuous signal involves sampling and quantization. In this thesis we assume the sampled signals are real-valued, ignoring the quantization process.

¹⁷ The ideal low-pass filter is $\tilde{\phi}_{LP}(t) = \text{sinc}(t) = \sin(\pi F_s t) / (\pi F_s t)$. The term sinc stands for *sinus cardinal* and was introduced by Philip M. Woodward in 1952 in [Woodward and Davies 1952], in which he said that the function "occurs so often in Fourier analysis and its applications that it does seem to merit some notation of its own"

finite time series or a vector,

$$\mathbf{x}_N \in \mathbb{R}^N, \quad (3.3)$$

with entries $x_N[n]$ for $n = 0, \dots, N - 1$.

The choice of F_s depends on the application since it is a trade-off between computational power, processing and rendering quality. Historically the two iconic values are 44.1 kHz for music distribution on CDs and 8 kHz for first-generation speech communication. Now multiples of 8 kHz are typically used in audio processing: (16, 48, 96, 128 kHz).

Audio signals are emitted by sources and are observed, received or recorded by microphones. A set of microphones is called a microphone *array*, whose signals are sometime referred to as *channels*. In this thesis, these objects are assumed to have been deployed in a indoor environment, called generically *room*. Let us provide some taxonomy, through some dichotomies, useful for describe the mixing process later:

- ⇒ SOURCES VS. MIXTURES. Sound sources emits sounds. When multiple sources are active at the same time, the sounds that reach our ears or are recorded by microphones are superimposed or *mixed* into a single sound. This resulting signal is denoted as *mixture*.
- ⇒ SINGLE-CHANNEL VS. MULTICHANNEL. The term *channel* is used here to indicate the output of one microphones or one source. A *single-channel* signal ($I = 1$) is represented by the scalar $\tilde{x}(t) \in \mathbb{R}$, while a *multichannel* ($I > 1$) is represented by the vector $\tilde{\mathbf{x}}(t) = [\tilde{x}_1, \dots, \tilde{x}_I]^\top \in \mathbb{R}^I$.
- ⇒ POINT VS. DIFFUSE SOURCES. *Point sources* are single and well-defined points in the space emitting single-channel signal. In certain application, human speakers or the sound emitted by a loudspeaker can be reasonably modeled as in this way. *Diffuse sources* refers for instance to wind, traffic noise, or large musical instruments, which emit sound in a large region of space. Their sound cannot be associate to a punctual source, but rather a distributed collection of them.
- ⇒ DIRECTIONAL VS. OMNIDIRECTIONAL. An *omnidirectional* source (resp. receiver) will in principle emit (resp. record) sound equally from all directions, both in time and in frequency. Although this greatly simplifies processing models and frameworks, this is not true in real scenario. The physical properties of real sources (resp. receivers) leads to *directivity patterns*, a.k.a. *polarity*, which may be different at different frequencies. In this thesis we will assume omnidirectional sources and receivers.

3.1.1 The mixing process

Let us assume the observed signal has I *channels* indexed by $i \in \{1, \dots, I\}$. Let us assume that there are J sources indexed by $j \in \{1, \dots, J\}$. Each microphone i and each source j have a well defined position in the space, $\underline{\mathbf{x}}_i$, $\underline{\mathbf{s}}_j$, respectively.

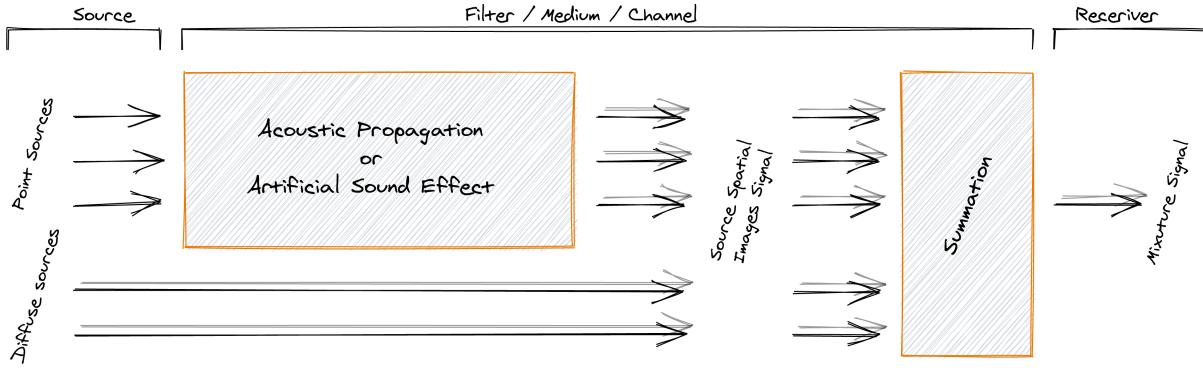


FIGURE 3.2: General mixing process, illustrated in the case of $J = 3$ sources, including three point sources and one diffuse source, and $I = 2$ channels.

The mixing process describes then the nature of the mixtures. In order to better formalized it, the authors of [Sturmel et al. 2012] introduced the intermediate representation called *source spatial images*: $\tilde{c}_{ij}(t)$ describes the contribution of the source j to the microphone i . Consequently, the *mixture* \tilde{x}_j is the combination of images associated to the source j . Depending on the “contribution” the image describes, the following type of mixture can be defined:

- ⇒ NATURAL VS. ARTIFICIAL MIXTURES. The former refers to microphone mixtures recorded simultaneously the same auditory scene, e. g. teleconferencing systems or hands-free devices. By contrast, the latters are created by mixing together different individual, possibly processed, recordings. This are the typical mixtures used professional music production where the usage of long-chain of audio effects typically “hide”, willingly or not, the recording environment of the sound sources.
- ⇒ INSTANTANEOUS VS. CONVOLUTIVE MIXTURES. In the first case, the mixing process boils down to a simple linear combination of the source signals, namely the mixing filters are just scalar factors. This is the typical scenario when sources are mixed using a mixing console. Convulsive mixtures, instead, denote the more general case where the each mixture is the sum of filtered signals. In between are the *anechoic* mixtures involving the sum of scaled and delayed source signals. Natural mixtures are convulsive by nature and ideal free-far-field natural recording are well approximated by anechoic mixtures.

In this thesis, we will particularly focus on natural mixture. The microphones listen to the propagation of sound in the room and this process is linear (cf. § 2.1) and time invariant provided a static scenario. Therefore, the resulting mixture is the simple summation of the sound images, which are the collections

instantaneous	$\tilde{c}_{ij} = a_{ij}\tilde{s}(t)$
anechoic	$\tilde{c}_{ij} = a_{ij}\tilde{s}_j(t - \tau_{ij})$
convulsive	$\tilde{c}_{ij} = (h_{ij} * \tilde{s}_j)(t)$

TABLE 3.1: Taxonomy of linear mixing models for a mixture channel x_i , sources s_j , impulse response \tilde{g}_{ij} , scaling factor a_{ij} and delay τ_{ij} . Notice that the anechoic cases is a particular case of the convulsive one, where $\tilde{h}_{ij} = a_{ij}\delta(t - \tau_{ij})$

of convolution between the RIRs and source signal:

$$\tilde{c}_{ij}(t) = (\tilde{h}_{ij} * \tilde{s}_j)(t) \quad (3.4)$$

$$\tilde{\mathbf{c}}_j(t) = [\tilde{c}_{1j}(t), \dots, \tilde{c}_{Ij}(t)]^\top$$

$$\tilde{\mathbf{x}}(t) = \sum_{j=1}^J \tilde{\mathbf{c}}_j(t). \quad (3.5)$$

Considering the time domain description of the RIR derived (and approximated) in the previous chapter, the time-domain *mixing filters* $\tilde{h}_{ij}(t)$ will be modeled as follows:

$$\tilde{h}_{ij}(t) = \sum_{r=0}^R \frac{\alpha_{ij}^{(r)}}{4\pi c \tau_{ij}^{(r)}} \delta(t - \tau_{ij}^{(r)}) + \tilde{\varepsilon}_{ij}(t) \quad (3.6)$$

where $\alpha_{ij}^r \in \mathbb{R}$ and $\tau_{ij}^r \in \mathbb{R}$ are the attenuation coefficient and the time delay of the reflection r . The noise term $\tilde{\varepsilon}_{ij}(t)$ collects later echoes ($r > R$) and the tail of the reverberation. We do not assume $\tilde{\varepsilon}_{ij}(t)$ to be known.

3.1.2 Noise, interferer and errors

In Eq. (3.5) no noise is included: all the sources are treated in the same way, including *target*, *interfering* and *noise* sources. While the definition of target sound source is quite self-explanatory (and will be denoted by default as the first source, that is $j = 1$), the term interfer and noise depends on the specific use case, problem, application, and research field. Notice that in Eq. (3.6) a noise term is added to gather unknown quantities.

Noise is a general term for unwanted (and, in general, unknown) modifications that a signal may suffer during capture, storage, transmission, processing, or conversion [Tuzlukov 2018].

Therefore, we will define and use the following type of noises:

- ▶ INTERFERS identifies undesired sources with properties similar to the target source. For instance, a concurrent speech source for speech application or concurrent music instrument in case of music. Later, in this thesis the interfer sources will be denoted as additional sources indexed by $j > 1$.
- ▶ NOISE collects all the remaining effects, typically nonspeech sources. Moreover we will make a further distinction between the followings.
 - *Diffuse Noise Field* describes the background diffuse sources present in the auditory scene, e.g. car noise, indistinct talking or winds. It can be recorded or approximated as Additive White Gaussian Noise (AWGN) with a specific spatial description as described in [Habets and Gannot 2007].
 - *Measurement and Model Noise* accounts for general residual mis-modeling error. In this manuscript, it will be denoted as $\tilde{\varepsilon}_{ij}(t)$ and will be modeled as AWGN. Examples of this term are the error due to approximation of the RIR with the Image Source Method (ISM) or sensor noise, respectively.

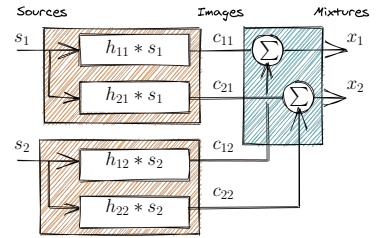


FIGURE 3.3: Graphical representation of the mixing model 3.5 for 2 sources and 2 microphones. The text is overlapping on purpose.

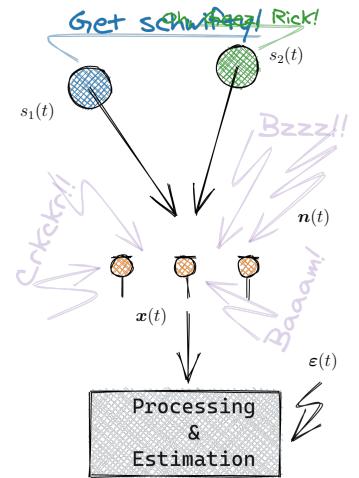


FIGURE 3.4: Graphical representation of the mixing model (3.5): $s_2(t)$ is the *interferer*, $n(t)$ contributes to the *diffuse noise field*, and $\varepsilon(t)$ model acquisition and modeling errors. With proper prior knowledge, it can be understood.

By making the noisy terms explicit, the mixing model in Eqs. (3.4) and (3.5) writes:

$$\tilde{c}_{ij}(t) = (\tilde{h}_{ij} \star \tilde{s}_j)(t) + \tilde{\varepsilon}_{ij}(t) \quad (3.7)$$

$$\begin{aligned} \tilde{\mathbf{c}}_j(t) &= [\tilde{c}_{1j}(t), \dots, \tilde{c}_{Ij}(t)]^\top \\ \tilde{\mathbf{x}}(t) &= \sum_{j=1}^J \tilde{\mathbf{c}}_j(t) + \tilde{\mathbf{n}}(t) \end{aligned} \quad (3.8)$$

3.2 SIGNAL MODEL IN THE SPECTRAL DOMAIN

The frequency, or spectral, representation is probably the most famous signal representation used in signal processing: Speech and music signals naturally exhibit pseudo-periodic behaviors (on well chosen time scale). In this domain, they are described as combination of sinusoids as function of their frequencies. This operation is achieved by the operator Fourier Transform (**FT**), \mathcal{F} , which projects a continuous-time-domain square-integrable signal \tilde{x} onto a space spanned by continuous-frequency complex exponentials:

$$\tilde{X}(f) = (\mathcal{F}\tilde{x})(f) = \int_{-\infty}^{+\infty} \tilde{x}(t) e^{-i2\pi ft} dt \in \mathbb{C}, \quad (3.9)$$

where $f \in \mathbb{R}$ are the *natural frequency* in Hz and i is the imaginary unit.

A part from providing a space where audio signal reveals their harmonic structures, the Fourier transforms benefits of two fundamental properties: it is linear and it converts time-convolution into element-wise products.

First, linearity allows to write Eq. (3.5) simply as:

$$\tilde{\mathbf{x}}(t) = \sum_{j=1}^J \tilde{\mathbf{c}}_j(t) \xrightarrow{\mathcal{F}} \tilde{\mathbf{X}}(f) = \sum_{j=1}^J \tilde{\mathbf{C}}_j(f) \quad (3.10)$$

Secondly, by the *convolution theorem*, the source spatial images in Eq. (3.4) writes as:

$$\tilde{c}_{ij}(t) = (\tilde{h}_{ij} \star \tilde{s}_j)(t) \xrightarrow{\mathcal{F}} \tilde{C}_{ij}(f) = \tilde{H}_{ij}(f) \tilde{S}_j(f). \quad (3.11)$$

As discussed in ??, assuming a pure echo model, the **FT** of **RIR**, a.k.a. the Room Transfer Function (**RTF**), can be computed exactly in closed-form as

$$\tilde{H}_{ij}(f) = \sum_{r=0}^R \alpha_{ij}^{(r)} e^{-i2\pi f \tau_{ij}^{(r)}}. \quad (3.12)$$

In practice, the filters \tilde{h}_{ij} are neither available in the continuous time domain nor in the continuous frequency domain directly. They must be estimated from the observation of the discrete-time mixtures $x_i[n]$, therefore, after the convolution with a source and the measurement process. In practice, we don't have access to continuous signals, neither in time nor in frequency domain. Every signal or spectrum the microphones capture are represented by finite- and discrete time signals for which the properties (3.11) are valid only with some precautions.

The frequency analysis was introduced by Joseph Fourier in his work on the heat equation [Fourier 1822]. His mathematical tool, named later Fourier Decomposition, aims at approximating any signal by a sum of sine and cosine waves.

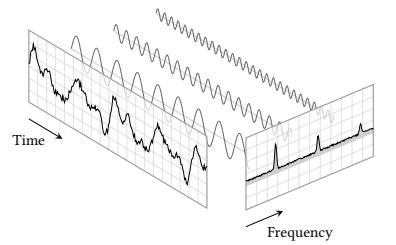


FIGURE 3.5: A signals resolved into its Fourier series: a linear combination of sines and cosines represented as peaks in the frequency domain.

3.2.1 Discrete time and frequency domains

In case of discrete-time signal, $x[n]$ with $n \in \mathbb{Z}$ and frequency support in $[-F_s/2, F_s/2]$, its spectral representation is given by the (forward) Discrete-Time Fourier Transform (**DTFT**), \mathcal{F}_{F_s} :

$$\tilde{X}_{F_s}(f) = (\mathcal{F}_{F_s} x)(f) = \sum_{n=-\infty}^{+\infty} x[n]e^{-i2\pi fn/F_s}, \quad (3.13)$$

which is a continuous function of f with period F_s . Notice that the term “discrete-time” refers to the fact that the transform operates on discrete signal. When these samples are uniformly spaced at rate F_s , it produces a function of continuous frequency that is a periodic summation of the continuous Fourier transform of the original continuous function. Under certain theoretical conditions, described by the *sampling theorem*, both the original continuous signal \tilde{x} and its sampled version x can be recovered perfectly from the **DTFT**.

The analogous convolution theorem for discrete-time domain signals $s[n]$ and $h[n]$ is

$$c[n] = (h \star s)[n] \xrightarrow{\mathcal{F}_{F_s}} \tilde{C}_{F_s}(f) = \tilde{H}_{F_s}(f)\tilde{S}_{F_s}(f), \quad (3.14)$$

where the dependency on the source and microphone index have been omitted for clarity.

The **DTFT** itself is a continuous function of frequency which require infinite discrete value to be computed. For these two reason, it is not accessible in practice or computed in the digital domain. Therefore the following representation is used instead.

For a discrete- and finite-time signal x_N , the spectral representation is given by its (forward) Discrete Fourier Transform (**DFT**)¹⁸, \mathbf{F} :

$$X_F[k] = (\mathbf{F} x_N)[k] = \sum_{n=0}^{N-1} x_N[n]e^{-i2\pi kn/F}. \quad (3.15)$$

where $k \in [0, F - 1]$ is the discrete *frequency bin* and F is the total number of bins. Again we use the subscript F and the brackets $[k]$ to stress the finite and discrete frequency support of the **DFT**. Note that now the sampling frequency term disappeared from the definition of the operators and, in general, operates directly on “sequences”. Therefore its link with the continuous **FT** for continuous-time domain signal is not trivial and it will be discussed in the following section.

Similarly to the **FT** and the **DTFT**, the convolution theorem can be defined for the **DFT** as well. However, it comes with following important modification. Let be two finite- and discrete-time domain *periodic* signals \ddot{s}_N and \ddot{h}_N of period N . Here $\ddot{\cdot}$ is used to denote periodicity. Let be S_F and H_F their **DFT**, respectively, then the convolution theorem for the **DFT** writes

$$\ddot{c}_N[n] = (\ddot{h}_N \circledast \ddot{s}_N)[n] \xrightarrow{\mathbf{F}} C_F[k] = H_F[k]S_F[k], \quad (3.16)$$

where \circledast denotes the *circular convolution*¹⁹. Note that now the results of the convolution, $\ddot{c}_N[n]$, is periodic with period N .

¹⁸ This can be interpreted as the projection onto the space spanned by a finite number of complex exponentials.

¹⁹ The finite-time circular convolution for two vectors $u_N, v_N \in \mathbb{R}^N$ is the $\ddot{q}_N = (u_N \circledast v_N)[n] = \sum_{m=0}^{N-1} u_N[m]\ddot{v}_N[n-m]$, where \ddot{v}_N is the periodic version of v_N and \ddot{q}_N is periodic.

3.2.2 The DFT as approximation of the FT

An important application of the DFT is to approximate numerically the FT. As mentioned at the beginning of the chapter, with the discretisation process the continuous signal is periodically sampled, low-passed and finally truncated. It can be proved that sampling in the time domain corresponds to periodizing the signal spectrum with period equal to the sampling frequency. In order to avoid aliasing artifact, it is common to limit the signal bandwidth before the sampling (and using the lowpass filter in Eq. (3.2)). By assuming sampling at rate F_s , in the continuous-frequency domain the spectrum $\tilde{X}(f)$ is repeated every intervals of size F_s Hz. By further assuming that the signal undergoes an ideal low-pass filter, no spectral leakage is present between each repetition.

So far, the sampled time domain signal, $x[n]$, is mapped to the continuous frequency domain $\tilde{X}(f)$. This particular case of the FT is called Discrete-Time Fourier Transform (DTFT) and it is denoted with $\tilde{X}_{F_s}[k]$.

$$\tilde{X}(f) = \int_{-\infty}^{+\infty} \tilde{x}(t) e^{-i2\pi f t} dt \rightarrow \tilde{X}_{F_s}(f) = \sum_{n=-\infty}^{\infty} x[n] e^{-i2\pi f \frac{n}{F_s}}. \quad (3.17)$$

Here the continuous integral the FT is approximated by Riemann sum over the discrete points $n \in \mathbb{Z}$: To be more rigorous, when computing a Riemann sum approximation, the length of the discretisation interval multiply the summation. In our application, this quantity always set to F_s and for readability reason such term is dropped. The quality of this approximation w.r.t. the original continuous spectrum is regulated by the choice of F_s : the higher F_s , the better the approximation. The upper bound to the possible value F_s is the results known as the Nyquist–Shannon’s sampling theorem.

Furthermore, we consider only the finite sequence x_N consisting of N samples. This would reduce the summation ranges the right part of Eq. (3.17). Instead, we can keep the infinite summation by multiplying the sampled signal by a discrete-time window function w selecting the non-zero portion of x , $x_N[n] = w[n]x[n]$. By the *convolution theorem*, the multiplication in the time domain translates in a convolution between the corresponding spectra. As a consequence, the spectrum of the truncated signal is distorted by the spectrum of the window function. In math,

$$\tilde{X}_{F_s}(f) = \sum_{n=-\infty}^{\infty} x[n]w[n]e^{-i2\pi f \frac{n}{F_s}} \leftrightarrow \tilde{X}_N(f) = \sum_{n=0}^{N-1} x_N[n]e^{-i2\pi f \frac{n}{F_s}}. \quad (3.18)$$

By the convolution theorem, we have that

$$x_N[n] = x[n]w[n] \leftrightarrow \tilde{X}_N(f) = (\tilde{X}_{F_s} \star \tilde{W}_{F_s})(f) \quad (3.19)$$

where \tilde{W}_{F_s} is the DTFT of the sampled window function $w[n]$.

Assuming the window function to be an ideal door function²⁰, its DTFT is a ideal low-pass filter, which acts on the original spectrum as a smoothing function. As a consequence, the quality of this approximation is then based on the spectral leakage of the chosen window function, $w[n]$. As a rule of thumb, here the longer the segment, the better the approximation²¹

²⁰door function here

²¹When short excerpt are considered instead (e.g. in case of the Short Time Fourier Transform (STFT)), particular types of window function are used but their analysis are out of the scope of this thesis.

Finally, we cannot access the DTFT directly because that involves an infinite number of frequencies $f \in \mathbb{R}$. Therefore, taking F uniformly-spaced frequency $f_k \in \mathbb{R}$ as in Eq. (3.24), we finally obtain the DFT as in Eq. (3.15), that is

$$\tilde{X}_N(f_k) = \sum_{n=0}^{N-1} x_N[n] e^{-i2\pi f_k \frac{n}{F_s}} \leftrightarrow X_F[k] = \sum_{n=0}^{N-1} x_N[n] e^{-i2\pi kn/F}. \quad (3.20)$$

Notice that the F_s term disappeared in the right part of the equation above as it cancels out when using Eq. (3.24). By increasing F , we can sample more densely $X_F[k]$ which leads to a better approximation to \tilde{X}_N . However this does not eliminate the distortion of the previous steps, due to \tilde{W}_{F_s} .

Again, we sampled a domain. Thus, according to the defined sampling process, this involves using an ideal low-pass filter. This filter acts now on the discrete spectrum, smoothing it and limiting the support of its transformation in the dual domain. Therefore, the inverse DFT of $X_F[k]$ is not properly $x_N[n]$, but its periodic version repeated every F samples. In fact, sampling in one of the two domains is equivalent to a periodization in the other domain while truncating lead to convolving with a window function. Moreover, the chain of operations (sampling in time and truncation in time and sampling in frequencies) are valid in both ways. Thus one can arbitrarily first sample and truncate frequency domain and finally sample in time. The only difference is in the interpretation of the windowing function, which in one case smooths the spectrum and in the other smooths the signal. All these relations and approximations that connect the FT to the DFT are well explained in this explanatory material Poisson Summation Formula, Revisited \square .

3.2.3 Signal models in the discrete Fourier domain

Conscious of the above approximations, we can now rewrite our signal model for the discrete case. Hereafter we will always consider finite-length sequences and the index N will be dropped to lighten the notation.

The DFT is linear, so the discrete version of Eq. (3.10) becomes

$$\mathbf{x}[n] = \sum_{j=1}^J c_j[n] \xrightarrow{\mathbf{F}} \mathbf{X}[k] = \sum_{j=1}^J \mathbf{C}_j[k] \quad (3.21)$$

Secondly, by using naïvely the discrete convolution theorem, one could translate Eq. (3.4) as

$$c_{ij}[n] = (h_{ij} * s)[n] \xrightarrow{\mathbf{F}} C_{ij}[k] = H_{ij}[k]S[k], \quad (3.22)$$

where $*$ is the finite-time linear convolution operator²².

The filter $H_{ij}[k]$ is the DFT of the room impulse response. As mentioned in the § 3.2.2, this just approximates the true RTF defined in the previous chapter. We recall that two approximations are involved: one concern that we model only a finite number of specular reflections; one concern sampling its FT. Nevertheless, we can use the closed-form frequency-domain echo model Eq. (2.16) to write the RTF's DFT in closed-form, that is

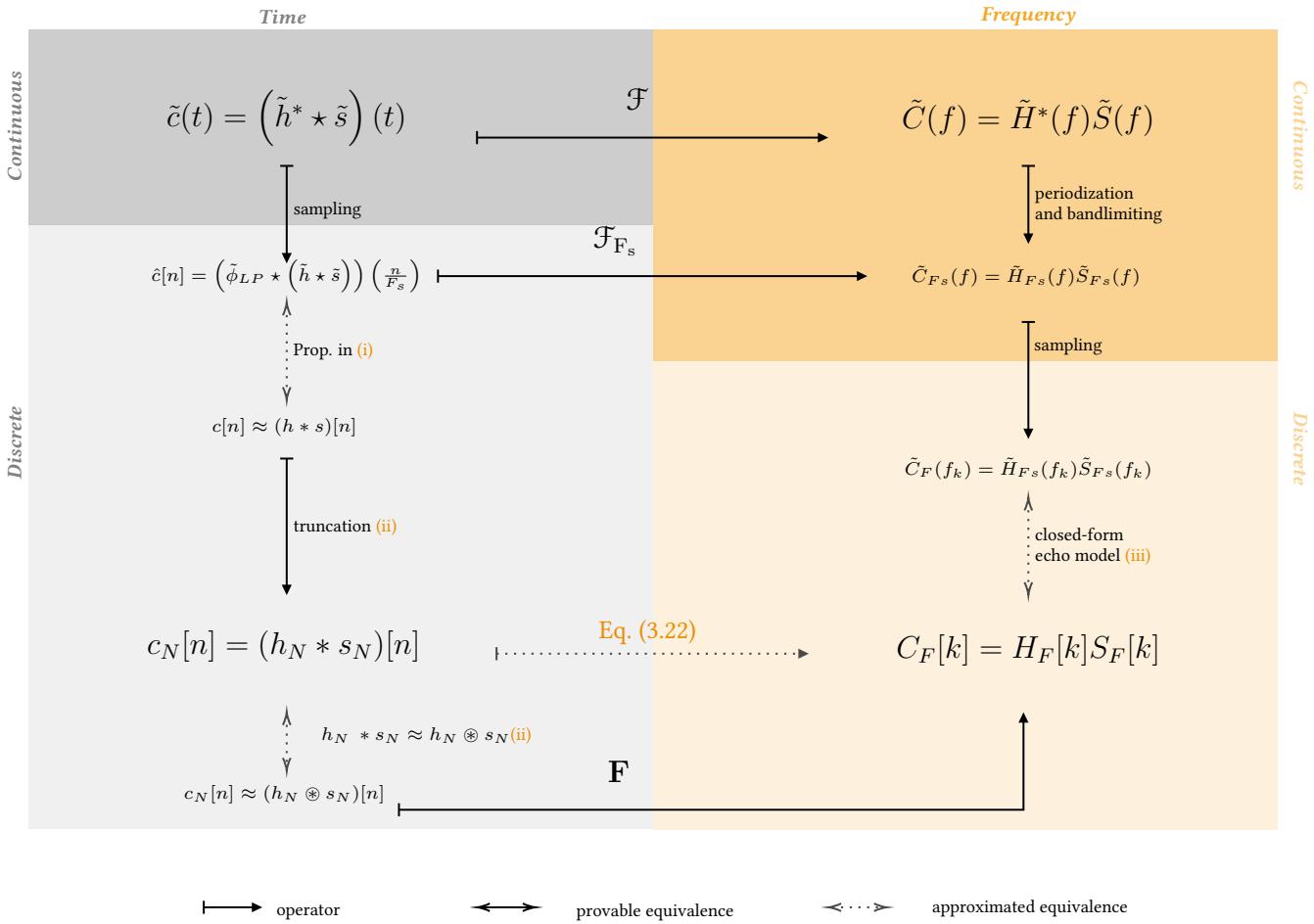
$$H_{ij}[k] = \sum_{r=0}^R \frac{\alpha_{ij}^r}{4\pi c \tau_{ij}^r} e^{-i2\pi k F_s \tau_{ij}^r / F}. \quad (3.23)$$

²² The finite-time linear convolution for two vectors $u \in \mathbb{R}^L$ and $v \in \mathbb{R}^D$ is $(u * v)[n] = \sum_{l=0}^{L-1} u[l]v[L-1+n-l]$ for $n = 0, \dots, D-L$.

Then, the natural frequency f_k in Hz corresponding to the k -th frequency bin can be computed as

$$f_k = \frac{k}{F} F_s. \quad (3.24)$$

- ALTHOUGH USED IN PRACTICE, the model (3.22) makes use of other approximations that are worth presenting. In particular, the work by [Tukuljac et al. 2018] properly discuss them in the context of the echo estimation problem. The paper mention three approximations, which are depicted in the following diagram. The diagram shows a chain of operators (sampling and transforms)



with provable and approximated equivalences that lead to Eq. (3.22) used in practice. In order,

- (i) In [van den Boomgaard and van der Weij 2001, Proposition 2] the authors show that if the signal $\tilde{s}(t)$ is band-limited by F_s , then sampling the continuous convolution is *exactly* equivalent to *linearly convolving* the infinite discrete signal $s[n]$ and the discrete and low-passed version of the filter. However, while the source signal is band-limited by nature, the true RIR $\tilde{h}^*(t)$ is not (in fact the RIR is modeled as a finite summation of spikes, which has spectrum with infinite support). Thus, the first

approximation (i) considers $h[n] \approx (\tilde{\phi}_{LP} * \tilde{h}^*)[n]$, in words we assume that the filter is band-limited by $\pm F_s/2$.

- (ii) As introduced in § 3.2.1, the discrete-time convolution theorem applies to the *circular convolution* involving periodic signals. This can be approximated by the *linear convolution*, that is $(h \circledast \hat{s})[n] \approx (h * \hat{s})[n]$ under some assumption. In particular such an approximation is reasonably good when many samples are available and when one of the two signals is quasi-periodic, which are typical cases for audio signals.
- (iii) The third approximation regards the closed-form of the **RTF**, $\tilde{H}^*(f)$. Here we make two strong assumptions: first, the **RTF** follows the echo model discussed in Eq. (2.16); second, we used its closed-form **DTFT** (Eq. (3.23)) at some frequency f_k as its **DFT**. In order to be computed exactly, this would require infinitely many samples and unlimited frequency support, which are not possible in practice.

Nevertheless, it is important to notice that approximations (ii) and (iii) become arbitrarily precise as the number of samples N grows to infinity. First, the **DFT** approaches the **DTFT** as $N \rightarrow \infty$ due to their definition (see § 3.2.1). Second, while the convolution theorem is exact for the **DTFT**, for the **DFT** it is only for considering the circular convolution on a finite-time signal (or, equivalently, considering periodic signals). Therefore, it is easy to notice that when the period becomes infinite, the circular convolution and the linear convolution become arbitrarily close. In the approximation (i), limiting the **RIR**'s spectrum with an ideal lowpass filter introduces smearing in the time-domain. Tukuljac et al. made an important observation in this regard: even if infinite number of samples are available, after the measurement process, the discrete-time filter $h[n]$ consists of infinite-length decimated combinations of sinc functions. In the context of this thesis, this observation tell us that even in ideal conditions, that is without noise, possibly knowing the transmitted signal, and processing infinitely many samples, the exact estimation of the echo properties of the **RIR** is a challenging task itself. This is a fundamental difference between **RIR** estimation and estimating the time of arrivals of the early echoes. Note, for instance, that we wrote the echo model only in the continuous-time domain or with its closed-form form discrete frequencies. The discrete-time domain was avoided on purpose since the echoes' arrival time are naturally off the sampling grid, namely they are not integer multiples of F_s . This fundamental limitation is subject of research in the field of *super-resolution* and it will be discussed in Chapter 5.

While the raw audio signal encodes the amplitude of a sound as a function of time, its spectrum represents it as a function of frequency. In order to jointly account for both temporal and spectral characteristic, joint time-frequency representations are used.

3.2.4 Time-Frequency domain representation

Time-Frequency (**TF**) representations aim to jointly describe the signal in the time and frequency domains. Instead of considering the entire signal, the main idea is to consider only a small section of the signal. To this end, one

fixes a so-called *window* function, $w_N[n]$, which is nonzero for only a period of time L_{win} shorter than the entire signal length, $L_{\text{win}} \ll N$. This function iteratively shifts and multiplies the original signal, producing consecutive *frames*. Finally, the frequency information are extracted independently from each frame. The choice of a window function $w[n]$ depends on the application since its contribution reflects in the **TF** representation together with the one of the signal.

The discrete **STFT** is the most commonly used **TF**-representation in audio signal processing. This representation encodes the time-varying spectra into a matrix $X[k, l] \in \mathbb{C}^{F, T}$ with frequency index k and time frame index l . More formally, the process to compute the complex **STFT** coefficients is given by

$$X[k, l] = \sum_{n=0}^{L_{\text{win}}-1} w[n]x[n + lL_{\text{hop}}]e^{-i2\pi kn/F} \in \mathbb{C} \quad (3.25)$$

where L_{win} is the window length and L_{hop} is the *hop size* which specifies how much the window needs to be shifted across the signal. Equivalently, Eq. (3.25) can be expressed as **DFTs** of windowed frames, $X[k, l] = \mathbf{F} x[n, l]$ where $x[n, l] = x[n + lL_{\text{hop}}]w[n]$.

Since each **STFT** coefficient $x[k, l]$ lives in the complex space \mathbb{C} , the squared magnitude of the **STFT**, $|X[k, l]|^2$ is commonly used for visualization and for processing. The resulting two-dimensional representation is called (log) *spectrogram*. It can be visualized by means of a two-dimensional image, whose axes represent time frames and frequency bins. In this image, the (log) value $|X[k, l]|^2$ is represented by the intensity or color in the image at the coordinate $[k, l]$. Throughout this work processing will be typically conducted in the **STFT** domain, unless specified. This is a common approach in the audio signal processing community, but it is not the only one: many algorithm are designed directly in the time domain or in alternatives **TF** representation, e. g., Mel-Scale, Filter-Banks.

As discussed [Vincent et al. 2018], the **STFT** has the following useful properties for audio processing:

- the frequency f_k is a linear function of the frequency bin k ;
- the resulting matrix allows the computation of the phase $\angle X[k, l]$, the magnitude $|X[k, l]|$ and the power $|X[k, l]|^2$;
- the **DFT** can be efficiently computed with the Fast Fourier Transform (**FFT**) algorithm;
- the process to invert the **STFT** is relatively easy (using the inverse of the **DFT**, overlap-add method with proper window functions);
- the **STFT** inherits the linearity and convolution property of the **DFT** under some condition about the length of the signals.

The **STFT** was introduced by Dennis Gabor in the 1946, the person behind Holography.

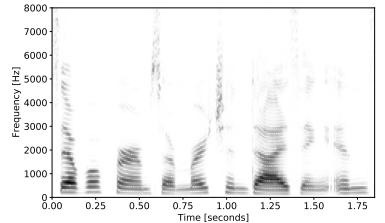


FIGURE 3.6: STFT spectrogram of an example speech signal. Higher energies are illustrated with darker colors.

For more mathematical detailed description on **DFT** and **STFT** can be found in [Oppenheim 1987]. For a audio-processing-oriented and music-processing-oriented explanation please refer to Chapter 2 of [Vincent et al. 2018] (Chapter2) and Chapter 2 of [Müller 2015], respectively.

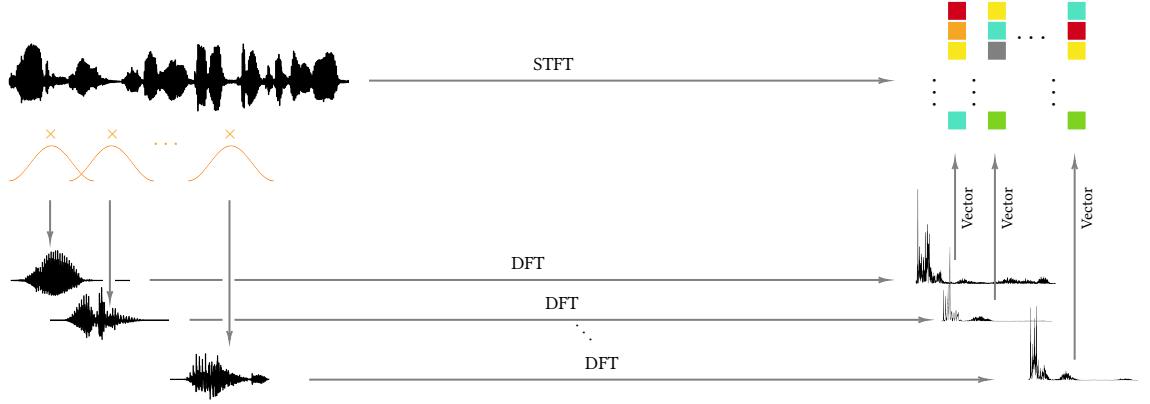


FIGURE 3.7: Illustration of the STFT.

3.2.5 The final model

The model (3.22) shows how in practice the RIRs are treated in the frequency-domain. However this does not generalize straightforwardly to the time-frequency domain: it depends on the length of the filter w. r. t. to the length of the analysis window on of the STFT. Issues arise with “long” filters, which are common in highly reverberant or time-varying scenarios. To circumvent this issue, the *convolutional STFT* for arbitrary window functions have been proposed²³ [Gilloire and Vetterli 1992]. Although mathematically exact, it is computationally and memory intensive.

In this thesis, we will assume that the filter length is shorter than the analysis window length. In the literature, this is known as the *narrowband approximation*, namely the time-domain filtering can be approximated by complex-valued multiplication in each time-frequency bin $[l, k]$:

$$C_j[l, k] \approx \mathbf{H}[k]S_j[l, k], \quad (3.26)$$

where the $\mathbf{H}_j[k] = [h_{1j}[k], \dots, h_{Ij}[k]]^\top$ is the $I \times 1$ vector of the RTFs for source j . It is sometimes practical to concatenate all these vectors into an $I \times J$ matrix $\mathbf{H}[k] = [\mathbf{H}_1(k), \dots, \mathbf{H}_J(k)]$ called *mixing matrix*.

With the above notation and considerations, mixing process including noise terms can be written in the STFT domain compactly as:

$$\mathbf{X}[l, k] = \mathbf{H}[l, k]\mathbf{S}[l, k] + \mathbf{U}[l, k] \quad (3.27)$$

where $\mathbf{U}[l, k] = \mathbf{N}[l, k] + \boldsymbol{\varepsilon}(l, k)$ includes the contribution of both diffuse noise sources, modeling and measurement errors.

3.3 OTHER (ROOM) IMPULSE RESPONSE SPECTRAL MODELS

RIRs are complicated quantities to model, compute and estimate. The representations of the RIR discussed so far explicitly models early echoes and reverberation deterministically. Furthermore, alternative models are common in the audio processing literature.

²³It translates the time-domain convolution into inter-frame and inter-band convolutions, rather than pointwise multiplication of Fourier transforms.

3.3.1 Steering vector model

In the absence of echoes and reverberation, namely assuming free-field propagation, the RIRs simplify to *steering vectors*, namely the DFT of Eq. (2.9):

$$\mathbf{D}_j[k] = \left[\frac{1}{4\pi q_{1j}} e^{-i2\pi f_k q_{1j}/c}, \dots, \frac{1}{4\pi q_{Ij}} e^{-i2\pi f_k q_{Ij}/c} \right] \quad (3.28)$$

Furthermore, assuming far-field regimes, the microphone-to-source distance q_{ij} is larger than the inter-microphone distance $d_{ii'}$ making the attenuation factors $1/4\pi q_{ij}$ approximately equal, hence ignored.

3.3.2 Relative transfer function and interchannel models

Let us consider now only two channels and only one source signal in the model Eq. (3.27). Dropping the dependency on j for readability and taking the first channel as reference, the Relative Transfer Function (ReTF)²⁴ associated to the i -th channel is defined as the element-wise ratio of the (D)FTs of the two filters [Gannot et al. 2001]

$$G_i[k] = \frac{H_i[k]}{H_1[k]}. \quad (3.29)$$

The continuous-time domain counterpart is called as Relative Impulse Response (ReIR) and can be interpreted as the filter “transforming” the i -th impulse response into the one of the reference channel. Considering the noisy observation \tilde{x}_i and \tilde{x}_1 , their signals can be re-written in term of \tilde{g}_i as follows

$$\begin{cases} \tilde{x}_1 = \tilde{h}_1 * \tilde{s} + \tilde{u}_1 \\ \tilde{x}_i = \tilde{h}_i * \tilde{s} + \tilde{u}_i \end{cases} \rightarrow \begin{cases} \tilde{x}_1 = \tilde{h}_1 * \tilde{s} + \tilde{u}_1 \\ \tilde{x}_i = \tilde{g}_i * \tilde{h}_i * \tilde{s} + \tilde{u}_i \end{cases}. \quad (3.30)$$

Notice that $\tilde{h}_i = \tilde{g}_i * \tilde{h}_1$ corresponds to Eq. (3.29) in the frequency domain. Moreover although the real-world RIRs h_1 and h_i are causal, their ReTF needs not be so.

The ReTF benefits of several interesting properties that will be of fundamental importance for this thesis. In particular:

- the ReTF associated to the reference channel ($i = 1$) is equal to 1 for each frequency bin k ;
- the problem of estimating the ReTF can be considered “easier” than RIRs estimation; In fact, in the noiseless case, it holds that $\tilde{x}_i = \tilde{g}_i * \tilde{x}_1$. Since \tilde{x}_i and \tilde{x}_1 are both available, the ReTF estimation boils down to transfer function estimation with known input signal;
- in the literature, many efficient and robust methods have been proposed to estimate ReTF [Gannot et al. 2001; Doclo and Moonen 2002; Markovich et al. 2009; Koldovsky and Tichavsky 2015; Kodrasi and Doclo 2017; Tammen et al. 2018];
- a RIR can be seen as a special case of ReTF where the non-reference microphone is a virtual one whose output is the original (non-spatial) source signal s . In fact, if $h_1 = \delta$ then $g_i = h_i$ ²⁵

²⁴In the literature of spatial filtering, the Relative Transfer Function is typically denoted as RTF, which collides with the notation proposed in the thesis. We reminds here that we use the following notation: Room Transfer Function (RTF) and Relative Transfer Function (ReTF).

²⁵In practice this virtual microphone is sometimes substituted by a microphone that is very close to the source.

- as discussed below, ReTF simplify to special steering vectors in free- and far-field conditions, which have interesting geometrical properties;
- the ReTF encodes acoustics properties of the related RIRs and is independent on the source signal. They can be used as *features* for the estimations of parameter of the acoustic propagation. Examples on this are many works in the Sound Source Localization domain (e.g., [Laufer et al. 2013; Deleforge et al. 2015; Di Carlo et al. 2019a].

In the general case of multiple microphone arrays ($I > 2$) and multiple sources, the vector of ReTF $\mathbf{G}_j[k] = [G_{1j}[k], \dots, G_{Ij}[k]]^\top$ for the j -th source is defined as

$$\mathbf{G}_j[k] = \frac{1}{G_{1j}[k]} \mathbf{G}_j[k]. \quad (3.31)$$

- ▶ THE RELATIVE STEERING VECTORS are obtained by combining Eqs. (3.28) and (3.29) as

$$\mathbf{D}_j[k] = \left[1, e^{-i2\pi f_k(q_{2j} - q_{i'j})/c}, \dots, e^{-i2\pi f_k(q_{Ij} - q_{i'j})/c} \right] \quad (3.32)$$

where $(q_{ij} - q_{1j})/c$ is the Time Difference of Arrival (TDOA) between the i -th and the reference microphones. The TDOAs will be the protagonists of ?? as they are fundamental quantities for sound source localization.

- ▶ IN THE CONTEXT OF SPATIAL AUDITORY PERCEPTION and Computational Auditory Scene Analysis (CASA), the ReTF is related to the *interchannel cues*²⁶. In fact, the ReTF encodes the so-called Interchannel Level Difference (ILD) and the Interchannel Phase Difference (IPD)

$$\begin{aligned} \text{ILD}_{ij}[k] &= 20 \log_{10} |g[k]| \quad [\text{dB}] \\ \text{IPD}_{ij}[k] &= \angle g[k] \quad [\text{rad}] \end{aligned} \quad (3.33)$$

As shown in Figure 3.8, the ILD and the IPD cluster around the direct path, associated to the direct path component. However early echoes and reverberation make them significantly diverge.

²⁶sometimes refers to as *interaural cues* when a stress is put on the fact that the two ears are considered as receivers

- * IN THIS CHAPTER, we presented the mathematical notions, notation, and tools to model audio scene analysis problems in signal processing. We particularly focused on the crucial approximations and assumptions that arise from the merging physical and signal processing models. As we conclude the part of the thesis dedicated to background knowledge, we will now focus on estimating acoustic echoes in Part II, and how to use them for solving audio scene analysis problems Part III.

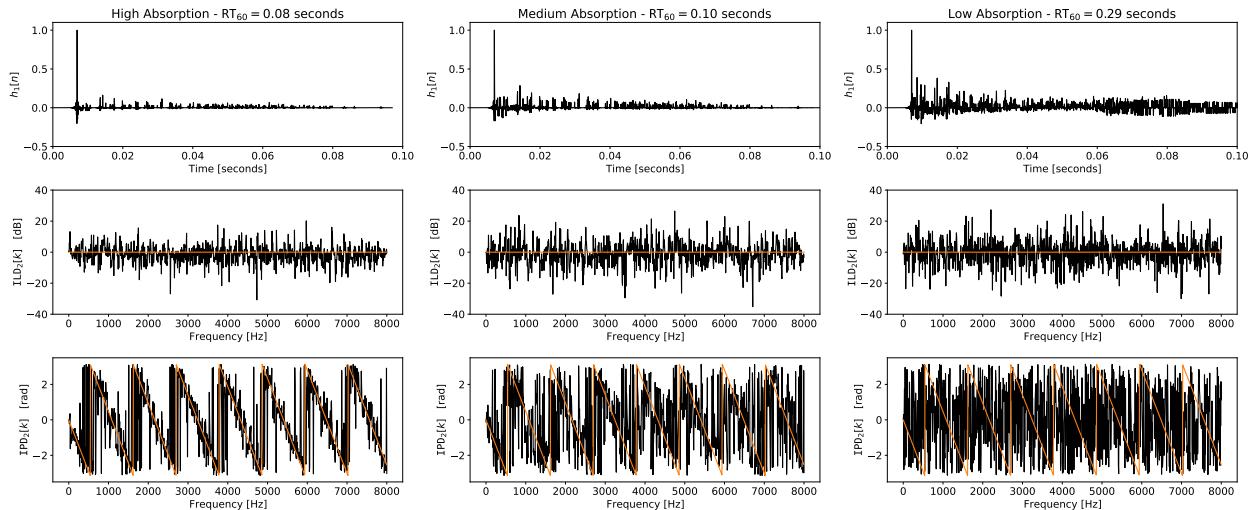


FIGURE 3.8: RIR, ILD and IPD corresponding to the pair of synthetic impulse responses of Figure 2.9 for different absorption conditions. Orange lines denote the theoretical far- and free- field ILD and IPD as defined by the relative steering vectors of Eq. (3.32)

Part II

ACOUSTIC ECHO RETRIEVAL

4 ACOUSTIC ECHO RETRIEVAL	
4.1 Problem Formulation	47
4.2 Taxonomy of Acoustic Echo Retrieval methods	48
4.3 Literature Review	49
4.3.1 Active and RIR-based method	49
4.3.2 Active and RIR-agnostic method	53
4.3.3 Passive and RIR-based method	54
4.3.4 Passive and RIR-agnostic methods	57
4.4 Data and Evaluation	57
4.4.1 Datasets	57
4.4.2 Metrics	58
4.5 Proposed Approaches	60
5 blaster: KNOWLEDGE-DRIVEN ACOUSTIC ECHO RETRIEVAL	
5.1 Introduction	62
5.2 Signal model	63
5.3 Background on On-grid Blind Channel Estimation	64
5.3.1 From cross-relation to LASSO	66
5.4 Proposed Approach	66
5.4.1 Echo localization with continuous dictionaries	67
5.4.2 From LASSO to BLASSO	69
5.4.3 The resulting algorithm	69
5.4.4 Homotopic path for λ estimation	70
5.5 Experiments	70
5.6 Conclusion	73
6 lantern: DATA-DRIVEN ACOUSTIC ECHO RETRIEVAL	
6.1 Introduction	74
6.2 Deep Learning Model	76
6.2.1 Multi-layer perceptions	76
6.2.2 Convolutional neural networks	76
6.2.3 Hybrid architectures	77
6.2.4 Learning and optimization	77
6.2.5 End-to-end and Hybrid approaches	78
6.3 Learning the first echo	78
6.3.1 Scenario and motivations	78
6.3.2 Proposed approach	79
6.3.3 Data, implementation and experimental results	80
6.4 Robustly learning the first echo	81
6.4.1 Gaussian- and Student's T-based CNNs	82

6.4.2	Experimental results	83
6.5	Conclusion	84
7	dechorate: DATASETS FOR ACOUSTIC ECHO ESTIMATION	
7.1	Introduction	86
7.2	Database realization	87
7.2.1	Recording setup	87
7.2.2	Measurements	88
7.3	Dataset annotation	89
7.3.1	RIRs annotation	89
7.3.2	Other tools for RIRs annotation	91
7.3.3	Limitations of current annotation	92
7.4	The dEchorate package	92

4

Acoustic Echo Retrieval

- ▶ **SYNOPSIS** This chapter aims to provide the reader with knowledge in the state-of-the-art of Acoustic Echo Retrieval (**AER**). After presenting the **AER** problem in § 4.1, the chapter is divided into three main sections: § 4.2 defines the categories of methods according to which the literature can be clustered and analyzed in details later in § 4.3. Finally, in § 4.4 some datasets and evaluation metrics for **AER** are presented.

4.1 PROBLEM FORMULATION

The continuous-time multi-channel signal model for one source signal $\tilde{s}(t)$ and I channels writes, for each channel i ,

$$\begin{aligned}\tilde{x}_i(t) &= (\tilde{h}_i \star \tilde{s})(t) \\ \tilde{h}_i(t) &= \sum_{r=0}^R \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \tilde{\varepsilon}_i(t),\end{aligned}\tag{4.1}$$

where $\tilde{h}_i(t)$ is the echo model for the **RIR** between the i -th channel and the source. The sum comprises the line-of-sight propagation (direct path) and the earliest R echoes we want to account for, while the error term $\tilde{\varepsilon}_i(t)$ collects later echoes and the reverberation tail.

THE ACOUSTIC ECHO RETRIEVAL (**AER**) PROBLEM CONSISTS in estimating the echoes' timings $\{\tau_i^{(r)}\}_{i,r}$ and attenuations (or gains) $\{\alpha_i^{(r)}\}_{i,r}$ of Eq. (4.1). Depending on the field of research, the echoes' timings are also known as time delays, Time of Arrival (**TOA**) or locations.

The term **AER** is not an established name for such problem and, depending on the field of research and the prior knowledge available, it can be referred to with different names. In fact **AER** can be seen as a generation of **TOAs estimation**, or an instance of *acoustic channel estimation* (or *shaping*), *spike retrieval* and *onset detection*. As opposed to **AER**, the task of **TOAs** estimation is only focused in estimating the echoes' timings $\{\tau_i^{(r)}\}_{i,r}$. The only knowledge of **TOAs** is sufficient for typical application related to **SSL** [Ribeiro et al. 2010a] and Room Geometry Estimation (**RooGE**) [Crocco et al. 2017].

Moreover knowing $\{\tau_i^{(r)}\}_{i,r}$, the attenuations $\{\alpha_i^{(r)}\}_{i,r}$ can be estimated in closed-form as showed in [Condat and Hirabayashi 2015].

TOAs estimation is sometimes called *time delays estimation*, when the origin of time is taken w. r. t. the first **TOA** and not to the time of emission. Hereafter

“[...] dicebat Bernardus Carnotensis nos esse
quasi nanos gigantium humeris insidentes.”
—Giovanni of Salisbury, *Metalogicon* (III, 4)

we will make distinction between the two.

The **AER** may be confused with the *acoustic echo cancellation* problem in telecommunication and telephony which refers to the problem of estimating and suppressing feedbacks due to loudspeakers being too close to microphone.

4.2 TAXONOMY OF ACOUSTIC ECHO RETRIEVAL METHODS

- ⇒ **ACTIVE VS. PASSIVE APPROACHES.** *Active* methods assume active scenarios, namely, they use one or more loudspeakers to probe the environment and one or more microphones to record the propagated probe sound. Therefore, they assume that the source reference signal is known. They falls into the big categories of *deconvolution problems* since a “clean” reference signal is used *deconvolve* the observed one. The main advantages of these approaches are twofold. First, provided a proper probe signal, a good estimation of the **RIR** can be achieved. Second, these methods can be used on single-channel recordings.

Instead, *passive* approaches use sets of passive sensors to record the sound field. To decouple the environment from the source, they rely either on prior knowledge about the source signal or by comparing the signals received at two (or more) spatially-separated microphones. When no prior information is available, they can be seen related to *blind problem* and are far more challenging. In the literature, the term *blind* is used as opposed to *informed*, accounting the amount of prior information available. With this taxonomy, the above active methods are informed by the exact knowledge of the reference signal. On the contrary, passive methods need to rely on other general type of information such as sparsity, nonnegativity, etc. In general the labels “active vs. passive” apply to the scenario, while “informed vs. blind” to the problem. Passive scenarios are more common in real applications and are sometimes preferable to active ones as they are non-intrusive since only already existing sounds are used in the estimation. Therefore, a great deal of efforts has been devoted to these problems and research is still active in topic.

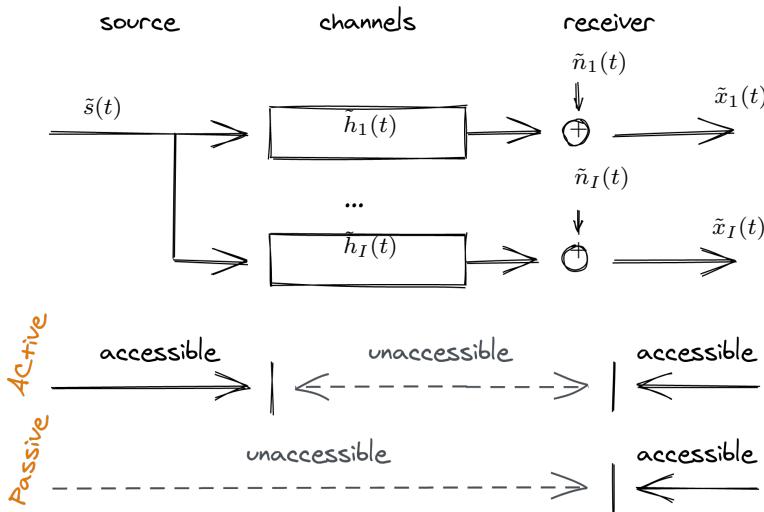


FIGURE 4.1: Schematic of active and passive approaches.

⇒ RIR-BASED VS. RIR-AGNOSTIC APPROCHES. *RIR-based* methods estimate the echoes' properties after estimating the (full or partial) **RIR**(s). By modeling the early part of the **RIR** as in Eq. (4.1), solving the **AER** problem can be seen as solving two subsequent tasks: **RIR** estimation followed by echo extraction. The former can be seen as an instance of *channel estimation* (a. k. a. *system identification*) problems, while the latter as a *spike retrieval*, *peak picking* or *onset detection* problem. Other methods estimate the **RIRs** partially using assumptions derived by the application. It is the case of *impulse response shaping* or *shortening*. In the context of room acoustics, the aim is to reduce the late reverberation allowing some few early reflections which are perceptually useful [Betlehem et al. 2012].

RIR-agnostic methods, instead, try to overcome the challenging task of estimating the acoustic channel and tuning peak-picking methods. They attempt to estimated echo properties directly in the parameter space of echos' **TOAs** and amplitudes.

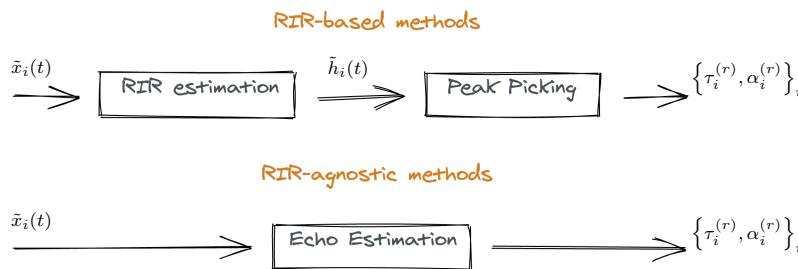


FIGURE 4.2: Schematic of **RIR**-based and **RIR**-agnostic approaches.

Given the above categories, we can now review the **AER** methods presented in the literature.

4.3 LITERATURE REVIEW

4.3.1 Active and RIR-based method

- THE **RIR ESTIMATION STEP** is typically modeled as a deconvolution problem whose performance depends on the type of transmitted signal. When the transmitted signal is arbitrary, several methods were developed to measure real **RIRs**. Since the **RIR** identifies the room response as a perfect impulse, one can measured it by producing an impulse sound, e. g. a clap, piercing a balloon, or a gun shot. Even though these methods are commonly used, they show clear limitations in term of reproducibility and safety. Instead, modern computational technique are used, involving the computation of deconvolution (or *cross-correlation*²⁷) between the known emitted signal and the recorded output. The Minimum Length Sequence (**MLS**) technique was first proposed in [Schroeder 1979] and it is based on the excitation of the acoustical space by a periodic pseudo-random signal, called **MLS**. The **RIR** is then calculated by circular correlation between the measured output and the original **MLS** signal. This method was further improved in order to achieve better **RIR** estimation in [Dunn and Hawksford 1993; Aoshima 1981]. Unfortunately this technique introduces several artifacts which yield to spurious peaks in

²⁷Cross-correlation and convolution operations are very similar, an mathematically they differ just by the inversion of the reference signal. While, the former measures the similarity between two signals as function of a translation, the latter measure the effect of one signal on the other signal. See § 4.3.2 for further explanation.

the estimation. Moreover, it is sensible to harmonic distortions generally introduced by playback devices, e. g., the loudspeakers.

To overcome these issues, the Exponential Sine Sweep (**ESS**) technique was introduced by Farina [Farina 2000; Farina 2007]. The probe signal is the **ESS** signal, a. k. a. *chirp signal*, which benefits of the following properties: the signal spans a user-defined frequency range; it is *self-orthogonal*, namely it compresses into Dirac's impulse during autocorrelation; and its Fourier inverse is available in closed form, allowing the user to not record and invert the probe signal. The reader can find a review of the presented techniques in [Szöke et al. 2019] applied to **RIR** measurements.

Sometimes the reference signal is known, but none of the above techniques can be used. For instance, when the source signal is narrowband or the acoustic condition are challenging. Therefore the **RIR** estimation problem needs to be addressed as a more general deconvolution problem, typically solved through optimization methods [Lin and Lee 2006]. This approach is well studied in literature and can be solved using standard Linear Least Squares with closed-form solution. However, in the case of signal with sparse frequency content (e. g. speech or music) or low SNR, it becomes ill-conditioned and prior knowledge about the **RIR** is used to improve the estimation.

- ▶ ECHO RETRIEVAL FROM RIR. As discussed in [Part I](#), acoustic echoes can be identified as peaks in the early part of the **RIR**. In general, due to the measurement process, such peaks are not necessarily positive, thus, to better visualize them, the *echogram* [Kuttruff 2016a], $|\tilde{h}(t)|$, or the energy envelope²⁸ [Schroeder 1979] are used instead.

Provided a good estimation of the **RIRs**, the echoes' location and amplitudes could be extracted manually by experts. However, even in ideal scenario, the automation of this process and the correct identification of such quantities are not straightforward tasks. As showed in [Tukuljac et al. 2018], since the **TOAs** are not necessarily multiple of the sampling grid, their true locations (and amplitudes) are blurred by spurious side peaks. This issue is referred to as *basis mismatch* in the *compressed sensing* literature. Although it can be alleviated by increasing the sampling frequency, it is bound to occur in practice. Moreover, the harmonic distortion due to the non-ideal source-receiver coupling may introduce other spurious peaks as well. Furthermore, as noticed in [Defrance et al. 2008b], even small errors of echoes' timing estimation may yield to significant differences in echo-based applications.

Existing methods for extracting echoes from **RIRs** can be further dichotomized into two broad categories: *on-grid* and *off-grid* approaches. The methods belonging to the former group are the most used in practice, and advanced techniques are used to cope with the presence of spurious peaks [Kuster 2008; Crocco et al. 2017; Remaggi et al. 2016; Defrance et al. 2008a; Bello et al. 2005; Cheng et al. 2016; Defrance et al. 2008a; Annibale et al. 2012; Kelly and Boland 2014; Usher 2010].

The most straightforward approach is to deploy iterative and adaptive thresholding algorithm on the **RIR**, followed by robust and manually tuned peak finders [Kuster 2008; Crocco et al. 2017]. To better inform the peak-picking, several strategies have been proposed. In the work of [Remaggi et al. 2016], based on a algorithm presented in [Naylor et al. 2006], peaks are clustered

²⁸ The energy envelope of a signal is computed as the magnitude of its analytic representation, computed with the Hilbert transform. For the signal x , its envelope is computed as

$$e[n] = \sqrt{x^2[n] + \mathcal{H}\{x\}^2[n]}, \quad (4.2)$$

where $\mathcal{H}\{\cdot\}$ denotes Hilbert transform.

according to changes in the phase slope of the **RIR** spectrum. Other works apply onset detection techniques used in music information retrieval and transcription based on edge-detection wavelet filters [Bello et al. 2005], non-negative matrix factorization [Cheng et al. 2016], or considering the **RIR**'s Kurtosis [Usher 2010].

By noticing that the reflection in the **RIRs** exhibit similar shape of the direct path, the author of [Defrance et al. 2008a] first proposed the use of *Matching Pursuit* (and improvements) to identify such shapes. Here the direct sound part was used as pattern (or atom) to be retrieved across the **RIR**. Unfortunately, in its pure form, this approach is unsuitable for **RIRs** because of the frequency dependent characteristic of the room absorption material. In order to improve the detection, [Kelly and Boland 2014] extends this approach employing *Dynamic Time Warping* to account for the non-uniform compression, dilation and concurrency of the echoes. Nevertheless, the idea of exploiting the direct path component to isolate the source-receiver coupling and thus identify first prominent reflection through deconvolution was used in [Annibale et al. 2012]. This technique is also referred to as *matching filter* or *direct-path compensation*. Alternatives approaches detect the echo timings in another transformed domain. In [Vesa and Lokki 2010] the echoes are localized in the Time-Frequency (**TF**) domain using the cross-wavelet transform based on previous works [Guillemain and Kronland-Martinet 1996; Loutridis 2005]. In [Peled and Rafaely 2013], the authors propose to identify the direction of arrival of the reflections in the spherical harmonic domain using plane wave decomposition. The works [Ristić et al. 2013; Pavlović et al. 2016] use (multi-)fractal analysis to detect echoes in the Time-Frequency (**TF**) domain. Alternatively, the authors of [Ferguson et al. 2019; Jia et al. 2017] propose to identify echoes properties in the cepstral domain. The *cepstrum* is usually defined as the inverse Fourier transform of the logarithmic magnitude spectrum²⁹. One interesting properties of cepstrum is to separate spectral envelop form line-spectra components in different “quefrency” ranges. Therefore, it is be used to detect periodicity in the spectrum of the observed signal, with many application in speech processing and in hydraulic and mechanic applications. This approach seems promising since time-domain peaks are mapped as complex sinusoids in frequency. However this representation is highly sensible to external and sampling noise and the accuracy is limited by the approximation of the **DFT** operator.

All the above mentioned active and **RIR**-based works aim at detecting echoes on the sampling grid. In order to face the inherent limitations of these approaches, off-grid frameworks have been proposed, e. g. [Condat and Hirabayashi 2013]. This approach can be related to classical Maximum Likelihood (**ML**) estimation approaches, which consist in selecting the model which is most likely to explain the observed noisy data. In this category fall classical spectral estimation techniques, e. g. Multiple Signal Classification (**MUSIC**) [Loutridis 2005], Estimation of Signal Parameters via Rational Invariance Techniques (**ESPRIT**) [Roy et al. 1986], which are fast but statistically suboptimal. The method presented in [Condat and Hirabayashi 2013] focuses on the general problem of estimating a finite stream of Dirac's impulses from uniform, noisy and lowpass-filtered samples. The authors showed that this particular problem can be reformulated as *matrix denoising*, from which the echoes' locations and amplitudes can be

²⁹ There are many concurrent definition of *cepsturm*. For a given time-domain signal \tilde{x} it is usually defined as $|\mathcal{F}^{-1}\{\log |\mathcal{F}x|^2\}|^2$, where \mathcal{F} is the **FT** operator.

retrieved in closed-form. Although this method reaches the statistical optimality in the ML sense, the exact knowledge of number of Diracs needs to be known in advance. If this number is unknown or approximated, errors in the estimation are observed. This results in a drawback since the exact number of echoes is difficult to know a priori and false-positive peaks are present even in clean RIRs due to source's harmonic distortions or limited bandwidth.

That being said, AER is far from trivial even when clean RIR estimates are provided. It is important to note that, for every TOA estimator, a practical trade off exists between the number of missed TOAs (false negative) and the number of spurious TOAs (false positive). This trade-off is only partially dependent on the noise level since many factors can provide spurious peaks. For instance, side lobes due to finite signal bandwidth, echo distortions due to frequency dependent attenuations and coalescing peaks due to close TOAs can affect peak estimation. This fact is often a source of unavoidable outliers that make the robustness of subsequent steps in echo-aware applications a delicate and very important issue. A way to overcome this trade-off is to allow spurious TOAs estimates and prune them using echo labeling afterwards.

- ▶ ECHO LABELING OR TOAs DISAMBIGUATION is the task of assigning acoustic echoes to different image sources or reflectors, as shown in Figure 4.3. Many methods have been proposed in the context of SSL [Scheuing and Yang 2006; Zannini et al. 2010], microphone calibration [Parhizkar et al. 2014; Salvati et al. 2016] and RooGE [Antonacci et al. 2010; Filos et al. 2011; Venkateswaran and Madhow 2012; Antonacci et al. 2012; Dokmanić et al. 2013; Crocco et al. 2014; Jager et al. 2016; El Baba et al. 2017]. A brief review of these methods is provided in [Crocco et al. 2017].

In the context of SSL, the disambiguation is typically performed in the TDOA space [Scheuing and Yang 2006; Zannini et al. 2010]. Moreover these works focus on actively localizing (the direction of arrival) of multiple sources while discarding reflections, rather than localizing the actual image sources.

Some other disambiguation schemes are typically used for RooGE. In [Venkateswaran and Madhow 2012] the pruning of the combinatorial candidate-image-source search is done through Bayesian inference. A similar approach can be found in [Dokmanić et al. 2013; Parhizkar et al. 2014] where the validity check is based on a particular structured matrix called *Euclidean Distance Matrix* and further improved using compatibility graphs in [Jager et al. 2016]. These methods rely on a combinatorial search with potentially high number of candidates, which leads to intractable computational complexity when multiple reflection are considered. Moreover these methods require that all the distances between each microphone are known with precision, which may not be true in practice. In the works of [Antonacci et al. 2010; Filos et al. 2011; Antonacci et al. 2012], the reflectors are modelled as planes tangent to the ellipsoids with foci given by each pair of microphone/source, as shown in Figure 4.4. By solving a non-convex optimization problem based on geometrical reasoning and the Hough transform³⁰, they are able to disambiguate TOAs and reconstruct reflectors position and orientation. However, they all require a very specific acquisition setup and the optimized non-convex cost functions are sensible to local minima.

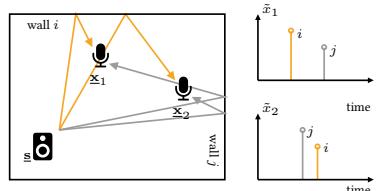


FIGURE 4.3: Illustration of different order of arrival of wall reflection

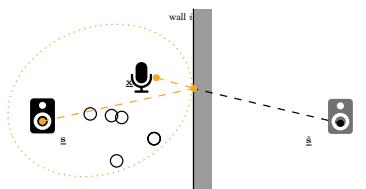


FIGURE 4.4: Illustration of the ellipsoid method: the length of the reflected path from the image source \hat{s} to the microphone at \hat{x} constraint the reflector line to be tangent to an ellipse.

³⁰A mathematical operator that maps points into curves in a 2-D space. If a set of points belongs to the same line, the corresponding curves will intersect in a single point. In computer vision, this transform is typically used as feature extractor to detect lines and edges in pictures.

In general, all the above methods do not have specific strategies to cope with missing or spurious echoes' estimates. This is due to malfunctioning of the peak finder or by erroneous selection of peaks corresponding to higher reflection order. A way to solve this issue is to exploit particular prior knowledge. For instance, the approach presented in [El Baba et al. 2017] exploits the shapes of linear and compact arrays of loudspeakers, which provides a natural ordering among the loudspeakers. By stacking side-by-side the measured RIRs in a matrix, they can be visualized as an image. Here the wavefront of each reflection draw specific pattern which can be identified easily and robustly even in presence of (a few) spurious and missing peaks. Moreover, this approach avoids the combinatorial search, but still requires a very specific setup for recordings.

In the work [Crocco et al. 2014] an iterative strategy is used. First the direct path arrivals are used to estimate a first guess of the microphones and source positions. Then, the whole set of extracted peaks are used to estimate the planar reflectors' positions which are then used to refine the microphones and source localization. Alternating between the geometrical space of microphones and source coordinates and the signal space of the echoes' TOAs, the ambiguous peaks are pruned during the optimization.

4.3.2 Active and RIR-agnostic method

This class of methods uses the signal at the microphones to directly estimate the echoes reflections, rather than estimating the RIRs. Here two different approaches can be identified: optimization-based approaches [Jensen et al. 2019; Saqib et al. 2020] and cross-correlation-based approaches [Crocco et al. 2014; Al-Karawi and Mohammed 2019].

The former approaches exploit the strong relation between the TOA of an echo with its Direction of Arrival (DOA). When multiple microphones are used and their geometry is known, the relation between these two quantities can be express in closed-form and used in a ML-based frameworks. By modeling the DOAs, such approaches are able to implicitly reduce the ambiguity of the estimated echoes. This idea is rooted in existing methods used in multipath communication systems, denoted as Joint Angle and Delay Estimation (JADE) [Vanderveen et al. 1997; Verhaevert et al. 2004].

Alternatively, the echoes contributions can be extracted from the cross-correlation between the observed and the reference signals. The cross-correlation analysis is a mathematical tool for the identification of repeated patterns in a signal as function of a certain time lag. Due to indoor sound propagation, the received signal consists in repeated copies of the emitted signal. Therefore, the received signal may correlate with the emitted one for certain time lags. Therefore, peaks in the cross-correlation function can be observed. By the extraction of these peaks, echoes' TOAs and relative amplitudes can be identified. This approach was used in [Tervo et al. 2011; Crocco et al. 2014; Al-Karawi and Mohammed 2019].

When the array geometry is known, the time lag axes of cross-correlation functions between channels can be mapped to possible 2D directions of ar-

rivals (elevation and azimuth), namely from **TOAs** to 2D-Directions of Arrival (**DOAs**). The identification of strong reflections can be then performed in the so-called *angular spectrum* domain [DiBiase et al. 2001]. With a proper clustering approach, the reflections can be inspected, disambiguated and their **TOAs** deduced. This approach is used in [O'Donovan et al. 2008; O'Donovan et al. 2010; Tervo and Politis 2015] and can be generalized by spatial filtering methods, such as steered-response power-based beamforming. In [O'Donovan et al. 2008], it was referred to as *acoustic camera* since it benefits of the following visual interpretation: as shown in Figure 4.5, the 2D-polar Coordinates can be mapped into cartesian ones so that the angular spectrum can be superimposed to a panoramic picture of the audio scene taken by the barycenter of the recording arrays.

4.3.3 Passive and RIR-based method

Passive approaches rely on external sound sources in the environment to conduct the estimation. In the literature, this problem belongs to the broad and deeply studied category of Blind Channel Estimation (**BCE**), or Blind System Identification (**BSI**), problems. In the particular case of a single source, it is referred to as **SIMO BCE**. Common to all these methods is the assumption that **RIRs** are discrete Finite Impulse Response (**FIR**) filters defined on the sampling grid, namely, vectors in the Euclidean space. In the general setting of arbitrary signals and filters, rigorous theoretical ambiguities under which the problem is ill-posed have been identified [Xu et al. 1995]. Some well-known limitations of these approaches are their sensitivity to the chosen length of the filters, and their intractability when the filters are too large. **FIR SIMO BCE** can be broadly dichotomized into the class of *statistical methods* and the class of *blind methods*.

- ▶ **STATISTICAL METHODS** exploit knowledge about the emitted signal. Since the nature of the source signal is by definition not deterministic, and their statistics can be modeled base on the signal category, e. g. speech or music, and modeled accordingly. Two main approaches can be identified [Tong and Perreau 1998]:
 - *Second Order Moments approaches* derive closed-form solution for which the knowledge of the source auto-correlation function is required.
 - *Maximum Likelihood approaches* require instead the source probability function. Even though they are optimal in the **ML** sense, they optimize non-convex cost functions, typically via Expectation Maximization (**EM**). In this category one may include the methods developed for multichannel blind source separation [Ozerov and Févotte 2009; Duong et al. 2010; Leglaive et al. 2016; Leglaive et al. 2018; Scheibler et al. 2018e]. These methods are built on the well-studied framework of Multichannel Nonnegative Matrix Factorization (**NMF**) [Ozerov and Févotte 2009] which lends itself to account for various type of side information. Here the source signals are typically modeled as Gaussian distributions centred in zero and with unknown variance. In general, they aims at estimating both the acoustic channels and the source contribution. In particular, the work of [Duong et al. 2010] extends this framework to

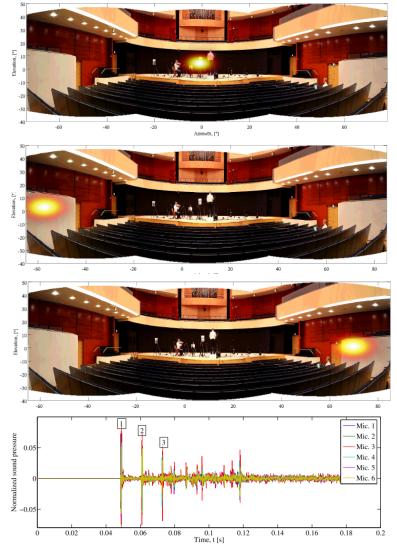


FIGURE 4.5: Visualization of the *audio camera*: The angular spectrum is overlapped to the corresponding images. Also shown are the impulse responses for 6 microphones. The numbered boxes indicate events shown in the audio camera. Images from [Tervo 2011].

The innovative idea of passive **Blind Channel Estimation (BCE)** can be traced back to [Sato 1975]. A review of the evolution of Single Input Multiple Output (**SIMO BCE**) can be found in [Huang and Benesty 2003].

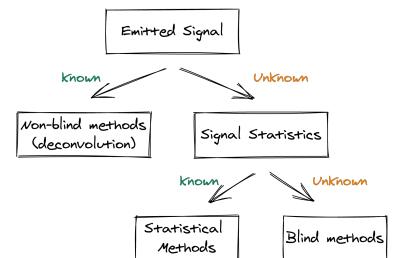


FIGURE 4.6: Classification of the State of the Art in channel estimation.

reverberant recordings using physics-based models for the late reverberation, while the works of [Leglaive et al. 2016] considers explicitly the contribution of early echoes, further improved in [Leglaive et al. 2018].

Even if statistical methods have reported a considerable success in the field of sound source separation, they play a minor role in RIR estimation. This is due to the difficulty in achieving reliable statistics of the emitted signals or a good initialization point required by the EM algorithm. Moreover, although the final estimated RIRs may match the real ones in the statistical sense, they lack of a sufficient details, indispensable for AER.

- ▶ BLIND METHODS comprises two main groups: *subspace* methods [Abed-Meraim et al. 1997] and *cross-relation methods* [Tong et al. 1994; Xu et al. 1995; Lin et al. 2007; Lin et al. 2008; Kowalczyk et al. 2013; Crocco and Del Bue 2015; Crocco and Del Bue 2016a]. The formers are based on the key idea that the channel (or part of it) vector spans a one-dimensional subspace of (a block of) noiseless observations. These methods have the attractive property that the channel estimates can often be obtained in a closed-form by optimizing a quadratic cost function. However they may be not robust, especially when the channel covariance matrix is close to being singular. The second disadvantage is that they are typically computationally expensive.

The second family of methods rely on the clever observation that in the noiseless case, for every pair of microphone (i, i') , it holds that

$$(\tilde{x}_{i'} \star \tilde{h}_i)(t) = (\tilde{x}_i \star \tilde{h}_{i'})(t) = ((\tilde{h}_{i'} \star \tilde{h}_i) \star s)(t), \quad (4.3)$$

by the commutativity of the convolution operator. This principle is called the *cross-relation* and it was first introduced by [Tong et al. 1994]. In this work, the RIR are estimated by solving a Least Square minimization of the sum of square cross-relation errors. In [Xu et al. 1995; Tong and Perreau 1998], sufficient and necessary conditions for channel identification are discussed. This approach has received significant attention as it does not require any assumption about the source signal. Later, the accuracy of estimated RIRs has been subsequently improved using prior knowledge of the filters: in particular, the authors of [Lin et al. 2007] have proposed to use sparsity penalty and non-negativity constraints to increase robustness to noise as well as Bayesian-learning methods to automatically infer the value of the hyper-parameters in [Lin et al. 2008]. Even if sparsity and non-negativity could be seen as a strong assumption, works in speech enhancement [Ribeiro et al. 2010a; Dokmanić et al. 2015b] and room geometry estimation [Antonacci et al. 2012; Crocco et al. 2017] have proven the effectiveness of this approach. On a similar scheme, in [Kowalczyk et al. 2013], the problem in Eq. (4.3) is solved using an adaptive time-frequency-domain approach while [Aissa-El-Bey and Abed-Meraim 2008] proposes to use the ℓ_p -norm instead of the ℓ_1 -norm. A successful approach has been presented by Crocco et al. in [Crocco and Del Bue 2015; Crocco and Del Bue 2016a], where the so-called *anchor constraint* is replaced by an *iterative weighted* ℓ_1 equality constraint to better balance sparsity penalty and the model constraints.³¹ Finally, the very recent work [Qi et al. 2019] extends cross-relation approaches under the umbrella of the Kalman filter which was previously used for echo-cancellation applications.

³¹ These approaches will be further formalized and detailed in Chapter 5.

An alternative approach is used in [Čmejla et al. 2019], where the **RIR** estimation problem is treated as special case of **RTF** estimation. As mentioned in § 3.3.2, in the noiseless case, the **RTF** identifies the **RIR** when the reference microphone is placed very close to the source. **RTF** estimation found its root in the field of Speech Enhancement (**SE**) [Gannot et al. 2001] and many techniques have been proposed since then [Gannot et al. 2001; Koldovský et al. 2015; Koldovsky and Tichavsky 2015; Kodrasi and Doclo 2017].³² By its definition, in noiseless case, **RTF** describes the relative filter between two observations and not directly their **RIRs**. The main limitation of this approach is that it is possible only in controlled scenarios, where the user has the possibility to place the microphone arbitrarily in the room and in presence of high Signal-to-Noise-Ratio (**SNR**) levels. Nevertheless, in this context, this particular setup is found to be useful not only for **RTF** estimation, but also for microphone calibration, since it allows to solve geometrical ambiguities, yielding a closed-form solution, as done in [Crocco et al. 2012].

In general, the main drawbacks of **FIR SIMO BCE** works is that they rely on on-grid estimation, sparsity-enforcing regularizers and peak-picking which need to be tuned manually. As described in § 3.2.3, due to the sampling process involving a sinc function, the filters are non-sparse and non-negative. This general bottleneck has been referred to as *basis mismatch* and was notably studied in the *compressed sensing* community [Chi et al. 2011]. In particular, the true peaks in the **RIR** do not necessarily correspond to the true echoes as shown in ?? Since these methods are fundamentally on-grid, the estimated

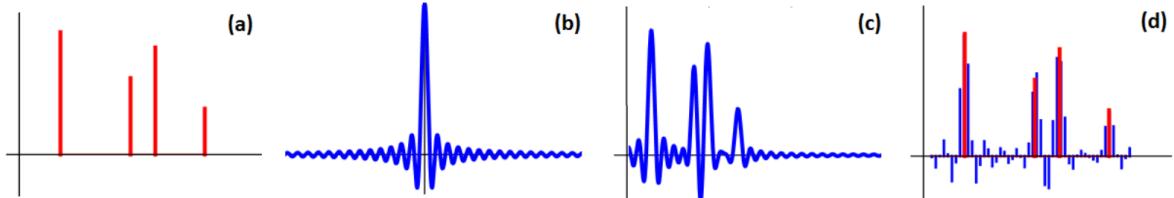


FIGURE 4.7: (a) Continuous-time stream of Diracs $\tilde{h}(t)$, (b) sinc kernel $\tilde{\phi}_{LP}(t)$, (c) smoothed stream of Diracs, (d) original stream of Dirac $\tilde{h}(t)$ (red) and its sampled (i.e., smoothed and discrete) version (blue). Image courtesy of [Tukuljac et al. 2018]

echo locations are integer multiples of the sampling period $1/F_s$. This prevents subsample resolution, which may be important in applications such as **RooGE** [Crocco et al. 2017] or acoustic parameter estimation [Defrance et al. 2008b]. Moreover, these methods strongly rely on the knowledge of the length of the filters. When this parameter is underestimated or overestimated, identifiability and computational issues may arise, affecting the estimation. Nevertheless, despite this slight mismatch between theoretical assumptions and real data, for some scenarios, the position of the estimated peaks by the methods [Crocco and Del Bue 2016a] reproduces the positions of the groundtruth peaks with remarkable precision as demonstrated in our own work [Di Carlo et al. 2020].

Di Carlo et al., “Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval”

³² Methods for **RTF** estimation will be detailed in ??.

4.3.4 Passive and RIR-agnostic methods

Methods in this category bypass the onerous task of estimating the (full or partial) acoustic channel and, to the best of our knowledge, only a few have been identified. As for the active and RIR-agnostic case, the audio camera based on the cross-correlation function can be used in passive settings. Exploiting the geometrical knowledge of the microphone array, **TDOAs** extracted from robust correlation function can be mapped to **DOAs** [DiBiase et al. 2001; O’Donovan et al. 2008; O’Donovan et al. 2010]. Assuming a single source scenario, difference **DOAs** can be disambiguated using geometrical prior knowledge and can be associated to image sources, hence reflectors. These methods typically ignore the echoes amplitudes and in general do consider only angles on the unit sphere, ignoring the distance from the source. Without proper prior knowledge, their application to **AER** is far from trivial as **RooGE** and *reflector estimation* methods need to be used to convert **DOAs** back to echoes timing.

Recently a fully blind, passive, off-grid and RIR-agnostic method was proposed by authors of [Tukuljac et al. 2018] for stereophonic recordings (i. e., $I = 2$). They proposed a method, called Multichannel Annihilation (**MULAN**), based on the properties of the *annihilation filter*³³, [Condat and Hirabayashi 2013] and the theory of Finite Rate of Innovation (**FRI**). If the source signal is known, starting from the cross-relation identity, the **AER** problem translates into finding the annihilation filter for the **RIRs**, which can be recasted into an eigenvalue problem. In the fully blind case, the problem is solved with non-convex optimization, iterating between the estimation of the two filters and the signal until convergence. The method was later extended to the multichannel case in [Tukuljac 2020] using the generalization of Cadzow denoising framework [Condat and Hirabayashi 2015]. This method is shown to outperform conventional approaches by several orders of magnitude in precision in noiseless case, with synthetic data and when the correct number of echoes is known *a priori*. However its effectiveness was not tested on challenging real scenarios featuring external noise and partial knowledge on the number of echoes.

4.4 DATA AND EVALUATION

AER is a relatively recent problem which is typically addressed in the context of much broader applications, e. g. **SE**, **RooGE**, **SSL**. Therefore the literature lacks of standard datasets as well as standard evaluation frameworks.

4.4.1 Datasets

As listed in [Szöke et al. 2019] and in [Genovese et al. 2019], a number of recorded **RIRs** corpora are freely available online, each of them meeting the demands of certain applications, usually **SE** and Finite Rate of Innovation (**ASR**). However, even if these datasets feature reverberation and strong early reflections, they lack of proper annotations, making them difficult to use for testing **AER** methods. For this reason, to bypass the complexity of recording real annotated RIR datasets, simulators based on the **ISM** are extensively used instead. While simulated datasets are more versatile, simple and quicker to

³³ For a sequence of Fourier coefficients $\mathbf{a} \in \mathbb{C}^N$ (describing a signal or a filter), its annihilation filter $\mathbf{b}_L \in \mathbb{C}^L$ is such that the linear convolution between the sequence and the filter coefficients is identically zero:

$$\sum_{l=0}^{L-1} \mathbf{b}[l]\mathbf{a}[n-l] = 0 \quad \forall n = -N + L, \dots, N.$$

Database Name	Annotated			Number of			Key characteristics	Purpose
	Pos.	Echoes	Rooms	RIRs	Rooms	Mic×Pos.		
[Dokmanić et al. 2013]	✓	~	~	15	3	5	1	Non shoebox room RooGE
[Crocco et al. 2017]	✓	~	✓	204	1	17	12	Accurate 3D calibration Many mic and src pos. RooGE
[Remaggi et al. 2016]	✓	~	✓	~1.5k	4	48×2	4-24	Circural dense array Circular placement of src. RooGE SE [†]
[remaggi2019modeling]	✓	~	✓	~1.6k	4	48×2 +2×2	3-24	Circural dense array Binaural Recordings RooGE [†] SE
BUT Reverb [Szöke et al. 2019]	✓	✗	~	~1.3k	8	(2-10)×6	3-11	Accurate metadata different device/arrays various rooms SE/ASR
VoiceHome [Bertin et al. 2019]	✓	✗	✗	188	12	8×2	7-9	Various rooms, real homes SE/ASR
DECHORATE (Ch. 7)	✓	✓	✓	~1.8k	1	30	6	Accurate annotation Different Echo-energy RooGE SE/ASR

TABLE 4.1: Comparison of some existing RIR databases which account for early acoustic reflections. Receiver positions are indicated in terms of number of microphones per array times number of different positions of the array (~ stands for partially available information). The reader is invited to refer to [Szöke et al. 2019; Genovese et al. 2019] for a more complete list of existing RIR datasets.

[†]The dataset in [Remaggi et al. 2016] is originally intended for RooGE and further extended for (binaural) SE in [Remaggi et al. 2016] with a similar setup.

obtain, they fail to fully capture the complexity and the richness of real acoustic environments. Due to this intrinsic issues, methods trained or validated on them may fail to generalize to real conditions, as will be shown in Chapter 7.

A good dataset for AER should include a variety of environments (room geometries and surface materials), of microphone placement (close to or away from reflectors, scattered or forming ad-hoc arrays) and, most importantly, precise annotations of the scene’s geometry and echo parameters. Moreover, in order to be versatile and used in echo-aware applications, the provided annotations should match the ISM, i.e., TOAs should be expressed in terms of image sources and vice-versa. Such data are difficult to collect since they require precise measurements of the positions and orientations of all the acoustic emitters, receivers and reflective surfaces inside the environment with dedicated planimetric equipment. We identified here two main classes of related RIR datasets in the literature: SE/ASR-oriented datasets, e.g. [Szöke et al. 2019; Bertin et al. 2019; Čmejla et al. 2019], and RooGE-oriented datasets, e.g. [Dokmanić et al. 2013; Crocco et al. 2017; Remaggi et al. 2016]. The formers include acoustic echoes as highly correlated interfering sources coming from close reflectors, (e.g. a desk in meeting rooms or a near wall), however annotations are not provided. The latter group deals with sets of distributed, synchronized microphones and loudspeakers in a room. These setups are not exactly suitable for SE methods, which typically involve compact or ad hoc arrays. To bridge this gap, we recorded a new dataset, dubbed Dataset dechorated by echoes (DECHORATE), that will be described in Chapter 7 and used for echo-aware application in Chapter 11. Table 4.1 summarizes some existing datasets that can be used in the context of AER.

4.4.2 Metrics

The metrics used in AER depend on the application and the methods used to estimate the echoes. When addressed as a FIR SIMO BCE problem, the groundtruth acoustic channels are considered as a discrete vector $h \in \mathbb{R}^L$,

and similarly their estimates, that is, $\hat{h} \in \mathbb{R}^L$. To assess the quality of the estimated discrete filters, the following metrics have been proposed in the literature:

- *The Root Mean Square Error (RMSE)* measures the distance between points in the Euclidean space, defined by vector coordinates:

$$\text{RMSE}(\hat{h}, h) \stackrel{\text{def}}{=} \sqrt{\sum_{n=0}^{L-1} |\hat{h}[n] - h|^2} \quad [\text{seconds (or, samples)}], \quad (4.4)$$

where $|\cdot|$ denotes the absolute value. This metric is known to be highly sensitive to scaling and translations.

- *The Normalized Projection Misalignment (NPM)* was originally proposed in [Morgan et al. 1998] to solve the limitation of the RMSE. In the formulation provided in [Huang and Benesty 2003; Ahmad et al. 2006], it writes as

$$\text{NPM}(\hat{h}, h) \stackrel{\text{def}}{=} 20 \log_{10} \left(\left\| h - \frac{h^\top \hat{h}}{\hat{h}^\top \hat{h}} \right\|_2 / \|h\|_2 \right) \quad [\text{dB}], \quad (4.5)$$

where $\|\cdot\|_2$ denotes the Euclidean norm. By projecting \hat{h} onto h and defining a projection error, only the intrinsic misalignment of the channel estimate is considered, disregarding an arbitrary gain factor and the length difference of both vectors. However it is not translation invariant.

- *The Hermitian angle* is similar to NPM and was used in the context of RTF estimation in [Varzandeh et al. 2017; Tammen et al. 2018]

$$\Delta\Theta(\hat{h}, h) = \arccos \left(\frac{h^\top \hat{h}}{\|h\|_2 \|\hat{h}\|_2} \right). \quad (4.6)$$

As for NPM, this metric is invariant to possible scaling factors and length differences between the groundtruth and the estimated vectors.

In the context of **RooGE**, **SSL** and microphone calibration, echoes' timings are typically mapped to reflectors or image source positions, either in cartesian or polar coordinates. Therefore, the models for **AER** are evaluated in the geometrical space, rather than in the space of echoes' parameters. For instance, for the task of reflectors localization, the accuracy is measured in terms of *plane-to-plane distance* between estimated and groundtruth surfaces and the *angular error* between their normals. In the case of **SSL** and microphone calibration, the *Euclidean distance* between the 3D coordinates is typically computed as **RMSE** between groundtruth and estimated **DOAs**. This metric considers only echoes' **TOA**, ignoring their amplitudes.

To our best knowledge, the literature lacks of metrics properly defined for **AER**. As for the application mentioned above, echoes' amplitudes in a single **RIR** or between them, are typically ignored or considered for peak picking only. More attention is paid on the echoes' timing which are evaluated using regression/classification metrics of *information retrieval* and *machine learning*. Let be $\hat{\tau} = \{\hat{\tau}_r\}_{r=0}^R$ and $\tau = \{\tau_r\}_{r=0}^R$ the sets of estimated and reference echoes' **TOAs**. The following metrics are used:

- the **RMSE** is defined as

$$\text{RMSE}(\hat{\tau}, \tau) \stackrel{\text{def}}{=} \sqrt{\sum_{r=0}^R |\hat{\tau}_r - \tau_r|^2} \quad [\text{seconds (or, samples)}], \quad (4.7)$$

This metric describes the mean error between estimated and reference of echoes' **TOAs**. Unfortunately, the **RMSE** is proportional to the size of the squared error, thus is sensitive to outliers. In the context of **AER**, the **RMSE** is computed only on the matched **TOAs**.

- the *Precision, Recall, and F-measure* are standard metrics used in information retrieval for evaluating classification problems, e. g. in onset detection [Böck et al. 2012]. Here the real valued estimates and groundtruth need to be converted into binary values indicating a *match*. Typically, hard thresholding is used to assess whether estimated **TOAs** match the reference one. In the context of **AER**, *precision* expresses the fraction of matching **TOAs** among all the estimated ones, while *recall* measure the fraction of matching **TOAs** that are correctly estimated. Finally, the *F-measure*, defined as the harmonic mean of precision and recall, is used to summarize precision and recall in one value.

Depending on the application, precision and recall can have different impact. **RooGE** methods are more sensible to missing **TOAs** than to their misalignment which can be redefined with geometrical reasoning. Thus they are more incline to privilege recall over precision and allow for some false-positive which can be pruned using echo labelling methods. Instead, echo-aware **SE** methods prefer to accurately select the relevant echoes, thus favoring higher precision.

Givent the number of true positive cases in the estimation TP, the false negative FN and the false positive PF, precision, recall and F-measure as:

$$\begin{aligned} \text{Precision} &= \text{TP}/(\text{TP} + \text{FP}), \\ \text{Recall} &= \text{TP}/(\text{TP} + \text{FN}), \\ \text{Fmeasure} &= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned}$$

Since these metrics rely on decision thresholds, their usage is not straightforward. In fact, in order to compare echoes, first both estimated and reference echoes need to be labeled, pruned and matched. As discussed at the end of § 2.3.3, echoes can be sorted differently according to their amplitudes, their **TOAs** or reflection order. **AER** tends to return echoes' parameter sorted by the echoes' amplitudes which can be distorted by the measurement process and modelling errors. This matching and labeling process introduces strong biases in the evaluation process which is currently unsolved without a proper echo labeling step.

4.5 PROPOSED APPROACHES

So far, we presented a view of current methods for solving the **AER** problem. In the following two chapters, we will explore two novel approaches which follow two paradigms occurring in the recent years of signal processing: *knowledge-driven* and *data-driven* methods.

- ↔ KNOWLEDGE-DRIVEN METHODS take advantage of prior information which may have deterministic (e. g. physical equation) or asymptotic behaviors (e. g. statistical models). In this context, **AER** exploits prior information about the sources, the mixing process and the physics laws of the acoustic propagation,

along with the audio. This knowledge is typically translated into mathematical models which lead solution computed through closed-forms or optimization-based algorithm to estimate latent variables. All the literature presented in this chapter follows this approach. In general, the advantages and the disadvantages of these approaches depend on the nature of available knowledge on the context which it is used.

Regarding our contributions, [Chapter 5](#) proposes a new knowledge-driven method for solving [AER](#) based on the theory of Continuous Dictionary ([CD](#)).

⇒ **DATA-DRIVEN METHODS**, instead, are based on machine learning algorithms where information is automatically “learned” in *supervised* ways, therefore the knowledge of the physical laws are need. Providing comprehensive and exhaustive annotated training datasets, such methods can learn function that maps an input to an output based on example input-output pairs. Due to its recent success, machine learning, and in particular *deep learning*, has been applied in many signal processing tasks. Along side with the huge benefits of having black-box models that are able to learn by their own, this paradigm hides a few limitations.

First, these models rely on the information encoded in training data which is sometimes are not representative enough of the real-world. In order to overcome this drawback, many strategies have been proposed, e.g. using *data augmentations* techniques or knowledge-driven generating models, based on simulators. In this case, prior knowledge is used to *generate* synthetic observation (i.e., from latent variables to the observations) and it is left to the learning model the task of invert the mapping (i.e., from observation to latent variables). This leads to a second limitation of these approaches, that is *overfitting* to the data obeying the generating model. Finally, machine learning models learn black-box functions. Although they can reach incredible performance, it is difficult to predict their behavior when facing new type of data. Despite these issues, data-driven methods are currently intensively studied and interlaced with knowledge-driven approaches. In this direction, we propose our contribution in [Chapter 6](#), as a new data-driven method for solving [AER](#) based on virtually supervised learning.

5

Blaster: Knowledge-driven Acoustic Echo Retrieval

- ▶ **SYNOPSIS** This chapter proposes a novel approach for *off-grid AER* from a stereophonic recording of an unknown sound source such as speech. In order to address some limitation of existing methods, we propose a new approach, named **BLASTER**. It builds on the recent framework of Continous Dictionary (**CD**), and it does not rely on parameter tuning nor peak picking techniques by working directly in the parameter space of interest. The method's accuracy and robustness are assessed on challenging simulated setups with varying noise and reverberation levels and are compared to two state-of-the-art methods. While comparable or slightly worse recovery rates are observed for recovering seven echoes or more, better results are obtained for fewer echoes, and the off-grid nature of the approach yields generally smaller estimation errors.

The material presented in this chapter was previously published in [Di Carlo et al. 2020] and results from a collaboration with the colleague Clement Elvira whose domain of expertise is in the Continous Dictionary (**CD**) framework. The section dedicated on the presentation of the **CD** framework applied for **AER** is extracted from the related publication and it is written with the help of my colleague. Here we briefly commented and expanded it and some attention is paid in grasping the motivation behind it. Finally, this chapter recall the main findings of the paper bringing additional insight in the existing models for **AER**.

5.1 INTRODUCTION

Let us recall from [Chapter 4](#) some knowledge-based methods addressing **AER**. Some existing approaches methods rely on a known source signal [Park et al. 2017; Jensen et al. 2019]. In contrast, when multiple receivers attend an unknown single source, **AER** can be seen as an instance of Single Input Multiple Output (**SIMO**) Blind Channel Estimation (**BCE**) problem, i. e. estimating the filters entailing an unknown input observed output of a system. A common approach for solving **AER** in the context of **SIMO-BCE** is to first blindly estimate a discrete version of the acoustic channels using the so-called cross-relation identity [Xu et al. 1995; Crocco and Del Bue 2016a]. The location of the echoes are then chosen among the strongest peaks with ad-hoc peak-picking techniques. Such methods are generally *on-grid* in the sense that the estimation

"My beautiful astronaut// With an impulsive world// A twisted scientist!"
—Venetian Snares, 1000 Years

Keywords: Blind Channel Identification, Super Resolution, Sparsity, Acoustic Impulse Response.

Resources:

- Paper
- Code
- Open-access paper with supplementary material
- Slides
- Presentation

Di Carlo et al., "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval"

relies on a fixed grid of points and *a priori* chosen filter lengths. However, in practice, the true timings of echoes rarely match the sampling grid, thus leading to pathological issues called basis-mismatch in the field of compressed sensing. To circumvent this issue, the authors of [Tukuljac et al. 2018] proposed to leverage the framework of finite-rate-of-innovation sampling to make one step towards off-grid approaches. Despite promising results in the absence of noise and with synthetic data, the quality of the estimation highly relies on the choice of a good initialization point.

Of particular interest the proposed approach is the recently proposed framework of Continuous Dictionary (**CD**) [Candès and Fernandez-Granda 2014]. By formulating an inverse problem as the recovery of a discrete measure over some parameter space, **CD** has allowed to overcome imaging device limitations in many applications such as super-resolution [Candès and Fernandez-Granda 2014] or PALM/STORM imaging [Denoyelle et al. 2019]. In this work, we formulate the problem of **AER** for stereophonic mixtures, i. e. using only one microphone pair, within the framework of continuous dictionaries. The resulting optimization problem is convex and thus not prone to spurious minimizers. The proposed method is coined *Blind and Sparse Technique for Echo Retrieval (**BLASTER**)* and requires no parameter tuning. The method is compared to state-of-the art on-grid approaches under various noise and reverberation levels using simulated data.

5.2 SIGNAL MODEL

We consider here the common setup of stereophonic mixtures, that is 2-channel microphone recordings. Using the notation presented [Chapter 3](#), recorded signal at microphone $i \in \{1, 2\}$ reads

$$\tilde{x}_i = \tilde{s} \star \tilde{h}_i^* + \tilde{u}_i \quad (5.1)$$

where \star denotes the (continuous) convolution operator, \tilde{u}_i models some additive noise in the measurement process and \tilde{h}_i^* denotes the Room Impulse Response (**RIR**). In the remainder of this chapter, the superscript $*$ refers to the ground truth. Assuming the echo model, the **RIRs** read

$$\tilde{h}^*(t) = \sum_{r=0}^{R_i} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) \quad (5.2)$$

where R_i is the (unknown) number of echoes.

In the noiseless case, that is when $\tilde{u}_i = 0$ for $i \in \{1, 2\}$, we have the cross-relation identity

$$\tilde{x}_1 \star \tilde{h}_2^* = \tilde{x}_2 \star \tilde{h}_1^* \quad (5.3)$$

by commutativity of the convolution operator.

However, in practice, only sampled versions of the two recorded signals are available. More precisely, we consider the measurement model introduced in [Chapter 3](#): the incoming signal undergoes a (ideal) low-pass filter $\tilde{\phi}_{LP}$ with frequency support $[-F_s/2, F_s/2]$ before being regularly sampled at the rate F_s . We denote $x_1, x_2 \in \mathbb{R}^{2N}$ the two vectors of $2N$ (consecutive) samples and

$i \in \{1, 2\}$ by

$$x_i[n] = \left(\tilde{\phi}_{LP} * \tilde{x} \right) \left(\frac{n}{F_s} \right) \quad \forall n \in \{0, \dots, 2N - 1\}. \quad (5.4)$$

5.3 BACKGROUND ON ON-GRID BLIND CHANNEL ESTIMATION

We now select and elaborate more on some of the methods mentioned in § 4.3.3 concerning optimization approaches for **FIR-SIMO-BCE**. Therefore, the following methods operates on discrete-time signal, which are denoted without the tilde according to the notation presented in Chapter 3. Starting from the identity Eq. (5.3), the common **SIMO-BCE** cross-relation framework aims to compute \tilde{h}_1, \tilde{h}_2 solving the following problem in the discrete-time domain [Lin et al. 2007]:

$$\hat{h}_1, \hat{h}_2 = \arg \min_{h_1, h_2} \frac{1}{2} \|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2 + \lambda \|h\|_1 \quad (5.5)$$

where

- x_i and h_i are the discrete, sampled version of \tilde{x}_i, \tilde{h}_i respectively,
- $\mathcal{T}(x_i)$ is the $(2N + L - 1) \times L$ Toeplitz matrix (build as shown in Figure 5.1) associated to convolution where $2N$ and L respectively denote microphone and filter signal length,
- and the ℓ_1 regularization term is used to enforce sparsity in the estimation, which is consistent with the “train of impulses” model for the early part of RIRs (but not for the tail).

This type of problem can be seen as an instance of Least Absolute Shrinkage and Selection Operator (**LASSO**) problem [Tibshirani 1996], written in the form:

$$\arg \min_u \frac{1}{2} \|v - Au\|_2^2 + \lambda \|u\|_1. \quad (5.6)$$

This type of well-known optimization problem are convex and, despite the non-differentiability of the ℓ_1 -norm, they can be easily tackled by standard optimization tool. Later, in this section, we show how to express Eq. (5.5) as a standard **LASSO** problem.

The accuracy of estimated RIRs has been subsequently improved using a priori knowledge on the filters. In particular, the authors of [Lin et al. 2007] have proposed to use non-negativity constraints to increase robustness to noise and avoid trivial solution. Therefore, let us define a more general formulation for Eq. (5.5), such that

$$\hat{h}_1, \hat{h}_2 = \arg \min_{h_1, h_2} \mathcal{J}(h_1, h_2) + \mathcal{P}(h_1, h_2) \text{ s.t. } \mathcal{C}(h_1, h_2) \quad (5.7)$$

where $\mathcal{J} = \frac{1}{2} \|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2$ is the cost function to optimize. $\mathcal{P}(h_1, h_2)$ and $\mathcal{C}(h_1, h_2)$ are respectively a regularization term used to promote sparse solution and a constrained set. Let us define $\mathbf{h} = [h_1^\top, h_2^\top]^\top$ as the concatenation of the two vectorized discrete filter. Thank to this formulation, current state of the art approaches can be summarized as in the Table Table 5.1.

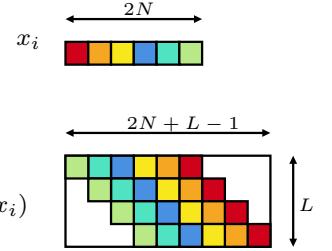


FIGURE 5.1: Graphical representation of the construction of $\mathcal{T}(x_i)$ from x_i

Reference	$\mathcal{P}(h_1, h_2)$	$\mathcal{C}(h_1, h_2)$	Note
[Tong et al. 1994]	\times	$\ \mathbf{h}\ _2^2 = 1$	Equivalent to a smallest-eigenvalue problem.
[Kowalczyk et al. 2013]	$\lambda\ \mathbf{h}\ _1$	$\ \mathbf{h}\ _2^2 = 1$	Non-convex due to the quadratic constraint.
[Lin et al. 2008]	$\lambda\ \mathbf{h}\ _1$	$ h_1[0] = 1$	With Bayesian learning for optimal λ .
[Lin et al. 2007]	$\lambda\ \mathbf{h}\ _1$	$h_1[0] = 1, \mathbf{h} \geq 0$	Non-negativity and anchor constraints.
[Aissa-El-Bey and Abed-Meraim 2008]	$\lambda\ \mathbf{h}\ _p^p$	$\ \mathbf{h}\ _2 = 1$	Sparsity enforced by ℓ_p -norm
[Crocco and Del Bue 2015]	$\lambda\left\ \mathbf{p}^{(z)} \odot \mathbf{h}\right\ _1$	$\ \mathbf{h}\ _1 = 1, \mathbf{h} \geq 0$	Iterative weighted ℓ_1 -norm
[Crocco and Del Bue 2015]	$\lambda\ \mathbf{h}\ _1$	$\mathbf{p}^{(z)\top} \mathbf{h} = 1, \mathbf{h} \geq 0$	Iterative weighted ℓ_1 constraint.

TABLE 5.1: Some state of the art penalties and constraint used in model Eq. (5.7).

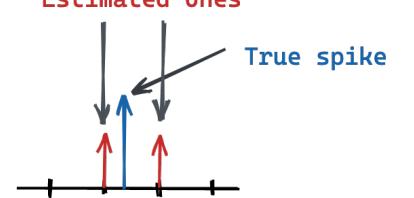
The constraint $h_i[0] = 1$ is called an *anchor constraint* and it is used to replace the ℓ_2 -norm while keeping the problem convex. The non-negativity $\mathbf{h} \geq 0$ constraint may not be satisfied due to effects such as measurement process, the filtering in the propagation media or the imperfect frequency response of a microphone. However, when those effects are common to both channels, they can be viewed as distortions to a common source. Therefore, the non-negativity assumption seems reasonable for real acoustic environments. Nevertheless, applications concerning **RooGE** require just the recovery of lower order reflections, i.e. the sparse portion of the **RIR**. Likewise works in speech enhancement have proven to work under such assumption, thus proving the effectiveness of this approach, such as in [Lin et al. 2008; Yu et al. 2011].

On a similar scheme, in [Kowalczyk et al. 2013], Eq. (5.5) is solved using an adaptive time-frequency-domain approach while [Aissa-El-Bey and Abed-Meraim 2008] proposes to use the ℓ_p -norm instead of the ℓ_1 -norm. Choosing $p < 1$, sparsity is enforced, however the problem become non-convex and ad-hoc optimization technique was proposed. A successful approach has been presented recently by the authors of [Crocco and Del Bue 2016a], where the anchor constraint is replaced by an *iterative weighted ℓ_1* equality constraint, i.e., such that at each iteration z , $\mathbf{p}^{(z)\top} \mathbf{h}^{(z)} = 1$.³⁴ In particular, the method is initialized using the solution of [Lin et al. 2007] and iterated enforcing sparsity using the solution of the previous problem, that is $\mathbf{p}^{(z)} = \hat{\mathbf{h}}^{(z-1)}$. The reader can find a comprehensive review of these methods in [Crocco and Del Bue 2015; Crocco and Del Bue 2016a].

³⁴ Note that when $\mathbf{p}^{(z)} = 1$, the constraint returns to the ℓ_1 penalty.

- ▶ THE LIMITATION OF THE DISCRETE-TIME METHODS described above are the followings:

- *Basis mismatch*: As explained in § 4.3.3, the filter are not sparse in practice due to the *basis mismatch*. This implies that the true peaks of the filter do not necessarily correspond to the true echoes and lead to followings drawbacks. As these methods operates fundamentally *on-grid*, they return echoes' timings which are integer multiples of the F_s .
- *Bodyguard effect*. In addition to affecting the **AER** performance, on-grid methods may converge slowly to suboptimal solutions. In fact, as show in Figure 5.2, instead of estimating the peak at its true location, two

FIGURE 5.2: Schematics of the *bodyguard effect* affecting on-grid approaches.

smaller “bodyguard” peaks are estimated around it instead. The estimation procedure may stop at this point returning two wrong peaks. Having smaller coefficients, this peaks may not be selected by the subsequent peak picking technique. Alternatively, the optimization procedure may continue, alternating tuning the weights of the two “bodyguards”, without converging to a solution.

- *Computational bottleneck.* A way to cope with the above limitations is to increase the F_s . However this results into a memory and computational bottleneck as several huge (Toeplitz) matrices needs to be built, one for each pair of microphones. In addition, this leads the risk that the optimization problem becomes ill-conditioned.

In the following section we will present a novel approach with aims at addressing the above limitation. It is based on a framework proposed for solving **LASSO** problems for continuos variables, hence the name **CD**. Before, let us show how to express to express the on-grid **BCE** problem proposed by [Lin et al. 2008] (See Table 5.1) as a standard **LASSO** problem.

5.3.1 From cross-relation to LASSO

Integrating the sparse penalty and the constraints proposed in Eq. (5.7), the Blind Sparse Nonnegative Channel Identification (**BSN**) problem proposed reads

$$\hat{h}_1, \hat{h}_2 = \arg \min_{h_1, h_2} \frac{1}{2} \|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2 + \lambda \|\mathbf{h}\|_1 \text{ s.t. } \begin{cases} \mathbf{h} \geq 0 \\ h_1[0] = 1 \end{cases}. \quad (5.8)$$

This cross-relation based optimization problem can rewritten in the **LASSO** formulation of Eq. (5.6) as

$$u = \arg \min_u \frac{1}{2} \|v - Bu\|_2^2 + \lambda \|u\|_1 \quad \text{s.t.} \quad u \geq 0,$$

where

$$v = T_2 e_1, \quad u = \begin{pmatrix} h_1[1:] \\ h_2 \end{pmatrix}, \quad A = \begin{pmatrix} -T_2[:, 1:] & T_1 \end{pmatrix},$$

where $T_i = \mathcal{T}(x_i)$. Here we used the light, yet common, Python notation for indexing the matrices and vectors. The matrix A is typically called *dictionary*.

5.4 PROPOSED APPROACH

The cross-relation identity Eq. (5.3) ensures that the relation

$$\tilde{\phi}_{LP} * \tilde{x}_1 * \tilde{h}_2^* = \tilde{\phi}_{LP} * \tilde{x}_2 * \tilde{h}_1^* \quad (5.9)$$

holds even during the introduced measurement process, hence

$$\mathcal{F}(\tilde{\phi}_{LP} * \tilde{x}_1) \cdot \mathcal{F}\tilde{h}_2^* = \mathcal{F}(\tilde{\phi}_{LP} * \tilde{x}_2) \cdot \mathcal{F}\tilde{h}_1^* \quad (5.10)$$

where \mathcal{F} denotes the Fourier Transform (**FT**) described in ??

In contrast with **SIMO-BCE** methods that operates in the time domain, here

we propose to use Eq. (5.10) in a penalized least-square problem. Such a formulation in the Fourier domain may even be considered as more convenient since the convolution operator is no longer involved. While the FT of \tilde{h}_i^* can be expressed in closed-form (see Eq. (5.13) below), the FT of $\tilde{\phi}_{\text{LP}} * \tilde{x}_i$ is not available due to the measurement process. To circumvent this issue, following approximation detailed in § 3.2.2, we consider the Discrete Fourier Transform (DFT) of the \tilde{x}_i :

$$\mathcal{F}(\tilde{\phi}_{\text{LP}} * \tilde{x}_i)\left(\frac{k}{2N}F_s\right) \approx X_i[k] \quad (5.11)$$

for all integers $k \in \{0, \dots, N\}$, where

$$X_i[k] = \sum_{n=0}^{2N-1} x_i[n] e^{-i2\pi \frac{kn}{2N}} \quad (5.12)$$

is the DFT of the real vector \tilde{x}_i as defined in ?? for positive frequencies only.

Let us define Δ_τ the following parametric vector of complex exponential

$$\Delta_\tau \stackrel{\text{def}}{=} \left(e^{-i2\pi \frac{k}{2N} F_s \tau} \right)_{0 \leq k \leq N} \in \mathbb{C}^{N+1}, \quad (5.13)$$

where we consider only the N positive frequency due to the Hermitian symmetry of the signal spectra in this application. Therefore, the Fourier-domain cross-relation of Eq. (5.10) evaluated at $f = \frac{k}{2N}F_s$ where $k \in \{0, \dots, N\}$ reads

$$\sum_{r=0}^{R_2-1} \alpha_2^{(r)} X_1 \odot \Delta_{\tau_2^{(r)}} = \sum_{r=0}^{R_1-1} \alpha_1^{(r)} X_2 \odot \Delta_{\tau_1^{(r)}} \quad (5.14)$$

where \odot denotes the component-wise Hadamard product.

With the above notation, in the following subsection we will present the CD framework for AER. This section is written with the help of the colleague Clement Elvira, co-author of a publication based on this work.

5.4.1 Echo localization with continuous dictionaries

By interpreting the FT of a Dirac as a parametric *atom*, we propose to cast the problem of RIR estimation into the framework of CD. To this aim, let the so-called *parameter set* be

$$\Theta \stackrel{\text{def}}{=} [0, T] \times \{1, 2\}, \quad (5.15)$$

where T is the length (in time) of the filter. Then, the two desired filters \tilde{h}_1^* , \tilde{h}_2^* given by Eq. (5.2) can be uniquely represented by the following discrete measure over Θ

$$\mu^* = \sum_{i=1}^2 \sum_{r=0}^{R_i-1} \alpha_i^{(r)} \delta_{(\tau^{(r)}, i)}. \quad (5.16)$$

where $\delta_{(\tau^{(r)}, i)}$ denotes the Dirac measure which is different from the Dirac function used when modeling the RIRs. The need of defining a measure over the parameter set Θ makes easier the parametrization of the problem in the context of CD. For instance, it is possible to better define operations which are

used in the algorithms and in the literature to solve the such type problems. Moreover, the rationale behind Eq. (5.15) and Eq. (5.16) is as follows. A couple of filters is now represented by a single stream of Diracs, where we have considered an augmented variable i indicating to which filter the spike belongs. For instance, a Dirac measure at $(\tau, 1)$ indicates that the filter 1 contains a Dirac at τ .

The set $\mathcal{M}_+(\Theta)$ of all unsigned and discrete Radon measures over Θ (i. e., the set of all couples of filters) is equipped with the total-variation norm (TV-norm) $\|\mu\|_{\text{TV}}$ ³⁵. We now define the *linear* observation operator $\mathcal{A}: \mathcal{M}_+(\Theta) \rightarrow \mathbb{C}^{N+1}$, which is such that

$$\mathcal{A}\delta_{(\tau,i)} = \begin{cases} -X_1 \odot \Delta_\tau & \text{if } i = 1 \\ +X_2 \odot \Delta_\tau & \text{if } i = 2. \end{cases} \quad \forall (\tau, i) \in \Theta. \quad (5.17)$$

³⁵See [Rudin 1987] for a rigorous construction of measures set and the TV-norm.

Therefore, by the linearity of the observation operator \mathcal{A} , the relation Eq. (5.14) can be rewritten as

$$\mathcal{A}\mu^* = \mathbf{0}_{N+1}, \quad (5.18)$$

where $\mathbf{0}_{N+1}$ is a $N + 1$ -length vector of zeros.

Before continuing our exposition, we note that the anchor constraint can be written in a more convenient way. Indeed, the constraint $\mu(\{(0, 1)\}) = 1$ ensures the existence of a Dirac at 0 in the filter 1. Then, the targeted filter reads

$$\mu^* = \delta_{(0,1)} + \bar{\mu}^* \quad (5.19)$$

where $\bar{\mu}^*$ is a (finite) discrete measure verifying $\bar{\mu}^*(\{(0, 1)\}) = 0$.

Denoting $y \stackrel{\text{def}}{=} -\mathcal{A}\delta_{(0,1)} \in \mathbb{C}^{N+1}$, the relation Eq. (5.18) becomes

$$\mathcal{A}\bar{\mu}^* = y. \quad (5.20)$$

For the sake of clarity, we use these conventions hereafter and omit the bar over μ . Now, following [De Castro and Gamboa 2012; Candès and Fernandez-Granda 2014], one can expect to recover the desired filter μ^* by solving

$$\hat{\mu} = \arg \min_{\mathcal{M}_+(\Theta)} \|\mu\|_{\text{TV}} \quad \text{s.t.} \quad \begin{cases} \mathcal{A}\mu = y \\ \mu(\{(0, 1)\}) = 0. \end{cases} \quad (5.21-\mathcal{P}_{\text{TV}}^0)$$

Note that (5.21- $\mathcal{P}_{\text{TV}}^0$) has to be interpreted as a natural extension of the well-known *basis pursuit* problem to the continuous setting. Indeed, for *any* finite discrete measure $\mu = \sum_{r=0}^{R-1} \alpha^{(r)} \delta_{(\tau^{(r)}, i)}$, the TV-norm of μ returns to the ℓ_1 -norm of the coefficients, i. e., $\|\mu\|_{\text{TV}} = \sum_{r=0}^{R-1} |\alpha^{(r)}|$.

Finally, Eq. (5.20) can be exploited to take into account noise during the measurement process (i. e., $n_i \neq 0$ in Eq. (5.1)), as well as approximation errors (see Eq. (5.11)-Eq. (5.14)). In that case, the first equality constraint in (5.21- $\mathcal{P}_{\text{TV}}^0$) is relaxed, leading to the so-called Beurling-LASSO (**BLASSO**) problem

$$\hat{\mu} = \arg \min_{\mu \in \mathcal{M}_+(\Theta)} \frac{1}{2} \|y - \mathcal{A}\mu\|_2^2 + \lambda \|\mu\|_{\text{TV}} \quad \text{s.t.} \quad \mu(\{(0, 1)\}) = 0. \quad (5.22-\mathcal{P}_{\text{TV}}^\lambda)$$

We emphasize that although continuous Radon measures may potentially be admissible, the minimizers of Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$) are guaranteed to be streams of Diracs [Bredies and Carioni 2020, Theorem 4.2]. In addition, although problem Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$) seems to depend on some regularization parameter λ , we describe in § 5.4.4 a procedure to automatically tune it to recover a desired number of spikes. Finally, note that problem Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$) is convex with linear constraints over the parameter set Θ . Therefore, theoretically, the problem can be solved exactly. However, in practice, optimization over space of measures, still complicated because many steps can only be done up to a prescribed precision.

5.4.2 From LASSO to BLASSO

In order to better understand the proposed approach based on the BLASSO algorithm, we can present it in light of the LASSO formulation.

$$\begin{aligned} \arg \min_u \frac{1}{2} \|v - Au\|_2^2 + \lambda \|u\|_1 & \text{ s.t. } u \geq 0 \\ \downarrow \\ \arg \min_u \frac{1}{2} \|y - \mathcal{A}\mu\|_2^2 + \lambda \|\mu\|_{\text{TV}} & \text{ s.t. } \mu \in \mathcal{M}_+(\Theta) \end{aligned}$$

Now, some parallelism can be envisioned:

- *From dictionary to operator:* The matrix A is typically referred to as *dictionaries*. Then selecting the l -th column of the dictionaries, i. e. Ae_l , means selecting an echo at location l -th w. r. t. the vector $u = \mathbf{h}[1 :]$. In the context of CD, the dictionary is translated into the operator \mathcal{A} thanks to the closed-form of the atom based on the Fourier theory. Therefore, $\mathcal{A}(\delta_{\tau})$ can be seen as the selection of an echo at location $\tau \in [0, T]$ ms.
- *Solution:* The LASSO-like approach promotes a solution $u = \mathbf{h}[1 :]$ which is sparse and non-negative vector. The last one, ensured by the non-negativity constraint. In the BLASSO, this is translated assuming the spike measure assumed for the channels, namely, $\mu = \sum_r \alpha^{(r)} \delta(t - \tau^{(r)})$.
- *Sparsity:* while in the initial case, the sparsity is enforced by the ℓ_1 -norm, in the second case it is pursued with the TV-norm.
- *Solver:* the former optimization problem can be solved with standard LASSO solvers, while for the latter a gradient-descent algorithm is used.

5.4.3 The resulting algorithm

The algorithm used to solve Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$) is an instance the sliding Frank-Wolfe algorithm proposed in [Denoyelle et al. 2019] to solve Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$). Detailed descriptions of the steps of the algorithm are given in § 12.2. In a nutshell, the algorithm iterative over the following steps until a condition on the cost function is met.

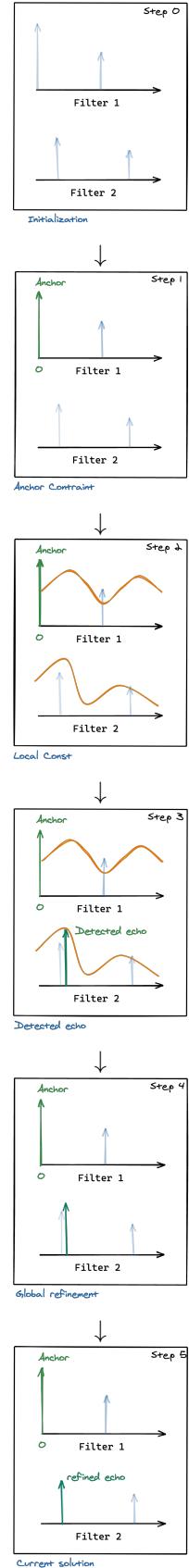


FIGURE 5.3: Illustration of the sliding Frank-Wolfe algorithm proposed in [Denoyelle et al. 2019] in BLASTER.

1. *Anchor constrain.* At first the anchor constraint is added arbitrarily on one of the two filters. This is used to initialize the two filters.
2. *Local cost based on Cross-relation.* For both the filters, a local cost function derived from the cross-relation for both the filters is computed. At this step either the initialization or previously found solution are used.
3. *Find the maximizer.* A new candidate echo's location is found as maximizer among the two local cost functions of the previous step.
4. *Update the amplitudes.* By solving a non-negative LASSO problem, all the echo's amplitude coefficients estimated until this point are updated.
5. *Joint refinement.* The position and the coefficient of the current solution are jointly refined to ease numeric resolution using the original cost function.
6. *Current solution and repeat.* The algorithm stops as soon as an iterate satisfies the first order optimality condition associated to the convex problem. if not, the algorithm iterates from step 2. using the current solution as input.

These steps are illustrated in Figure 5.3.

5.4.4 Homotopic path for λ estimation

Existing works, as well as the proposed one, relies of the regularization λ parameter to weight the sparsity penalty. However, this becomes an hyperparameter that needs to be carefully tuned according to the input data. Instead, we propose to compute a *path of solutions* to automatically estimate it in Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$). In the context of sparse optimization this technique is also referred to as *homotopic path*. More precisely, let λ_{\max} be the smallest value of λ such that the null measure is the solution to Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$). It can be shown that λ_{\max} is upper bounded by $\max_{\theta \in \Theta} |y^T \mathcal{A} \delta_\theta|$ (See § 12.2). Starting from $z = 1$ and the empty filter, we consider a sequential implementation where the solution of Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$) is computed for $\lambda^{(z)} = 10^{-0.05z} \lambda_{\max}$ until the desired number of spikes is found in each channel when incrementing z . For each $\lambda^{(z)}$, we search for a solution of Eq. (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$) with the solution obtained for $\lambda^{(z-1)}$ as a warm start.

5.5 EXPERIMENTS

The proposed method (BLASTER) is compared against the non-negative ℓ_1 -norm method (BSN) of [Lin et al. 2007] and the iterative ℓ_1 -norm approach (IL1C) described in [Crocco and Del Bue 2016a].³⁶ The problem is formulated as estimating the time locations of the first $R = 7$ strongest components of the RIRs for 2 microphones listening to a single sound source in a shoebox room. It corresponds to the challenging task of estimating first-order early reflections. The robustness of the methods is tested against different level of noise (SNR) and reverberation time (RT_{60}).

³⁶ A reference implementation for IL1C and BSN were provided by the authors of [Crocco and Del Bue 2016a].

The quality of the AER estimation is assessed in terms of precision³⁷ in percentage as in the literature of onset detection [Böck et al. 2012] and the RMSE in samples. Both metrics evaluate only the *matched* peaks, where a *match* is defined as being within a small window τ_{\max} of a reference delay. These two metrics are similar to the ones used in [Crocco and Del Bue 2015].

For this purpose we created three synthetic datasets of 1000 observations each, which are summarized in Table 5.2.

Dataset	Signals	SNR [dB]	RT ₆₀ [s]
$\mathcal{D}^{(\text{valid})}$	broadband noise	☒	☒
\mathcal{D}^{SNR}	broadband noise, speech	☒	400 ms
$\mathcal{D}^{\text{RT}_{60}}$	broadband noise, speech	20 dB	☒

³⁷ Since only K time locations are considered in both the ground truth and the estimation, precision and recall are equal.

TABLE 5.2: Summary of the dataset used for evaluation. ☒ and ☓ stands for randomly sampled from a continuous and discrete set of values respectively with uniform low.

$\mathcal{D}^{(\text{valid})}$ is used for tuning the hyperparameter λ and the peak-picking parameters for IL1C and BSN using RT₆₀ and SNR randomly drawn from $\mathcal{U}[0, 1]$ (sec) and $\mathcal{U}[0, 20]$ (dB) respectively; \mathcal{D}^{SNR} features SNR value uniformly sampled in $[0, 6, 14, 20, \infty]$ while the RT₆₀ is kept fixed to 400 ms; akin the $\mathcal{D}^{\text{RT}_{60}}$ is built sampling RT₆₀ value uniformly in $[200, 400, 600, 800, 1000]$ ms keeping SNR to 20 dB. Moreover, while for $\mathcal{D}^{(\text{valid})}$ broadband signals (white noise) are used as the source, for \mathcal{D}^{SNR} and $\mathcal{D}^{\text{RT}_{60}}$ speech utterances from the TIMIT dataset are also included. The signal duration is kept fixed to 1 s with sampling frequency $F_s = 16$ kHz. For a given RT₆₀ value and room with random dimensions, a unique absorption coefficient is assigned to all surfaces based on the Sabine's formula (Eq. (2.17)). Then, the two microphones and the source are randomly positioned inside the room. The parameters of such audio scene are then passed as input to the `pyroomacoustics` simulator [Scheibler et al. 2018b], which returns the corresponding RIRs as well as the *off-grid* echo delays and attenuation coefficients computed with the Image Source Method (ISM) [Allen and Berkley 1979]. Note that when generating the data, no samples have been pruned to match any minimal separation condition. To generate the microphone signals, an over-sampled version of the source signal is convolved with ideal RIRs at high frequency ($F_s = 1024$ kHz) made up of on-grid Diracs. The results are later resampled to meet the original F_s and Gaussian white noise is added to meet the given SNR value.

Finally, as described throughout Chapter 4, IL1C and BSN uses tuned peak picking step to identity the echoes. Here the same peak picking technique provided with reference implementation of these methods was used and tuned on a small validation set.

- ▶ QUANTITATIVE RESULTS are reported in Figure 5.4, Figure 5.5 and Table 5.3. Here, for both RMSE and precision and for both broadband and speech signal, the metrics are displayed against the dataset parameters. We observe that BSN performs worst in all tested conditions, possibly due to its strong reliance on the peak picking step. For $R = 7$ or higher, BLASTER yields similar or slightly worse performance than IL1C for the considered noise and reverberation levels, with decreasing performance for both as these levels increase. Using speech rather than broadband signals also yields worse results for all methods.

However, the echo timing RMSE is significantly smaller using BLASTER due to its off-grid advantage. We also note that BLASTER significantly outperforms IL1C on the task of recovering $R = 2$ echoes. As showed in Tab. 5.3, in mild conditions ($\text{RT}_{60} = 200$ ms, $\text{SNR} = 20$ dB), up to 68% of echoes can be retrieved by BLASTER with errors lower than half a sample in that case. This is promising since the practical advantage of knowing the timing of two echoes per channel has been demonstrated in [Di Carlo et al. 2019b] (See Chapter 10), and in [Scheibler et al. 2018e] (See Chapter 9).

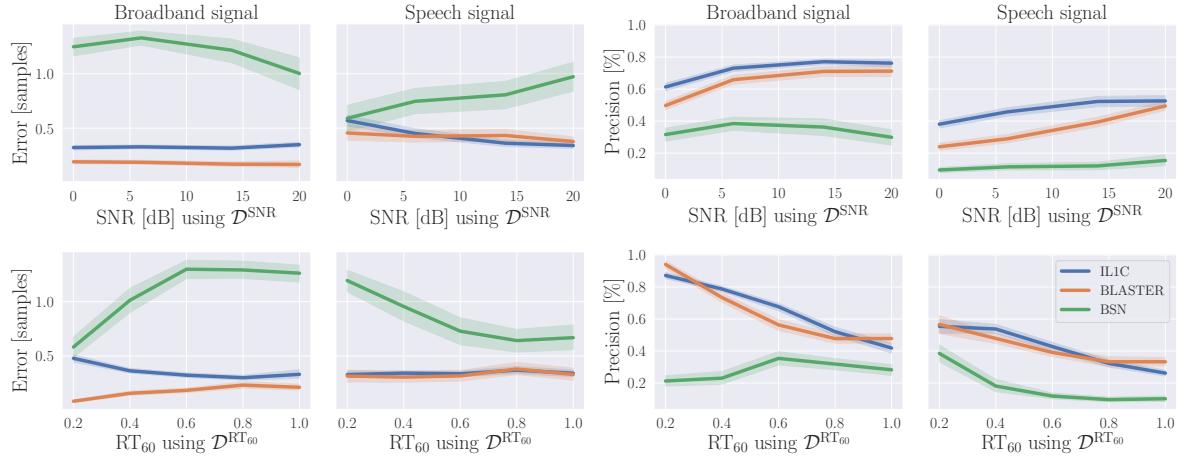


FIGURE 5.4: Mean error (left) and precision (right) versus SNR level (top) and RT_{60} level (bottom) using broadband and speech signals for the task of recovering $R = 7$ echoes. A threshold of $\tau_{\max} = 2$ samples is used to compute the precision. Error bands denotes 95% confidence intervals.

τ_{\max}	Precision [%]									
	R = 2 echoes					R = 7 echoes				
	0.5	1	2	3	10	0.5	1	2	3	10
BSN	8	9	27	46	62	5	8	38	54	73
IL1C	51	55	55	56	58	42	53	55	56	58
BLASTER	68	73	74	75	75	46	53	56	57	61

TABLE 5.3: Precision for different threshold τ_{\max} in samples for the recovery of $R = 2$ and 7 echoes, $\text{RT}_{60} = 200$ ms and $\text{SNR} = 20$ dB.

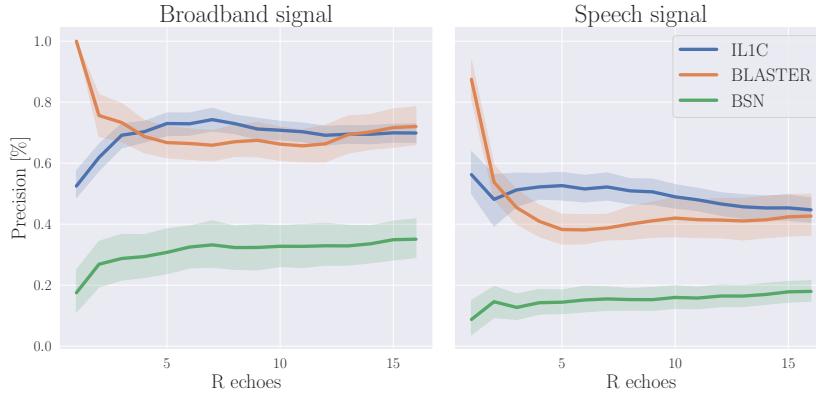


FIGURE 5.5: Precision versus number of echoes R to be retrieved for broadband (left) and speech (right) signals with $RT_{60} = 400$ ms and $SNR = 20$ dB. Error bands denotes 95% confidence intervals.

5.6 CONCLUSION

In this chapter we presented a novel knowledge-driven blind, off-grid echo retrieval method, based on the framework of continuous dictionaries. In particular “knowledge” we used is the echo model for the early part of the RIRs. The main motivation behind this approach is to overcome the pathological limitation of classic methods for BCE, discussed in the previous chapter. Despite an heavy mathematical formulation, it can be as the continuos extension of an LASSO problem used for addressing BCE. Comparisons with state-of-the-art approaches on various noise and reverberation conditions show that this method performs best when the number of echoes to retrieve is small. Future works will include many exciting directions, such as:

- extending the framework to multichannel recordings using the Multichannel cross-relation, as already envisioned by the related works [Crocco and Del Bue 2015; Lin et al. 2008];
- compare this approach with other off-grid AER methods [Tukuljac 2020, Chapter 6];
- adapt this approach for dealing with ReTF which allow for source-independent acoustic features (See Chapter 11);
- use deep learning approaches to estimate the level of sparsity (a. k. a. number of the most relevant echoes) in the RIRs;
- and, validate the approach on real-world recordings, such as the one provided by the DECHORATE dataset (See Chapter 7).



6

Lantern: Data-driven Acoustic Echo Retrieval

- ▶ **SYNOPSIS** In this chapter, we will use virtually supervised deep learning models to learn the mapping from microphone recordings to the echoes' timings. After presenting a quick overview of deep learning techniques (6.2), we will present a first simple model to estimate the echo coming from a close surface (6.3). This case is motivated by an application in Sound Source Localization (SSL), which will be discussed in detail in Chapter 10. Finally, we will present a possible way to achieve a more robust estimation (??) and discuss a possible way to scale this approach to multiple echoes (??).

Deep learning notions are personal digestion of materials available in standard machine learning textbooks, such as [Bishop 2006; Goodfellow et al. 2016], and the recent comprehensive work of [Purwins et al. 2019] oriented to audio signal processing. Instead, in the remaining sections, we will present part of the previously published work [Di Carlo et al. 2019b] and a technical report [*HRI-JF collaboration - Final Phase II Deliverable*] and on going research.

It is essential to say that the approaches presented here are part of a first investigation in echo-aware SSL. Therefore they are profoundly interconnected with Chapter 10, which can be considered a following up of the earlier work [Gaultier et al. 2017] authored by colleagues. With that being said, we will mainly focus on a simple yet common scenario where we estimate only the first and strongest echo. The generalization to multiple echoes will only be discussed as future work.

6.1 INTRODUCTION

Recently Deep Neural Networks (DNNs) has captured the attention of many research fields due to recent advances that allow for faster implementation and training, as well as for achieving considerable performance improvement. Therefore, DNNss have been widely used in many domains. They belong to the class of *supervised* learning or *data-driven* models, where the training step is performed on labeled data, namely input-output pairs. The inclusion of these models in audio signal processing tasks has meant a considerable improvement in the performance, mainly audio and music source separation, speech enhancement, and source localization (see related sections in Chap-

Keywords: Acoustic Echo Retrieval, TDOA Estimation, Supervised Learning, Deep Learning, Robust Regression.

Resources:

- ICASSP2018 Paper
- Code
- ICASSP2018 Poster

Di Carlo et al., "Mirage: 2D source localization using microphone pair augmentation with echoes"

Di Carlo and Deleforge, *HRI-JF collaboration - Final Phase II Deliverable*

ter 8).

Moreover, with respect to traditional machine learning methods, the use of deep learning models presents other advantages. They are flexible and adaptable across tasks; for instance, Convolutional Neural Network (**CNN**)-based models from Computer Vision can be adapted for audio source separation (e.g., [Xiao et al. 2016; Sainath et al. 2017; Perotin et al. 2018]), and for source localization (e.g., [Chakrabarty and Habets 2017; Vesperini et al. 2018; Nguyen et al. 2018; Salvati et al. 2018]). Furthermore, DNNs reduce – even remove – the step of designing hand-crafted features, by including the feature learning as part of the learning process.

Other supervised learning methods have been proposed in audio processing literature, such as Gaussian Mixture Model (**GMM**)-based frameworks [Laufer et al. 2013; Deleforge et al. 2015; Gaultier et al. 2017]. This framework allows to build power models based on probabilistic prior knowledge and are competing with DNN models in case of a small training dataset. However, they often require a strong assumption of the nature of the mapping between input-output pairs, such as (local-)linearity or independence between the variables.

The use of supervised learning models (not only the DNN), presents some disadvantages. One of the most relevant ones is their dependence on annotated data. It results in a significant bottleneck in audio signal processing tasks because collecting and annotating comprehensive real-world acoustic data is not trivial, since it requires proper tools, expertise, and time. To overcome this issue, a standard approach nowadays is to use training data generated by simulators. We will refer to this approach as *virtually supervised learning* and it provides great versatility, for instance:

- many different acoustic conditions can be included in the training data, e.g., different Reverberation Time (**RT₆₀**) and Signal-to-Noise-Ratio (**SNR**) levels;
- a precise (up to computer precision) annotation of the data come directly, e.g. spatial, temporal information of the audio scene events;
- and finally, the data can be potentially re-used for different applications, e.g., localization, separation, diarization, etc.;

Besides, the solutions obtained in a data-driven fashion suffer from bias depending on the dataset used and may not generalize to real-world scenarios.

Interestingly, the work in [Nguyen et al. 2018] discusses an automatic approach for collecting, annotating data automatically as a pre-calibration step. However, this approach is possible only with specific technologies and applications (the authors programmed a humanoid robot equipped with a binaural hearing system that automatically builds a training set). The following section will give a quick overview of relevant models used in our work.

*At first, the Gaussian Locally-Linear Mapping (**GLLiM**) framework was considered to address the AER problem. Then, we oriented our attention towards DNN models, which gave already satisfactory results. It seemed that the local-linearity assumption of **GLLiM** was in contrast with the highly non-linear nature of the mapping between echoes and observation. Nevertheless, we recognize other benefits in using the **GLLiM** framework, which cannot be found in the DNN-based approach, such as, the ability to generalize to missing data at the input or including prior statistical knowledge in the model.*



6.2 DEEP LEARNING MODEL

6.2.1 Multi-layer perceptions

Multi-layer Perceptron (MLP) are simple and basic modules of DNN and are also known as *fully-connected* or *dense layers*. They consist of a sequence of layer, each of which defined by an affine transformation followed by a non-linearity function:

$$\mathbf{y} = f(\mathbf{W}\mathbf{x} + \mathbf{b}),$$

where

- $\mathbf{x} \in \mathbb{R}^{D_{\text{in}}}$ is the input,
- $\mathbf{y} \in \mathbb{R}^{D_{\text{out}}}$ is the output,
- and $\mathbf{b} \in \mathbb{R}^{D_{\text{out}}}$ and $\mathbf{W} \in \mathbb{R}^{D_{\text{out}} \times D_{\text{in}}}$ are the *bias vector* and the *weight matrix*, respectively.

The function $f(\cdot)$ is a non-linear *activation* function, which allows the model to learn non-linear structures. Models built with these layers are typically used to map the input to a representation space where problems (e.g., classification, and regression) can be addressed more easily. The main drawback of these simple models is that they are not invariant to scaled or shifted input. Since in audio data, gain, temporal, and frequency variations are common, they are not suitable. Nevertheless, MLPs are used in combination with other type of layers, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs).

- ▶ NON-LINEAR ACTIVATION FUNCTIONS are one of the key feature of DNNs. Thanks to these, the model can achieve more *expressive power* with respect to linear models [Goodfellow et al. 2016]. Without these functions, it can be shown that the composition of affine transformations is equivalent to a single affine transformation. Therefore, the activation functions make the model capable of account for more complex relationships between the input and the output than an affine transformation. Typical examples of these functions are, for instance, hyperbolic tangent, the Sigmoid function, and Rectified Linear Unit (ReLU).

6.2.2 Convolutional neural networks

The CNN consists of convolutional layers that have been introduced to overcome some limitations of the simple linear ones. In a nutshell, they consist of learnable kernel functions that are convolved with their input. Therefore, they are characterized by

- shift-invariance property, and can be used to reduce the model complexity since the same kernel can be used at different input location;
- the ability to detect local structures at a different level of abstraction in the network (see ??).

Using a Python-like notation, a 2D-dimensional convolutional layer is defined by the following equation:

$$\mathbf{Y}[:, :, k] = f\left(\sum_{i=0}^{I-1} \mathbf{W}[:, :, i, k] * \mathbf{X}[:, :, i] + \mathbf{b}[:, :, k]\right),$$

where

- $\mathbf{X}_k \in \mathbb{R}^{F \times T \times I}$ and $\mathbf{Y}_c \in \mathbb{R}^{A \times B \times K}$ are input and output tensors of dimension, respectively.
- $i \in [0, I]$ denotes the channel dimensions,³⁸
- $k \in [0, K]$ denotes output;
- the tensors \mathbf{W} and \mathbf{b} denotes the weight and bias tensors, respectively;
- f is an activation function;
- and, $*$ denotes discrete convolution;

³⁸ In DNN community is common refer the input dimensions as channels. Based on the application, such channel do not necessarily corresponds to the channel of microphone recordings.

CNN can be designed for either 1D or 2D convolution, or combination of them. In general, in audio 1D convolution layer are used to process the time-domain or frequency-domain input, whereas the 2D convolutional layer are used for time-frequency representation, such as spectrograms. The output of a convolutional layer consists of a collection of local convolution operations, which are typically referred to as *feature map*. It is common to use CNN architectures combining multiple convolutional layers with *pooling layers* in between. Pooling layers are used to down-sample the features map. In addition, to reduce the model complexity (i.e., number of free-parameters), pooling layers subsequently reduce the size of the data as the model goes “deeper”. In this way, deeper layers can integrate larger extents of the data and extract spatial (in the sense of the input dimensions) features at a larger scale. The typical pooling operator is the *max-pooling* and *mean-pooling*, which samples non-overlapping patches by keeping the biggest and the average value in the region, respectively.

6.2.3 Hybrid architectures

As mentioned earlier, modern DNN architectures consist of a combination of different types of layers. For instance, CNNs are used to overcome the lack of shift and scale invariance that the MLP suffers of, and allows to extract spatial features. In contrast, MLPs offers a simple mapping from big-dimensional space to smaller ones, suitable for classification and regression problems. Therefore hybrid architectures are now the standard for deep learning models.

6.2.4 Learning and optimization

By its own, a DNN model consists of thousands of parameters θ . In order to learn a specific task, such as regression or classification, they need to be optimized. To optimize the parameters θ , a variant of the gradient descent algorithm is usually implemented. A *loss function* $\mathcal{L}_\theta(\hat{\mathbf{Y}}, \mathbf{Y})$ measure the

difference between the predicted values $\hat{\mathbf{Y}}$ and the desired \mathbf{Y} . The optimization processes then iteratively update the parameters θ , so the loss function is minimized, that is

$$\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}_{\theta}(\hat{\mathbf{Y}}, \mathbf{Y}),$$

where η is referred to as *learning rate* and accounts for how much to update the θ at each iteration. Due to the composite structure of the DNN, the gradient $\nabla_{\theta} \mathcal{L}_{\theta}(\hat{\mathbf{Y}}, \mathbf{Y})$ is computed via the chain rule, using the *back propagation* algorithm.

The *Stochastic Gradient Descent* is a variant of the gradient descent algorithm which has become the standard for training DNNs today. It was introduced to solve computational and memory load issues due to large training since it approximates the gradient at each step on a mini-batch of data samples. Finally, we would like to mention the following common techniques used to avoid overfitting and speed up the convergence.

- *early stopping* allows to stop the training upon some condition on the loss function (or other metrics), for instance, when it is not improving after certain amount of iterations;
- *batch normalization* consists of scaling the data by estimating its statistics during training, which usually leads to better performance and faster convergence;
- *dropout* is a regularization technique for reducing overfitting and consists in randomly omit neural network units during training.

6.2.5 End-to-end and Hybrid approaches

The general scheme of DNNs-based approaches is one of the reasons for their success. Therefore they have been applied to solve many problems and at a different level of complexity. We can then identify main tendencies in signal processing: end-to-end and hybrids approaches. The former attempts to address signal processing problems by getting rid of signal processing entirely as opposed to the latter. Although end-to-end architecture performs only marginally better than hybrid or other models, it is a common belief that soon they will outperform them [Vincent et al. 2018, Chapter 19].

Nevertheless, in this chapter, we will propose a framework based on a hybrid approach: we will use DNN to estimate echoes' properties, and later, in Chapter 10, we will use such quantities to address SSL problem.

6.3 LEARNING THE FIRST ECHO

6.3.1 Scenario and motivations

As a first step, we will use DNN to estimate the timings of the first and strongest echo from stereophonic recordings, i.e. $R = 1$. To this end, we consider a simple yet common scenario: two microphones placed close to a surface, as illustrated in Figure 6.1. Such a scenario is motivated by the echo-aware Sound Source Localization (SSL) application that will be discussed in Chapter 10. The

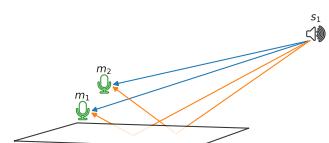


FIGURE 6.1: Typical setup with one source source recorded by two microphones. The illustration shows direct sound path (blue lines) and resulting first-order echoes (orange lines).

reason why we consider only single microphone pairs is that this approach can be generalized to microphone array using data-fusion-line approaches. In fact, by considering all the pairs, it is possible to aggregate all their estimation a second moment.³⁹ In the remaining part of this chapter, we will discuss how to achieve robust first-echo estimation and how it could be generalized to multiple echoes.

6.3.2 Proposed approach

Our approach is to train a DNN on a dataset simulating the considered close-surface scenario. Under this assumption and the Short Time Fourier Transform (STFT) signal model presented in Eq. (3.25), we can consider the following simplified model for the Room Transfer Functions (RTFs),

$$H_i[k] = \sum_{r=0}^{R=1} \alpha_i^{(r)}[k] e^{-i2\pi f_k \tau_i^{(r)}} + \varepsilon_i[k], \quad (6.1)$$

where $\tau_i^{(r)}$ and $\alpha_i^{(r)}[k]$ are the echoes's time of arrival and amplitude, respectively. The f_k denotes k -th frequency bin and the error term $\varepsilon_i[k]$ collects later echoes, the reverberation tail, diffusion, and noise.

We model the AER problem as multi-target regression problem, namely the output on the model are multiple real-valued parameters. Following the approaches suggested in [Deleforge et al. 2015; Gaultier et al. 2017], we consider the instantaneous Interchannel Level Difference (ILD) and the IPD as input features. As discussed in § 3.3.1, the ILD and IPD can be consider as a particular case of Relative Transfer Function (ReTF). The ILD and IPD can be estimated from the STFT of the microphone signals, such as,



$$\begin{cases} \text{ILD}[k] = \frac{1}{T} \sum_{l=1}^T \log \left| \frac{X_2[k, l]}{X_1[k, l]} \right|, \\ \text{IPD}[k] = \frac{1}{T} \sum_{l=1}^T \frac{X_2[k, l] / |X_2[k, l]|}{X_1[k, l] / |X_1[k, l]|}, \end{cases} \quad (6.2)$$

where $X_i[k, l]$ is the STFT of the i -th microphone.

More precisely, the input of the network is

$$\xi = [\text{ILD}, \text{Re}\{\text{IPD}\}, \text{Im}\{\text{IPD}\}],$$

namely the concatenation of the above features for all the frequencies. Here $\text{Re}\{\cdot\}$ and $\text{Im}\{\cdot\}$ denote real and imaginary part operators, respectively. Note that for the IPD, the frequency $k = 0$ is discarded because it is constant for every observation.

 Initial investigation shown that learning directly the echoes arrival time and their amplitude yield huge estimation error, probable to the complexity of the task. Motivated by the application in SSL discussed in Chapter 10, we consider only the Time Difference of Arrivals (TDOAs) between the direct path propagation.

$$\tau_{\text{TDOA}} = \frac{1}{c} \|\underline{x}_2 - \underline{s}\| - \frac{1}{c} \|\underline{x}_1 - \underline{s}\| = \tau_2^{(0)} - \tau_1^{(0)} \quad [\text{s}], \quad (6.3)$$

$$\tau_{\text{iTDOA}} = \frac{1}{c} \|\dot{\underline{x}}_2 - \underline{s}\| - \frac{1}{c} \|\dot{\underline{x}}_1 - \underline{s}\| = \tau_2^{(1)} - \tau_1^{(1)} \quad [\text{s}], \quad (6.4)$$

$$\tau_{\text{TDOE}, 1} = \frac{1}{c} \|\dot{\underline{x}}_1 - \underline{s}\| - \frac{1}{c} \|\underline{x}_1 - \underline{s}\| = \tau_1^{(1)} - \tau_1^{(0)} \quad [\text{s}], \quad (6.5)$$

³⁹ An example of this approach is the Steered Response Power with Phase Transform (SRP-PHAT) [DiBiase et al. 2001] algorithm used in SSL which uses the knowledge of the microphone array geometry to aggregate the contribution of each microphone pair.

where $\hat{\mathbf{x}}_i$ denotes the position of the image of the microphone at position \mathbf{x}_i with respect to the reflector. Note that $\tau_{\text{TDOE},2} = \tau_{\text{iTDOA}} + \tau_{\text{TDOE},1} - \tau_{\text{TDOA}}$. These three quantities are directly connected to RIRs, as illustrated in Figure 6.2. Let $V = [\tau_{\text{TDOA}}, \tau_{\text{iTDOA}}, \tau_{\text{TDOE},1}] \in \mathbb{R}^3$ be the vector of the parameters of interest and $\mathcal{V} = \{\text{TDOA}, \text{iTDOA}, \text{TDOE}_1\}$ the set of corresponding labels.

Since we use ReTF-based feature to estimate echoes, the proposed approach is dubbed as *LeArNing echo regression from room TransFer functioNs (LANTERN)*.

In general, the mapping between V and the proposed feature is not unique. In particular, this happens when $\tau_2^{(r)} = \tau_1^{(r)}$ for $r = \{0, 1\}$. In order to avoid this, we preventively pruned all the entries with $|\tau_2^{(r)} - \tau_1^{(r)}| < 10^{-6}$ for $r = \{0, 1\}$ from the dataset.

Here we use the simple MLP architecture described in § 6.2.1. This model consists of a D -dimensional input layer, a 3-dimensional output layer, and 3 fully connected hidden layers with respective input sizes 500, 300, and 50. ReLU activation functions are used except at the output layer, and each hidden layer has a dropout probability $p_{\text{do}} = 0.3$ to prevent overfitting.

We use the Mean Squared Error (MSE) loss function for training, that is,

$$\mathcal{L}_{\theta}(V, \hat{V}) = \frac{1}{|\mathcal{V}|} \sum_{\xi \in \Xi} \sum_{v \in \mathcal{V}} |\tau_{v,\xi} - \hat{\tau}_{v,\xi}|^2 \quad (6.6)$$

where v indexes one of TDOA in \mathcal{V} , ξ denotes a sample in the mini-batch Ξ , and θ are the model parameters. 

 The normalized Root Mean Squared Error (nRMSE)⁴⁰ is taken as validation metric for assessing the quality of the estimation \hat{V} . The network is manually tuned on a validation set to find the best combination of number of hidden layers, their sizes and p_{do} .

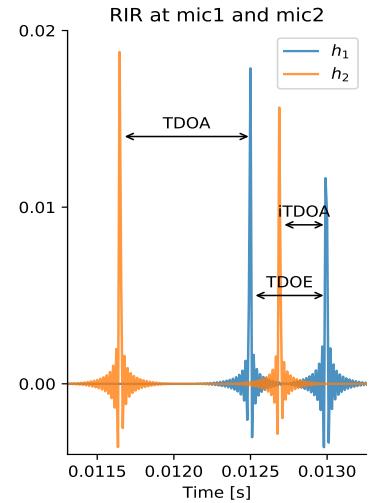


FIGURE 6.2: Superposition of two RIRs and visualization of time difference of arrival between direct paths (TDOA), first echoes (Image TDOA (iTDOA)) and direct path and first echo (Time Difference of Echo (TDOE)).

⁴⁰ The acnRMSE takes values between 0 (perfect fit) and ∞ (bad fit). If it is equal to 1, then the prediction is no better than a constant. It is typically chosen because it is more robust to outliers than RMSE.

6.3.3 Data, implementation and experimental results

For training and validation of the MLP we generate many random, shoe-box room configurations using the software presented in [Schimmel et al. 2009]. This software implements both the Image Source Method (ISM) for simulating reflections and a ray-tracing algorithm for diffusion. Room widths are uniformly drawn at random in [3, 9] m, heights in [2, 4] m. Random source and microphones positions are used, respecting the close-surface scenario.⁴¹ In particular, the microphones are at most 30 cm from the close-surface, placed 10 cm from each other, and the other walls' absorption coefficients are uniformly sampled in (0.5, 1) and the one of the close-surface is in (0, 0.5). The same realistic diffusion profile [Gaultier et al. 2017] is used for all surfaces. Around 90,000 audio scenes are generated this way, yielding a RT₆₀ between 20 ms and 250 ms.

The RIRs are convolved with 1 sec of white-noise (wn) with no additional noise. All signals and RIRs are sampled at 16 kHz. The STFT is performed on 1024 point with 50% overlap. Finally the features are computed as in (6.2) yielding a vector of size $D = 1534$ for each observation.

⁴¹ A rejection-sampling strategy was used to approximate uniform distributions.

While we validate the **MLP** on a portion of the dataset in a *holdout* fashion, the test is conducted on 200 new **RIRs** convolved with both **wn** and speech (**sp**) utterances. This set is generated similarly to the training and validation sets. Moreover, the test recordings are perturbed by external white noise at 10 dB Signal-to-Noise-Ratio (**SNR**) (**wn+n**, **sp+n**). The speech signals are normalized speech utterances of various lengths (from 1 s to 6 s), randomly selected from the TIMIT corpus.

► EXPERIMENTAL RESULTS

To check the validity of **TDOA** estimation with the proposed **MLP** model, we compare it to a baseline algorithm Generalized Cross Correlation with Phase Transform (**GCC-PHAT**) (see § 10.3.1). **GCC-PHAT** is a popular method for estimating **TDOA** between two microphone recordings. It is known to achieve good performance in the case of **the broadband source signal**. However, it was shown that as soon as the acoustic condition **becomes** challenging due to strong early echoes, high reverberation level, and noise, the method performances decrease [Chen et al. 2006].

	Input		nRMSE	
		TDOA	iTDOA	TDOE
MIRAGE	wn	0.18	0.28	0.25
MIRAGE	wn+n	0.68	0.69	0.89
MIRAGE	sp	0.31	0.34	0.56
MIRAGE	sp+n	0.99	0.98	1.48
GCC-PHAT	wn	0.21	-	-
GCC-PHAT	wn+n	0.68	-	-
GCC-PHAT	sp	0.32	-	-
GCC-PHAT	sp+n	1.38	-	-

TABLE 6.1: Normalized root mean squared error for TDOA estimation using the **MLP** architecture.

TDOA estimation errors using the proposed approach and **GCC-PHAT** are presented in Table 6.1. Training a **MLP** to estimate **TDOAs** brings similar performances as **GCC-PHAT** in terms of nRMSE. However, the estimation of **iTDOA** and **TDOE** seems to be a more challenging task for such a simple **MLP** model. When some external noise is added, the performances of both methods degrades. This is a well-known and expected behavior for **GCC-PHAT**, and it suggests that noise should also be considered in the training phase. **When** Nevertheless, our results confirm the possibility of retrieving the strongest echoes from only two-microphone recordings.



6.4 ROBUSTLY LEARNING THE FIRST ECHO

The above model was proposed in **the** our published work [Di Carlo et al. 2019b]. To overcome the limitations of **such a simple approach** with introduced the following three modifications:

- use state-of-the-art deep learning architecture;
- include corrupting noise in the training data;
- **and** use a robust loss function which is able to return a level of confidence of the estimation.

Motivated by their success, we consider a CNN architecture (see § 6.2.2). To this end, we inspire by the works [Chakrabarty and Habets 2017; Nguyen et al. 2018] for the task of SSL. As shown in Figure 6.3, the architecture consists of

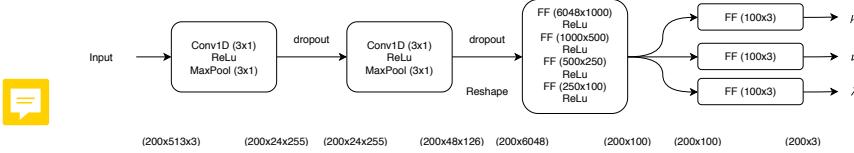


FIGURE 6.3: Architecture of the proposed deep neural network. Input and output dimensions for each stage are reported. The first dimension is the batch size $B = 200$.

two convolutional modules made of a one-dimensional convolutional layer (1DConv) followed by max-pooling along the frequencies, followed by ReLU activation function and batch-normalization. The second part consists of a cascade of fully connected feed-forward (FF) layers (basically the MLP model discussed above). In order to perform 1D-convolution on the right dimension, we re-arranged the input so that the each of features [ILD, Re{IPD}, Im{IPD}] is considered as channel for the 1DConv. After each layer a dropout probability $p_{\text{do}} = 0.3$ is applied to prevent overfitting.

6.4.1 Gaussian- and Student's T-based CNNs

In the MLP model presented in the previous section, the output consisted only in the TDOAs V . As mention earlier, our final goal is to generalize this approach to microphone array in a data-fusion-like fashion. To this end, we need to provide a confidence measure on the estimate. In our work [Di Carlo et al. 2019b], we proposed to use predicted values and the error on the validation set as mean and variance to build a Gaussian prior. Instead, here, we explicitly modify the CNN model to estimated these parameters. This idea recalls Bishop's Mixture Density Network (MDN) in [Bishop 1994], where the output of a neural network parametrizes a Gaussian mixture model. The rationale behind this design choice is to allow the learning model to assess its prediction quality. In order to achieve this, both loss functions and network outputs need to be modified accordingly.

- THE GAUSSIAN-BASED LOSS FUNCTION is derived as follows. First, we modify the network to return mean and variances, namely $V_N = \{\mu_v(\xi; \theta), \sigma_v^2(\xi; \theta)\}_v$ for $v = \{\text{TDOA}, \text{iTDOA}, \text{TDOE}_1\}$. Moreover, we assume that the probability of observing each of the τ_v for each $v \in \{\text{TDOA}, \text{iTDOA}, \text{TDOE}_1\}$, given the observation ξ in the mini-batch Ξ , follows a Gaussian distribution, namely

$$p(\tau_v | \xi; \theta) \sim \mathcal{N}(\tau_v; \mu_v(\xi; \theta), \sigma_v^2(\xi; \theta)). \quad (6.7)$$

The general form of the Gaussian probability density function is

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

From now on, we omit the dependency on $(\xi; \theta)$ for the sake of clarity. Finally, the corresponding training loss function is the negative Gaussian log-likelihood, that is,

$$\mathcal{L}_{\theta}^N(V, \hat{V}) = \frac{1}{|\mathcal{V}|} \sum_{\xi \in \Xi} \sum_{v \in \mathcal{V}} \log \sigma_v^2 + \frac{|\tau_v - \mu_v|^2}{\sigma_v^2}. \quad (6.8)$$

It follows that for each input ξ , the network will return the parameters of 3 Gaussian distributions, (μ_v, σ_v^2) for each variable $v \in \mathcal{V}$. Moreover, it is

easy to see that the estimated TDOAs values identify with the $\{\mu_v\}_v$ for each $v \in \mathcal{V}$. Hereafter we denote with $\hat{V}_{\mathcal{N}}$ the set of the six network outputs.

This approach can then be generalized to other this of distribution. For instance, we can consider a Student's *t-distribution*, which is supposed to be more robust to outliers. Using such distribution has already been proposed in the context of binaural SSL, such as [Zohny and Chambers 2014; Deleforge and Forbes 2016].

- ▶ THE STUDENT'S *T*-BASED LOSS FUNCTION can be derived similarly in the Gaussian case assuming a *t*-distribution prior on the prediction. We will reminds to the ?? for the detailed derivation. With that being said, the corresponding loss function is the negative *t*-distribution log-likelihood writes

$$\begin{aligned} \mathcal{L}_{\theta}^T(V, \hat{V}) = & \sum_{\xi \in \Xi} \sum_{v \in \mathcal{V}} \frac{1}{2} \log(\nu_v \pi_v) + \frac{1}{2} \log(\lambda_v^2) - \log \Gamma\left(\frac{\nu_v + 1}{2}\right) \\ & + \log \Gamma\left(\frac{\nu_v}{2}\right) + \frac{\nu_v + 1}{2} \log\left(1 + \frac{|\mu_v - \tau_v|^2}{\nu_v \lambda_v^2}\right) \end{aligned} \quad (6.9)$$

where Γ is the Gamma function. Similarly to the Gaussian case, at each each input ξ , the network will return the parameters of 3 *t*-distribution $(\mu_v, \nu_v, \lambda_v)$ for each variable $v \in \mathcal{V}$. The set of there nine network output is denoted as $\hat{V}_{\mathcal{T}}$.

6.4.2 Experimental results

In order to validate the proposed approach, it is compared to the vanilla MLP model proposed above in the previous section. Rather than training on noiseless observation, we consider microphone recordings featuring background noise. To this end, we perturbed the observed microphone signals with AWGN, leading to SNR values uniformly distributed in [0, 30] dB. While, the test set include observation featuring SNR level of [0, 10, 20] dB. The protocol to generate the observations and the feature extraction steps are the same (?? 41), but the investigation is restricted to broadband source signal (white noise) only. The network is manually tuned on a validation set to find the best combination of the number of hidden layers and their sizes.

The general form of the Student's *t* probability density function is

$$\mathcal{T}(x; \mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where $\Gamma(n) = (n - 1)!$ is the Gamma function, defined for positive integers.

Figure 6.4 shows the performances of the proposed models $\text{CNN}_{\mathcal{V}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{N}}}$ and $\text{CNN}_{\mathcal{V}_{\mathcal{T}}}$ with respect the MLP $_{\mathcal{V}}$ and the baseline GCC – PHAT of the described in the section for the task of estimating TDOA, iTDOA and TDOE. Again we report the performance in terms of hRMSE computed on the test samples. It can be seen that the proposed modifications (CNN, robust loss function, and considering noise in training) yield considerable improvement in the performance, with respect to the vanilla and baseline methods. In particular, they are more robust to noise, as indicate by the error bar. Nevertheless, while both $\text{CNN}_{\mathcal{V}_{\mathcal{N}}}$ and $\text{CNN}_{\mathcal{V}_{\mathcal{T}}}$ outperform the $\text{CNN}_{\mathcal{V}}$ for TDOA and iTDOA estimation, it seems there is no significant differences between them.

Deeper insight about this can be envisioned in Figure 6.5, where the performances are reported for different SNR levels considered at the test phase. These results suggest that using a simple MSE-based loss function yield to performance more sensitive to external noise. Besides, again it seems that

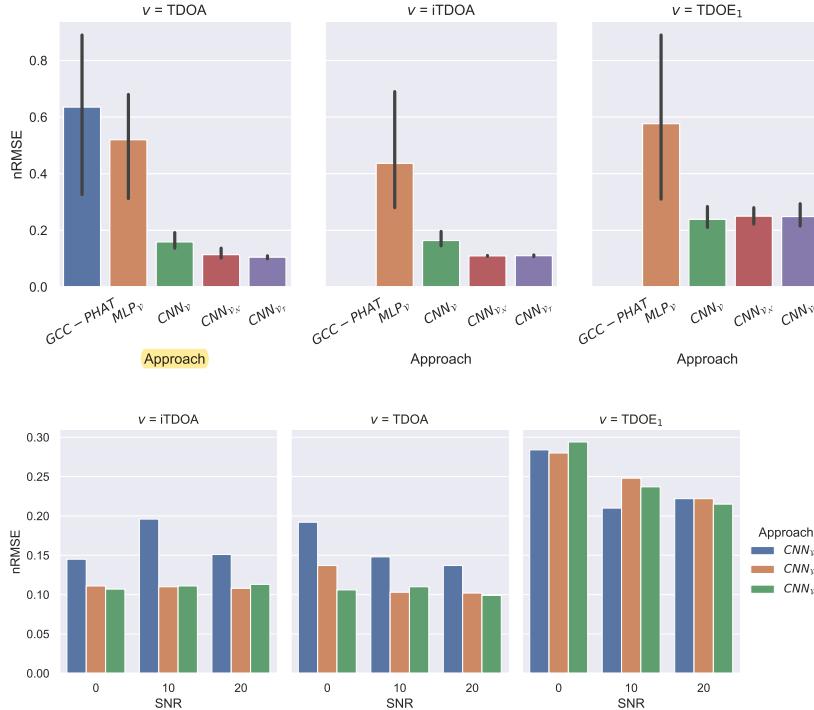


FIGURE 6.4: Normalize root mean squared error (the lower, the better) for **TDOA** estimation using the proposed architectures.



FIGURE 6.5: Normalize root mean squared error (the lower, the better) for **TDOA** estimation using the proposed architectures for different **SNR** levels.



considering the *i*-distribution do not bring significant difference over the Gaussian-case. Nevertheless, the choice of modeling the output of the network according to different distributions may lead to differences at the data-fusion step.

6.5 CONCLUSION

This chapter presented an attempt to use deep learning models to learn the complex mapping from stereophonic recordings to echoes' timings. As part of a first investigation, we consider a simple yet common scenario where two microphones are placed close to a surface and listen to a single source's sound. Moreover, the proposed approach is limited to estimating the time delays between the two microphones, the direct propagation path, and the first-strongest reflection. This simplification is motivated by an application to echo-aware SSL, which will be presented in Chapter 10.

We studied two standard DNN architectures, which are trained in virtually supervised fashion with data generated by an acoustic simulator. Finally, we discuss three ways to increase such performances, i.e., training on noisy data, using the CNN architecture, and use robust loss functions.

The investigation so far was limited to microphone pairs. Taking inspiration from data-fusion approaches, in chapter Chapter 10, we will propose to aggregate multiple pairs' contributions together to improve the final estimation. To this end, we modify the output of the DNN in order to return a confidence measure.

A nature follow-up of this work could consider the following directions.

- The first direction consider the generalization to more reflections.

Unfortunately this step is not trivial. Current informal investigation spanned different training paradigm, such as, *curriculum learning* [Bengio et al. 2009], *teacher-student learning* [Tu et al. 2019], where the network is subsequently trained to estimate more echoes. Other approaches are currently subject of investigation, such as, using a combination of DNN pre-trained for single echo estimation.

- Due to the drop of performances when using speech signals, we could consider more robust audio features (e.g., ReTF using state-of-the-art ReTF estimators) or DNN architecture designed to handle missing input data.
- Finally, another interesting direction comes from physics-driven neural networks [Nabian and Meidani 2020; Rao et al. 2020; Jin et al. 2020], where physical-based layers or physically-motivated regularizers are implemented facilities the learning.

7

dEchorate: Datasets for Acoustic Echo Estimation

- ▶ **SYNOPSIS** This chapter presents dEchorate: a new database of measured multichannel room impulse response (RIRs) including annotations of early echoes and 3D positions of microphones, real and image sources under different wall configurations in a cuboid room. These data provide a tool for benchmarking recent methods in *echo-aware* speech enhancement, room geometry estimation, RIR estimation, acoustic echo retrieval, microphone calibration, echo labeling, and reflectors estimation. The database is accompanied by software utilities to easily access, manipulate, and visualize the data and baseline methods for echo-related tasks.

The material presented in the chapter is the result of a work done while visiting prof. Sharon Gannot and ing. Pinchas Tandeitnik at the Bar’Ilan University, Israel. The work described here, together with its continuation described in ?? will be submitted as a journal article to the EURASIP special edition *Data-driven Audio Signal Processing: Methods and Apps*.

7.1 INTRODUCTION

As discussed § 4.4.1, many **RIRs** datasets are available online. However, most of them are specifically designed for applications either to Speech Enhancement (**SE**) or to Room Geometry Estimation (**RooGE**). The main common drawback of these datasets is that they can not be easily used for other tasks than the one which they were designed for. In particular, **SE**-oriented datasets lack of a proper annotation of echoes in the **RIRs** or the absolute position of objects inside the room. Alternatively, the datasets for **RooGE** typically focus on scenarios which are not suitable for **SE** application. dEchorate was designed to fill this gap: a fully calibrated multichannel **RIR** database with accurate annotation of the geometry and echoes in different configurations of a cuboid rooms with varying wall acoustic profiles. The database currently features 1800 annotated **RIRs** obtained from 6 arrays of 5 microphones each, 6 sound sources in 10 different acoustic conditions. All the measurements were realized in the acoustic lab at Bar-Ilan university following a consolidated protocol previously established for the realization of two other multichannel **RIRs** databases: the BIU’s Impulse Response Database [Hadad et al. 2014] gathering **RIRs** of different reverberation levels sensed by uniform linear arrays

“I’m here ‘cause I [messed up]
Somebody jump in my computer server
And take the information out
Gentleman, gentleman”
—Venetian Snares, Gentleman

Keywords: Room impulse response, Early reflection, Acoustic echoes, Audio database, Microphone arrays.

Resources:

- [Code repository](#)
- [Dataset](#)

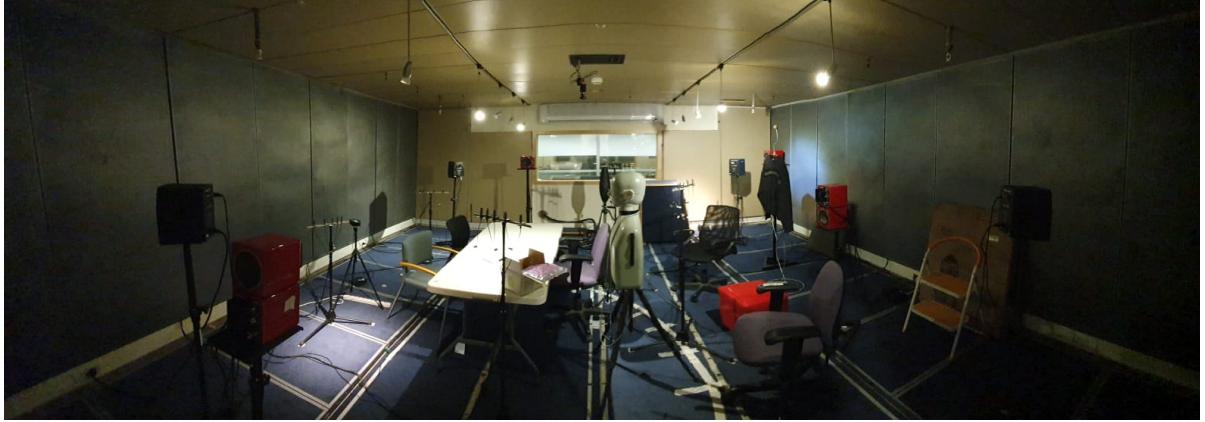


FIGURE 7.1: Broad-view picture of acoustic lab at Bar-Ilan university.

(ULAs); and MIRaGE⁴² [Čmejla et al. 2019] providing a set of measurements of a source placed on a dense position grid. dEchorate is designed for Acoustic Echo Retrieval (AER) with linear arrays, and is more generally aimed at analyzing and benchmarking RooGE and echo-aware signal processing methods on real data. In particular, it can be used to assess robustness against the number of reflectors, the reverberation time, additive spatially-diffuse noise and non-ideal frequency and directive characteristics of microphone-source pairs and surfaces in a controlled way. Due to the amount of data and recording conditions, it could also be used to train machine learning models or as a reference to improve RIR simulators. The database is accompanied with a Python toolbox that can be used to process and visualize the data, to perform analysis or to annotate new datasets.

7.2 DATABASE REALIZATION

7.2.1 Recording setup

The recording setup is situated in a cuboid room with dimension $6 \text{ m} \times 6 \text{ m} \times 2.4 \text{ m}$. The 6 facets of the room (walls, ceiling, floor) are covered by acoustic panels allowing controllable reverberation time (RT_{60}). We placed 4 directional loudspeakers (direct sources) facing the center of the room and 30 microphones mounted on 6 non-uniform linear arrays (nULA) of 5 sensors each. An additional channel is used for the loop-back signal, which serves to compute the time of emission and detect errors. Each loudspeaker and each array was positioned close to one of the walls in such a way that the nature of the strongest echo can be easily identified. Moreover, their positioning was chosen to cover a wide distribution of source-to-receiver distances, hence, a wide range of direct-to-reverberant ratios (DRR). Further, 2 more loudspeakers were positioned pointing towards the walls (indirect sources). This was done to study the case of early reflections being stronger than the direct-path. Each linear microphone array consists of 5 microphones with non-uniform inter-microphone spacings of $[4, 5, 7.5, 10] \text{ cm}^{42}$. Each array is steered towards a different vertical edge of the room for calibration and reproducibility purposes.

⁴² i.e. $[-12.25, -8.25, -3.25, 3.25, 13.25]$ cm w.r.t the barycenter

Loudspeakers	(directional, direct) 4× Avanton
(directional, indirect) 2× Avanton	
(omnidirectional) 1× B&G	
(babble noise) 4× 6301bx Fostex	
Microphones	30× AKG CK32
Array	6× nULA (5 mics each, handcrafted)
A/D Converter	ANDIAMO.MC
Indoor Positioning	Marvelmind Starter Set HW v4.9

TABLE 7.1: Technical specification of the measurements equipment used in the recordings.

7.2.2 Measurements

The main feature of this room is the capability to change the acoustic profile of each of its facet by flipping double-sided panels with one reflective and one absorbing face. This allows to achieve precise values of RT_{60} that range from 0.1 to almost 1 second. In this dataset the panels of the floor were kept always absorbent.

Two types of sessions were considered, namely, *one-hot* and *incremental*. For the first type, a single facet was placed in reflective mode while all the others were kept absorbent. For the second type, starting from fully-absorbent mode, facets were progressively switched to reflective mode one after the other until all but the floor were reflective, as shown in Table 7.2.

The dataset features an extra recording session. For this session, office furnitures were positioned in the room to simulate a typical meeting room with chairs and tables (See Figure 7.1). These recordings will be used in future works for asserting the robustness of echo-aware methods in a more realistic scenario?.

For each room configuration and loudspeaker, three different excitation signals were played and recorded in sequence: chirps, white noise and speech utterances. The former consists in a repetition of 3 Exponential Sine Sweep (ESS) signals of duration 10 seconds and frequency range from 100 Hz to 14 kHz interspersed with 2 seconds of silence. Such frequency range was chosen to match the characteristics of the loudspeakers. To prevent rapid phase changes

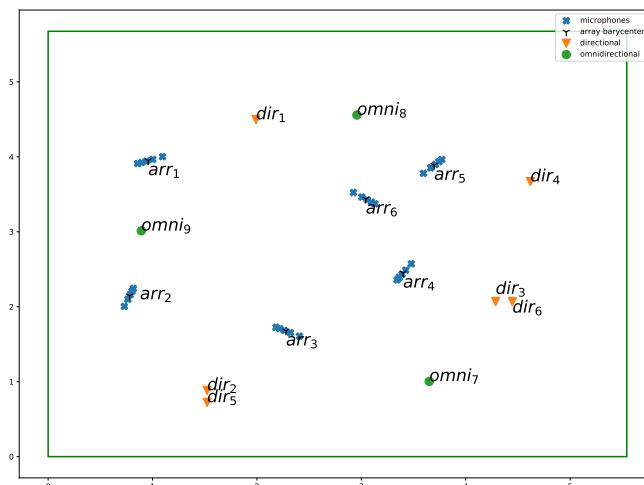


FIGURE 7.2: Illustration of the recording setup - top view.

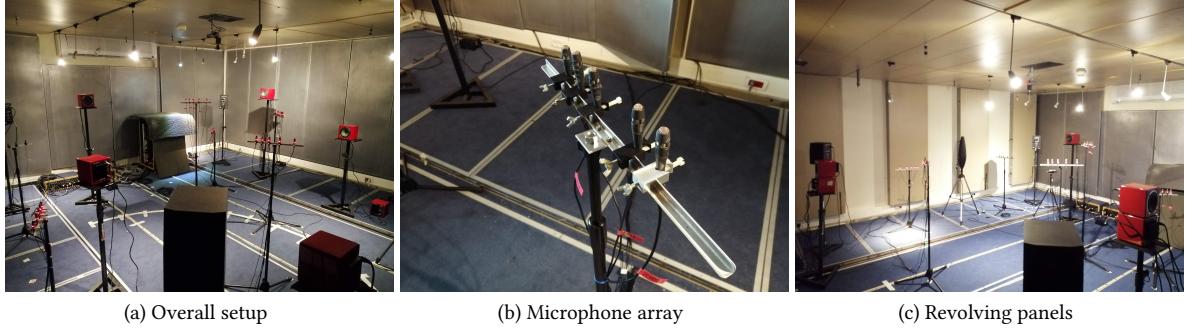


FIGURE 7.3: Picture of the acoustic lab. From left to right: the overall setup, one microphone array, the setup with revolved panels.

Surfaces:	Floor	Ceiling	West	South	East	North
000000	X	X	X	X	X	X
010000	X	✓	X	X	X	X
001000	X	X	✓	X	X	X
...				...		
000001	X	X	X	X	X	✓
011000	X	✓	✓	X	X	X
011100	X	✓	✓	✓	X	X
...				...		
011111	X	✓	✓	✓	✓	✓

TABLE 7.2: Surface coding in the dataset: each binary digit indicates if the surface is absorbtive (0, X) or reflective (1, ✓).

and “popping” effects, the signals were linearly faded in and out over 0.2 seconds with a Tuckey taper window⁴³. Secondly, 10 seconds bursts of white noise and 3 anechoic speech utterances from the Wall Street Journal (**WSJ**) dataset [Paul and Baker 1992] were reproduced in the room. Through all the recordings, at least 40 dB of sound dynamic range was asserted and a room temperature of $24^\circ \pm 0.5^\circ$ and humidity of 80% were registered. Moreover, 1 minute of *room tone* (silence) and 4 minutes of diffuse babble noise were recorded for each session. The latter was simulated by transmitting different chunks of the same single-channel babble noise recording from additional loudspeakers facing the four corners of the room.

⁴³ The code to generate the reference signals and to process them is available together with the data. Such code is based on the `pyrirtool` Python library.

All the microphone signals were synchronously acquired and digitally converted to 48 kHz with 32 bits/sample using the equipment listed in Table 7.1. The polarity of each microphone was registered by clapping a book in the middle of the room.

7.3 DATASET ANNOTATION

RIRs are estimated with the ESS technique [Farina 2007]: the signal of a microphone recording an **ESS** source is deconvolved by division in the frequency domain. Notice that the **FT** of the **ESS** signal is available in closed form and we used its **DFT** approximation.

7.3.1 RIRs annotation

The objective of this database is to feature annotations in the “geometrical space”, namely microphone, wall and source positions, *fully consistent* with

annotations in the “signal space”, namely the echo timings within the RIRs. This results is achieved as follows:

- (i) First, the ground truth positions of array and source centres are acquired via a Beacon indoor positioning system (bIPS). This system consists in 4 stationary bases positioned at the corners of the ceiling and a movable probe used for measurements which can be located within errors of ± 2 cm. The elements of this system are shown in Figure 7.4.
- (ii) The estimated RIRs are superimposed on synthetic RIRs computed with the Image Source Method (ISM) from the geometry obtained in the previous step. A Python GUI⁴⁴ (showed in Figure 7.5), was used to manually tune a peak finder and label the echoes corresponding to found peaks, that is, annotate their positions and their corresponding image source position and wall label..
- (iii) By solving a simple Multi-Dimensional Scaling (MDS) problem [Dokmanić et al. 2015; Crocco and Del Bue 2016b; Plinge et al. 2016], refined microphone and source positions were computed. The non-convexity of the problem was alleviated by using a good initialization (obtained at the previous step), by the high SNR of the measurements and, later, by including the additional image sources in the formulation. The prior information about the arrays’ structures reduced the number of variables of the problem, leaving the 3D positions of the sources and of the arrays’ barycenters in addition to the the arrays’ tilt on the azimuthal plane.
- (iv) By employing a multilateration algorithm [Beck et al. 2008], where the positions of one microphone per array serve as anchors and the TOAs are converted into distances, it was possible to localize image sources along side with the real sources. This step will be further discussed in ??.

Knowing the geometry of the recording room, we were able to manually label the echoes by iterating through steps (ii), (iii) and (iv).

- ▶ THE FINAL GEOMETRICAL AND SIGNAL ANNOTATION was chosen as a compromise between the bIPS measurements and the MDS output. While the former ones are noisy but consistent with the scene’s geometry, the latter ones matches the TOAs but not necessarily the physical world. In particular, geometrical ambiguities such as global rotation, translation and up-down flips were observed. Instead of manually correcting this error, we modified the original problem from using only the direct path distances (dMDS) to considering the image sources’ TOA of the ceiling as well in the cost function (dcMDS). Table 7.3 shows numerically the *mismatch* (in cm) between the geometric space (defined by the bIPS measurements) and the signal space (the one defined by the echo timings, converted in cm). To better quantify it, we introduce here a *goodness of match* (GoM) metric: it measures the fraction of (first-order) echo timings annotated in the RIRs matching the annotation produced by the geometry within a threshold. Including the ceiling information, dcMDS produces a geometrical configuration which has a small mismatch (0.41 cm in average) in both the signal *and* geometric spaces with 98.1% of matching first



FIGURE 7.4: Picture of the Beacon indoor positioning system used for measuring array and loudspeaker 3D position.

⁴⁴This GUI is available in the dataset package.

	Metrics	bIPS	dMDS	dcMDS
Geom.	Max.	0	6.1	1.07
	Avg. \pm Std.	0	1.8 ± 1.4	0.39 ± 0.2
Signal	Max.	5.86	1.20	1.86
	Avg. \pm Std.	1.85 ± 1.5	0.16 ± 0.2	0.41 ± 0.3
Mismatch	GoM (1.0 ms)	97.9%	93.4%	98.1%
	GoM (0.1 ms)	26.6%	44.8%	53.1%
	GoM (0.05 ms)	12.5%	14.4%	30.2%

order echoes within within a 1 ms threshold. Nevertheless, it is interesting to see that the bIPS measurements produce a good but less precise annotation.

7.3.2 Other tools for RIRs annotation

Finally, we want to mention that the following tools and techniques were found helpful in annotating the echoes.

- THE SKYLINE VISUALIZATION consists in presenting multiple RIRs as an image, such that the wavefronts corresponding to echoes can be highlighted [Baba et al. 2018]. More precisely, it is the visualization of the $L \times N$ matrix \mathbf{H} created by stacking column-wise N normalized echograms⁴⁵, that is $\mathbf{H}_{l,n} = \bar{\eta}_n(l) = |h_n(l)| / \max |h_n(l)|$, where $l = 0, \dots, L - 1$ is the sample index and n is an arbitrary indexing of all the microphones for a fixed room configuration. 4 RIR SKYLINES for 4 directional sources for the full reflective scenario are shown in Figure 7.6, stacked horizontally, preserving the order of microphones within the arrays. The reader can notice several clusters of 5 adjacent bins of similar color (intensity) corresponding to the arrivals at the array’s sensors. Thanks to the usage of linear arrays, this visualization allowed us to identify both TOAs and their labeling.

TABLE 7.3: Mismatch between geometric measurements and signal measurements in terms of maximum (Max.), average (Avg.) and standard deviation (Std) of absolute mismatch in centimeters. The *goodness of match* (GoM) between the signal and geometrical measurements is reported as fraction of matching echo timing for different threshold in milliseconds.

⁴⁵ The echogram is defined either as the absolute value or as the squared value of the RIR.

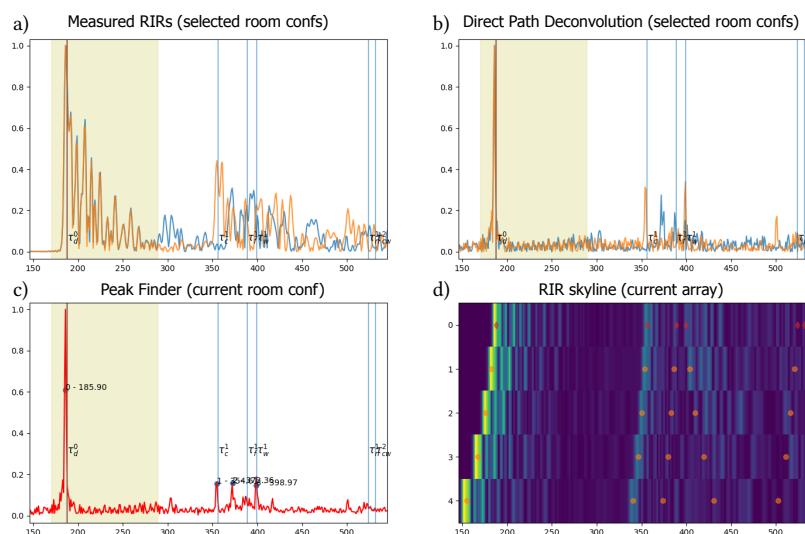


FIGURE 7.5: Detail of the GUI used to manually annotate the RIRs. For a given source and microphone, a) and b) shows 2 RIRs for 2 different room walls configuration (blue and orange) before and after the direct path deconvolution respectively. c) shows the results of the peak finder for one of the deconvolved RIRs, and d) is a zoom on the RIR skyline (See Figure 7.6).

- DIRECT PATH DECONVOLUTION was used to compensate the frequency response of the source loudspeaker and microphone [Antonacci et al. 2012; Eaton et al. 2016]. In particular, the direct path of the RIR was manually isolated and used

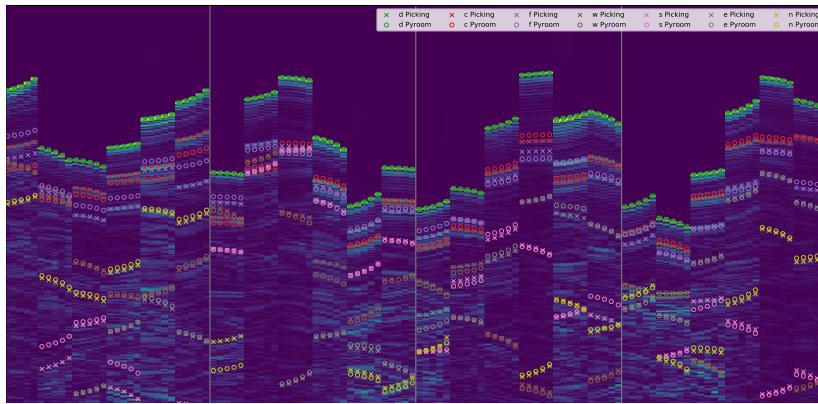


FIGURE 7.6: **RIR** Skyline annotated with observed peaks (\times) together with their geometrically-expected position (\circ) computed with the Pyroomacoustic simulator. As specified in the legend, different colors are used to indicate the room facets responsible for the reflection: direct path (d), ceiling (c), floor (f), west wall (w), . . . , north wall (n).

as an equalization filter to enhance early reflections from their superimposition and from background noise before proceed with peak picking. Each RIR was equalized with its relative direct path. As depicted in Figure 7.5, in some situation this process was necessary for correctly identifying the underlying **TOAs**' peaks.

- ▶ DIFFERENT WALL CONFIGURATIONS for the same geometry influenced the peaks' predominance in the **RIR**, hence facilitating its echo annotation. An example of **RIRs** corresponding to 2 different surface configurations is shown in Figure 7.5: the reader can notice how the peak predominance changes for the different configurations.
- ▶ AN INTERPOLATION-BASED PEAK FINDER⁴⁶ was used the equalized echograms $\bar{\eta}_n(l)$ to provide an initial guess on the peak positions, which was manually correct afterwards.

⁴⁶ In this work, peaks are found using the Python `peakutils` library.

7.3.3 Limitations of current annotation

As stated in [Defrance et al. 2008b], we want to emphasize that annotating the correct **TOAs** of echoes and even the direct path in “clean” real **RIRs** is far from straightforward. The peaks can be blurred out by the loudspeaker characteristics or the concurrency of multiple reflections. However as showed in Figure 7.6, the proposed annotation was found to be sufficiently consistent both in the geometric and in the echo space. Thus, no further refinement was done. This database can be used as a first basis to develop better **AER** methods which could be used to iteratively improve the annotation, for instance including 2nd order reflections.

7.4 THE dEchorate PACKAGE

The dataset comes with both data and code to parse and process it. The data are presented in 2 modalities: the raw data, that is, the collection of recorded wave files, are organized in folders and can be retrieved by querying a simple database table; the processed data, which comprise the estimated **RIRs** and the geometrical and signal annotations, are organized in tensors directly importable in Matlab or Python (e.g. all the **RIRs** are stored in a tensor of dimension $L \times I \times J \times D$, respectively corresponding to the RIR length in

scr	mic	signal	floor	...	filename
		chirp	False	...	2020-01-22_22-50-36.wav
		speech	False	...	2020-01-22_22-59-36.wav
:	:	:	:	:	:

FIGURE 7.7: Sample view of the database table to retrieve the raw wave file and its attributes.

samples, the number of microphones, of sources and of room configurations). Together with the data a Python package is available on the same website. This includes wrappers, GUI, examples as well as the code to reproduce this paper. In particular, all the scripts used for estimating the RIRs and annotating them are available and can be used to further improve and enrich the annotation or as baselines for future works.

- * THIS WORK introduced a new database of Room Impulse Response (RIR) featuring accurate annotation of early echoes and microphone positions. These data can be used to test methods in the room geometry estimation pipeline and in echo-aware audio signal processing. We will show some application in **SE** and **RooGE** in [Chapter 11](#)

Part III

ECHO-AWARE APPLICATION

8	AUDIO SCENE ANALYSIS MEETS SIGNAL PROCESSING	
8.1	Audio Scene Analysis Problems	97
8.1.1	Common scenario and model	97
8.1.2	Problem formulation	98
8.2	Overview on Multichannel Sound Source Separation	99
8.2.1	Multichannel NMF source separation	101
8.3	Overview on Spatial filtering	102
8.3.1	Beamforming	102
8.4	Overview on Sound Source Localization	104
8.4.1	Knowledge-based approaches	105
8.4.2	Data-driven approaches	105
8.5	Conclusion	106
9	separake: ECHO-AWARE SOUND SOURCE SEPARATION	
9.1	Literature review in Echo-aware Audio Source Separation . .	107
9.2	Modeling	110
9.3	Source Separation by NMF	111
9.3.1	NMF using Multiplicative Updates (MU-NMF)	111
9.3.2	NMF using Expectation Maximization (EM-NMF) . . .	112
9.3.3	Supervised NMF with pre-trained Dictionaries	113
9.4	Echo-aware Source Separation	113
9.5	Numerical Experiments	113
9.5.1	Setup	114
9.5.2	Dictionary Training, Test Set	114
9.5.3	Implementation:	115
9.5.4	Results	116
9.6	Conclusion	117
10	mirage: ECHO-AWARE SOUND SOURCE LOCALIZATION	
10.1	Literature review in Echo-aware Sound Source Localization . .	119
10.2	Proposed Approach	119
10.3	Background in microphone array SSL	120
10.3.1	2-channel 1D-SSL	121
10.3.2	Multichannel 2D-SSL	122
10.4	Microphone Array Augmentation with Echoes	123
10.5	Experimental Results	124
10.5.1	2-channel scenario	124
10.5.2	Multi-channel synthetic-data scenario	126
10.5.3	Multi-channel real scenario	128
10.6	Conclusion	130

11 ECHO-AWARE APPLICATIONS OF dechorate

11.1 Echo-aware Spatial Filtering	132
11.1.1 Literature review	132
11.1.2 Background in spatial filtering	134
11.1.3 Elements of Beamforming	135
11.1.4 Noise, steering vectors, rake filters, and relative transfer functions	136
11.1.5 Considered beamformers	137
11.1.6 Experimental evaluation	138
11.2 Room Geometry Estimation	140
11.2.1 Room Geometry Estimation through multilateration .	141
11.2.2 Using the dEchorate dataset for RooGE	141
11.3 Conclusions and Perspectives	142

8

Audio Scene Analysis meets Signal Processing

- ▶ **SYNOPSIS** In this chapter, we will present a selection of algorithms and methodologies which we identified as potential beneficiaries of echo-aware additions. At first, in section § 8.1, we present a typical scenario that highlights some cardinal problems. Then, in section ??, state-of-the-art approaches to address these problems are listed and commented, highlighting the relationship with some acoustic propagation models. The content presented here serves as a basis for a deeper investigation conducted in each of the following chapters.

Here, we present some audio scene analysis problems that will be later discussed in their echo-aware extension. Following the last part's structure, this introductory chapter gathers the common knowledge shared across the following ones. Here we make a strong transition: we assume the echo properties are known *a priori*, so that the our focus is on the benefit of their knowledge. The literature for each of them is reviewed, but since it is vast and spans diverse scientific research decades, we do not aim to cover it entirely. Moreover, since the following chapters are dedicated to each of these problems under the echo-aware perspective, this specific literature is not considered here.

The material presented here results from the personal synthesis of concepts and references available in the literature. Furthermore, some definitions are digested from classical textbooks already used for this thesis, such as [Vincent et al. 2018].

8.1 AUDIO SCENE ANALYSIS PROBLEMS

As mentioned in the first chapter, audio scene analysis aims to extract relevant information in the audio scene. Different types of information are estimated or inferred by solving specific problems. Despite their diversity, most of these problems can be defined with a common model.

8.1.1 *Common scenario and model*

Let there be a meeting room with well-defined geometry. In it, J sound sources are located at determined positions, such as some speakers chatting while standing in the room. As an indoor scenario, all the elements of reverberation (in particular echoes) are present. Diffuse background noise is present as

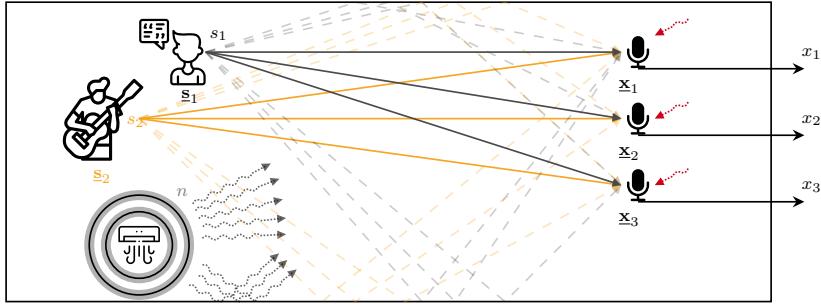


FIGURE 8.1: Illustration of an indoor audio scene recorded by two microphones.

well, for instance, due to the air conditioner or car traffic outside. This whole audio scene is recorded by a device featuring a microphone array of I sensors. Furthermore we assume a static far field scenario and we model each j sources and i microphone as well-defined points with coordinate \underline{s} and \underline{x} , respectively. This is a reasonable assumption in the context of table-top devices, such as smart home devices. Recalling the (discrete) time-domain signal model already discussed in § 3.1.1, the signal recorded at the i -th microphones reads

$$x_i[n] = \sum_{j=1}^J (h_{ij}(\underline{x}_i | \underline{s}_j) * s_j)[n] + n_i[n], \quad (8.1)$$

or alternatively, using the source spatial image signals,

$$\begin{aligned} x_i[n] &= \sum_{j=1}^J c_{ij}[n] + n_i[n], \\ c_{ij}[n] &= (h_{ij}(\underline{x}_i | \underline{s}_j) * s_j)[n]. \end{aligned} \quad (8.2)$$

Note that the filter $h_{ij}(\underline{x}_i | \underline{s}_j)$ denotes the **RIR** where we intentionally highlight the dependencies on geometry, namely, accounting for the whole sound propagation for the source position \underline{s}_j to the microphone position \underline{x}_i . In fact, as discussed throughout Part I, we can decouple the information of indoor microphone natural recordings into two orthogonal contributions: the **RIRs** (thus the mixing matrix) accounting for only the sound propagation, and the source signals describes only the semantic content.

8.1.2 Problem formulation

The Audio Scene Analysis Problems presented already in the introductory chapter (See § 1.2) can now be extended and rewritten in terms of the above notation. Furthermore, we will consider here the only ones directly addressed in this thesis: room impulse response estimation, audio source separation, Spatial filtering, sound source localization, and room geometry estimation.

As introduced in § 4.1, these problems can be said either *informed* or *blind* and the related scenario *active* or *passive*. These two dichotomies emphasize the amount of prior knowledge available for solving them. As opposed to the active scenario, where the source signal is known, transmitted, and available, the passive one considers only the microphone measurements. For instance, when addressing the active echo estimation problem or **RIR** measurement, the exact time of emission of the source signal is known, as well as the source

Audio scene analysis problems	<i>from the mixtures $\{x_i\}_i$, can we estimate...</i>	Chapter
Audio Source Separation	the source signals $\{s_j\}_j$ and the filters $\{h_{ij}(\underline{x}_i \underline{s}_j)\}_{ij}$?	Chapter 9
Spatial filtering	the source signals $\{s_j\}_j$, knowing the filters $\{h_{ij}(\underline{x}_i \underline{s}_j)\}_{ij}$?	?? § 11.1
Sound Source Localization	the source positions $\{\underline{s}_j\}_j$?	Chapter 10
Room Geometry Estimation	the shape of the room?	?? § 11.2

TABLE 8.1: List of audio scene analysis problems considered in this thesis accompanied by their mathematical description.

signal itself.

The second dichotomy refers to the possibility of exploiting prior knowledge to solve the problem more easily. This information may derive from annotations and meta-data. In the community of audio source separation, the following definitions were proposed in [Vincent et al. 2014]: as opposed to informed problems, for solving the blind ones, absolutely no information is given about the source signal or the mixing process. In between, there are *semi-blind* or *guided* problems: here general information is available, such as on the nature of the source signal (speech, music, environmental sounds), microphone position, recording scenario (indoor, outdoor, professional music), mixing process, etc. In some books and works other categories of problems are defined, such as *weakly-guided*, *strongly-guided*. Here we do not consider these distinctions.

In the considered echo-aware applications, the echoes properties build our prior knowledge on the problem. Therefore, according to the above taxonomy, the addressed problems are necessarily guided. In general and unless specified, this is the only knowledge we will assume to have. Based on this, we will now review some classical works for solving the above problems.

- ▶ IN THE FOLLOWING SECTIONS we will present the general overview of the literature related to the problems considered in this thesis: multichannel audio source separation, Spatial filtering, and sound source localization. We will limit the discussion to the most relevant techniques adopted nowadays with respect to the acoustic propagation modeling. Later in the thesis, dedicated sections on echo-aware method to address these problems will be provided in each related chapter. Since room geometry estimation is mainly based on echo estimation and labeling discussed in § 4.3.1, its description is postponed to § 11.2.

8.2 OVERVIEW ON MULTICHANNEL SOUND SOURCE SEPARATION

Multichannel audio source separation refers to the process of extracting acoustic signals from multichannel mixtures featuring targets, interfering, and noisy sounds. In psychoacoustics, this problem is known as *the cocktail party problem* [Cherry 1953], referring to the human ability to focus on a particular stimulus in a audio scene. This problem has interested mainly in two research fields in the audio signal processing community: speech and music process-

ing. Both share many methods, which are accordingly modified, taking into account scenarios and applications. In the context of the multichannel speech recordings, some of the most successful and popular methods used nowadays include Spatial filtering, **TF** masking, and end-to-end regression. In this thesis, we deliberately distinguish between the Spatial filtering, which will be discussed in the following subsection, and **TF** masking.

TF masking relies on **TF** diversity of the sources and processes each mixture channel separately. In a nutshell, it involves computing the **STFTs** of the mixture channels, multiplying them by masks containing gains between 0 and 1. One of the most popular masking rules is adaptive Wiener filtering. For each time-frequency bin, the **STFTs** of the estimated source spatial images C_{ij} of the j -th source at the i microphone, writes

$$\hat{C}_{ij} = \underbrace{\frac{|C_{ij}|^2}{\sum_{j=0}^J |C_{ij}|^2}}_{\text{Wiener filter}} X_i \quad (8.3)$$

where X_i is the **STFT** of the microphone channel. Here the **TF** bin is omitted for clarity.

In order to be computed, the Wiener filter requires estimating all the source spatial images, or equivalently, the mixing filters and the source signals. Therefore, this approach has been generalized in several ways to account for both these unknowns. In these thesis we stress the difference between source separation and Spatial filtering. In the former, all the source signals and mixing filters are indispensable to weigh each of the **TF** bins of the microphone channel's **STFT**. As opposed to, in the latter problem, the "masks", that is the spatial filters, are estimate only based the mixing filters and noise statistics. In multichannel recordings, a clear overlap exist between the two problems, and some techniques can be used reciprocally. Furthermore the related research trends are now converging under the same umbrella of the so-called *speech enhancement*. The work [Gannot et al. 2017] provide an unified framework merging source separation and Spatial filtering. Nevertheless, we will treat them as separated problems.

Moreover, the benefit of the **TF** masking approach is that the masks can be estimated in various ways. For instance, clustering and classification techniques [Rickard 2007] can be used to assign each **TF**-bin to each of the sources. Recently learning-based methods have been used in this sense on the same task [Hershey et al. 2016; Wang et al. 2018]. Alternatively, deep learning techniques can used to directly estimated the sources' **TF**, as done in one of the reference implementation [Stöter et al. 2019]. The work of [Nugraha et al. 2016], instead, uses a deep learning model build by unfolding the the popular Multichannel **NMF** source separation framework of [Ozerov and Févotte 2010; Sawada et al. 2013].

The Gaussian-based Multichannel **NMF** [Ozerov and Févotte 2010] is one of the most successful framework for source separation using **TF** masks. It combines the **NMF** and narrowband spatial model (discussed in § 3.2.5) and deploys optimization-based framework for estimating both the mixing matrix and the sources. One of the main advantage of this approach is that it allows to easily

Many other methods have been proposed in the literature. The reader can refer to [Vincent et al. 2018; Makino 2018]

incorporated prior knowledge on the problem. Thanks to the narrowband approximation (Eq. (3.26)), spatial and semantic content are orthogonal and can be treated disjointedly. This opens to many possibilities, such as, using pre-trained dictionaries to model source content [Schmidt and Olsson 2006; Smaragdis et al. 2009], or using proper model for the mixing filters, such as RIRs, steering vectors, ReTF, etc.

However, it has been shown that even with oracle TF [Luo and Mesgarani 2019], the estimation is still affected by artifacts. This limitation affects all the approaches operating in the TF domain. To overcome this, end-to-deep deep learning models [Luo and Mesgarani 2019; Tzinis et al. 2020] were developed and now hold the record in source separation. These models work directly in the time domain: both input and output are time-domain waveforms. This approach has proved to reach good separation quality, especially in terms of perceived sounds at the listener. Nevertheless, all deep learning methods rely on trained black-box models for which it is hard to inject prior knowledge. Instead, Multichannel NMF-based frameworks account for this freedom.

8.2.1 Multichannel NMF source separation

Multichannel NMF source separation methods can be grouped according to how they model sound propagation of the mixing process:

- those that simply ignore it [Le Roux et al. 2015];
- (*free field propagation*) those that assume a single anechoic path [Rickard 2007; Nesta and Omologo 2012];
- (*reverberant propagation*) those that model the RTFs entirely [Ozerov and Févotte 2010; Duong et al. 2010; Li et al. 2019];
- (*reverberant propagation*) and those that attempt to separately estimate the contribution of the early echoes and the contribution of the late tail [Leglave et al. 2015].

Therefore, these existing approaches either ignore sound propagation or aim at estimating it fully, which affect the quality of the separation. In the first case, strong echoes and reverberant constitute a low bound in the separation capability. In fact, these elements of the sound propagation blur and spread the energy of the source over multiple TF bins, for which the assignment is harder. When computing the TF masking operation, these bins may introduce strong artifacts. In the second case, the algorithm needs to estimate more parameters with consequences in complexity and estimation accuracy.

- ▶ ECHO-AWARE SOURCE SEPARATION METHODS have been introduced as a possible solution to overcome some of these limitations. More details will be given in ??, where a new method for speech source separation based on the Multichannel NMF framework and echoes is described.

8.3 OVERVIEW ON SPATIAL FILTERING

Spatial filtering aim at the enhancement of a desired signal while suppressing the background noise and/or interfering signals. It is a large and active research field that has interested the signal processing and telecommunication communities since several decades. It has produced an vast literature including several reference books dedicated to the topic. For more details in this direction, the reader can refers to, e.g., the book [Van Trees 2004]. In audio, this topic was been extensively review in the context of speech enhancement in a recent publication [Gannot et al. 2017]⁴⁷

For a comprehensive review on Spatial filtering methods, the reader can refer to the book [Van Trees 2004].

In Spatial filtering, the RIRs (and related models, e. g., RTFs, steering vectors or ReTF) play a central role. Intuitively, giving the mixing model in Eq. (8.1), the enhancement of a target source can be achieved by merely denoising the recordings and filtering by the inverting RIRs. However, this is not always possible as the inversion of the RIRs is not a straightforward operation. The work in [Neely and Allen 1979] discuss explicitly the issues of inverting RIRs. Later, several techniques were investigated, which are sometimes referred to as Room Response *Equalization* [Cecchi et al. 2018].

⁴⁷ The content of this work has been extended in the book [Vincent et al. 2018].

8.3.1 Beamforming

Beamforming is one of the most famous techniques used in Spatial filtering. One of the most famous beamformer is the Delay-and-Sum (DS). The intuitive idea behind it is to sum the microphone channels constructively by compensating the time delays between the sound source and the spatially separated microphones. Thus, the target source signal is enhanced, while noise, interferences, and reverberation being suppressed. Figure 8.2 illustrate this ideas. Later, this idea has been extended to Frequency and Time-Frequency processing with direct modeling of the noise sources. More formally, beamformers design mathematical *optimization criterion*, namely objective function, defining the desired shape of the estimated signal and return a filter to be applied to the microphone recordings. For instance, one may want to keep a unit gain towards the desired sound source's direction while minimizing the sounds from all the other directions. The literature on beamformers spans in two directions: different optimization criteria and how to estimate the parameters required by their computation. We will now elaborate the two directions in turn.

- ▶ MANY BEAMFORMERS CRITERIA have been proposed. Among all, some of the most famous are the DS, the Minimum-Variance-Distortionless-Response (MVDR) [Capon 1969], the Maximum SNR (MaxSNR) [Cox et al. 1987], the Maximum SINR (MaxSINR) [Van Veen and Buckley 1988], and the Linearly-Constrained-Minimum-Variance (LCMV) [Frost 1972]. These criteria are designed to satisfy different constraints and model prior knowledge, as discussed in § 11.1.3. The reader can also refer to the above-suggested book for more details.
- ▶ PARAMETER ESTIMATION is a crucial step for beamformers. We can identify two

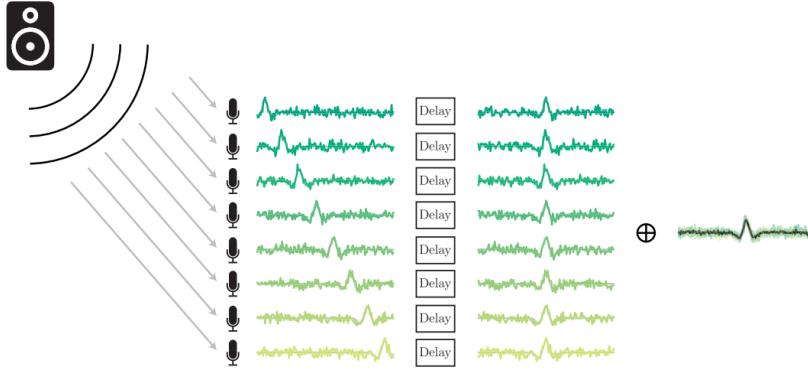


FIGURE 8.2: Illustration of the **DS** Beamforming in the time-domain applied to a pulse source signal. The delays of the recorded signals at each microphone are compensated. By summing the shifted signals, the source signal is enhanced. Image courtesy of Robin Scheibler, author of [Scheibler et al. 2015].

main categories of parameters: the one related to the **RIRs** and the one related to the source and noise statistics. In the former case fall all the methods that model the acoustic propagation of sound. Therefore, similarly to the methods for separation, we can group existing methods in the following groups:

- (*free and far field propagation*) methods based on relative steering vectors build on **DOA** [Takao et al. 1976; Applebaum and Chapman 1976; Cox et al. 1987; Van Veen and Buckley 1988];
- (*multipath propagation*) methods based on rake rake receiver [**Jan 1995 matched**; Flanagan et al. 1993; Dokmanić et al. 2015; Peled and Rafaely 2013; Scheibler et al. 2015; Kowalczyk 2019];
- (*reverberant propagation*) methods based on full acoustic channel estimation (see [Chapter 4](#));
- (*reverberant propagation*) methods based on **DOAs** and the statistical modeling of the diffuse sound field, [Thiergart and Habets 2013; Schwartz et al. 2014];
- (*reverberant propagation*) methods based on **ReTF** [Gannot et al. 2001; Doclo and Moonen 2002; Cohen 2004; Markovich et al. 2009];
- (*reverberant propagation*) methods based on (deep) learning [Li et al. 2016a; Xiao et al. 2016; Sainath et al. 2017; Ernst et al. 2018];

The **DOAs**-based methods exploit the closed-form mapping between **DOAs** and the steering vectors in far-field scenarios. Thus, good performances are possible only upon a reliable estimation of the **DOAs** (see next section), a challenging problem in noisy and reverberant environments. The steering vectors' computation depends on the array geometry, which is unknown in some practical cases. Alternatively, one can estimate the full acoustic channels, which is a cumbersome task by itself.

The **ReTF**-based approaches have been introduced to overcome these two limitations. They automatically encode the **RIRs**, the geometrical information, and are “easier” to estimate than the **RIRs**. The main limitation of these methods is that they return *spatial source image* at the reference microphone, rather than the “dry” source signal. Therefore, when reverberation is detrimentally affecting the speech signal’s intelligibility, post-processing is necessary [Schwartz et al. 2016].

Recently, DNN have been proposed for solving this task, either to estimate the beamformer filter [li2016neural directly; Xiao et al. 2016; Sainath et al. 2017] or in an end-to-end task [Ernst et al. 2018]. Moreover, DNN has been used to estimate some of parameters, such as the DOAs [Salvati et al. 2018; Chazan et al. 2019], ReTF estimation [Chazan et al. 2018].

- ▶ EARLY ECHOES are neither considered nor modeled as noise terms in the literature discussed thus far. This direction is taken by the echo-aware methods accounting specifically for the multipath propagation. We will discuss these methods in more detail in chapter Chapter 11 together with their implementation.

8.4 OVERVIEW ON SOUND SOURCE LOCALIZATION

Sound Source Localization (SSL) consists in determining the distant and direction (or position) of sound sources from microphone (array) in the 3D space, typically in a passive scenario. As discussed above, the information on the sources' and microphones' position in the room is encoded in the RIRs. Therefore, assuming the uniqueness of the mapping between locations to a RIR, it is theoretically possible to retrieve the absolute position of microphones and sources, as show in [Ribeiro et al. 2010a; Crocco and Del Bue 2016a]. However, this is yet a very challenging task, which typically involves the solution of several sub-problems. Therefore, it is more common to relax the SSL problem as follows. First, rather than operating in the 3D cartesian coordinate system, most of the existing methods aim at estimating 2D-DOA, namely the angles for on the unit sphere with the center in a reference point. This reference point is usually the barycenter of the microphone array. These angles are called *azimuth* and *elevation* as shown in Figure 8.3. When only a single microphone pair is considered, the problem is simplified to 1D-SSL, estimation the 1D-DOA, named *angle of arrival*, with respect the microphone axis. Second, they assume far-field scenarios. The main reasons for adopting such simplifications are the followings:

- estimating the distance is known to be a much more challenging task than estimating the DOAs [Vesa 2009];
- the far-field scenario is a reasonable assumption when using a compact array recording distant sounds;
- the problem can be independent to room geometry, whose estimation is an ambitious task;

Finally, in far-field settings, sometimes DOAs are sufficient to achieve reasonable speech enhancement performance as demonstrated by the literature in spatial filtering.

Despite these approximations, the SSL problem still challenges today's computational methods. Popular approaches consists in two main components: *feature extraction* and *mapping*. First, the audio data are represented as features, as independent as possible from the source's content while preserving spatial

The reader can find more details in the SSL in the recent review articles [Rascon and Meza 2017; Argentieri et al. 2015] as well as in [Vincent et al. 2018, Chapter 4].

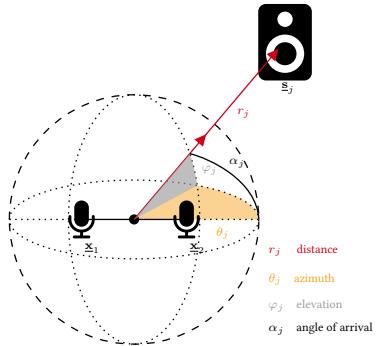


FIGURE 8.3: Geometrical illustration of the Sound Source Localization, showing the azimuth θ_j , the elevation φ_j and angle of arrival α_j for the source j at s_j and two microphones at x_1 , x_2 respectively.

As an example, consider the TDOA-based SSL methods: the feature extraction step consist into extract time delays from audio recordings, then these quantities are mapped to direction of arrivals, thanks to the acoustic propagation model and the array geometry.

information. Second, the features are mapped to the source position. Two lines of research have been investigated to obtain such mappings: knowledge-driven and data-driven approaches.

8.4.1 Knowledge-based approaches

Knowledge-based approaches rely on a physical model for sound propagation. These models rely on closed-form mapping from the sound's direct path Time Differences of Arrival at the microphone pair and the source's azimuth angle in this pair. If multiple microphone pairs are available and form a non-linear array, their TDOAs can be aggregated to obtain 2D-DOAs [DiBiase et al. 2001]. Furthermore, what differentiates the approaches in this categories lies in their ability to localize either single sources or multiple ones, their robustness to noise and reverberation, and the particular methods they used. We can identify the following approaches based on:

- subspace methods, such as **MUSIC** [Dmochowski et al. 2007];
- **TDOA**-based techniques, which uses Generalized Cross Correlation (**GCC**) functions [**Blandin2012**; Knapp and Carter 1976; Lebarbenchon et al. 2018b] to estimate **TDOA** and then compute the most reliable **DOA** from them; These methods are related to beamforming-based techniques, such as SRP-PHAT**dibiase2001robust!** (**dibiase2001robust!**), which search the direction that maximizes the power of the output of a beamformer.
- methods based of **RIRs** estimation and blind system identification [Chen et al. 2006],
- methods based on probabilistic framework solved with Maximum Likelihood optimization [Stoica and Sharman 1990; Li et al. 2016b].

The main limitations of these approaches arise from the approximation considered in the models. In particular, common to all of them is the assumption that sound propagation being typically free-field. Thus, they intensely suffer in environments it is violated, e. g., in the presence of strong acoustic echoes and reverberation as discussed in [Chen et al. 2006].

8.4.2 Data-driven approaches

Data-driven approaches have been proposed to overcome the challenging task of modeling sound propagation. This is done using a supervised-learning framework, that is, using annotated training dataset to implicitly learn the mapping from audio features to source positions Such data can be obtained from annotated real recordings [Deleforge et al. 2015; Nguyen et al. 2018] or using acoustic simulators [Laufer et al. 2013; Vesperini et al. 2018; Adavanne et al. 2018; Chakrabarty and Habets 2017; Perotin et al. 2018; Gaultier et al. 2017].

In comparison to knowledge-driven methods, these methods have the advantage that they can be adapted to different acoustic conditions by including challenging scenarios in the training dataset. Therefore, these methods were showed to overcome some limitations of the free-field model. Under this

perspective, in the data-driven literature two main trends can envisioned: end-to-end learning models and two-step models. In the former case, all the **SSL** pipeline is encapsulated into a single robust learning framework, taking as input the microphone recordings and returning the source(s) **DOAs**. Examples of these approaches are the works in [Chakrabarty and Habets 2017; Perotin et al. 2018; Adavanne et al. 2018], where the task is performed with **DNNs** models. In the latter, learning models are used as a substitute for either feature extraction or the mapping. For instance, in [Laufer et al. 2013; Deleforge et al. 2015; Gaultier et al. 2017; Nguyen et al. 2018], **GMMs**-based models were used to learning the mapping from features derived from the **ReTF** of pair of microphones. In [Vesperini et al. 2018], the author proposes to use Neural Network (**NN**) models to estimate source location using features computed through **GCC-PHAT**.

Despite the considerable benefit of data-driven approaches in learning complex functions, their main limitation lies in the training data. First, these data are typically tuned for specific microphone arrays and fail whenever test conditions strongly mismatch training conditions. Moreover, due to the cumbersome task of collecting annotated datasets that cover as many possible scenarios as possible, synthetic data generated by simulators are used. However, these data may be too artificial or simplified, with the consequence that the models may not be able to generalize to real-world conditions.

"If you dont pay attention, your learing methods risk to "learn a model from model""
—Zybek Koldowsky

- ▶ REVERBERATION AND, IN PARTICULAR, ACOUSTIC ECHOES are typically treated as nuances in methods developed for **SSL** and **DOAs** estimation. Moreover, while knowledge-based approaches operate under strong assumptions which are typically violated in multipath scenario, for data-driven methods it is not trivial to inject prior information about sound propagation. Based on these limitations, our contribution propose to combines the best of the two worlds to effectively exploits the echo knowledge: More specifically, to deploy a **DNN** to estimate few echoes [Chapter 6](#) and use well-understood knowledge-based method to map them to source **DOAs** [Chapter 10](#).

8.5 CONCLUSION

This chapter presented some fundamental audio signal processing problems and an overview of related approaches to address them. These problems will be considered in their echo-aware settings in the following chapters, in particular:

- Multichannel NMF-based Audio Source Separation in [Chapter 9](#),
- SRP-PHAT-based Sound Source Localization in [Chapter 10](#),
- and in [Chapter 11](#), we will discuss some applications of the dEchorate dataset ([Chapter 7](#)) for some beamformers-based Spatial filtering methods ([§ 11.1](#)) and room geometry estimation ([§ 11.2](#)).

9

Separake: Echo-aware Sound Source Separation

- ▶ **SYNOPSIS** In this chapter, echoes are used for improving performance of classical Audio Source Separation methods. First, we describe existing methods that either ignore the acoustic propagation or attempt to estimate it fully. Instead, this work investigates whether sound separation can benefit from the knowledge of early acoustic echoes. These echoes are derived from the known locations of a few *image microphones*. The improvements are shown for two variants of a method based on non-negative matrix factorization: one that uses only magnitudes of the transfer functions and one that uses the phases as well. The experimental part shows that the proposed approach beats its vanilla variant by using only a few echoes and that with magnitude information only, echoes enable separation where it was previously impossible.

The material presented in this chapter results from a collaboration with Robin Scheibler and Ivan Dokmanić and was published in [Scheibler et al. 2018e]. This chapter recalls the main findings of the paper and brings additional insights on the literature and on the proposed model, which has been re-written using this thesis' notations. The personal contribution to this collaboration, done in the early months of the Ph.D., was adapting and implementing in Python of the proposed Expectation Maximization (**EM**)-**NMF** method accounting for echoes and using a pre-trained dictionary.

9.1 LITERATURE REVIEW IN ECHO-AWARE AUDIO SOURCE SEPARATION

The scientific literature on audio source separation is vast, still active, and spans decades. The problem covers a huge number of scenarios and use-cases, such as, number of microphones (single- vs. multi-channel recordings), the number of sources with respect to the number of channels (under- vs. over- vs. determined), the type of observed signals (speech vs. music), type of recordings (artificial vs. microphone recordings) etc., and, of course, combination of them. In this chapter we will consider only the case of multichannel convolutive recordings featuring reverberant speech data in overdetermined settings. Even selecting this narrow case, the literature remain vast and we will not review it in the context of this thesis.⁴⁸

*“Please could you stop the noise,
I’m trying to get some rest
From all the unborn chicken voices in my
head.”*

—Radiohead, *Paranoid Android*

Keywords: Blind Channel Identification, Super Resolution, Sparsity, Acoustic Impulse Response.

Resources:

- Paper
- Code
- Slides

Scheibler et al., “Separake: Source separation with a little help from echoes”

⁴⁸ The interested reader can refers to [Vincent et al. 2018; Makino 2018].

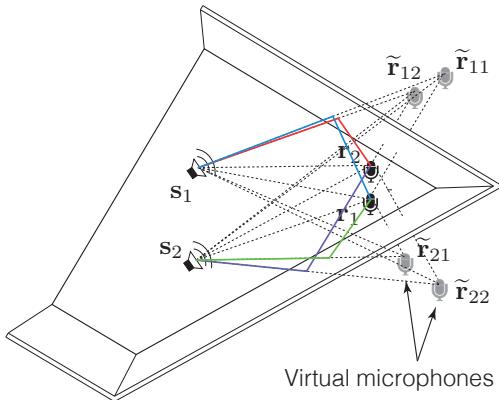


FIGURE 9.1: Typical setup with two speakers recorded by two microphones. The illustration shows the virtual microphone model (grey microphones) with direct sound path (dotted lines) and resulting first-order echoes (colored lines).
Image taken from our paper [Scheibler et al. 2018e].

Here, we will focus on multichannel source separation based on Nonnegative Matrix Factorization (**NMF**). **NMF** refers to a set of techniques to model spectra of complex sounds by a sum of basic components. Modeling sound structure is beneficial for source separation, since it makes separation possible in many challenging scenarios. Moreover, this approach allows to easily incorporate side information on both the sources and the acoustic propagation as will be shown later in the chapter.

Currently, in the literature, only few works can be found that incorporate the knowledge of echoes into sound source separation. In the work [Huang et al. 2005], the authors propose a decomposition of the source separation problem into different steps. First they estimate the **RIRs** by extending the **SIMO-BCE** framework from Multiple Input Multiple Output (**MIMO**) systems. Here the **RIRs** are modeled as **FIR** filters following the multipath echo model. Secondly, the estimated filters are used to build the demixing matrix, and then to separate the sources with an inverse-filtering approach. However, this method exhibits a high computational cost, which was addressed later in [Rotili et al. 2010]. Nevertheless these approaches were shown to lack robustness in low SNR conditions.

Alternatively, the work in [Asaei et al. 2014] proposes geometry-based approach embedded in a sparse optimization framework. First, by localizing the image sources and estimating the room geometry, the echoes' timings are estimated. Then, after computing the echoes' coefficients in a convex optimization framework, the individual speech signals are separated with either inverse-filtering or sparse recovery. The performance of this approach relies on the **RIR** and geometry estimation steps, which are very sensitive to the challenging acoustic condition, e. g. low SNR or high RT_{60} .

Instead, the work in [Leglaive et al. 2015] proposes to tackle the convolutive model by imposing a probabilistic prior on the early part of the **RIRs**, namely, modeled as an autoregressive process in the frequency domain. Later, the same authors extended this work in [Leglaive et al. 2016] accounting for both early and late part of the mixing filters.

- ▶ **THE PROPOSED APPROACH** is yet different from those presented above. First, rather than fitting the echo model as in [Leglaive et al. 2015; Leglaive et al. 2016], or estimating the mixing filters as in [Huang et al. 2005; Asaei et al. 2014], we aim to show that separation in the presence of known echoes is

better than separation without echoes. Second, we conduct this investigation in the context of source separation with non-negative source models with a deterministic model for the mixing filters. Third, we propose to solve the problem from the point of view of *image microphones*, already used in [1]. The image microphone model is equivalent to the Image Source Method (**ISM**) [Allen and Berkley 1979], where virtual receivers are placed outside of the room (See Figure 9.1). Even if the **ISM** is more common and implemented in practice in acoustic simulators, the two models are strictly equivalent. This approach is based on the acoustic *rake receivers*⁴⁹ previously proposed in [Dokmanić et al. 2015a] and is thus dubbed Separake.

⁴⁹ Rake receivers are a particular type of beamformers accounting for the effect of multipath propagation (see § 11.1)

The considered setup is illustrated in Figure 9.1. We assume that the array is placed close to a wall or a corner. This is useful for the following reasons: first, it makes echoes from the nearby walls significantly stronger than all other echoes; second, it ensures that the resulting image array (real and image microphones) is compact, allowing to assume the far field regime. This assumption is motivated by typical echo-aware signal processing applications, such as for smart-home devices, typically placed on a table or close to a wall.

- ▶ TRANSLATING ECHOES INTO IMAGE ARRAYS provides an interesting geometrical interpretation in light of beamforming theory. Real and virtual microphones form dipoles with diverse frequency-dependent directivity patterns. By integrating more and more virtual microphones, the directivity patterns change and higher spatial selectivity can be achieved [Dokmanić et al. 2015b]. This effect is shown in Figure 9.2. Therefore, the goal of this work is to design audio source separation algorithms which benefit from this known spatial diversity.

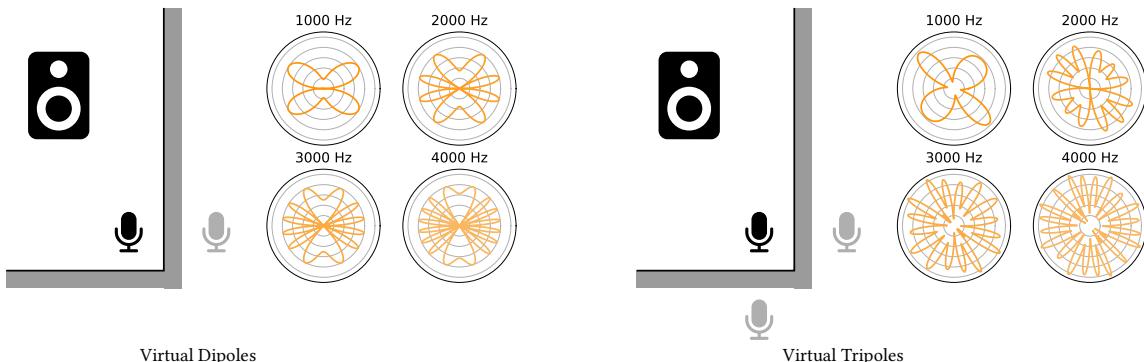


FIGURE 9.2: The frequency-dependent directivity pattern for an virtual array built with a real and one (left) and two (right) image microphones. Image taken from

9.2 MODELING

Recalling the echo model for the RIRs, and assuming R echoes per source are known, the approximate Room Transfer Function (RTF) from source j to microphone i writes

$$\tilde{H}_{ij}(f) = \sum_{r=0}^R \alpha_{ij}^{(r)} e^{-i2\pi f \tau_{ij}^{(r)}}. \quad (9.1)$$

Absolute TOAs relate to the source's distance which is not assumed to be known here. Instead, we will assume that only the TDOAs are known, by arbitrarily fixing the delay of the direct path to zero. This is equivalent to assume a far field scenario, thanks to which we can relate the TDOAs to the DOAs of the source signal with respect to the center of the virtual array. In addition, we assume all walls to be spectrally flat in the frequency range of interest and that $\alpha_{ij}^{(r)}$ are known up to a scaling (i.e. $\alpha_{ij}^{(0)} = 1$). In this work the echoes properties are assumed to be known.

Assuming the narrowband approximation, the mixing process can be modeled as in § 3.2.5. That is, the Short Time Fourier Transform (STFT) of the i -th microphone signal reads

$$X_i[k, l] = \sum_{j=1}^J H_{ij}[k] S_j[k, l] + N_i[k, l] \quad (9.2)$$

with $k \in [0, \dots, F]$ and $l \in [0, \dots, T]$ being the frequency and frame index, $H_{ij}[k]$ is the DFT approximating the RTF of (9.1), $X_j[k, l]$ the STFT of the j -th source signal, and $N_i[k, l]$ a term including noise and model mismatch. It is convenient to group the microphone observations in vector-matrix form,

$$\mathbf{X}[k, l] = \mathbf{H}[k] \mathbf{S}[k, l] + \mathbf{N}[k, l], \quad (9.3)$$

where $\mathbf{X}[k, l], \mathbf{N}[k, l] \in \mathbb{C}^{I \times 1}$, $\mathbf{S}[k, l] \in \mathbb{C}^{J \times 1}$ and $\mathbf{H}[k, l] \in \mathbb{C}^{I \times J}$.

Let the squared magnitude of the spectrogram of the j -th source be $\mathbf{P}_j = [|S_j|^2]_{kl} \in \mathbb{R}^{F \times T}$. As depicted in Figure 9.3, the spectrogram can be modeled as the product of 2 non-negative matrices:

$$\mathbf{P}_j = \mathbf{D}_j \mathbf{Z}_j, \quad (9.4)$$

where \mathbf{D}_j is the non-negative *dictionary* whose columns are called *atoms* and can be interpreted as spectral templates of the source, while the latent variables \mathbf{Z}_j , called *activations*, indicating when and how these templates are activated.

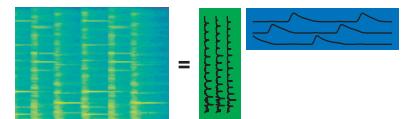


FIGURE 9.3: Spectrogram of a sound source signal decomposed into dictionary and activation. Image taken from the slides accompanying the paper of this work [Scheibler et al. 2018e].

- NMF-BASED AUDIO SOURCE SEPARATION can then be cast as an inference problem in which we minimize the error between the observed \mathbf{X} over all possible non-negative factorizations (9.4). This normally involves learning the channels, namely the frequency-domain mixing matrices \mathbf{H} . Instead of learning them, we build the channels based on the prior knowledge of the earliest few echoes.

9.3 SOURCE SEPARATION BY NMF

In this work we consider two classical, well-understood multi-channel source separation algorithms which, by default, estimate the channels together with sources' dictionaries and activations. The first algorithm is Nonnegative Matrix Factorization (**NMF**) via Multiplicative Updates (**MU**) and consider only the magnitudes of the transfer functions. The second one is the multichannel **NMF** via Expectation Maximization (**EM**), which instead explicitly models the phases of the mixing filters. In this work, we considered only the (over)determined case ($J \leq I$). In the following we briefly describe the idea behind the two algorithms. We remind to the work of [Ozerov and Févotte 2010] for further details.

9.3.1 NMF using Multiplicative Updates (MU-NMF)

MU for **NMF** involves the observed magnitude spectrograms only and the updates rules guarantee non-negativity as long as the initialization is non-negative. This model has been originally proposed by in [Lee and Seung 2001], however we will consider its formulation as it appears in [Ozerov and Févotte 2010]. The observed multi-channel squared magnitude spectra are denoted $\mathbf{V}_i = [|X_i[k, l]|^2]_{kl}$ and their non-negative factorizations

$$\widehat{\mathbf{V}}_i = \sum_{j=1}^J \text{diag}(\mathbf{Q}_{ij}) \mathbf{D}_j \mathbf{Z}_j, \quad i = 1, \dots, I \quad (9.5)$$

where $\mathbf{Q}_{ij} = [|H_{ij}[k]|^2]_k$ is the vector of squared magnitudes of the approximate **RTF** between microphone i and source j .

- ▶ THE MU RULES MINIMIZE the *divergence*⁵⁰ between the observed spectrogram $\mathbf{V}_i[k, l]$ and the model $\widehat{\mathbf{V}}_i[k, l]$. In case of *Itakura-Saito* divergence [Févotte et al. 2009], the cost function writes

$$\mathcal{C}_{\text{MU}}(\Theta_{\text{MU}}) = \sum_{jkl} \mathcal{D}_{\text{IS}}(\mathbf{V}_i[k, l] | \widehat{\mathbf{V}}_i[k, l]) + \gamma \sum_j \|\mathbf{Z}_j\|_1, \quad (9.6)$$

where $\mathcal{D}_{\text{IS}}(v|\hat{v}) = \frac{v}{\hat{v}} - \log \frac{v}{\hat{v}} - 1$ and $\Theta_{\text{MU}} = \{\mathbf{Q}_{ij}, \{\mathbf{D}_j, \mathbf{Z}_j\}_j\}_{ij}$ is the set of parameters. We add an ℓ_1 -penalty term to promote sparsity in the activations due to the potentially large size of the dictionary [Sun and Mysore 2013].

- ▶ THE MU RULE for each scalar parameter of interest θ is obtained by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w. r. t. this parameter, namely,

$$\theta \leftarrow \theta \frac{[\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_-}{[\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_+}$$

where $\mathcal{C}_{\text{MU}}(\Theta_{\text{MU}}) = [\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_+ - [\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_-$ and the summands are both nonnegative. Following the **MU** rule derivations as in Ozerov and Févotte,

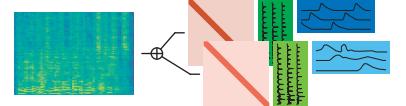


FIGURE 9.4: Schematics of the signal model used for **MU-NMF**. Image taken from the slides accompanying the paper of this work [Scheibler et al. 2018e].

⁵⁰ The divergence is an asymmetric “distance” measure. The Itakura–Saito divergence was proposed to account for perceptual (dis)similarity between spectra. Other types of divergence are used in Audio Source Separation, such as the Kullback–Leibler divergence.

we obtain:

$$\mathbf{Q}_{ij} \leftarrow \mathbf{Q}_{ij} \odot \frac{\left[\widehat{\mathbf{V}}_j^{-2} \odot \mathbf{V}_j \odot (\mathbf{Z}_j \mathbf{D}_j) \right] \mathbf{1}_{1 \times T}}{\left[\widehat{\mathbf{V}}_j^{-1} \odot (\mathbf{Z}_j \mathbf{D}_j) \right] \mathbf{1}_{1 \times T}} \quad (9.7)$$

$$\mathbf{Z}_j \leftarrow \mathbf{Z}_j \odot \frac{\sum_i (\text{diag}(\mathbf{Q}_{ij}) \mathbf{D}_j)^T (\mathbf{V}_j \odot \widehat{\mathbf{V}}_j^{-2})}{\sum_i (\text{diag}(\mathbf{Q}_{ij}) \mathbf{D}_j)^T \widehat{\mathbf{V}}_j^{-1} + \gamma}, \quad (9.8)$$

$$\mathbf{D}_j \leftarrow \mathbf{D}_j \odot \frac{\sum_i \text{diag}(\mathbf{Q}_{ij})^T (\mathbf{V}_j \odot \widehat{\mathbf{V}}_j^{-2}) \mathbf{Z}_j^T}{\sum_i \text{diag}(\mathbf{Q}_{ij})^T \widehat{\mathbf{V}}_j^{-1} \mathbf{Z}_j^T}, \quad (9.9)$$

where multiplication \odot , power, and division are element-wise and $\mathbf{1}_{1 \times T}$ is a N -vector of ones.,

9.3.2 NMF using Expectation Maximization (EM-NMF)

Unlike the MU algorithm that independently maximizes the log-likelihood of spectral magnitudes, the EM-NMF maximizes the joint log-likelihood over all complex-valued channels [Ozerov and Févotte 2010]. Hence, the model takes explicitly into account observed phases. In this approach, each source j is modeled as a complex Gaussian in the form of

$$S_j[k, l] \sim \mathcal{N}_c(0, (\mathbf{D}_j \mathbf{Z}_j)_{kl}), \quad (9.10)$$

and the magnitude spectrum \mathbf{P}_j of (9.4) can be understood as the variance of source j .

Under this model, and assuming uncorrelated noise, the microphone signals also follow a complex Gaussian distribution with covariance matrix

$$\Sigma_{\mathbf{X}}[k, l] = \mathbf{H}[k] \Sigma_S[k, l] \mathbf{H}^H[k] + \Sigma_N[k, l], \quad (9.11)$$

where Σ_S and Σ_N are the covariance matrices of the sources and noise, respectively.

- ▶ THE EM COST FUNCTION corresponds to the negative log-likelihood of the observed signal, that is,

$$\mathcal{C}_{\text{EM}}(\Theta_{\text{EM}}) = \sum_{kl} \text{trace} \left(\mathbf{X}[k, l] \mathbf{X}[k, l]^H \Sigma_{\mathbf{X}}^{-1}[k, l] \right) + \log \det \Sigma_{\mathbf{X}}[k, l]. \quad (9.12)$$

where the $\Theta_{\text{EM}} = \left\{ \mathbf{H}, \{ \mathbf{D}_j, \mathbf{Z}_j \}_j, \Sigma_N \right\}$ is the set of parameters.

- ▶ THE EM ALGORITHM estimates all the parameters Θ by alternating between the so-called E-step and M-step. In a nutshell, one iteration of the E-step consists in computing the *conditional expectation* of the the “complete” log likelihood⁵¹ with respect to the current parameter estimates, and the M-step re-estimates the parameters by maximizing the conditional expectation of the complete log-likelihood. This quantity can be efficiently minimized using the EM algorithm proposed in [Ozerov and Févotte 2010]. Since adding sparsity priors is not straightforward in the EM framework, it was not included in the proposed method.

⁵¹ The complete data log-likelihood includes both observed variables \mathbf{X} and latent variables S

9.3.3 Supervised NMF with pre-trained Dictionaries

Pre-trained *dictionaries* are a typical way to informing the **NMF** algorithm, which is sometimes referred to as *supervised NMF*. The idea to run NMF on training sets containing examples from desired sound classes and collect the atoms of the estimated non-negative matrices [Schmidt and Olsson 2006]. At test phase, these atoms are used as basis vectors for the dictionary matrix (i.e., \mathbf{D}) and can be used as a good initialization point or kept fixed in the algorithm⁵². This can be seen as an instance of the problem of *dictionary learning* which also exists in many other research fields. For audio source separation, this idea has been studied extensively since promising results were obtained, even in single channel scenarios [Smaragdis et al. 2009]. As discussed later in § 9.5.2, in this work we will use two different dictionaries: one *universal*, and the other *speaker-specific*.

⁵² In the context of **NMF**-based music transcription applied to piano music, the dictionary can be a collection of spectral templates, each of which is associated to a piano note [Müller 2015]

9.4 ECHO-AWARE SOURCE SEPARATION

To evaluate the usefulness of echoes in source separation, we modified the multi-channel **NMF** framework of Ozerov and Févotte [Ozerov and Févotte 2010]. The knowledge of the echoes is embedded in the model by approximating the entries of mixing matrices with (9.1), that is,

$$H_{ij}[k] = \sum_{r=0}^R \alpha_{ij}^{(r)} e^{-i2\pi f_k \tau_{ij}^{(r)}}, \quad (9.13)$$

$$\mathbf{H}[k] = [H_{ij}[k]]_{ij},$$

where $f_k = kF_s/F$ are the discretized frequencies in Hz corresponding to the k -th bin in the **DFT**.

Futhermore, the early-echo channel model is kept fixed throughout the iterations. Moreover, instead of updating both sources' dictionaries and activations, we adapted pre-trained dictionaries to better guide the source separation.

- ▶ NEGLECTING THE REVERBERATION (or working in the anechoic regime) leads to a constant \mathbf{Q}_{ij} for all j and i . A consequence is that the **MU-NMF** framework breaks down with a unique, universal dictionary, namely, $\mathbf{D} = \mathbf{D}_j \forall j$. Indeed, (9.5) becomes the same for all i ,

$$\widehat{\mathbf{V}}_i = \sum_j \mathbf{DZ}_j = \mathbf{D} \sum_j \mathbf{Z}_j,$$

so even with the correct atoms identified, we can assign them to any source without changing the value of the cost function. Therefore, anechoic multi-channel separation with a universal dictionary cannot work well. This intuitive reasoning is corroborated by numerical experiments in Section 9.5.4. The problem is overcome by the **EM-NMF** algorithm which keeps the channel phase and is thus able to exploit the phase diversity across the array. Of course, as showed in this work, it is also overcome by using echoes.

9.5 NUMERICAL EXPERIMENTS

We test our hypotheses through computer simulations. In the following, we describe the simulation setup, dictionary learning protocols, and we discuss

		Number of echoes R						
		0	1	2	3	4	5	6
$\gamma =$	anechoic	10	10^{-1}	10	10^{-3}	0	0	0
								0

the results.

9.5.1 Setup

An array of three microphones arranged on the corners of an equilateral triangle with edge length 0.3 m is placed in the corner of a 3D room with 7 walls. We select 40 sources at random locations at a distance ranging from 2.5 m to 4 m from the microphone array. Pairs of sources are chosen so that they are at least 1 m apart. The floor plan and the locations of microphones are depicted in Figure 9.5. The scenario is repeated for every two active sources out of the 780 possible pairs.

The sound propagation between sources and microphones is simulated using the image source model implemented in the `pyroomacoustics` Python package [Scheibler et al. 2018a]. The wall absorption factor is set to 0.4, leading to a RT_{60} of approximately 100 ms.⁵³ The sampling frequency is set to 16 kHz, STFT frame size to 2048 samples with 50% overlap between frames, and we use a cosine window for analysis and synthesis. Partial RTFs are then built from the R nearest image microphones. The global delay is discarded, and only the relative amplitudes between echoes are kept.

With this setup, we perform three different experiments. In the first one, we evaluate MU-NMF with a universal dictionary. In the other two, we evaluate the performance of MU-NMF and EM-NMF with speaker-specific dictionaries. We vary R from 1 to 6 and use the following three baseline scenarios:

1. *anechoic*: Anechoic conditions, no model mismatch.
2. *learn*: The RTFs are learned from the data along the activations as originally proposed [Ozerov and Févotte 2010] and the full reverberation is present in the observed data.
3. *no echoes*: Reverberation is present but ignored (i.e. $R = 0$).

With the universal dictionary, the large number of latent variables warrants the introduction of sparsity-inducing regularization. The value of the regularization parameter γ was chosen by a grid search on a holdout set with the signal-to-distortion ratio (SDR) as the figure of merit [Vincent et al. 2007] (See Table 9.1).

9.5.2 Dictionary Training, Test Set

First, we introduce a dictionary learned from available training data. We explore both speaker-specific and universal dictionaries [Sun and Mysore 2013]. Speaker-specific dictionaries can be beneficial when speakers are known in advance. Universal dictionary is more versatile but gives a weaker regularization prior. All dictionaries were trained on samples from the TIMIT corpus

TABLE 9.1: Value of the regularization parameter γ used with the universal dictionary.

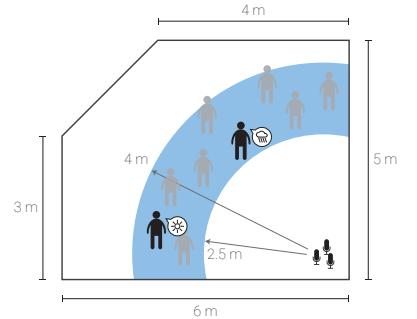


FIGURE 9.5: The simulated scenario. Image taken from the paper of this work [Scheibler et al. 2018e]

⁵³ The resulting RT_{60} is short. Realistic values for office and house rooms are typically around 500 ms. At the time of the publication we did not investigate different reverberation levels as the focus was on the early echoes.

[Garofolo et al. 1993] using the NMF solver in scikit-learn Python package [Pedregosa et al. 2011].

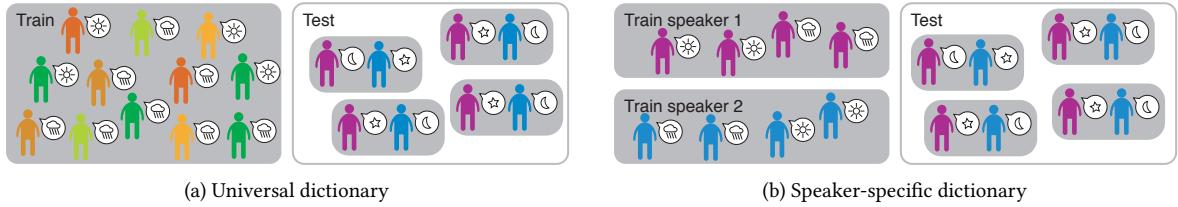


FIGURE 9.6: Schematic representation of the dictionary protocol used in this work. Image taken from the slides accompanying the paper of this work [Scheibler et al. 2018e].

- ▶ **UNIVERSAL DICTIONARY:** Following the methodology of [Sun and Mysore 2013] we select 25 male and 25 female speakers and use all available training sentences to form the universal dictionary $\mathbf{D} = [\mathbf{D}_1^M \cdots \mathbf{D}_{25}^M \mathbf{D}_1^F \cdots \mathbf{D}_{25}^F]$. The test signals were selected from speakers *and* utterances outside the training set. The number of latent variables per speaker is 10 so that with STFT frame size of 2048 we have $\mathbf{D} \in \mathbb{R}^{1025 \times 500}$.
- ▶ **SPEAKER-SPECIFIC DICTIONARY:** Two dictionaries were trained on one male and one female speaker. One utterance per speaker was excluded to be used for testing. The number of latent variables per speaker was set to 20.

9.5.3 Implementation:

Authors of [Ozerov and Févotte 2010] provide a Matlab implementation⁵⁴ of MU-NMF and EM-NMF methods for stereo separation. We ported their code to Python and extended it to arbitrary number of input channels. However this software features some ad-hoc decisions which do not fit our scenario. Thus, we provide a Python adaptation with the following modifications:

- first the original code was restricted to the 2-channel case, i.e. $I = 2$, thus, in order to embrace the specificities of our scenario and for the sake of generalization, we extend it to the multi-channel case, that is $\forall I \geq 1$;
- the MU-NMF was modified to handle sparsity constraint as described in 9.3.1;
- since the EM method degenerates when zero-valued entries are present in the dictionary matrix, \mathbf{D} , all these entries are initially set to a small constant value of 10^{-6} ;
- the code was further modified to deal with fixed dictionary and channel models matrices, which are normalized in order to avoid indeterminacy issues [Ozerov and Févotte 2010].

⁵⁴ Multichannel nonnegative matrix factorization toolbox (in Matlab)

Finally, no “simulated annealing” strategy [Ozerov and Févotte 2010, Section III.A.6] was used in the final experiments. In fact in some preliminary and informal investigations we noticed that this yielded better results than using

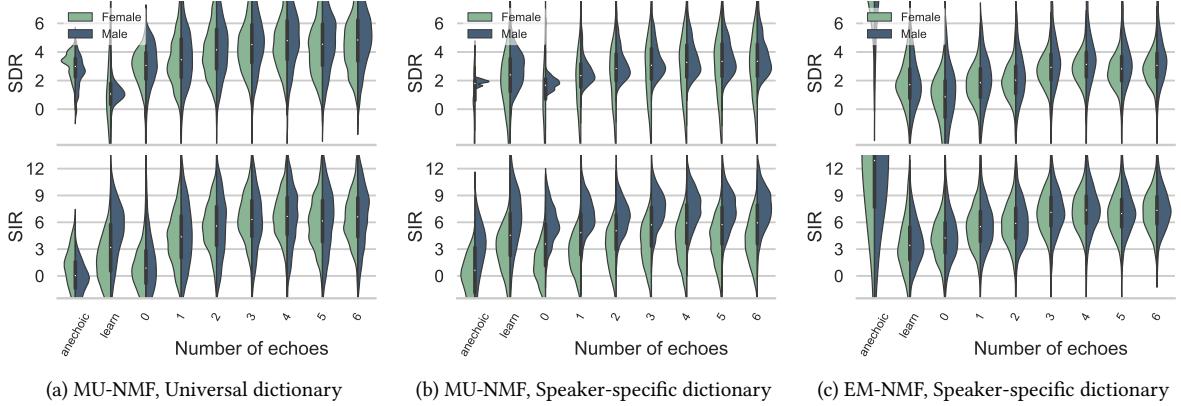


FIGURE 9.7: Distribution of **SDR** and **SIR** for male and female speakers as a function of the number of echoes included in modeling, and comparison with the three baselines.

annealing. In the experiments, the number of iterations for **MU-NMF** (resp. **EM-NMF**) was set to 200 (resp. 300).

9.5.4 Results

We evaluate performance in terms of Signal to Distortion Ratio (**SDR**) and Signal to Interference Ratio (**SIR**) as defined in [Vincent et al. 2007]. We compute these metrics using the `mir_eval` toolbox [Raffel et al. 2014].

The distributions of **SDR** and **SIR** for separation using **MU-NMF** and a universal dictionary are shown in Figure 9.7a, with a summary in Figure 9.8. We use the median performance to compare the results from different algorithms. First, we confirm that separation fails for flat RTFs (anechoic and $R = 0$) with **SIR** at around 0 dB. Learning the RTFs performs somewhat better in terms of **SIR** than in terms of **SDR**, though both are low. Introducing approximate RTFs significantly improves performance: the proposed approach outperforms the learned approach even with a single echo. With up to six echoes, gains are +2 dB **SDR** and +5 dB **SIR**. Interestingly, with more than one echo, non-negativity and echo priors are already sufficient for achieving good separation, overtaking the ℓ_1 regularization.

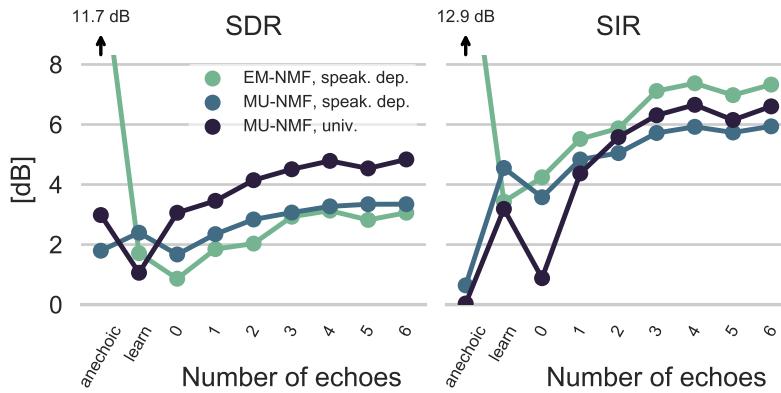


FIGURE 9.8: Summary of the median **SDR** and **SIR** for the different algorithms evaluated.

Separation with speaker-dependent dictionaries is less challenging since we have a stronger prior. Accordingly, as shown in Figures 9.7b and 9.8, MU-NMF now achieves a certain degree of separation even without the channel information. The gains from using echoes are smaller, though one echo is still sufficient to match the median performance of learned RTFs. Using an echo, however, results in a smaller variance, while adding more echoes further improves SDR (SIR) by up to +2 dB (+3 dB).

In the same scenario, EM-NMF (Figure 9.7c) has near-perfect performance on anechoic signals which is expected as the problem is overdetermined. For MU, a single echo suffices to reach the performance of learned RTFs and further improve it. Moreover, echoes significantly improve separation quality as illustrated by up to 3 dB improvement over *learn*. It is interesting to note that in all experiments the first three echoes nearly saturate the metrics. This is good news since higher order echoes are hard to estimate.

9.6 CONCLUSION

In this work, we investigated the potential benefit of early echo knowledge for the problem of sound source separation. Unlike earlier work, instead of fitting an echo model or trying to estimate blindly the acoustic channels, we investigate the potential of including the properties of known echoes in well established NMF-based source separation algorithms. In particular, we modified the MU approach (which considers only spectral magnitudes) and the EM (which accounts for complex spectra) by integrating a simple echo model. Despite its simplicity, such echo model lends itself to an interesting interpretation by revising the ISM model: to each echo corresponds an image microphone (instead of image source as in ISM). It follows that real and image microphones can be considered as microphones arrays with specific directivity pattern.

Numerical results show that echoes seem to play an essential role in magnitude-only algorithms, like the MU-NMF. In general, they show that using knowledge of a few echoes significantly improve results with respect to an anechoic model. This improvement is measured by the standard metrics even when compared to approaches that learn the transfer functions.

Finally, this work confirms the potential of including echoes in sound source separation framework.

► FUTURE WORK

on echo-aware source separation could include:

- integrating the blind estimation of the echoes properties, e.g. using the algorithm blaster, proposed in Chapter 5;
- including the late reverberation part in the model for the RTFs;
- experiment with more microphones, more room configurations, more sources on real data, e.g., using the ones offered by the dEchorate dataset, described in Chapter 7; In fact, such dataset was designed and recorded also for echo-aware application. In this sense, the method discussed in this chapter can be perfectly tested on such kind of dataset.

10

Mirage: Echo-aware Sound Source Localization

- ▶ **SYNOPSIS** This chapter addresses the problem of audio source localization in the context of strong acoustic echoes. Classic **SSL** methods are deceived by strong acoustic echoes: rather than estimating the location of a true source, they might be fooled by its strong correlated echos. Instead, we show early-echo characteristics can, in fact, benefit **SSL**. To this end, we introduce the concept of microphone array augmentation with echoes **MIRAGE**, using the model of image microphones presented in the previous chapter. In particular, we show that in a simple scenario involving two microphones close to a reflective surface and one source, the proposed approach can estimate both azimuthal and elevation angles, an impossible task assuming an ideal propagation, as classical approaches do. Later, the proposed approach is extended to multichannel recording and tested on real data scenario.

Together with [Chapter 6](#), this chapter describes methods and results published in [Di Carlo et al. 2019b], which considers only stereophonic recordings. In this sense, this chapter provides an application for the data-driven echo estimation **LANTERN** presented in the above mentioned chapter. Subsequently, the proposed approach was extended to multi-microphone recordings in collaboration with Randy Gomez from Honda Research Institute. In particular, the method was tested on an autonomous robot platform called HARU[Ackerman 2018; Gomez et al. 2018], consisting of a base and two screens mimicking a face with two eyes (See [Figure 10.1](#)) The robot is fitted with cameras and a microphone array for visual and auditory sensing. In particular the such array is circular with radius 6.5 cm featuring 7 sensors. The partner agreed on using its technology to see the impact of echo-aware sound source localization. Therefore, the multichannel extention of this method consider the circular microphone array featuring 7 sensor of the HARU robot. The results of this study were described in an internal technical report [*HRI-JF collaboration - Final Phase II Deliverable*].

Keywords: Sound Source Localization, Image Microphones, Acoustic Echoes, TDOA Estimation.

Resources:

- Paper
- Code
- Poster
- HARU Robot presentation

Di Carlo et al., “Mirage: 2D source localization using microphone pair augmentation with echoes”



FIGURE 10.1: The HARU Robot.

10.1 LITERATURE REVIEW IN ECHO-AWARE SOUND SOURCE LOCALIZATION

Common to most sound source localization approaches reviewed in § 8.4 is the challenge posed by environment reverberation. It is typical to observe that DOAs estimation degrades with increasing acoustic reflection [Chen et al. 2006]. For these reasons, most sound source localization methods regard reverberation and, in particular, acoustic echoes as a nuisance. Room reverberation is considered in the works [Rui and Florencio 2004; Chen et al. 2006; Zhang et al. 2007] while the authors of [Weinstein et al. 1994; Taghizadeh et al. 2015; Salvati et al. 2016] attempt to solve SSL by estimating the full RIRs. However, both the cases have drawbacks: in the former, the generic model for reverberation does not reduce strong early echoes, and in the latter, RIRs estimation is a difficult task.

The echo-aware sound source localization methods take another direction: they exploit the closed-form relation between echoes timings and audio scene geometry expressed by the Image Source Method (ISM). Early works such as [Korhonen 2008; Ribeiro et al. 2010a; Ribeiro et al. 2010b; Svaizer et al. 2011] uses knowledge from the room geometry to estimate the position of the sound source with respect to the arrays. This idea was subsequently extended in later works, reducing the amount of prior knowledge required or addressing different applications. The authors of [Nakashima et al. 2010] study the SSL problem in binaural recordings. To improve localization, they propose to use ad-hoc reflectors as artificial *pinnae* and a simple reflection model. In the work [Kreković et al. 2016], the author addresses the problem Source Localization and Mapping (SLAM)⁵⁵ using echoes. The authors of [An et al. 2018] propose to use cameras, depth sensors, and laser sensors to identify reflectors and build a corresponding acoustic model that is used for echo-aware SSL. Finally, in a very recent work, the well-known MUSIC framework for localizing multiple sources is modified for accounting an echo model for the spherical harmonic representation [Birnie et al. 2020]

⁵⁵ SLAM enables the estimation of a moving robot's position in relation to a number of external acoustic sources.

All the above mentioned echo-aware methods are explicitly knowledge-driven, namely, using closed-form solutions based on physics, acoustics, and signal processing models. As explained in the previous chapter, data-driven methods, especially DNN, have been successfully applied to address SSL. The main benefit is in their ability to learn complex mapping functions based on simple input-output pairs. However, they are typically trained for specific applications and use-cases (e.g., arrays geometry, acoustic conditions, etc.) and fail whenever test conditions strongly mismatch training conditions.

10.2 PROPOSED APPROACH

We propose here to combine the best of the two worlds: using a deep learning model to estimate challenging acoustic parameters and a physically-motivated model to map such parameters to source's DOAs. To this end, we introduce the framework of Microphone Augmentation with Echoes (**MIRAGE**) for SSL, based on the *image microphones* model [Bergamo et al. 2004; Korhonen 2008] (See § 9.1).

Let us consider a simple yet common scenario to illustrate this idea: two microphones, one source, and a nearby reflective surface, as illustrated in Fig. [Figure 10.2](#). This may occur when the sensors are placed on a table or next to a wall. Striking examples of these scenarios are the smart table-top devices, such as Amazon Echo, Google Home, etc. The reflective surface is assumed to be the most reflective and closest one to the microphones in the environment, generating the strongest and earliest echo in each microphone. Under this stereophonic *close-surface* model, we ask the following questions.

- ? CAN EARLY ECHOES BE ESTIMATED FROM TWO-MICROPHONE RECORDINGS OF AN UNKNOWN SOURCE? This question was already addressed in [Part II](#). In particular, here, we proposed to use the [DNN](#) models discussed in [Chapter 6](#). The models are trained on a simulated close-surface dataset to estimate early echoes properties from audio features.
- ? CAN EARLY ECHOES BE USED TO ESTIMATE BOTH THE AZIMUTH AND ELEVATION ANGLES OF THE SOURCE? To answer the second question, we propose the [MIRAGE](#) framework. It exploits echoes' time of arrival by expressing them as [TDOAs](#) in the *virtual 4-microphone array* formed by the true microphone pair and its image with respect to the reflective surface. Note that this is considered as an *impossible* task in free field stereophonic conditions.

10.3 BACKGROUND IN MICROPHONE ARRAY SSL

In this section, we briefly review some necessary background in microphone array [SSL](#). Let us assume a microphone array of i sensors is placed inside a room and records the sound emitted by one static point sound source ($J = 1$). In all generality, the relationship between the signal x_i recorded by the i -th sensor placed at fixed position \underline{x}_i and the signal s emitted by the source at fixed position \underline{s} is defined by:

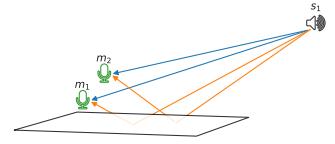
$$x_i[n] = (h_i * s)[n] + n_i[n], \quad (10.1)$$

where the convolution with [RIR](#) $h_i[n]$ embodies the fact that sensor i receives a spatial image of the source and n_i denotes possible measurement noise. As fully described in [Chapter 2](#), the [RIR](#) depends on the spatial parameters of the scene: microphone positions, source position w.r.t. the room, as well as the room acoustic properties (size, absorption, and diffuseness of the wall materials).

Let us assume that [RIRs](#) follows the echo model under the narrowband approximation presented in [§ 3.2.4](#). Therefore, in the discrete-frequency domain, this leads to

$$H_i[k] = \sum_{r=0}^R \alpha_i^{(r)}[k] e^{-i2\pi f_k \tau_i^{(r)}} + \varepsilon_i[k], \quad (10.2)$$

where f_k is the k -th frequency bin and the error term $\varepsilon_i[k]$ collects later echoes, the reverberation tail, diffusion, and noise. In this work, we will consider only the first strongest echo, therefore $R = 1$. Note that for $r = 0$ denotes the ideal propagation path, being $\tau_i^{(0)}$ the ideal propagation path from the source to the



[FIGURE 10.2](#): Typical setup with one source recorded by two microphones. The illustration shows direct sound path (blue lines) and resulting first-order echoes (orange lines).

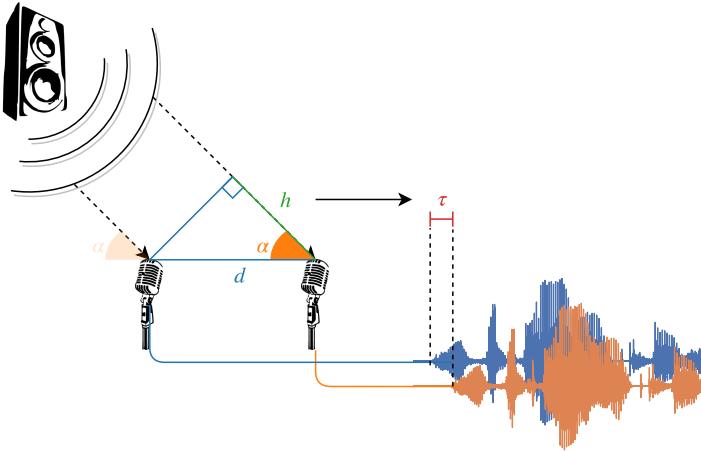


FIGURE 10.3: Illustration of the relation between DOA and TDOA with one source and two microphones. Knowing the distance d between the two microphones, simple trigonometry yields the AOA ϑ according to Eq. (10.3).

i -th microphone, and $\alpha_i^{(0)}$ the air attenuation. In the remainder of this work, we make the approximation of $\alpha_i^{(r)}$ being frequency-independent.

10.3.1 2-channel 1D-SSL

Let us first consider the case of stereophonic recordings ($I = 2$). Under the far-and free-field assumption, traditional SSL methods use the Time Difference of Arrival (TDOA),

$$\tau_{\text{TDOA}} \stackrel{\text{def}}{=} \tau_2^{(0)} - \tau_1^{(0)} \quad [\text{second}],$$

as a proxy for the estimation of the Angle of Arrival (AOA), ϑ , since:

$$\vartheta = \arccos(c \tau_{\text{TDOA}} / d) \quad [\text{rad}], \quad (10.3)$$

where c is the speed of sound and d the inter-microphone distance.

Then, SSL reduces to estimating the TDOA, which can be done by Cross Correlation (CC)-based methods such as the widely used and well performing GCC-PHAT method [Knapp and Carter 1976; Blandin et al. 2012]. Given STFT X_1 and X_2 of the two microphones signals, the CC and GCC-PHAT angular spectra are defined as:

$$\Psi_{\text{CC}}(\tau) = \sum_{k,l} X_1[k, l] X_2^H[k, l] e^{-i2\pi f_k \tau}, \quad (10.4)$$

$$\Psi_{\text{PHAT}}(\tau) = \sum_{k,l} \frac{X_1[k, l] X_2^H[k, l]}{|X_1[k, l] X_2^H[k, l]|} e^{-i2\pi f_k \tau}, \quad (10.5)$$

where $|\cdot|$ denotes the absolute value.

The weighting function $1/|X_1[k, l] X_2^H[k, l]|$ is called *phase transform*, aiming at reducing the source's autocorrelation component from the angular spectrum.

The “generalized” cross-correlation methods adds weighting functions (e.g. the phase transform (PHAT), or the smoothed coherence transform (SCOT)) to the CC. Their purpose is to improve the estimation of the time delay on specific characteristic of the signal and noise. See [Chen et al. 2006] for an overview.

Then, the TDOA estimate is given by

$$\hat{\tau}_{\text{TDOA}} = \arg \max_{\tau} \Psi(\tau),$$

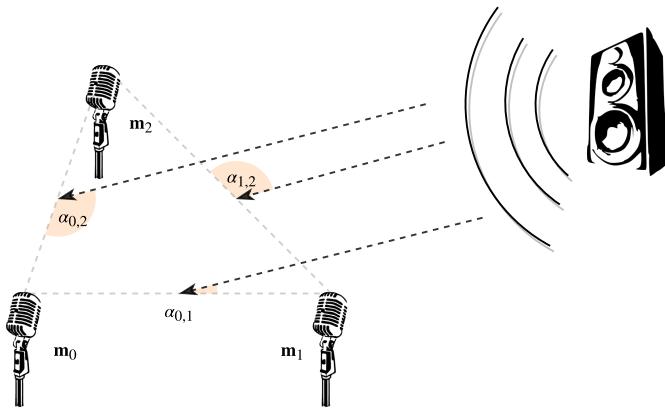


FIGURE 10.4: Illustration of the different DOAs at each microphone pairs listening one sound source. Knowing the position of the microphone, the angle with respect to a reference point can be deduced in closed-form.

with Ψ begin either $\Psi_{\text{CC}}(\tau)$ or $\Psi_{\text{PHAT}}(\tau)$. Note that these functions can also be expressed directly as a function of the AOA using (10.3), hence the term *angular spectrum*. Despite the theoretical limits of this method, discussed in [Chen et al. 2006], this method is known to work well practical non-challenging scenarios. Moreover, it was showed to be state-of-the-art for SSL in a large benchmark study [Blandin et al. 2012].

10.3.2 Multichannel 2D-SSL

When more microphones are available and the microphones array is compact and not linear⁵⁶, 2D-SSL can be envisioned. A possible approach is to use 1D-SSL on all pairs and combine their results, a principle which was successfully applied in the Steered Response Power with Phase Transform (SRP-PHAT) method [DiBiase et al. 2001].

The SRP-PHAT methods returns the source's DOA, namely the pair azimuth and elevation (θ, ϕ) , by estimating TDOAs from each microphone pairs. In order to achieve this, it requires the geometry of the microphone array to be known. In a nutshell, this algorithm aims to estimate a *global angular spectrum* $\Psi_{\text{SRP}}(\theta, \phi)$ in the polar coordinates system with respect to reference point in the array, typically its barycenter. This function will exhibit a local maximum in the direction of the active source.

The algorithmic can be exemplified in the following steps:

1. a global grid of DOAs candidates is defined according to a desired resolution and computational load;
2. for each pair of microphones, a local set of AOA (hence, TDOAs) is defined based on the above chosen DOAs and the input geometry;
3. a TDOA-based algorithm (e. g. GCC-PHAT) is used to compute the associated local angular spectrum;
4. all the local contributions (a collection of local $\Psi_{\text{GCC}}(\tau)$) are geometrical aggregated and interpolated back to the global DOA grid to form $\Psi_{\text{SRP}}(\theta, \phi)$;

⁵⁶ In case of coplanarity, the angle can be estimated up to “up-down” ambiguity.

See [MBSSLocate](#) for a free MATLAB implementation and comprehensive documentation of this algorithm.

5. the **DOA**(s) maximizing Ψ_{SRP} is (are) used as estimate (in case of multiple sources).

This algorithm can be seen as an application of the divide-and-conquer paradigm to **TDOA**-based methods: “at the leaves”, the **GCC-PHAT** method provide **TDOA** for each microphone pair; the “merge” operation consists in aggregating **TDOA** defined on a different axis based on the knowledge of the array geometry. Finally, we stress that this algorithm is independent of the method used to estimate the **TDOA**.

10.4 MICROPHONE ARRAY AUGMENTATION WITH ECHOES

We now introduce the proposed concept of Microphone Augmentation with Echoes (**MIRAGE**). The Eq. (10.2) corresponds to the well known Image Source Method (**ISM**), where reflections are treated as mirror images of the true source with respect to reflective surfaces, emitting the same signal. We will employ here a less common but equivalent interpretation of **ISM**, namely, the image-microphone (**IM**) model. As illustrated in Fig. Figure 10.5, image microphones are mirror images of the true microphones with respect to reflective surfaces. In this view, the echoic signal received at a true microphone is the sum of the anechoic signals received at this microphone and its images. If we consider the virtual array consisting of both true and image microphones, multiple microphone pairs are now available. For each of them, it is then possible to define a corresponding time difference of arrival. Among them, we will refer to the one between the two real microphones as **TDOA**, the one between the two image microphones as **iTDOA** and the one between the first microphone and its image as **TDOE**. Therefore, we have:

$$\tau_{\text{TDOA}} = \frac{1}{c} \|\underline{x}_2 - \underline{s}\| - \frac{1}{c} \|\underline{x}_1 - \underline{s}\| = \tau_2^{(0)} - \tau_1^{(0)}, \quad (10.6)$$

$$\tau_{\text{iTDOA}} = \frac{1}{c} \|\dot{\underline{x}}_2 - \underline{s}\| - \frac{1}{c} \|\dot{\underline{x}}_1 - \underline{s}\| = \tau_2^{(1)} - \tau_1^{(1)}, \quad (10.7)$$

$$\tau_{\text{TDOE},1} = \frac{1}{c} \|\dot{\underline{x}}_1 - \underline{s}\| - \frac{1}{c} \|\underline{x}_1 - \underline{s}\| = \tau_1^{(1)} - \tau_1^{(0)} \quad (10.8)$$

where $\dot{\underline{x}}_i$ denotes the position of the image of \underline{x}_i with respect to the reflector. Note that $\tau_{\text{TDOE},2} = \tau_{\text{iTDOA}} + \tau_{\text{TDOE},1} - \tau_{\text{TDOA}}$. These three quantities are directly connected to **RIRs**, as illustrated in Figure 10.6. Notice that they are the same estimated by the **LANTERN** learning-approach presented in Chapter 6.

Let $V = [\tau_{\text{TDOA}}, \tau_{\text{iTDOA}}, \tau_{\text{TDOE},1}] \in \mathbb{R}^3$

Following the 2D-SSL scheme described in § 10.3.2 and given the virtual microphone-array geometry (which depends on the relative position of microphones to the surface), V could in principle be used to estimate the 2D directional arrival of the source.

As stated in § 10.3.2, given a microphone pair, the local maximum of angular spectrum Ψ_{PHAT} corresponds to the **TDOA**. Moreover, peaks corresponding to the early reflection are also present. Figure 10.7 shows the Ψ_{CC} 's and the Ψ_{PHAT} 's angular spectra for synthetic data where the source signal is noise or speech for all the pairs of the HARU's circular microphone array. The location of the quantities in V are highlighted with vertical dotted lines. Theoretically,

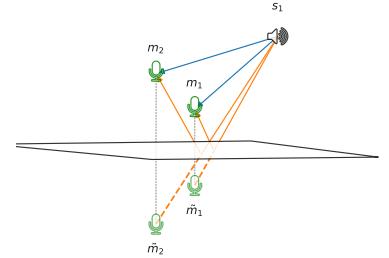


FIGURE 10.5: Illustration of the images $\dot{\underline{x}}_1$ and $\dot{\underline{x}}_2$ of microphones \underline{x}_1 and \underline{x}_2 in the presence of a reflective surface and a source. Blue lines correspond to direct paths, orange lines correspond to echo paths.

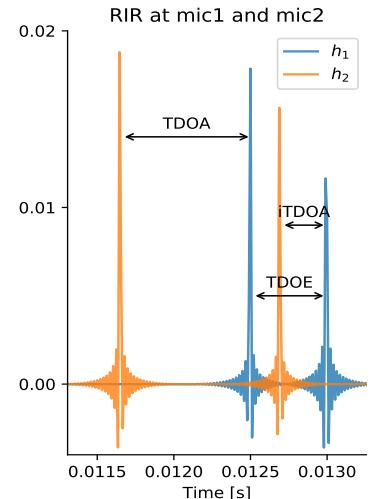


FIGURE 10.6: Superposition of two **RIRs** and visualization of time difference of arrival between direct paths (**TDOA**), first echoes (**iTDOA**) and direct path and first echo (**TDOE**).

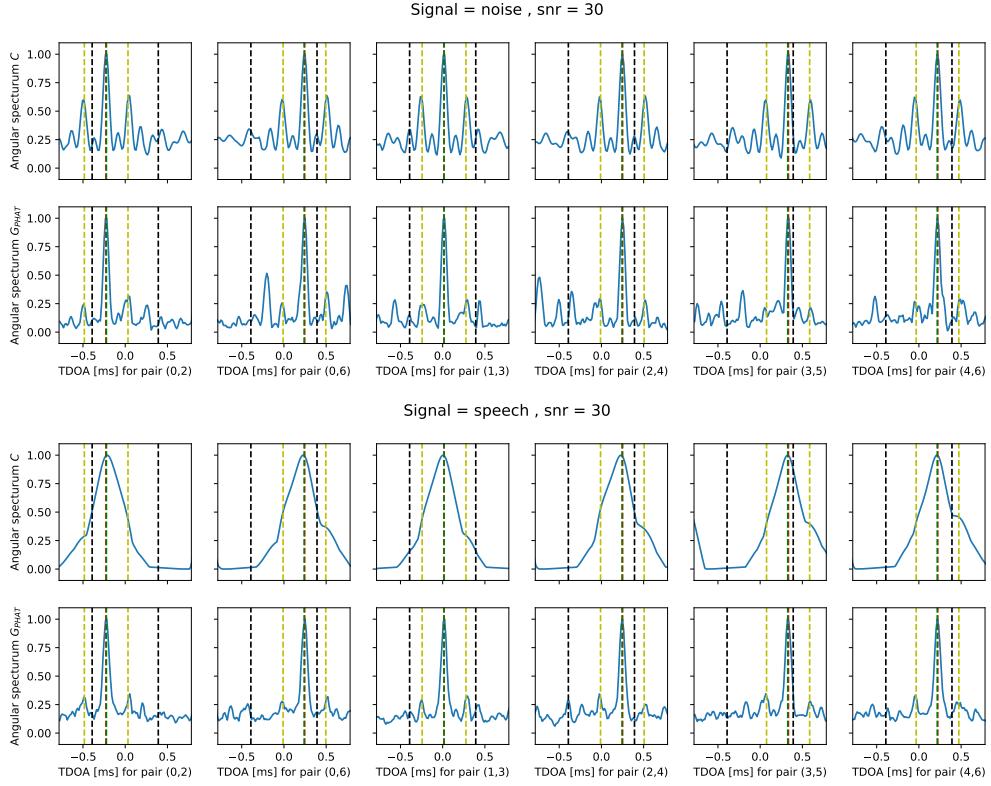


FIGURE 10.7: Angular Spectra Ψ_{CC} and Ψ_{PHAT} for different pairs of microphone in the HARU array using synthetic RIRs and *white noise* (top) and *speech* (bottom) signal. Vertical lines mark the positions of τ_{TDOA} (red), τ_{iTDOA} (green), $\tau_{TDOA} - \tau_{TDOE,1}$ (yellow) and $\tau_{TDOA} + \tau_{TDOE,2}$ (yellow) are marked with vertical lines. The black vertical lines correspond to the maximum TDOA given the pair distance, i.e. corresponding to the AOA = $\{0, 2\pi\}$

when only the first reflection are considered ($K = 1$), the position of the peaks in the angular spectra correspond to τ_{TDOA} , τ_{iTDOA} , $\tau_{TDOA} - \tau_{TDOE,1}$, and $\tau_{TDOA} + \tau_{TDOE,2}$. It is important to note that for speech signals, Ψ_{PHAT} removes the auto-correlation part in order to promote a sharp peak at the position of the TDOA. Since acoustic echoes increase the auto-correlations of the signal in one microphones, the PHAT transform tends to lower their contribution, so that their peaks are not distinguishable from spurious ones.

10.5 EXPERIMENTAL RESULTS

10.5.1 2-channel scenario

To the best of our knowledge, no reference implementation of algorithms for 2D-SSL using only two microphones is available to date. To check the validity of TDOAs estimation, our approach is compared to **GCC-PHAT** using only the true microphones (see Sec. § 10.3.1).

In order to estimate the quantities of interest V , we proposed in [Di Carlo et al. 2019b] to use the Multi-layer Perceptron (MLP) neural network presented in § 6.2.1, which was specifically designed for this task. In order to use these estimates in the **SRP-PHAT** algorithm, synthetic local angular spectra are built. They consist of Gaussians centered at \hat{V} and with variances equal to the prediction errors made by the DNN on the validation set. For data,

nRMSE	Input	ACCURACY	
		$\theta < 10^\circ$	$\theta < 20^\circ$
MIRAGE	wn	4.10 (77)	5.97 (97)
MIRAGE	wn+n	5.00 (26)	9.89 (54)
MIRAGE	sp	4.83 (63)	7.26 (82)
MIRAGE	sp+n	4.60 (16)	9.88 (35)
GCC-PHAT	wn	4.22 (81)	6.19 (97)
GCC-PHAT	wn+n	4.03 (65)	5.34 (83)
GCC-PHAT	sp	4.08 (82)	5.34 (97)
GCC-PHAT	sp+n	4.70 (19)	8.38 (32)

DOA	Input	ACCURACY		ACCURACY	
		$< 10^\circ$		$< 20^\circ$	
		θ	ϕ	θ	ϕ
MIRAGE	wn	4.5 (59)	3.9 (71)	6.8 (79)	5.9 (88)
MIRAGE	wn+n	4.4 (18)	5.5 (26)	9.4 (35)	11.1 (66)
MIRAGE	sp	4.6 (45)	4.8 (59)	8.1 (71)	7.2 (83)
MIRAGE	sp+n	5.2 (17)	5.9 (12)	10.7 (38)	12.3 (43)

implementation, and training, the reader is invited to refer to § 6.2.1.

A free and open-source Matlab implementation of **SRP-PHAT**⁵⁷ is used to aggregate local angular spectra obtained from the **DNN**'s output. The same toolbox is used for the implementation of **SRP-PHAT** with **GCC-PHAT** for which only real pairs are used. A sphere sampling with 0.5° resolution and coordinates $\theta \in [-179, 180]$ and $\phi \in [0, 90]$ is used for the DQA search.

The methods are tested on the same synthetic data described in § 6.2.1. Such test data feature white noise (wn), and speech (sp) signals convolved with **RIRs** are generated by the acoustic simulator [Schimmel et al. 2009]. Moreover, some recordings are perturbed by external white noise at 10 dB **SNR** (wn+n, sp+n, respectively).

TABLE 10.1: Mean angular error in $^\circ$ (with accuracies (%)) for AOA estimation with 10° and 20° angular tolerance.

TABLE 10.2: Mean angular error in degree (with accuracies (%)) for 2D SSL (azimuth and elevation) with 10° and 20° tolerance.

⁵⁷MBSSLocate^a

► RESULTS FOR AOA ESTIMATION

Mean angular error for **AOA** using the proposed approach and **GCC-PHAT** are presented in Table 10.1.

It can be seen that the two methods yield the same accuracy within a 20° threshold. When a smaller tolerance is considered, **GCC-PHAT** outdoes the proposed approach in accuracy, with comparable errors. This behavior is due to two aspects: first, the synthetic angular spectrum is might be a too simple model; second, since **nRMSE** was chosen as validation metrics, accuracy is not directly optimized. When some external noise is added, performances of both methods degrades. This is a well-known and expected behavior for **GCC-PHAT**.

In Table Table 10.2, the performance of the full 2D-SSL pipeline is shown. Within a tolerance of 20° , the **MIRAGE** model allows estimation of both azimuth and elevation of the target source. However, since the two microphones were free to move in our data, the inclinations of the true and image pairs are rarely flat. While this helps elevation estimation, it reduces the accuracy of predicting the right azimuth. While external noise is again decreasing the accuracy dramatically, it is interesting to notice that our **DNN** model trained and validated with white noise sources somewhat generalizes to speech sources.

With this work, we demonstrated how a simple echo model could allow 2D SSL with only two microphones, using simulated data. In the following section, we will discuss a real-world application of the **MIRAGE** framework to larger microphone arrays.

10.5.2 Multi-channel synthetic-data scenario

In this section, we compare the **SRP-PHAT** algorithm (using GCC-PHAT for TDOA estimation) with the proposed approach, **MIRAGE**, on multichannel synthetic data generated with the Python library `pyroomacoustics`⁵⁸. This time we use the more robust **DNN** model based on **CNN** trained with the Gaussian-based log-likelihood of Eq. (6.8), dubbed here as **MIRAGE – CNN_{V_N}**. The data are generated to match the design of the HARU’s microphone array placed on top of a table: the 7 microphones are at most 30 cm from the close-surface, placed 13 cm from each other; the other walls’ absorption coefficients are uniformly sampled in (0.5, 1), and the one of the close-surface is in (0, 0.5). In the next paragraph, the two methods are compared for **TDOA** estimation task, while in the following discussion, the performances for 2D-SSL.

⁵⁸[pyroomacoustics](#)

► TDOA ESTIMATION ON SYNTHETIC DATA

For this comparison, 200 different audio scenes have been generated, similarly to § 6.2.1. The following metrics have been used: **nRMSE**, **RMSE** and Standard Deviation (**STD**). In Table 10.3 TDOA estimation errors are presented. From these results, the proposed approach outperforms the baseline for both speech and noise data. For better understanding, all the error for **TDOA** estimation are computed in samples, that is they are all multiplied by the constant sampling frequency of the signals (16 kHz). As an additional metric, the empirical computational time of the proposed approach is more than 10 times smaller than the baseline.

	signal	nRMSE	RMSE	STD	time
MIRAGE – CNN_{V_N}	noise	0.09	0.29	0.26	0.19
GCC-PHAT	noise	0.26	0.26	1.04	2.40
MIRAGE – CNN_{V_N}	speech	0.37	1.03	1.09	0.20
GCC-PHAT	speech	0.88	2.08	2.86	2.51

TABLE 10.3: Evaluation metrics (nRMSE, RMSE, STD) for TDOA estimation and empirical computational time for different source signals. In bold the best records.

In Figure 10.8 the **TDOA** estimation error is illustrated against the **SNR** level of the recordings. When the source signal is noise, both methods yield similar mean error independently of the noise level. However, **GCC-PHAT** seems to produce more outliers lower **SNR**. On the other hand, when data is speech, for both the approach, the performance decrease as the SNR reduces; however, **MIRAGE** gives statistically better results in term of standard deviation and outliers.

► 2D-SSL ESTIMATION ON SYNTHETIC DATA

DOA estimation errors using the proposed approach and **SRP-PHAT** are presented in Table 10.4. For white noise source signals, **SRP-PHAT** has better performances for both elevation and azimuth but with comparable errors. However, the proposed approach outperforms the baseline when the emitted

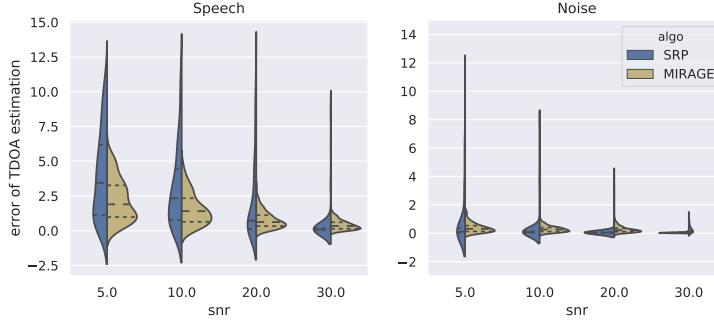


FIGURE 10.8: Violin-plots of the TDOA errors (in samples) versus SNR ranges (in dB) for speech and noise source signal. Ticks in the SNR axis indicates the upper limit of the following ranges: (0, 5], (5, 10], (10, 20], (20, 30] dB.

signal is speech.

	signal	Error θ	Error ϕ
MIRAGE – CNN \mathcal{V}_N	noise	1.29 ± 1.17	2.30 ± 3.35
SRP-PHAT	noise	0.49 ± 0.61	1.70 ± 1.42
MIRAGE – CNN \mathcal{V}_N	speech	9.51 ± 15.84	12.26 ± 12.20
SRP-PHAT	speech	35.27 ± 54.57	15.10 ± 16.67

TABLE 10.4: Mean squared errors and standard deviation in degrees for azimuth (θ) and elevation (ϕ). In bold the best records.

In Figures 10.9 and 10.10b, the DOA estimation results are reported as scatter-plots in a prediction-vs-ground-truth plane. When the test data contains noise, both the methods perform the same regardless of the SNR level. When speech data are considered, MIRAGE outperforms the baseline, which suffers at a low level of SNR. However, the performance of both of the method drops for elevation estimation. Unfortunately, this seems to contradict the good performances on TDOE estimation, and further investigation is needed to explain this observation.

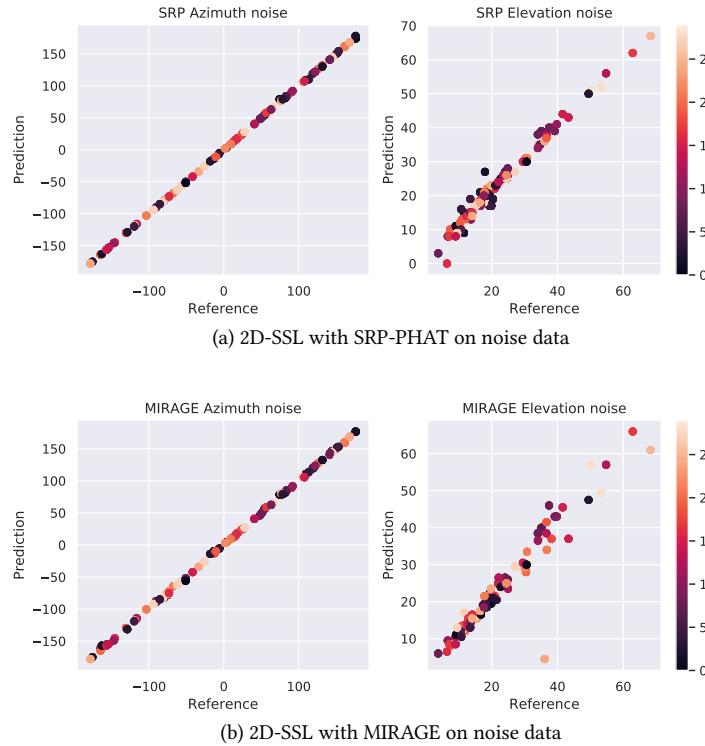


FIGURE 10.9: Scatter-plots (predictions-vs-ground-truth) for DOA estimation (azimuth and elevation) on synthetic data when the source signal is noise. The color map corresponds to different SNR level [dB] in the data.

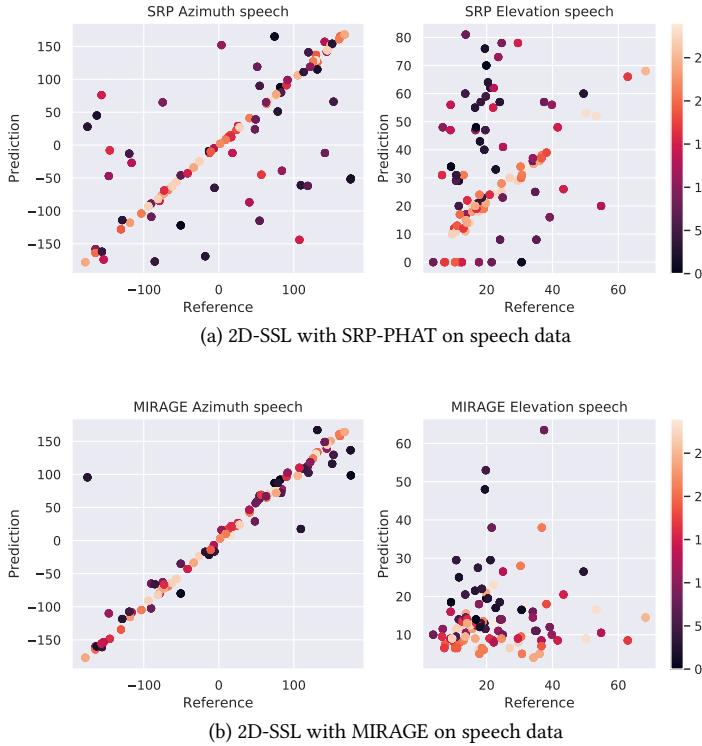


FIGURE 10.10: Scatterplots (predictions-vs-ground-truth) for DoA estimation (azimuth and elevation) on synthetic data when the source signal is **speech**. The color map corresponds to different SNR level [dB] in the data.

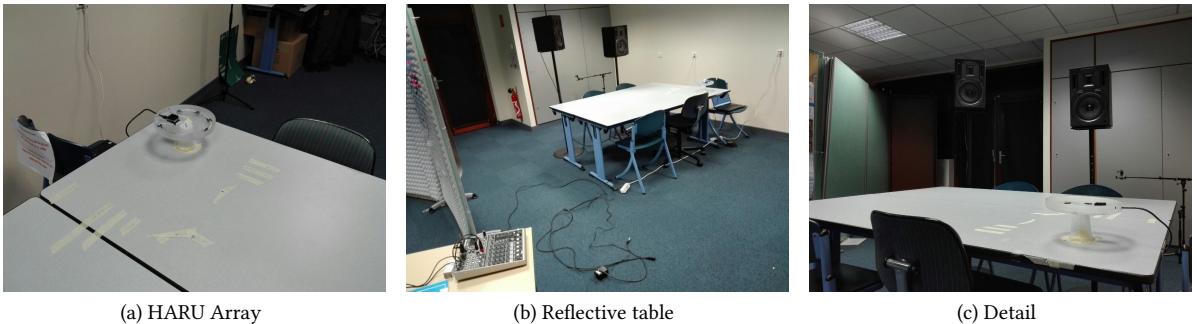


FIGURE 10.11: Picture of the room and setup for recording real multichannel data with the HARU circular microphone array.

10.5.3 Multi-channel real scenario

In this section, we will analyze the two methods on real recordings. The real multichannel data were recorded with the HONDA's HARU microphone circular array ($I = 7$). For the sake of simplicity, hereafter, we will denote this array as HARU. The experiments were performed in a big office room $10 \text{ m} \times 15 \text{ m} \times 3 \text{ m}$ with a reverberation time around 0.2 seconds. The HARU was placed on top of a table with a height of 0.10 m to simulate the close-reflector scenario, used to train the DNN model. The room setup is shown in § 10.5.3. Two loudspeakers were used to emit one anechoic and normalized utterance from the TIMIT dataset and 10 seconds of white noise. The dataset consists of 5 different azimuthal positions and for each of them 2 different elevations yielding 10 different locations in space. The geometry of the setup was annotated using metric tape measures.

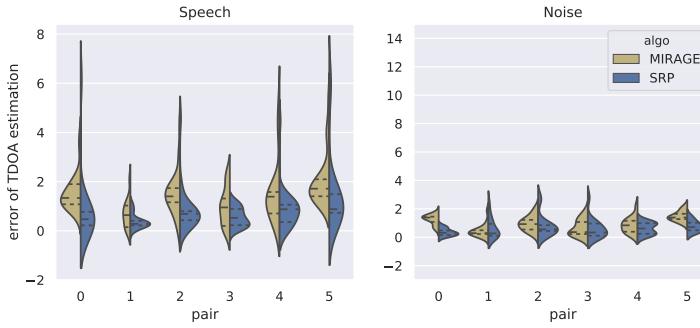


FIGURE 10.12: Violin-plots of the TDOA estimation errors (in samples) versus the microphone pairs in HARU for two different source signals on real data

► TDOA ESTIMATION AND 2D-SSL PERFORMANCES ON REAL DATA

As reported in Table 10.5, the two methods are comparable both for speech and noise emitted signals even if SRP-PHAT performs slightly better. In Figure 10.12 the error on TDOA estimation is shown of each microphone pair of the HARU. It can be seen that the error and the deviation are not homogeneous among the pairs. This might be due to some perturbations of the array’s microphone positioning: the SRP-PHAT method is only successful if the array’s geometry is perfectly known *a priori*. However, little misplacement leads to local distortion in the input angular spectra⁵⁹. Moreover, the proposed approach needs the height of the robot (of each pair) as additional information, and again some perturbation can affect performances.

	signal	nRMSE	RMSE	STD	time
MIRAGE – CNN γ_N	noise	0.26	0.94	0.57	0.26
SRP-PHAT	noise	0.20	0.62	0.57	6.48
MIRAGE – CNN γ_N	speech	0.40	1.38	0.99	0.24
SRP-PHAT	speech	0.32	0.82	1.08	4.02

⁵⁹ In[Salvati et al. 2018], the authors address this problem integrating SRP-PHAT with a CNN together.

TABLE 10.5: Evaluation metrics (nRMSE, RMSE, STD) for TDOA estimation and empirical computational times for different source signals. Boldness denotes the best records.

Finally, we evaluate the performance of the methods for the 2D-SSL task. The results in terms of RMSE and standard deviation are shown in Table 10.6. When the signal emitted is speech, the CNN-based method outperforms the SRP-PHAT. However, the latter seems to perform better for noise signals, even if comparable error margins are observed.

	signal	Error θ	Error ϕ
MIRAGE – CNN γ_N	noise	1.29 ± 1.17	2.30 ± 3.35
SRP-PHAT	noise	0.49 ± 0.61	1.70 ± 1.42
MIRAGE – CNN γ_N	speech	9.51 ± 15.84	12.26 ± 12.20
SRP-PHAT	speech	35.27 ± 54.57	15.10 ± 16.67

TABLE 10.6: Mean squared errors and standard deviations in degrees for estimation of azimuth (θ) and elevation (ϕ). In bold the best records.

Figures 10.13 and 10.14 illustrate the distribution of the prediction of the methods with respect to the ground-truth in the azimuth-vs-elevation planes. We see that the predictions do not match the ground-truth properly. SRP-PHAT seems to overestimate elevation while predicting well the azimuth, especially for noise signals. MIRAGE seems to return more reasonable azimuth-elevation pairs. However, the elevation prediction seems to be almost constant across the points. Moreover, it seems that there is a constant offset or deviation,

especially for azimuth prediction, suggesting that our real-data annotation was perhaps no accurate enough.

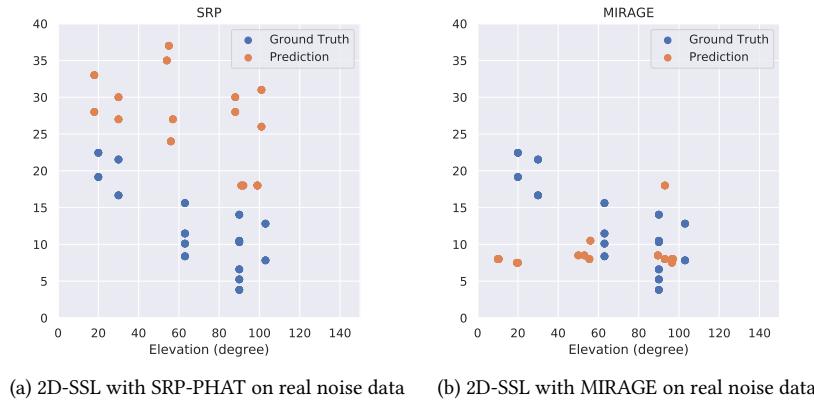


FIGURE 10.13: Scatterplots (azimuth-vs-elevation) for DOA estimation on real data when the source signal is **noise**. The colors indicate reference points (blue) and predicted ones (orange).

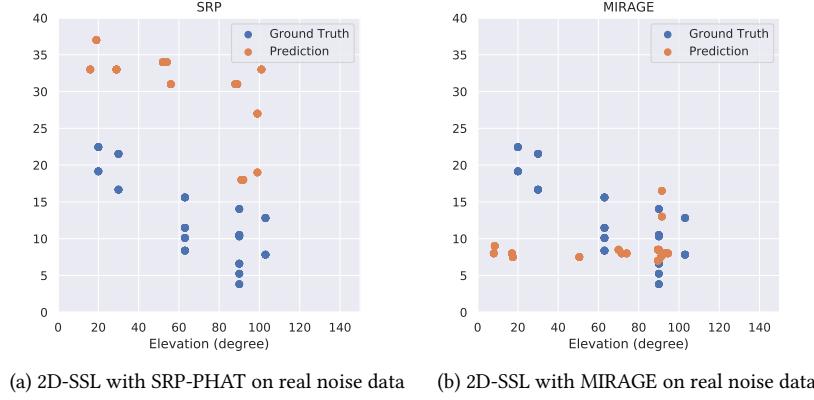


FIGURE 10.14: Scatterplots (azimuth-vs-elevation) for DoA estimation on real data when the source signal is **speech**. The colors indicate reference points (blue) and predicted ones (orange).

10.6 CONCLUSION

This chapter demonstrated how a simple echo model could boost an SSL algorithm in strongly echoic scenarios when microphones are placed close to a reflector. Instead of integrating the physical equation into their algorithm, we proposed to use a successful algorithm for multichannel SSL on the virtual array created accounting for image microphone. In order to create such an array, the echoes' parameters need to be estimated. To this end, we use the learning-based acoustic echo retrieval methods proposed in [Chapter 6](#).

Preliminary results on synthetic data for stereophonic recordings prove the effectiveness of the proposed approach. However, results obtained on real data reveal that the task is still very challenging for both the proposed and baseline methods. Considering the current knowledge, this is the first time an echo-aware method combines both knowledge-driven and data-driven for sound source localization. The learning approach could still be significantly improved by considering other acoustic features (such as advance [ReTF](#) methods), other architectures, and other challenges. For instance, handling the missing frequencies of the speech while training on a broadband signal such

as in [Gaultier et al. 2017], using physical-driven regulation such as in [Nabian and Meidani 2020] and make the learning step independent to the array geometry. Moreover, further investigation is needed to strengthen these results and further improve the robustness of the learned mapping, for which the dEchorate may be used as a valuable testing dataset.

11

Echo-aware Applications of dEchorate

- ▶ **SYNOPSIS** This chapter presents two echo-aware applications that can benefit from the dataset dEchorate. In particular, we exemplify the utilization of these data considering two possible use-cases: echo-aware speech enhancement (§ 11.1) and room geometry estimation (§ 11.2). This investigation is conducted using state-of-the-art algorithms described and contextualized in the corresponding sections. In the final section (§ 11.3), the main results are summarized, and future perspectives are presented.

This chapter is the continuation of the work presented in [Chapter 7](#). Therefore, it is the results of the collaboration with prof. Sharon Gannot and ing. Pinchas Tandteinik at the Bar’Ilan University, Israel. The algorithms presented here are straightforward extensions of the one available in the literature. Nevertheless, they are presented according to the thesis notation. In addition, they are gathered and implemented in the following Python library available online: **dEchorate** related to the **DECHORATE** dataset, **Risotto** for **RIR** estimation and **Brioche** for echo-aware beamforming. A description of these libraries is reported in the ??.

11.1 ECHO-AWARE SPATIAL FILTERING

In the previous chapters, we showed how to integrate echoes for sound source separation ([Chapter 9](#)) and sound source localization ([Chapter 10](#)). In this section, we investigate this in the context of spatial filtering. To this end, we compare two types of spatial filters: echo-agnostic and echo-aware beamformers. In order to study their empirical potential, we will evaluate their performances on both synthetic and measured data, as available in the dEchorate dataset ([Chapter 7](#)). For all the methods presented in this part of the thesis, we assume that echoes are known. To this end, we used the annotations that come with the considered dataset.

11.1.1 *Literature review*

The following paragraphs provide a broad overview of existing beamforming methods, with a specific focus on how they handle echoes. Spatial filtering methods exist in many forms, one of the most popular of which being *beamforming*.

Keywords: Early reflection, Speech Enhancement, Beamforming, Room Geometry Estimation, Reflector Localization.

Resources:

- [dEchorate](#)
- [Risotto](#)
- [Brioche](#)

- ▶ ECHO-AGNOSTIC BEAMFORMERS do not need any echo-estimation step: they either ignore their contributions, such as the direct-path beamformers [Van Trees 2004], or they consider coupling filters between pairs of microphones, using so-called Relative Impulse Response (**ReIR**) [Gannot et al. 2001]. In their vanilla form, neither approaches compute the full acoustic channels explicitly. In case of direct-path Delay-and-Sum (**DS**) beamformers, only the **DOA** of the target source is used to build the so called (relative) steering vector. Then, in order to cope with distortions due to reverberation, external noise or interfering speakers, the statistical description of such forms of noise can be included in extended beamformer design, such as the **MVDR** beamformer [Van Trees 2004].

The **ReIRs** (and their frequency counterparts, **ReTF**) have been introduced with the explicit purpose of avoiding the computation of the acoustic channel related to each microphone [Gannot et al. 2001]. **ReIRs**-based beamformers instead of returning the dry source signal, return the reverberant source spatial image as it is recorded at a reference microphone. Compared to the difficult task of estimating the acoustic channels and of relying on bad channel estimates, this is typically sufficient for achieving good enhancement performances in many practical scenarios.⁶⁰ Since then, **ReIRs** have been incorporated in powerful beamforming algorithms, used for both dereverberation and noise reduction (e.g. [**Schwartz2014multi**; Kodrasi and Doclo 2017]).

- ▶ ECHO-AWARE BEAMFORMERS explicitly model multipath sound propagation instead. They can be seen as *rake receivers*, borrowing the idea from telecommunication where an antenna *rakes* (*i.e.* combines) coherent signals arriving from different propagation paths. In particular, they consider “extended” steering vectors, whose formulation uses known echo delays and attenuations [**Jan1995matched**; Flanagan et al. 1993]. The underlying motivation is two-fold: on the one hand, they better describe the acoustic propagation of the source signal; on the other hand, the early echoes’ energy is included in the estimated signal and not considered noise to be removed. Later, this approach has been extended to enhance desired signals in the context of the *cocktail party problem* in [Dokmanić et al. 2015] and for noise and reverberation reduction [Peled and Rafaely 2013; Javed et al. 2016; Kowalczyk 2019]. In particular, the work [Peled and Rafaely 2013] estimates early echoes in the spherical harmonic domain and uses their **DOAs** to build **ReIR**. However, this approach is not generalizable to all microphone array configurations as it requires the deployment of (3D) spherical arrays. Alternatively, the authors of [Dokmanić et al. 2015b] (with its extension to the time-domain [Scheibler et al. 2015]) propose to modify the original formulation of the **DS** and **MVDR** beamformer designs to include the knowledge of the echoes as image sources. While that study covers the case of interferer and noise reduction, the late reverberation reduction is not considered, which was instead addressed by the work in [Kowalczyk 2019].

In this section, we compare the beamformer designs proposed in [Kowalczyk 2019] for both noise and late reverberation reduction. Besides, we take a different perspective: we investigate the benefit of echo knowledge when using either synthetic or measured impulse responses.

⁶⁰ Note that, as opposed to channel estimation, estimating the **ReTF** is a non-blind problem (See ??)

11.1.2 Background in spatial filtering

Given the narrowband STFT signal model presented in § 3.2.5, the signals captured by I microphones listening to a single sound source ($J = 1$) in a noisy reverberant room reads:

$$\mathbf{X}[k, l] = \mathbf{H}[k]S[k, l] + \mathbf{N}[k, l], \quad (11.1)$$

where $\mathbf{X}[k, l], \mathbf{H}[k, l], \mathbf{N}[k, l] \in \mathbb{C}^I$ and $S[k, l] \in \mathbb{C}$. Note that since only one sound source is considered ($J = 1$), for a given TF bin, the mixing matrix reduces to the vector $\mathbf{H}[k]$ and the source contribution reduces to the complex scalar $S[k, l]$. Hereafter, we omit the dependency on the discrete time-frequency bin $[k, l]$ for the sake of clarity.

The filter vectors can be decomposed in order to highlight the sound propagation components, that is,

$$\mathbf{H} = [H_i^{\text{dp}} + H_i^{\text{ee}} + H_i^{\text{lr}}]_i \quad (11.2)$$

where the summands correspond to the direct-path (dp), early echoes (ee) and late reverberation (lr), respectively. We can now model the early part of the RIR associated to the i -th channel according the echo model, that is,

$$H_i^{\text{dp}} + H_i^{\text{ee}} = \sum_{r=0}^R \alpha_i^{(r)} e^{-i2\pi f_k \tau_i^{(r)}}, \quad (11.3)$$

where the $r = 0$ is the index of the direct propagation. Given such a decomposition of the RIRs, the vector \mathbf{X} can be expressed accordingly as:

$$\mathbf{X} = \mathbf{X}^{\text{dp}} + \mathbf{X}^{\text{ee}} + \mathbf{X}^{\text{lr}} + \mathbf{N}. \quad (11.4)$$

In the context of echo-aware spatial filtering, the source signal of interest includes both the direct path and the R early reflections, as done in [Dokmanić et al. 2015b; Kowalczyk 2019]. This assumption is motivated by psychoacoustics studies as discussed in § 2.4: the first early echoes are shown to contribute to increasing speech intelligibility, as they are fully integrated into the direct sound, increasing its perceived intensity. Based on this, we define as the signal of interest the following

$$\mathbf{X}_s = \mathbf{X}^{\text{dp}} + \mathbf{X}^{\text{ee}} = [H_i^{\text{dp}} + H_i^{\text{ee}}]_i S. \quad (11.5)$$

The noise and the late reverberation are typically described as random processes for which priors are typically available. Therefore, it is common to express the microphone signal model of Eq. (11.4) from a statistical point of view. Under the assumption of source and noise signals being statistically independent, we can define the covariance matrix of the microphone signals, $\Sigma_x = \mathbb{E}\{xx^H\} \in \mathbb{C}^{I \times I}$, as

$$\Sigma_x = \mathbf{H} \Sigma_s \mathbf{H}^H + \Sigma_s^{\text{lr}} + \Sigma_n, \quad (11.6)$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator. Here Σ_s^{lr} and Σ_n denote the Power Spectral Density (PSD) matrices of the late reverberation and noise, respectively. We will describe each term in turn.

- ▶ THE SOURCE covariance matrix Σ_s is assumed here to be diagonal, since we assume that all the spatial information is expressed by the filters \mathbf{H} and the source signals are independent to each other, that is,

$$\Sigma_s = \sigma_s^2 \mathbf{I}, \quad (11.7)$$

where \mathbf{I} is the identity matrix of dimension $I \times I$ and $\sigma_s^2 = \mathbb{E}\{|s|^2\}$.

- ▶ THE LATE REVERBERATION covariance matrix can be estimated using the time-invariant spatial coherent matrix model proposed in [Kuster 2012], based a diffuse sound field model [Kuttruff 2016b].

$$\Sigma_x^{lr} = \sigma_{lr}^2 \mathbf{\Gamma}, \quad (11.8)$$

where σ_{lr}^2 denotes the power of the late reverberation and $\mathbf{\Gamma}$ is the $I \times I$ spatial coherence matrix, which is available in closed-form.⁶¹ This approach has been found successful in many dereverberation application [Naylor and Gaubitch 2010; Cauchi et al. 2014; Tammen et al. 2018].

- ▶ THE ADDITIVE NOISE is assumed to be stationary over the recording. Therefore its covariance matrix can be easily estimated from the observation using non-speech segments. In a blind setting, this would require the usage of a voice activity detector. Alternatively, the noise component can be modeled as a stationary random process whose spatial covariance matrix can be estimated with advanced techniques.

⁶¹ Given the distance $d_{ii'}$ between to microphone i and i' , the interchannel coherence function in the continuous-frequency domain writes

$$\tilde{\gamma}_{ii'}(f) = \frac{\sin(2\pi d_{ii'}/c)}{2\pi d_{ii'}/c}.$$

Then, the matrix $\mathbf{\Gamma}$ is built by computing the $\tilde{\gamma}$ for all the pairs of channel on a discrete set of frequencies.

11.1.3 Elements of Beamforming

In the STFT domain, a beamformer forms a linear combination of the microphone channels to yield the desired output $Y \in \mathbb{C}$:

$$Y = \mathbf{W}^H \mathbf{X} = \mathbf{W}^H \mathbf{H} S + \mathbf{W}^H \mathbf{N},$$

where the vector $\mathbf{W} \in \mathbb{C}^I$ contains the beamformer weights. These weights can be computed by optimizing different design criteria which will be described below.

- ▶ THE DELAY-AND-SUM is the simplest and often a quite effective beamformer. In its vanilla version, the DS is designed to only compensate the propagation delay from the source to the microphones along to the ideal propagation path. Assuming the far field scenario and $i = 0$ to be the reference microphone, this is typically achieved using the direct-path relative steering vector \mathbf{D}' defined in Eq. (3.32), that is,

$$\mathbf{D}' = \left[1, e^{-i2\pi f_k \tau_{i+1}^{(0)}/c}, \dots, e^{-i2\pi f_k \tau_I^{(0)}/c} \right] \quad (11.9)$$

where f_k is the k -th frequency bin in Hz, $\tau_i^{(0)}$ is the TOA of the direct-path for channel i , and c is the speed of sound.

Therefore, the beamformer weights reads

$$\mathbf{W}_{dp} = \frac{\mathbf{D}'}{\|\mathbf{D}'\|}, \quad (11.10)$$

where $\|\cdot\|$ denotes the Euclidean norm.

- ▶ THE MVDR beamformer optimizes the following criterion⁶²

$$\mathbf{W}_{\text{MVDR}} = \arg \min_{\mathbf{W}} \left\{ \mathbf{W}^H \boldsymbol{\Sigma}_u \mathbf{W} \text{ s.t. } \mathbf{W}^H \mathbf{H} = 1 \right\}. \quad (11.11)$$

It aims at minimizing the total output energy (i. e., minimum variance) while simultaneously keeping the unit gain of the array on the desired signal fixed (i. e. distortionless response). Therefore, the reduction of the output energy suppresses any external noise modeled by $\boldsymbol{\Sigma}_u$.

The Least-Square minimization through the Lagrangian multiplier method yields the following closed-form optimal solutions

$$\mathbf{W}_{\text{MVDR}} = \frac{\boldsymbol{\Sigma}_u^{-1} \mathbf{H}}{\mathbf{H}^H \boldsymbol{\Sigma}_u^{-1} \mathbf{H}}. \quad (11.12)$$

Note that these techniques do not require any reference signal, only the knowledge of the source's filter \mathbf{H} and an estimate of the observed signal's covariance matrix. This criterion design can be easily extended to work with any type of noise and acoustic transfer functions modeled by $\boldsymbol{\Sigma}_u$ and \mathbf{H} , respectively.

11.1.4 Noise, steering vectors, rake filters, and relative transfer functions

The vectors \mathbf{H} between the source and the I microphones account for the RTFs. To overcome the complexity of estimating them, three main directions have been pursued: steering vectors, rake receivers and relative transfer functions.

- ▶ STEERING VECTORS are the RTFs of the ideal propagation path component, namely, the two are equivalent in anechoic scenarios. They can be easily built on the knowledge of the TOAs of the source signal and their integration in the MVDR criterion is straightforward: the \mathbf{H} simply identifies with the corresponding relative steering vector \mathbf{D} , that is,

$$\mathbf{H}_{\text{DP}}[k] = \mathbf{D}[k] = \left[e^{-i2\pi f_k \tau_i^{(r)}} \right]_i. \quad (11.13)$$

In distant-talking scenarios, relative time delays, or TDOAs, with respect to a reference microphone, are used. Moreover, knowing the microphone array geometry, the TDOAs may be derived from the source's DOA⁶³, thus, the steering vectors can be easily computed. However, DOAs need to be estimated aside, using Sound Source Localization (SSL) methods (See § 8.4).

- ▶ THE RAKE FILTERS are beamformers that explicitly deal with the multipath propagation model in Eq. (11.3). To this end, the MVDR design is modified by integrating the spatial information of R reflections into the distortionless constraint. This turns out to be equivalent to extending the definition of (relative) steering vectors as follows:

$$\mathbf{H}_{\text{RAKE}}[k] = \left[\sum_{r=0}^R \alpha_i^{(r)} e^{-i2\pi f_k \tau_i^{(r)}} \right]_i. \quad (11.14)$$

As before, both the echoes' delays and amplitudes are considered relatively to the reference microphone.

⁶² The MVDR design is equivalent to the Minimum Power Distortionless Response (MPDR) beamformer which minimize the following

$$\mathbf{W}_{\text{MPDR}} = \arg \min_{\mathbf{W}} \left\{ \mathbf{W}^H \boldsymbol{\Sigma}_x \mathbf{W} \text{ s.t. } \mathbf{W}^H \mathbf{H} = 1 \right\}.$$

However, it exhibit higher sensitivity to misalignment errors than the MVDR beamformer [Gannot et al. 2017, Section V.A].

⁶³ The mapping between TDOAs and DOAs is not always unique. It depends on the microphone array geometry, such as its compactness, its shape, and the number of sensors.

- THE RELATIVE TRANSFER FUNCTION (**ReTF**) was originally proposed in the work of [Gannot et al. 2001] to overcome the limitation of accessing the full **RTF** for each channel. Given the **RTF** of the i -th channel \mathbf{h}_i , its **ReTF** with respect to the first channel is given by

$$\mathbf{H}_{\text{ReTF}}[k] = \frac{\mathbf{H}[k]}{H_1[k]} \quad (11.15)$$

In a reverberant environment, it contains both direct-path information and information representing early and late reverberations. More details are given in § 3.3.2.

- THE NOISE CONTRIBUTION is taken into account through the covariance matrix Σ_u . This term could potentially include every source of “noise”, such as interfering sources, measurement noise as well as diffuse background noise. As long as the power spectral density of the modelled noise source is available, its suppression can be achieved by including it in Σ_u and using it in Eq. (11.12).

⁶⁴ For instance,

$$\Sigma_u = \Sigma_{s_q} + \Sigma_n, \quad (11.16)$$

where the summands are the covariance matrix of an interfering source q and of the an independent background noise n .

Assuming stationarity of the noise sources, a naïve approach to estimate their contribution is to use voice activity detection or speaker diarization⁶⁵ tools. If these excerpts are long enough, the whole covariance matrix Σ_u , including both noise and interferer, is well approximated by its sampled version whose calculation is straightforward. Alternatively, other designs can be used, for instance, in the **LCMV** beamformer, interference reduction is achieved using steering vectors for all interfering sources of the interfering source instead of their **PSD** matrices.

- THE LATE REVERBERATION contribution Σ_s^{lr} is not directly available in isolated audio segments as it is “glued” to the target signal. A common way to achieve dereverberation in a beamformer design [Schwartz et al. 2014; Thiergart et al. 2014; Kowalczyk 2019] is by adding the **PSD** of the late diffuse part of Eq. (11.8) to the noise covariance matrix, that is,

$$\Sigma_{\text{noise-late}} = \Sigma_u + \Sigma_s^{\text{lr}} = \Sigma_u + \sigma_{\text{lr}}^2 \boldsymbol{\Gamma}. \quad (11.17)$$

11.1.5 Considered beamformers

In this work we evaluate the performance of echo-agnostic and echo-aware beamformers for noise and late reverberation suppression. Table 11.1 summarizes the considered beamformers designs.

In this work, the elements used to build the beamformers are estimated as follows:

- the noise covariance matrix, Σ_u , is estimated from 0.5 second excerpt of diffuse noise only audio segments (this is equivalent to having access to an ideal voice activity detector);

⁶⁴ This type of criterion design is more properly known as **MaxSINR** or **MaxSNR** beamformers. In general, they aim at optimizing directly the **SNR** or the Signal-to-Interference-plus-Noise-Ratio (**SINR**) metrics. Nevertheless, it can be shown that the any **MaxSINR** (or **MaxSNR**) can be identified with an **MVDR** if the distortionless constraint is satisfied [Gannot et al. 2017].

⁶⁵ Speaker diarization is the problem of partitioning an audio signal into segments according to the source identities. In other words, it addresses the problem of “who is speaking when”.

Acronym	Description	Beamforming weights
DS	Align delayed copies of the signal at microphones	$\mathbf{W} = \mathbf{H}_{\text{DP}} / \ \mathbf{H}_{\text{DP}}\ $
MVDR-DP	$\min \mathbf{W}^H \Sigma_u \mathbf{W}$ s.t. $\mathbf{W}^H \mathbf{H}_{\text{DP}} = 1$	$\mathbf{W} = (\mathbf{H}_{\text{DP}}^H \Sigma_u^{-1} \mathbf{H}_{\text{DP}})^{-1} \Sigma_u^{-1} \mathbf{H}_{\text{DP}}$
MVDR-ReTF	$\min \mathbf{W}^H \Sigma_u \mathbf{W}$ s.t. $\mathbf{W}^H \mathbf{H}_{\text{ReTF}} = 1$	$\mathbf{W} = (\mathbf{H}_{\text{ReTF}}^H \Sigma_u^{-1} \mathbf{H}_{\text{ReTF}})^{-1} \Sigma_u^{-1} \mathbf{H}_{\text{ReTF}}$
MVDR-Rake*	$\min \mathbf{W}^H \Sigma_u \mathbf{W}$ s.t. $\mathbf{W}^H \mathbf{H}_{\text{RAKE}} = 1$	$\mathbf{W} = (\mathbf{H}_{\text{RAKE}}^H \Sigma_u^{-1} \mathbf{H}_{\text{RAKE}})^{-1} \Sigma_u^{-1} \mathbf{H}_{\text{RAKE}}$
MVDR-DP-Late	$\min \mathbf{W}^H (\Sigma_u + \Sigma_{1r}) \mathbf{W}$ s.t. $\mathbf{W}^H \mathbf{H}_{\text{DP}} = 1$	$\mathbf{W} = (\mathbf{H}_{\text{DP}}^H (\Sigma_u + \Sigma_{1r})^{-1} \mathbf{H}_{\text{DP}})^{-1} (\Sigma_u + \Sigma_{1r})^{-1} \mathbf{H}_{\text{DP}}$
MVDR-ReTF-Late	$\min \mathbf{W}^H (\Sigma_u + \Sigma_{1r}) \mathbf{W}$ s.t. $\mathbf{W}^H \mathbf{H}_{\text{ReTF}} = 1$	$\mathbf{W} = (\mathbf{H}_{\text{ReTF}}^H (\Sigma_u + \Sigma_{1r})^{-1} \mathbf{H}_{\text{ReTF}})^{-1} (\Sigma_u + \Sigma_{1r})^{-1} \mathbf{H}_{\text{ReTF}}$
MVDR-Rake-Late*	$\min \mathbf{W}^H (\Sigma_u + \Sigma_{1r}) \mathbf{W}$ s.t. $\mathbf{W}^H \mathbf{H}_{\text{RAKE}} = 1$	$\mathbf{W} = (\mathbf{H}_{\text{RAKE}}^H (\Sigma_u + \Sigma_{1r})^{-1} \mathbf{H}_{\text{RAKE}})^{-1} (\Sigma_u + \Sigma_{1r})^{-1} \mathbf{H}_{\text{RAKE}}$

TABLE 11.1: Summary of the considered beamformers. (*) denotes echo-aware beamformers.

- the ReTF, \mathbf{H}_{ReTF} , is estimated from the observed signal using the Generalized Eigenvalue Decomposition (GEVD) method described in [Doclo and Moonen 2003];
- \mathbf{H}_{DP} , \mathbf{H}_{RAKE} are computed using known relative delays and amplitudes available in the DECHORATE dataset;
- the late reverberation power, σ_{1r}^2 , is estimated in closed-form knowing the filters \mathbf{H} and the noise PSD, as suggested in [Schwartz et al. 2016]. Recently, in [Tammen et al. 2018] it was proposed an iterative Least-Square approach to estimate the late PSD together with the Relative Early Transfer Functions (ReETFs). This work is based on [Kodrasi and Doclo 2017], where the effectiveness of such approach was shown..

11.1.6 Experimental evaluation

The performances of the different designs are compared on the task of enhancing a target speech signal in a 5-channel mixture using a linear array from the dEchorate dataset. In particular, they are tested in scenarios featuring high reverberation and diffuse babble noise, appropriately scaled to given pre-defined signal-to-noise ratio ($\text{SNR} \in \{0, 10, 20\}$). Using the dEchorate data, we considered the room configuration 011111 ($\text{RT}_{60} \approx 600$ ms) and all the possible combinations of target/array's positions. Both real and matching synthetic RIRs are used, which are then convolved with anechoic utterances from the WSJ corpus and corrupted by recorded diffuse noise. The synthetic RIRs are computed with the `pyroomacoustics` Python library, based purely on the Image Source Method (ISM).

The evaluation is conducted similarly to the one in [Kowalczyk 2019]. Here we consider the first microphone as the reference one and the clean target signal as the clean signal convolved with the early part of the RIR (up to the R -th echo), namely, \mathbf{X}_s in Eq. (11.5). For evaluating the performances, we consider the following metrics:

- THE SIGNAL-TO-NOISE-PLUS-REVERBERATION IMPROVEMENT (ΔSNRR) in [dB] is the difference between the input SNRR_i at the reference microphone and the SNRR_o at the filter output. Denoting with X_1 the signal at the reference microphone and $X_{s,1}$ the target speech at the same microphone (See Eq. (11.5)),

these quantities are defined as follows:

$$\begin{aligned} \text{SNRR}_i &= 10 \log_{10} \left(\frac{\sigma_{X_{s,1}}^2}{\sigma_{X_1}^2 - \sigma_{X_{s,1}}^2} \right) \quad [\text{dB}] \\ \text{SNRR}_o &= 10 \log_{10} \left(\frac{\sigma_{W^H X_s}^2}{\sigma_{W^H X - \sigma_{W^H X_s}^2}^2} \right) \quad [\text{dB}] \\ \Delta \text{SNRR} &= \text{SNRR}_o - \text{SNRR}_i \quad [\text{dB}] \end{aligned} \quad (11.18)$$

- ▶ THE SPEECH-TO-REVERBERATION-ENERGY-MODULATION RATIO (**SRMR**) IMPROVEMENT (ΔSRMR) is an adimensional and absolute (i. e., it does not require the reference signal) measure of dereverberation and was initially proposed in [Falk et al. 2010]. It is based on the *modulation spectrum*, which allows for reliable characterization of the speech envelope smearing due to reverberation [Santos and Falk 2014].⁶⁶ Later it has been applied in several works addressing dereverberation as well as in some challenges on speech processing, such as the ACE challenge [Eaton et al. 2015]. It is computed as follows:

- the processed speech signals filtered by a 23-channel Gammatone filterbank to emulate the processing performed by the cochlea.
- the envelope of each output of the filter is computed though Hilbert transform in order to extract the temporal dynamics information (See Eq. (4.2)).
- the modulation spectrum is computed as the 8-bands **STFT** on a selected frequency range (up to 8 Hz) of each envelope.
- the **SRMR** is obtained as the ratio of the average modulation energy content available in the first four modulation bands, to the average modulation energy content available in the higher frequency modulation bands.

An implementation of these metrics is available at the `speechmetrics` Python library.

- ▶ THE PERCEPTUAL EVALUTATION OF SPEECH QUALITY (**PESQ**) is a relative adimensional metrics presented in [Rix et al. 2001] and outputs a score ranging from 1 (bad) to 5 (excellent). This measure, assumed to cover several speech degradation and distortion, was promoted as a standard in the ITU-T recommendation P.862. It was originally used for telecommunication and telephony and it is now considered one of the most reliable metrics to predict the overall speech quality. Practically, after applying an auditory model to the reference and distorted (i. e., the estimated) signals (based on a Bark frequency scale) the loudness spectra are estimated. From the loudness spectra differences, a Mean Opinion Score (MOS)⁶⁷ is inferred using a linear regression model.

An implementation of this metrics is available at the `speechmetrics` Python library.

- ▶ NUMERICAL RESULTS are reported in Figure 11.1. When using synthetic data, the known echo timings perfectly match the components in the simulated **RIRs**, and, likewise, the echo model matches the **RIRs**' early part. Here, one can see

⁶⁶ The modulation spectral domain quantifies the envelop fluctuations that can be attributed to perceptive room acoustics characteristics, such as late reverberation levels and coloration (See § 2.4).

⁶⁷ The MOS is a measure used in the domain of Quality of Experience and telecommunications engineering, representing overall quality of a stimulus or system.

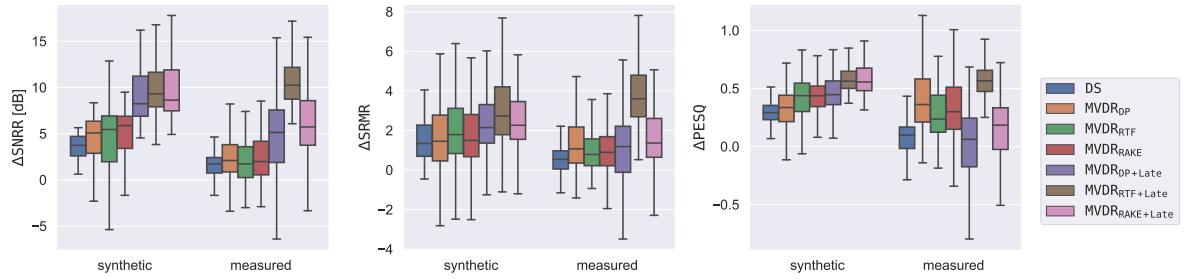


FIGURE 11.1: Comparison of echo-aware beamforming for the room configuration 011111 ($RT_{60} \approx 600$ ms) on measured and synthetic data for all combinations of source-array positions in the dEchorate dataset.

that more information is used, the better performances are. ReTF- and Rake-beamformers outperform the simple designs based on the direct path, and including the late reverberation statistics considerably boosts performance in all cases. Interestingly, ReTF have a slight edge over Rake-versions in terms of mean Signal-to-Noise-plus-Reverberation-Ratio (SNRR). This can be explained by the fact that GEVD methods tend to robustly consider the stronger and more stable components of the ReTF, which in reverberant and noisy static scenarios may identify with the earlier portion of the RIRs. Moreover, since it is not constrained by a fix echo model, the ReTF can capture more information which, in turn, yields to slightly better enhancement. Nevertheless, the PESQ metrics suggest that for this ideal scenario both echo-aware (Rake) and echo-agnostic (ReTF) design are comparable.

When it comes to measured RIRs, the trends are different. Here, the errors in echo estimation, due to calibration mismatch and the richness of the acoustic propagation, lead to a drop in the performances for echo-aware methods, both in terms of means and variances. This is even clearer when considering the Δ PESQ metric, as it also accounts for artifacts. Here, the echo-agnostic beamformer considering late reverberation MVDR_{ReTF+Late} outperforms the other methods, maintaining the trend exhibited on simulated data. In general, it looks like the MVDR_{Rake+Late} has more variance than MVDR_{ReTF+Late}, suggesting that in some situations, it performs better, while in other it performs less well. This is probably due to the tiny annotation mismatch and the complexity of the RIRs and future work will be devoted in a deeper understanding of the underlying factors.

11.2 ROOM GEOMETRY ESTIMATION

In this section, we shortly present another application of the dEchorate dataset: Room Geometry Estimation (RooGE), namely, the task of estimating the shape of a room knowing the positions of first-order image sources. This problem is typically addressed by solving multiple instances of *reflector localization*, aiming at estimating the position of a single surface (e.g. wall, floor, etc.). Several methods have been proposed which take into account different levels of prior information and noise. They were briefly discussed in the context of echo labeling in § 4.3.1. In general, these methods can be decomposed into three successive steps:

1. echo labeling, in order to associate the echoes to image sources using one of the methods mentioned in § 4.3.1;
2. estimation of the image source position either through multilateration [Dokmanić et al. 2013; Dokmanić et al. 2015], Maximum Likelihood (ML) [Tervo 2011] or convex optimization [Crocco et al. 2012];
3. and finally, the *image-source-reversion*, in order to localize the reflector, based on the geometrical assumption of the Image Source Method (ISM).

More advance techniques have been proposed in the literature of reflector localization for different setups and scenarios. A comprehensive review can be found in [Remaggi et al. 2016; Crocco et al. 2017].

Nonetheless, when the echoes' TOAs and their labeling are known for 4 spatially-separated non-coplanar microphones, one can perform this task using closed-form multilateration algorithms.

11.2.1 Room Geometry Estimation through multilateration

Multilateration is the problem of recovering the position of a point in the space from multiple distances between the point and known spatially-separated locations. It is the 3D extension of the *trilateration* problem, namely, determining an unknown position based on the distance to two other known vertices of a triangle. In the context of *RooGE*, the distances from the source to the microphones can be obtained by converting the TOAs [seconds] into distances [meters]. Then, the 3D coordinates of each image source can be retrieved, solving a convex problem as described in [Beck et al. 2008]. Ideally, this problem can be solved in closed-form. However, due to measurement error (e.g. errors in estimating the image's TOAs), it may be ill-conditioned. To overcome this, the algorithm proposed in [Beck et al. 2008] relies on a robust iterative approach yielding accurate solutions. Finally, the position and orientation of each wall can be easily derived from the ISM as the plane bisecting the line joining the real source position and the position of its corresponding image (see Figure 11.2).

11.2.2 Using the dEchorate dataset for *RooGE*

In dEchorate, the annotation of all the first-order images of sound sources is available. We used the Beck et al.'s multilateration method (available in the Python library dEchorate) to estimate the image source position of each of the direct source using 6 non-coplanar microphones (one for each of the 6 arrays). Then, room facets are estimated using each of the sources as a probe. Table 11.2 shows the results of the estimation of the wall positions in terms of distance error (in centimeters) and surface orientation error (in degrees). These metrics were previously used in the literature of reflector estimation, such as in [Annibale et al. 2012; Crocco et al. 2017]. Figure 11.2b depicts an example of reflector estimation using the dEchorate data.

Despite a few outliers, the majority of the facets are estimated correctly in terms of their placement and orientation with respect to the coordinate system. For instance, for source #4, all 6 surfaces were localized within less than 6 cm

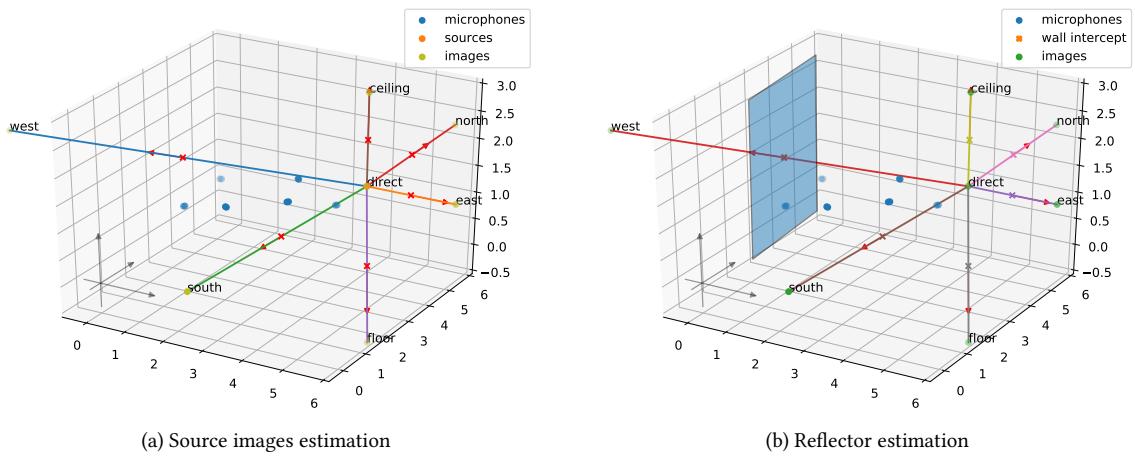


FIGURE 11.2: Images source estimation (right) and corresponding reflector estimation (left) for one of the sound sources in the [DECHORATE](#) dataset.

source id	1	2	3	4
wall	DE	AE	DE	AE
west	0.74	8.99°	4.59	8.32°
east	0.81	0.08°	0.9	0.50°
south	3.94	<i>16.08°</i>	0.18	1.77°
north	1.34	0.76°	1.40	8.94°
floor	5.19	1.76°	7.27	2.66°
ceiling	1.16	0.28°	0.67	0.76°

TABLE 11.2: Distance errors (DE) in centimeters and angular errors (AE) in degrees between ground truth and estimated room sides using each of the sound source (#1 to #4) as a probe. For each facet, bold font is used in correspondence to the source yielding the best DE and AE; while, italic font highlights outliers.

and 2.5° errors. Small errors are due to concurrency of multiple factors, such as tiny offsets in the annotation and the ideal shoebox approximation⁶⁸. It is also possible that for some source-receiver pairs, the far-field assumption is not verified, causing the inaccuracy of *reverting* the **ISM**. Finally, the 2 outliers for the source #3 are due to a wrong annotation caused by the source directivity and misclassification. In particular, when a wall is “behind” the source, the energy of the related reflection is very small and might not appear in the **RIRs**. This happened for the eastern wall, and a second-order image was taken instead. Secondly, the contribution of multiple reflections arriving at the same time can be merged into signal spikes in estimated **RIRs**. This effect is particularly amplified when the microphones and loudspeakers exhibit long impulse responses. As a consequence, some spikes were probably miss-classified. This can be noticed for the southern-wall were again a second-order image was taken instead. Nevertheless, this second type of error can be manually corrected, and the annotations updated.

⁶⁸ In the real recording room, some gaps were present between revolving panels in the walls

11.3 CONCLUSIONS AND PERSPECTIVES

In this chapter, we presented two applications of the dEchorate dataset described in [Chapter 7](#): echo-aware spatial filtering and room geometry estimation.

The first one deals with the possibility of using early echoes to enhance a target speech signal corrupted by diffuse noise and a high level of reverberation. To this end, two types of state-of-the-art spatial filtering criteria are considered:

echo-agnostic and echo-aware beamformers. Experimental results on real and synthetic data, both available in the proposed dataset, led to the following findings. The synthetic data were computed using **ISM**-based simulation; thus, the early parts of **RIRs** match the early echo model. Therefore, replacing the acoustic vectors with few known echoes gives significant enhancement performance gains compared with baseline methods, which consider only the direct ideal propagation. In this scenario, both echo-aware and state-of-the-art **ReTF**-based echo-agnostic perform similarly, suggesting the effectiveness of echo-aware approaches. However, when using the corresponding real data available in the dataset, performances drop in terms of perceptual quality as predicted by the **PESQ** score. This may be due to the small mismatches between real and annotated echoes and the richness of the acoustic field, which impact the echo-aware methods. The best-performing method is the echo-agnostic one based on **ReTF**, which does not suffer from any echo mismatch and can include other information about the acoustic propagation.

The relatively lower performance of rake-based filters on real scenarios than simulated ones emphasizes the importance of having precise enough **AER** algorithms and encourages further studies on these echo-aware methods. Moreover, the knowledge of the very same echoes is limited to spatial filtering and can be used to retrieve the entire room geometry, as demonstrated in the second section of this chapter.

By using standard approaches based on geometrical reasoning and robust multilateration algorithms, it is possible to revert the **ISM** and map echoes' **TOAs** to source and image-source position. Here, we showed this on the **dEchoratedata** both as an application and as a way to validate the dataset. Although the results highlight that some echo **TOAs** have not been correctly classified, the overall annotation is consistent with the actual room planimetry. Finally, we would like to mention that the best of our knowledge, this is the very first dataset of this kind, that is, featuring real data with full echo/image-source annotations.

Future works will explore several directions.

- By making this dataset freely available to the audio signal processing community, we hope to foster research in **AER** and echo-aware to improve the performance of existing methods on real data.
- The dataset could be updated by including more robust annotations derived from more advanced algorithms for calibration and **AER**.
- regarding the applications, the echo-aware methods presented above could be validated over more challenging scenarios than the one presented. Such scenarios, e. g. the presence of interfering sound sources and challenging levels of **SNR** and RT_{60} , are already included in the dataset, but not used in the evaluation above.
- By the amount of the data collected, they can be used to train learning-based signal processing methods or for data augmentation.

In a more long-term perspective, the data analysis conducted in this chapter brings the attention to exploring the impact of mismatch between simulated and real RIR on audio signal processing methods. Moreover, by using the pairs of simulated vs. real RIRs available in the dataset, develop techniques to convert one to the other, using style transfer and domain adaptation techniques.

Part IV

EPILOGUE

12 ECHO-AWARE REFLECTIVE REFLECTION	
12.1 Looking Back	147
12.2 Looking Ahead	148

SLIDING FRANK-WOLFE ALGORITHM & NON-NEGATIVE BLASSO

APPENDICES

DERIVATION OF THE ACOUSTIC WAVE

RIR AND RT60 MEASUREMENTS

RIR estimation	158
RT60 estimation	158

BIBLIOGRAPHY

BIBLIOGRAPHY

12

Echo-aware Reflective Reflection

- ▶ IN THIS THESIS, we studied acoustic echoes for audio scene analysis and signal processing. The two main lines of work can be briefly summarized as follows:
 - A. We investigated new methodologies for *acoustic echo retrieval* (AER) in case of passive stereophonic recordings.
 - B. We re-proposed some fundamentals *audio scene analysis problems* under an echo-aware perspectives.

12.1 LOOKING BACK

After reviewing some useful acoustic notions and presenting signal precessing modeling in **Part I**, the contributions of this thesis were presented in **Parts II** and **III**, developing the two direction above. The support these two goals takes the form of the following methods and artifacts.

- A KNOWLEDGE-DRIVEN METHOD FOR AER dubbed **BLASTER**. This approach enables direct and *off-grid* estimation of echoes' properties in stereophonic passive recordings. In particular, "knowledge" we used is the echo model for the early part of the **RIRs**. Due to its off-grid natures, it should overcome some theoretical limitation of on-grid methods. Although it is currently not outperforming the state-of-the-art, this investigation is motivated by theoretical guaranies.
- A DATA-DRIVEN METHOD FOR AER based on deep learning, dubbed **LANTERN**. Thanks to the availability of powerful acoustic simulators, the properties of the first echoes are estimated using state-of-the-art architectures which are trained in virtually supervised fashion. The proposed model combines results in spatial filtering and understandable deep learning using self-confident measures.
- A ECHO-AWARE DATASET designed for both AER and echo-aware application, dubbed **DECHORATE**. These annotated data should fill the gap between existing dataset and it is designed for validating future echo-aware research. The dataset are accompanied by software utilities to easily access, manipulate, and visualize the data and baseline methods for echo-related tasks.
- A ECHO-AWARE AUDIO SOURCE SEPARATION METHOD, dubbed **SEPARAKE**. It is based on the popular Multichannel NMF framework, which allows

"Some say the end is near.
Some say we'll see Armageddon soon.
Certainly hope we will.
I sure could use a vacation from this [...]"
—Tool, AEnema

simple yet effective integration of the echoes properties. Assuming their knowledge, we can reformulate such a framework in term of image microphones and virtual arrays. Therefore results show how this leads to enough spatial diversity to get a performance boost over the vanilla version of two classic NMF-based algorithm.

- A ECHO-AWARE SOUND SOURCE LOCALIZATION METHOD, dubbed **MIRAGE**. By converting echoes into image microphones, this method allows for source's azimuth and elevation estimation in passive stereophonic recordings. Therefore, the strong echo coming from a close reflective table, can be used to create a virtual array on which powerful array processing techniques can be applied. This methods of simple extention to multi-channels recordings as long as the geometry of the array is available. To this end, we conducted some preliminary studies on a real-world recordings using the microphone arrays of the Honda's Haru robots.
- the following LIBRARIES for echo-aware processing:

↳ dEchorate – code for DECHORATE , Room Impulse Response estimation and annotation.	dEchorate ↗
↳ Risotto – a collection of state-of-the-art methods for estimation of Relative Impulse Response.	Risotto ↗
↳ Brioche – a collection of state-of-the-art beamforming including, but are not limited to, echoes.	Brioche ↗
↳ Blaster – code for BLASTER , its results and related state-of-the-art methods.	Blaster ↗
↳ Separake – code for SEPARAKE including an Python implementation of the Matlab toolbox Multichannel NMF [Ozerov and Févotte 2009] for audio source separation.	Separake ↗
↳ pyMBSSLocate – Python implementation of the Matlab toolbox MBSSLocate [Lebarbenchon et al. 2018b] for sound source localization.	pyMBSSLocate ↗

Taken together, these contributions make a step forward in our ability to estimate and use acoustic echoes in audio signal processing. But much remains to be done.

12.2 LOOKING AHEAD

The work presented in this dissertation took a few steps toward echo-aware audio signal processing. Nevertheless, there are many issue and many potential areas of improvements in the methods presented here. Besides, it is clear that we have only scratched the surface of many problem related to echoes processing. In the following points, we elaborate on some short and long term research possibilities that arise as natural follow-ups to the topics discussed thus far.

► ESTIMATING ACOUSTIC ECHOES

Recently it becomes clear to use that estimating the echoes accurately for arbitrary scenarios is a very challenging task. Using useful prior information, the problem can be easily relaxed, for instance, knowing the microphone array's geometry. The impact of the microphone array (and antenna) geometry is well studied in signal processing communities, such as in telecommunication and array processing. Therefore, the knowledge of array geometry can extend the echo model to closed-form steering vectors, depending on both the echo timing and direction of arrival. This might simplify the parameter space where the echoes are defined. Examples of this approach are the works in [Jensen et al. 2019] for active Acoustic Echo Retrieval (**AER**) and in [Kowalczyk 2019] where such a steering vectors are used.

Besides this, we also noticed that a critical parameter for **AER** is the *number* of such echoes. Therefore the *echo counting* problem seem as reasonable direction to investigate. Estimating reflections knowing theirs number in a received signal is then a different the problem, which can be referred to *sparse coding* problems [Tropp and Wright 2010]. Unfortunately, problems in this class are not necessarily easier to solve, but it was shown they produce more accurate solutions [Bourguignon et al. 2015; Nadisic et al. 2020].

Another important line of study which we did not pursue here is to consider robust acoustic features or enhance the recorded sound with respect to noise and late reverberation, instead of the raw microphone signals. The most promising direction in this regard is to use the recent results on relative early transfer functions estimation. Relative early transfer functions are transfer functions accounting for the early part of the **RTF** and are used for speech enhancement and dereverberation.

Alternatively, statistical models for reverberation that are parametrized by **RT₆₀** or Direct to Reverberant Ratio (**DRR**) can be used. Particularly appealing are the works of Leglaive et al. and Badeau, which define a framework to deal with early reflection with a statistical model. This exploratory direction is also motivated by recent results in estimating these parameters blindly for microphone recordings [Looney and Gaubitch 2020; Bryan 2020]. As these parameters depend on the spatial characterization of the audio scene, using them as priors may seem natural. Finally, statistical model has the potential to solve issues of the strong approximation used with the image source model, such as flat walls.

Learning-based approach and especially Deep Neural Network (**DNN**) are potent tools. Unfortunately, their potential was not fully exploited in this work, but many exciting directions can be perused. Of particular interest is the recent development in physics-driven neural networks [Nabian and Meidani 2020; Rao et al. 2020; Jin et al. 2020], where physics-based layers and regularizers are used to facilitate the learning.

► USING ECHOES

In this thesis, we presented a few selected applications of acoustic echoes in audio scene analysis problems. This approach can be extended to other related problems such as auditory scene analysis, acoustic measurements, and source tracking. Auditory scene analysis puts the focus on humans and their bodies. Humans have two hears and complex auditory systems that provide natural

filtering of the incoming sounds. Moreover, reflection from the *pinnae*, the torso and shoulders are integrated into the processing to provide localization cues [Rascon and Meza 2017]. Therefore, one could imagine studying these reflections and setting us echo-aware processing to design better hearing aids, smart noise-canceling headphones, and auditory models. The latter ones are used to conduct experimental studies on sound perception [Barumerli et al. 2018], build perceptual evaluation metrics.

During the data collection for dEchorate, we experienced some limitations of existing tools for measuring room impulse response. Recent acoustic measurement tools and software started to provide insight into reflection patterns for which **AER** method could improve accuracy.

Finally, the **MIRAGE** framework presented in [Chapter 10](#) can be extended in the long term to sound source tracking, as well as to acoustic echo cancellation or self-calibration. Under this light, the focus will be on smart table-top devices, whose the two most striking examples are Google Home, Amazon Echo. These smart speakers could benefit from echo-aware processing, not only in recording sounds but also in sound production. The former case was deeply discussed in these as it concern echo-aware speech enhancement techniques. Instead, in the latter case, echo processing could provide dynamic *sweet-spots* that adapts to the user and device positions. The device could then produce an immersive sound field for music listening or better directivity in hand-free phone-calls.

Finally, we can expand the frameworks proposed in the thesis for acoustical data to other non-acoustic scientific domains. For example, other research fields put relevant focus on reflections, such as, submarine navigation [Kleeman and Kuc 2016], in seismology [Sato et al. 2012], in ultrasound imaging [Achim et al. 2010] and in radioastronomy [Pan et al. 2016].

► USING THE dEchorate DATASET AND THE OTHER LIBRARIES

The main idea behind this dataset is to foster research in **AER** and echo-aware on real data. A natural step is to use this data for the applications discussed in this thesis and particularly on **AER**, for which this dataset was designed. Moreover, new line of research with deep neural networks and optimal transport, such as style transfer and domain adaptation could be envisioned. For instance, by using the pairs of simulated vs. real **RIRs** available in the dataset, one could develop techniques to convert one to the other. Ideally, this approach could be at the basis on new type of learning-based acoustic simulators.

► CROSSING THE DIRECTIONS

Ultimately, the two parts, estimation and application, of this dissertation should plug together. So far we only showed how from audio features is possible to estimate echoes and how from echoes is possible to estimated audio scene analysis information, e. g. source content and location. This problem have an innate uroboric nature: where, what, when and how are connected — the knowledge of one helps the estimation of the others, in a vicious (or maybe virtuous) circle. Therefore it should be possible to build iterative schemes linking echo-estimation and echo-applications.

So long and thanks for all the echoes.



Sliding Frank-Wolfe algorithm & Non-negative Blasso

Among all the methods that address the resolution of (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$), a significant number of them are based on variations of the well-known Frank-Wolfe iterative algorithm, see, e.g., [Bredies and Carioni 2020; Denoyelle et al. 2019]. In this paper, we particularize the *sliding Frank-Wolfe* (SFW) algorithm proposed in [Denoyelle et al. 2019].

Starting from an initial guess (e.g., the null measure), SFW repeats the four following steps until convergence:

1. add a parameter (position of echo) to the support of the solution,
2. update all the coefficients solving a (finite dimensional) Lasso,
3. update jointly the position of the echoes and the coefficients,
4. eventually remove parameters (echoes) associated to coefficients equal to zero.

Finally, SFW stops as soon as an iterate satisfies the first order optimality condition associated to the convex problem (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$). More particularly, denoting $\mu^{(t)}$ the estimated filters at iteration t , SFW stops as soon as $\mu^{(t)}$ satisfies [Bredies and Carioni 2020, Proposition 3.6]

$$\sup_{\theta \in \Theta} \lambda^{-1} \left| \langle \mathcal{A}\delta_{\theta}, \mathbf{y} - \mathcal{A}\mu^{(t)} \rangle \right| \leq 1. \quad (1)$$

The complete SFW method for echo estimation is described by Algorithm 1. We now provide additional details about the implementation of each step.

► NON-NEGATIVE BLASSO

To take into account the non-negative constraint on the coefficients, the authors of [Denoyelle et al. 2019] have proposed to slightly modify the SFW algorithm by *i*) removing the absolute value in (1) and *ii*) adding the non-negativity constraints at step 2 and 3 (see lines 14 and 15 of Algorithm 1). The reader is referred to [Denoyelle et al. 2019, remark 8 in Section 4.1] for more details.

► REAL PART IN (1).

We have shown earlier that SFW stops as soon as an iterate $\mu^{(t)}$ satisfies (1) at some iteration t . Since the estimated coefficients $\left\{ c_r^{(t)} \right\}_{r=1}^R$ are (non-negative) scalars, (1) can be rewritten as

$$\sup_{\theta \in \Theta} \lambda^{-1} \operatorname{Re}(\langle \mathcal{A}\delta_{\theta}, \mathbf{y} - \mathcal{A}\mu^* \rangle) \leq 1. \quad (2)$$

In particular, using the real part in the implementation allows to remove the imaginary part that may appear due to the imprecision.

Algorithm 1: Sliding Frank-Wolfe algorithm for solving (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$).

Input: Observation operator \mathcal{A} , positive scalar λ , precision ε

Output: Channels represented as a measure $\hat{\mu}$

```

// Initialization
1  $\mathbf{y} \leftarrow -\mathcal{A}\delta_{(0,1)} // observation vector$ 
2  $\mu^{(0)} = 0_{\mathcal{M}} // estimated filters$ 
3  $\mathcal{E}^{(0)} = \emptyset // estimated echoes$ 
4  $x_{\max} = (2\lambda)^{-1}\|\mathbf{y}\|_2^2;$ 

// Starting algorithm
5 repeat
6    $t \leftarrow t + 1 // Iteration index$ 
   // 1. Add new element to the support
7   Find  $\theta^{\text{new}} \in \arg \max_{\theta \in \Theta} \operatorname{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^{(t-1)} \rangle)$ ;
8    $\eta^{(t)} \leftarrow \lambda^{-1} \operatorname{Re}(\langle \mathcal{A}\delta_{\theta^{\text{new}}}, \mathbf{y} - \mathcal{A}\mu^{(t-1)} \rangle)$ ;
9   if  $\eta^{(t)} \leq 1 + \varepsilon$  then
10    | Stop and return  $\hat{\mu} = \mu^{(t-1)}$  is a solution ;
11   end
12    $\mathcal{E}^{(t-\frac{1}{2})} \leftarrow \mathcal{E}^{(t-\frac{1}{2})} \cup \{\theta^{\text{new}}\};$ 
13    $R^{(t)} \leftarrow \operatorname{card}(\mathcal{E}^{(t-\frac{1}{2})}) // Number of detected echoes$ 

   // 2. Lasso update of the coefficients
14    $\mathbf{c}^{(t-\frac{1}{2})} \leftarrow \arg \min_{\mathbf{c} \in \mathbf{R}_+^{R^{(t)}}} \frac{1}{2} \left\| \mathbf{y} - \sum_{\theta \in \mathcal{E}^{(t-\frac{1}{2})}} c_\theta \mathcal{A}\delta_\theta \right\|_2^2 + \lambda \|\mathbf{c}\|_1$ 
   approximated using a proximal gradient algorithm ;

   // 3. Joint update for a given number of spikes
15    $\mathcal{E}^{(t)}, \mathbf{c}^{(t)} \leftarrow$ 
      
$$\arg \min_{\theta \in \Theta^{R^{(t)}}, \mathbf{c} \in [0, x_{\max}]^{R^{(t)}}} \frac{1}{2} \left\| \mathbf{y} - \sum_{r=1}^{R^{(t)}} \mathbf{c}_r \mathcal{A}\delta_{\theta_r} \right\|_2^2 + \lambda \|\mathbf{c}\|_1$$

   approximated using a non-convex solver initialized with
    $(\mathcal{E}^{(t-\frac{1}{2})}, \mathbf{c}^{(t-\frac{1}{2})})$  ;

   // 4. Eventually remove zero amplitude Dirac masses
16    $\mathcal{E}^{(t)} \leftarrow \left\{ \theta_r^{(t)} \in \mathcal{E}^{(t)} \mid \mathbf{c}_r^{(t)} \neq 0 \right\};$ 
17    $\mathbf{c}^{(t)} \leftarrow \left\{ \mathbf{c}_r^{(t)} \mid \mathbf{c}_r^{(t)} \neq 0 \right\};$ 
18    $\mu^{(t)} \leftarrow \sum_{r=1}^{\operatorname{card}(\mathcal{E}^{(t)})} \mathbf{c}_r^{(t)} \delta_{\theta_r^{(t)}};$ 
19 until until convergence;

```

► PRECISION OF THE STOPPING CRITERION.

Unfortunately, condition (1) cannot be met due to the machine precision, *i.e.*, the solution of (5.22- $\mathcal{P}_{\text{TV}}^{\lambda}$) is computed up to some prescribed accuracy. In this paper, we say that the algorithm stops as soon as

$$\sup_{\theta \in \Theta} \lambda^{-1} \operatorname{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^* \rangle) \leq 1 + \varepsilon \quad (3)$$

where ε is a positive scalar set to $\varepsilon = 10^{-3}$.

► FINDING NEW PARAMETERS (LINE 7).

The new parameter is found by solving

$$\arg \max_{\theta \in \Theta} \operatorname{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\hat{\mu} \rangle). \quad (4)$$

To solve this optimization problem, we first find a maximizer on a thin grid made of 20000 points. We then proceed to a local refinement using the `scipy` optimization library⁶⁹.

⁶⁹ <https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html>.

► NONNEGATIVE LASSO (LINE 14).

The nonnegative Lasso is solved using a custom implementation of a proximal gradient algorithm. In particular, the procedure stops as soon as a stopping criterion in terms of duality gap is reached (10^{-6}).

► JOINT UPDATE (LINE 15).

In order to ease the numerical resolution, we show that given a positive integer R , the solution of

$$\arg \min_{\theta \in \Theta^R, \mathbf{c} \in \mathbf{R}^R} \frac{1}{2} \left\| \mathbf{y} - \sum_{r=1}^R \mathbf{c}_r \mathcal{A}\delta_{\theta_r} \right\|_2^2 + \lambda \|\mathbf{c}\|_1 \quad (5)$$

is equivalent to the solution of

$$\arg \min_{\theta \in \Theta^R, \mathbf{c} \in [0, x_{\max}]^R} \frac{1}{2} \left\| \mathbf{y} - \sum_{r=1}^R \mathbf{c}_r \mathcal{A}\delta_{\theta_r} \right\|_2^2 + \lambda \|\mathbf{c}\|_1 \quad (6)$$

where

$$x_{\max} = \frac{1}{2\lambda} \|\mathbf{y}\|_2^2. \quad (7)$$

Indeed, let us denote θ^* , \mathbf{c}^* the minimizers of (5). For any $\theta \in \Theta^R$, the couple $\theta, \mathbf{0}_R$ is admissible for (5) so we have by definition

$$\frac{1}{2} \left\| \mathbf{y} - \sum_{r=1}^R \mathbf{c}_r^* \mathcal{A}\delta_{\theta_r^*} \right\|_2^2 + \lambda \|\mathbf{c}^*\|_1 \leq \frac{1}{2} \|\mathbf{y}\|_2^2. \quad (8)$$

Hence

$$0 \leq c_r^* \leq \|\mathbf{c}^*\|_1 \leq \frac{1}{2\lambda} \|\mathbf{y}\|_2^2 \triangleq x_{\max}. \quad (9)$$

Finally, the joint update of the coefficients and parameters is performed using the Sequential Least SQuares Programming (SLSQP) implemented in the `scipy` optimization library, see Sidenote 69.

Derivation of the Acoustic Wave

► THE DERIVATION OF THE ACOUSTIC WAVE EQUATION

Let's starts by considering a infinitesimal volume unit \mathcal{V} of a fluid or gas (such as air), whose center of gravity is located at $\underline{x} \in \mathbb{R}^3$. Let m be the mass of such volume. By the well-known Newton's second law, applying a force \mathbf{F} to the fluid, its acceleration increase proportionally to m , namely:

$$\mathbf{F} = m \frac{\partial \mathbf{v}(\underline{x}, t)}{\partial t} \quad (10)$$

where $\mathbf{v}(\underline{x}, t)$ denotes the volume velocity and t the time [s]. The force can be expressed in terms of difference of acoustic pressure p at \underline{x} on a surface of the volume, S , namely

$$\mathbf{F} = -S[\nabla p(\underline{x}, t)] \quad (11)$$

where ∇ is the gradient operator.

By combining Eqs. (10) and (11), we obtain the famous *Euler's equation of motion*:

$$\nabla p(\underline{x}, t) = -\rho_0 \frac{\partial \mathbf{v}(\underline{x}, t)}{\partial t} \quad (12)$$

where $\rho_0 = \frac{m}{S}$ is the static density of the medium¹.

► BY THE CONSERVATION OF MASS PRINCIPLE

states that, in a deformable medium, the total mass must remain constant. This principle translates into the *continuity equation*, which its differential form writes

$$\frac{\partial \nu(\underline{x}, t)}{\partial t} = V[\nabla \cdot \mathbf{q}(\underline{x}, t)] \quad (13)$$

where

- $\nu(\underline{x}, t)$ is the volume variation due to the pressure changing, and
- $\mathbf{q}(\underline{x}, t)$ is the *flux* of mass m per unit volume (a. k. a. flux)⁷⁰.

► THE POLYTROPIC PROCESS RELATION

assumed properties on the propagation medium. Since the exchange of heat in negligible in the acoustic frequencies range, the whole process can be considered thermodynamically *adiabatic*. In such scenario, the relation between the total pressure \mathcal{P} and the total volume \mathcal{V} is given by

$$\mathcal{P}\mathcal{V}^\gamma = const. \quad (14)$$

where γ is the adiabatic index of the medium⁷¹. The total pressure and the total volume consist in a sum of a constant and a variable term, that is $\mathcal{P} = P_0 + p$, $\mathcal{V} = V_0 + \nu$ respectively. Considering that $p \ll P_0$ and $\nu \ll V_0$, the time-differential of Eq. (14) with respect to time reads

$$\frac{\partial p(\underline{x}, t)}{\partial t} = -\gamma \frac{P_0}{V_0} \frac{\partial \nu(\underline{x}, t)}{\partial t} \quad (15)$$

Newton's Law II: The alteration of motion is ever proportional to the motive force impress'd; and is made in the direction of the right line in which that force is impress'd.
Original: *Lex II: Mutationem motus proportionalem esse vi motrici impressae, et fieri secundum lineam rectam qua vis illa imprimitur.*

¹ Selected physical quantities for air:
Air Density $\rho_{air} = 1.18 \frac{\text{kg}}{\text{m}^3}$.

Air Gas constant $R_{air} = 286.9 \frac{\text{J}}{\text{kgmol}}$.
Air Adiabatic index $\gamma_{air} = 1.4$.
Speed of sound in air $c_{air} = 343.1 \frac{\text{m}}{\text{s}}$.

$\nabla \cdot = \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial z}$ is the divergence operator.

⁷⁰flux is by definition equal to the density times the velocity. In math, $\mathbf{q}(\underline{x}, t) = \rho_0 \mathbf{v}(\underline{x}, t)$.

The dependency upon space and time (\underline{x}, t) is here omitted for sake of compactness and readability.

⁷¹Cfr. Footnote 1

- FINALLY, THE ACOUSTIC WAVE EQUATION can be now derived by combining together the equation of motion (12), the continuity equation (13), and the thermodynamic balance of the medium (15). In particular the combination of Eqs. (13) and (15),

$$\frac{\partial p(\underline{x}, t)}{\partial t} = -\gamma P_0 [\nabla \cdot \mathbf{q}(\underline{x}, t)], \quad (16)$$

can be differentiated with respect to time t yielding to

$$\frac{\partial^2 p(\underline{x}, t)}{\partial t^2} = -\gamma P_0 \left[\nabla \cdot \frac{\partial \mathbf{q}(\underline{x}, t)}{\partial t} \right]. \quad (17)$$

Taking the divergence of each side of the Eq. (12) and remembering the definition of flux, we get

$$\nabla^2 p(\underline{x}, t) = -\rho_0 [\nabla \cdot \mathbf{q}(\underline{x}, t)] \quad (18)$$

The above two equation can be combined leading to

$$\nabla^2 p(\underline{x}, t) = \frac{1}{c^2} \frac{\partial^2 p(\underline{x}, t)}{\partial t^2}, \quad (19)$$

which is equivalent to Eq. (2.2). The constant c is the wave speed, in our case the speed of sound. Notice that it is related to the medium properties through

$$c^2 = \frac{\gamma P_0}{\rho_0}. \quad (20)$$

The *Laplacian* of a function is equivalent to the divergence of the gradient of that function.

In math, $\nabla^2 x = \nabla \cdot \nabla x$

—
The conventional way to solve Eq. (2.3) is to find a set of Functions $\Psi_l(f, \underline{x})$ for $l = 0, \dots, \infty$ which satisfy the homogenous equation Eq. (2.3) for a certain interval and its boundary conditions \mathcal{B} .

This type of functions are called *characteristic function* or *eigenfunction*⁷² and depends on \mathcal{B} . In general, such kind of function are too difficult to be computed in closed form and needs to be approximated with numerical methods. Their analytical expression is known only for a few room shapes combined with simple boundary conditions.

Subsequently, the general expression of the Green's function $H(f, \underline{x} | \underline{s})$ can be expressed as a sum of the eigenfunction weighted on a coefficient $C_l(f, \underline{s})$ dependent on the source position [Habets 2006]:

$$H(f, \underline{x} | \underline{s}) = \sum_{l=0}^{\infty} C_l(f, \underline{s}) \Psi_l(f, \underline{x}) \quad (21)$$

⁷²Notice that that Eq. (2.3) can be written as eigenfunction/eigenvalue equation, i.e. $\nabla^2 H = -k^2 H$

—
Let us assume the simplest possible 3D enclosure: a *shoebox*, i. e. a cuboid room with perfectly smooth and rigid facets. Lets define the domain \mathcal{D} of the problem: the cuboid length L , width W and height H , that is

$$\mathcal{D} = \{\underline{x} = (x, y, z) \mid 0 \leq x \leq L_x, 0 \leq y \leq L_y, 0 \leq z \leq L_z\} \quad (22)$$

Given the boundaries \mathcal{B} of \mathcal{D} , the frequency-domain Green's function associated to Eq. (2.3) is given by

$$H(f, \underline{x} | \underline{s}) = -\frac{1}{V} \sum_{\mathbf{m} \in \mathcal{M}} \frac{\Psi_{\mathbf{m}}(\underline{x}) \Psi_{\mathbf{m}}(\underline{s})}{\kappa_{\mathbf{m}}^2 - k^2} \quad (23)$$

where

$$\begin{aligned} k &= 2\pi f/c, \\ \mathbf{m} &= (n_x, n_y, n_z), \\ n_i &\in \mathbb{N} \quad \forall i \in \{x, y, z\}, \\ V &= L_x L_y L_z \text{ is the room volume,} \\ \mathcal{M} &= \{(n_x, n_y, n_z) \mid n_i \in N\}. \end{aligned}$$

$$k_i = n_i \pi / L_i \quad \forall i \in \{x, y, z\},$$

$$\begin{aligned}\kappa_{\mathbf{m}} &= \left(\frac{n_x \pi}{L}, \frac{n_y \pi}{W}, \frac{n_z \pi}{H} \right) = (k_x, k_y, k_z) \\ \kappa_{\mathbf{m}}^2 &= |\kappa_{\mathbf{m}}|^2 = k_x + k_y + k_z \\ \Psi_{\mathbf{m}}(\underline{\mathbf{r}}) &= \cos(k_x x) \cos(k_y y) \cos(k_z z)\end{aligned}\quad (24)$$

is the eigenfunction⁷³ for the specific shoebox boundaries \mathcal{B} [Kuttruff 2016a].

⁷³Cfr Eq. (21)

Using the exponential expansion of the cosine for Eq. (24) and using it in Eq. (23), we obtain

$$H(f, \underline{\mathbf{x}} | \underline{\mathbf{s}}) = -\frac{1}{8V} \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{m} \in \mathcal{M}} \frac{\Psi_{\mathbf{m}}(\underline{\mathbf{x}}) \Psi_{\mathbf{m}}(\underline{\mathbf{s}})}{\kappa_{\mathbf{m}}^2 - k^2} \quad (25)$$

—

► RELATION WITH THE HELMHOLTZ EQUATION

Finally, the equation becomes

$$H(f, \underline{\mathbf{x}} | \underline{\mathbf{s}}) = \sum_{p=1}^8 \sum_{\underline{\mathbf{r}}=-\infty}^{\infty} \frac{1}{4\pi \|\mathbf{R}_p + \mathbf{R}_{\underline{\mathbf{r}}} \|} \quad (26)$$

by taking the inverse Fourier Transform, the echo structure becomes explicit.

We can write the final Room Impulse Response $h_{ij}(t)$ as follows:

$$\tilde{x}(t) = (h_{ij} * \tilde{s})(t) \quad (27)$$

$$h_{ij}(t) = \sum_{r=0}^R \frac{\alpha_r}{4\pi \tau_r / c_{\text{air}}} \delta(t - \tau_r) \quad (28)$$

where

- $\alpha_r \in [0, 1]$ is the attenuation coefficient of the r -th reflection
- $\tau_r = \|\underline{\mathbf{x}}_i - \underline{\mathbf{s}}_r\|$ is the distance between the microphone and the r -th image of source j .

RIR and RT60 measurements

RIR ESTIMATION

RT60 ESTIMATION

Bibliography

- Abed-Meraim, Karim, Philippe Loubaton, and Eric Moulines (1997). “A subspace algorithm for certain blind identification problems”. In: *IEEE transactions on information theory* 43.2, pp. 499–511 (cit. on p. 55).
- Achim, Alin, Benjamin Buxton, George Tzagkarakis, and Panagiotis Tsakalides (2010). “Compressive sensing for ultrasound RF echoes using a-stable distributions”. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, pp. 4304–4307 (cit. on p. 150).
- Ackerman, Evan (2018). “Haru: An Experimental Social Robot From Honda Research”. In: *IEEE Specturm* (cit. on p. 118).
- Adavanne, Sharath, Archontis Politis, and Tuomas Virtanen (2018). “Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1462–1466 (cit. on pp. 105, 106).
- Ahmad, Rehan, Andy WH Khong, and Patrick A Naylor (2006). “Proportionate frequency domain adaptive algorithms for blind channel identification”. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 5. IEEE, pp. V–V (cit. on p. 59).
- Aissa-El-Bey, Abdeldjalil and Karim Abed-Meraim (2008). “Blind SIMO channel identification using a sparsity criterion”. In: *2008 IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*. IEEE, pp. 271–275 (cit. on pp. 55, 65).
- Al-Karawi, Khamis A and Duraid Y Mohammed (2019). “Early reflection detection using autocorrelation to improve robustness of speaker verification in reverberant conditions”. In: *International Journal of Speech Technology* 22.4, pp. 1077–1084 (cit. on p. 53).
- Allen, Jont B and David A Berkley (1979). “Image method for efficiently simulating small-room acoustics”. In: *The Journal of the Acoustical Society of America* 65.4, pp. 943–950 (cit. on pp. 23–25, 71, 109).
- An, Inkyu, Myungbae Son, Dinesh Manocha, and Sung-eui Yoon (2018). “Reflection-aware sound source localization”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 66–73 (cit. on p. 119).
- Annibale, P., F. Antonacci, P. Bestagini, A. Brutti, A. Canclini, L. Cristoforetti, E. Habets, W. Kellermann, K. Kowalczyk, A. Lombard, E. Mabande, D. Markovic, P. Naylor, M. Omologo, R. Rabenstein, A. Sarti, P. Svaizer, and M. Thomas (2011). “The SCENIC project: Environment-aware sound sensing and rendering”. In: *Procedia Computer Science* 7, pp. 150–152. ISSN: 18770509. doi: [10.1016/j.procs.2011.09.039](https://doi.org/10.1016/j.procs.2011.09.039). URL: <http://dx.doi.org/10.1016/j.procs.2011.09.039> (cit. on p. 5).
- Annibale, Paolo, Jason Filos, Patrick A Naylor, and Rudolf Rabenstein (2012). “Geometric inference of the room geometry under temperature variations”. In: *2012 5th International Symposium on Communications, Control and Signal Processing*. IEEE, pp. 1–4 (cit. on pp. 50, 51, 141).
- Antonacci, Fabio, Augusto Sarti, and Stefano Tubaro (2010). “Geometric reconstruction of the environment from its response to multiple acoustic emissions”. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 2822–2825 (cit. on p. 52).
- Antonacci, Fabio, Jason Filos, Mark RP Thomas, Emanuël AP Habets, Augusto Sarti, Patrick A Naylor, and Stefano Tubaro (2012). “Inference of room geometry from acoustic impulse responses”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.10, pp. 2683–2695 (cit. on pp. 52, 55, 91).
- Aoshima, Nobuharu (1981). “Computer-generated pulse signal applied for sound measurement”. In: *The Journal of the Acoustical Society of America* 69.5, pp. 1484–1488 (cit. on p. 49).
- Applebaum, S and D Chapman (1976). “Adaptive arrays with main beam constraints”. In: *IEEE Transactions on Antennas and Propagation* 24.5, pp. 650–662 (cit. on p. 103).

- Argentieri, Sylvain, Patrick Danès, and Philippe Souères (2015). “A survey on sound source localization in robotics: From binaural to array processing methods”. In: *Computer Speech & Language* 34.1, pp. 87–112 (cit. on p. 104).
- Asaei, Afsaneh, Mohammad Golbabaei, Herve Bourlard, and Volkan Cevher (2014). “Structured sparsity models for reverberant speech separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.3, pp. 620–633 (cit. on p. 108).
- Baba, Youssef El, Andreas Walther, and Emanuël A.P. Habets (2018). “3D room geometry inference based on room impulse response stacks”. In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 26.5, pp. 857–872. ISSN: 23299290. doi: 10.1109/TASLP.2017.2784298 (cit. on p. 91).
- Badeau, Roland (2019). “Common mathematical framework for stochastic reverberation models”. In: *The Journal of the Acoustical Society of America* 145.4, pp. 2733–2745 (cit. on pp. 22–24, 149).
- Barron, Michael (1971). “The subjective effects of first reflections in concert halls—the need for lateral reflections”. In: *Journal of sound and vibration* 15.4, pp. 475–494 (cit. on p. 27).
- Barumerli, Roberto, Michele Geronazzo, and Federico Avanzini (2018). “Localization in Elevation with Non-Individual Head-Related Transfer Functions: Comparing Predictions of Two Auditory Models”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 2539–2543 (cit. on p. 150).
- Beck, Amir, Petre Stoica, and Jian Li (2008). “Exact and approximate solutions of source localization problems”. In: *IEEE Transactions on Signal Processing* 56.5, pp. 1770–1778. ISSN: 1053587X. doi: 10.1109/TSP.2007.909342 (cit. on pp. 90, 141).
- Bello, Juan Pablo, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler (2005). “A tutorial on onset detection in music signals”. In: *IEEE Transactions on speech and audio processing* 13.5, pp. 1035–1047 (cit. on pp. 50, 51).
- Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). “Curriculum learning”. In: *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48 (cit. on p. 85).
- Bergamo, Pierpaolo, Shadnaz Asgari, Hanbiao Wang, Daniela Maniezzo, Len Yip, Ralph E Hudson, Kung Yao, and Deborah Estrin (2004). “Collaborative sensor networking towards real-time acoustical beamforming in free-space and limited reverberance”. In: *IEEE Transactions on Mobile Computing* 3.3, pp. 211–224 (cit. on p. 119).
- Bertin, Nancy, Ewen Camberlein, Romain Lebarbenchon, Emmanuel Vincent, Sunit Sivasankaran, Irina Illina, and Frédéric Bimbot (2019). “VoiceHome-2, an extended corpus for multichannel speech processing in real homes”. In: *Speech Communication* 106, pp. 68–78 (cit. on p. 58).
- Betlehem, Terence, Paul D Teal, and Yusuke Hioka (2012). “Efficient crosstalk canceler design with impulse response shortening filters”. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 393–396 (cit. on p. 49).
- Birnie, Lachlan I, Thushara D Abhayapala, and Prasanga N Samarasinghe (2020). “Reflection Assisted Sound Source Localization Through a Harmonic Domain MUSIC Framework”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, pp. 279–293 (cit. on p. 119).
- Bishop, Christopher M (1994). “Mixture density networks”. In: (cit. on p. 82).
- (2006). *Pattern recognition and machine learning*. Springer (cit. on p. 74).
- Blandin, Charles, Alexey Ozerov, and Emmanuel Vincent (2012). “Multi-source TDOA estimation in reverberant audio using angular spectra and clustering”. In: *Signal Processing* 92.8, pp. 1950–1960 (cit. on pp. 121, 122).
- Böck, Sebastian, Florian Krebs, and Markus Schedl (2012). “Evaluating the Online Capabilities of Onset Detection Methods.” In: *ISMIR*, pp. 49–54 (cit. on pp. 60, 71).
- Bourguignon, Sébastien, Jordan Ninin, Hervé Carfantan, and Marcel Mongeau (2015). “Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance”. In: *IEEE Transactions on Signal Processing* 64.6, pp. 1405–1419 (cit. on p. 149).

- Bredies, Kristian and Marcello Carioni (2020). "Sparsity of solutions for variational inverse problems with finite-dimensional data". In: *Calculus of Variations and Partial Differential Equations* 59.1, p. 14 (cit. on pp. 69, 152).
- Bregman, Albert S (1990). "Auditory scene analysis". In: *McAdams and Bigand, editors, Thinking in Sound*, pp. 10–36 (cit. on p. 3).
- Bryan, Nicholas J (2020). "Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 149).
- Candès, Emmanuel J and Carlos Fernandez-Granda (2014). "Towards a mathematical theory of super-resolution". In: *Communications on pure and applied Mathematics* 67.6, pp. 906–956 (cit. on pp. 63, 68).
- Capon, Jack (1969). "High-resolution frequency-wavenumber spectrum analysis". In: *Proceedings of the IEEE* 57.8, pp. 1408–1418 (cit. on p. 102).
- Cauchi, Benjamin, Ina Kodrasi, Robert Rehr, Stephan Gerlach, Ante Jukic, Timo Gerkmann, Simon Doclo, and Stefan Goetze (2014). "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme". In: *Proc. REVERB challenge workshop*. Vol. 1, pp. 1–8 (cit. on p. 135).
- Cecchi, Stefania, Alberto Carini, and Sascha Spors (2018). "Room response equalization—A review". In: *Applied Sciences* 8.1, p. 16 (cit. on p. 102).
- Chakrabarty, Soumitro and Emanuel AP Habets (2017). "Broadband DOA estimation using convolutional neural networks trained with noise signals". In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 136–140 (cit. on pp. 75, 82, 105, 106).
- Chazan, Shlomo E, Jacob Goldberger, and Sharon Gannot (2018). "DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6712–6716 (cit. on p. 104).
- Chazan, Shlomo E, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot (2019). "Multi-microphone speaker separation based on deep DOA estimation". In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1–5 (cit. on p. 104).
- Chen, Jingdong, Jacob Benesty, and Yiteng Arden Huang (2006). "Time delay estimation in room acoustic environments: an overview". In: *EURASIP Journal on Advances in Signal Processing* 2006.1, p. 026503 (cit. on pp. 81, 105, 119, 121, 122).
- Cheng, Tian, Matthias Mauch, Emmanouil Benetos, Simon Dixon, et al. (2016). "An attack/decay model for piano transcription". In: *ISMIR* (cit. on pp. 50, 51).
- Cherry, Colin (1953). "Cocktail party problem". In: *Journal of the Acoustical Society of America* 25, pp. 975–979 (cit. on p. 99).
- Chi, Yuejie, Louis L Scharf, Ali Pezeshki, and A Robert Calderbank (2011). "Sensitivity to basis mismatch in compressed sensing". In: *IEEE Transactions on Signal Processing* 59.5, pp. 2182–2195 (cit. on p. 56).
- Čmejla, Jaroslav, Tomáš Kounovský, Sharon Gannot, Zbyněk Koldovský, and Pinchas Tandeitnik (2019). "MIRaGe: Multichannel Database Of Room Impulse Responses Measured On High-Resolution Cube-Shaped Grid In Multiple Acoustic Conditions". In: *arXiv preprint arXiv:1907.12421* (cit. on pp. 56, 58, 87).
- Cohen, Israel (2004). "Relative transfer function identification using speech signals". In: *IEEE Transactions on Speech and Audio Processing* 12.5, pp. 451–459 (cit. on p. 103).
- Condat, Laurent and Akira Hirabayashi (2013). "Robust spike train recovery from noisy data by structured low rank approximation". In: *Int. Conf. Sampl. Theory Appl. (SAMPTA), Bremen, Germany* (cit. on pp. 51, 57).
- (2015). "Cazdow denoising upgraded: A new projection method for the recovery of Dirac pulses from noisy linear measurements". In: (cit. on pp. 47, 57).
- Cox, Henry, Robert M Zeskind, and Markm Owen (1987). "Robust adaptive beamforming". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35.10, pp. 1365–1376 (cit. on pp. 102, 103).
- Crocchetto, Marco and Alessio Del Bue (2015). "Room impulse response estimation by iterative weighted l 1-norm". In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1895–1899 (cit. on pp. 55, 65, 71, 73).

- Crocco, Marco and Alessio Del Bue (2016a). "Estimation of TDOA for room reflections by iterative weighted l1 constraint". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3201–3205 (cit. on pp. 55, 56, 62, 65, 70, 104).
- (2016b). "Estimation of TDOA for room reflections by iterative weighted l1 constraint". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2016-May. 2. IEEE, pp. 3201–3205. ISBN: 978147999880. doi: [10.1109/ICASSP.2016.7472268](https://doi.org/10.1109/ICASSP.2016.7472268) (cit. on p. 90).
- Crocco, Marco, Alessio Del Bue, Matteo Bustreo, and Vittorio Murino (2012). "A closed form solution to the microphone position self-calibration problem". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2597–2600 (cit. on pp. 56, 141).
- Crocco, Marco, Andrea Trucco, Vittorio Murino, and Alessio Del Bue (2014). "Towards fully uncalibrated room reconstruction with sound". In: *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 910–914 (cit. on pp. 52, 53).
- Crocco, Marco, Andrea Trucco, and Alessio Del Bue (2017). "Uncalibrated 3D room geometry estimation from sound impulse responses". In: *Journal of the Franklin Institute* 354.18, pp. 8678–8709 (cit. on pp. 5, 47, 50, 52, 55, 56, 58, 141).
- Davis, AH and N Fleming (1926). "Sound pulse photography as applied to the study of architectural acoustics". In: *Journal of Scientific Instruments* 3.12, p. 393 (cit. on p. 18).
- De Castro, Yohann and Fabrice Gamboa (2012). "Exact reconstruction using Beurling minimal extrapolation". In: *Journal of Mathematical Analysis and Applications* 395.1, pp. 336–354 (cit. on p. 68).
- Defrance, Guillaume, Laurent Daudet, and Jean-Dominique Polack (2008a). "Detecting arrivals within room impulse responses using matching pursuit". In: *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland*. Vol. 10. Citeseer, pp. 307–316 (cit. on pp. 50, 51).
- (2008b). "Finding the onset of a room impulse response: Straightforward?" In: *The Journal of the Acoustical Society of America* 124.4, EL248–EL254 (cit. on pp. 50, 56, 92).
- Deleforge, Antoine and Florence Forbes (2016). "Rectified binaural ratio: A complex T-distributed feature for robust sound localization". In: *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1257–1261 (cit. on p. 83).
- Deleforge, Antoine, Florence Forbes, and Radu Horaud (2015). "Acoustic space learning for sound-source separation and localization on binaural manifolds". In: *International journal of neural systems* 25.01, p. 1440003 (cit. on pp. 42, 75, 79, 105, 106).
- Deleforge, Antoine, **Di Carlo, Diego**, Martin Strauss, Romain Serizel, and Lucio Marcenaro (2019). "Audio-Based Search and Rescue With a Drone: Highlights From the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]". In: *IEEE Signal Processing Magazine* 36.5, pp. 138–144 (cit. on p. 9).
- Denoyelle, Quentin, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies (2019). "The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy". In: *Inverse Problems* 36.1, p. 014001 (cit. on pp. 63, 69, 152).
- Di Carlo, Diego and Antoine Deleforge. *HRI-JF collaboration - Final Phase II Deliverable*. Tech. rep. Inria Nancy - Grand Est (cit. on pp. 74, 118).
- Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019a). "Mirage: 2D Source Localization Using Microphone Pair Augmentation with Echoes". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2019-May, pp. 775–779. ISBN: 9781479981311. doi: [10.1109/ICASSP.2019.8683534](https://doi.org/10.1109/ICASSP.2019.8683534) (cit. on p. 42).
- (2019b). "Mirage: 2D source localization using microphone pair augmentation with echoes". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 775–779 (cit. on pp. 72, 74, 81, 82, 118, 124).
- Di Carlo, Diego, Clement Elvira, Antoine Deleforge, Nancy Bertin, and Rémi Gribonval (2020). "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 156–160 (cit. on pp. 56, 62).

- DiBiase, Joseph H, Harvey F Silverman, and Michael S Brandstein (2001). "Robust localization in reverberant rooms". In: *Microphone Arrays*. Springer, pp. 157–180 (cit. on pp. 54, 57, 79, 105, 122).
- Dmochowski, Jacek P, Jacob Benesty, and Sofiene Affes (2007). "Broadband MUSIC: Opportunities and challenges for multiple source localization". In: *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 18–21 (cit. on p. 105).
- Doclo, Simon and Marc Moonen (2002). "GSVD-based optimal filtering for single and multimicrophone speech enhancement". In: *IEEE Transactions on signal processing* 50.9, pp. 2230–2244 (cit. on pp. 41, 103).
- (2003). "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments". In: *EURASIP Journal on Advances in Signal Processing* 2003.11, p. 495250 (cit. on p. 138).
- Dokmanić, Ivan, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli (2013). "Acoustic echoes reveal room shape". In: *Proceedings of the National Academy of Sciences* 110.30, pp. 12186–12191 (cit. on pp. 5, 52, 58, 141).
- Dokmanić, Ivan, Robin Scheibler, and Martin Vetterli (2015). "Raking the Cocktail Party". In: *IEEE Journal on Selected Topics in Signal Processing* 9.5, pp. 825–836. ISSN: 19324553. doi: [10.1109/JSTSP.2015.2415761](https://doi.org/10.1109/JSTSP.2015.2415761) (cit. on pp. 5, 103, 133).
- Dokmanić, Ivan, Robin Scheibler, and Martin Vetterli (2015a). "Raking the Cocktail Party". In: *IEEE J. Sel. Top. Signal Process.* 9.5, pp. 825–836 (cit. on p. 109).
- (2015b). "Raking the cocktail party". In: *IEEE journal of selected topics in signal processing* 9.5, pp. 825–836 (cit. on pp. 55, 109, 133, 134).
- Dokmanić, Ivan, Juri Ranieri, and Martin Vetterli (2015). "Relax and unfold: Microphone localization with Euclidean distance matrices". In: *European Signal Processing Conference, (EUSIPCO)*, pp. 265–269. ISBN: 9780992862633. doi: [10.1109/EUSIPCO.2015.7362386](https://doi.org/10.1109/EUSIPCO.2015.7362386) (cit. on pp. 90, 141).
- Duffy, Dean G (2015). *Green's functions with applications*. CRC Press (cit. on pp. 13, 15).
- Dunn, Chris and Malcolm J Hawksford (1993). "Distortion immunity of MLS-derived impulse response measurements". In: *Journal of the Audio Engineering Society* 41.5, pp. 314–335 (cit. on p. 49).
- Duong, Ngoc QK, Emmanuel Vincent, and Rémi Gribonval (2010). "Under-determined reverberant audio source separation using a full-rank spatial covariance model". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, pp. 1830–1840 (cit. on pp. 54, 101).
- Eaton, James, Nikolay D Gaubitch, Alastair H Moore, and Patrick A Naylor (2015). "The ACE challenge—Corpus description and performance evaluation". In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 1–5 (cit. on p. 139).
- Eaton, James, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor (Oct. 2016). "Estimation of Room Acoustic Parameters: The ACE Challenge". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 24, pp. 1681–1693. ISSN: 23299290. doi: [10.1109/TASLP.2016.2577502](https://doi.org/10.1109/TASLP.2016.2577502) (cit. on p. 91).
- El Baba, Youssef, Andreas Walther, and Emanuël AP Habets (2017). "Time of arrival disambiguation using the linear Radon transform". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 106–110 (cit. on pp. 52, 53).
- Ernst, Ori, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger (2018). "Speech dereverberation using fully convolutional networks". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 390–394 (cit. on pp. 103, 104).
- Falk, Tiago H, Chenxi Zheng, and Wai-Yip Chan (2010). "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, pp. 1766–1774 (cit. on p. 139).
- Farina, Angelo (2000). "Simultaneous measurement of impulse response and distortion with a swept-sine technique". In: *Audio Engineering Society Convention 108*. Audio Engineering Society (cit. on p. 50).
- (2007). "Advancements in impulse response measurements by sine sweeps". In: *Audio Engineering Society Convention 122*. Audio Engineering Society (cit. on pp. 50, 89).

- Ferguson, Eric L, Stefan B Williams, and Craig T Jin (2019). "Improved multipath time delay estimation using cepstrum subtraction". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 551–555 (cit. on p. 51).
- Févotte, Cédric, Nancy Bertin, and Jean-Louis Durrieu (2009). "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis". In: *Neural computation* 21.3, pp. 793–830 (cit. on p. 111).
- Filos, Jason, Antonio Canclini, Mark RP Thomas, Fabio Antonacci, Augusto Sarti, and Patrick A Naylor (2011). "Robust inference of room geometry from acoustic measurements using the Hough transform". In: *2011 19th European Signal Processing Conference*. IEEE, pp. 161–165 (cit. on p. 52).
- Flanagan, James L, Arun C Surendran, and Ea-Ee Jan (1993). "Spatially selective sound capture for speech and audio processing". In: *Speech Communication* 13.1-2, pp. 207–222 (cit. on pp. 5, 103, 133).
- Fourier, Jean Baptiste Joseph (1822). *Théorie analytique de la chaleur*. F. Didot (cit. on p. 33).
- Frost, Otis Lamont (1972). "An algorithm for linearly constrained adaptive array processing". In: *Proceedings of the IEEE* 60.8, pp. 926–935 (cit. on p. 102).
- Gannot, Sharon, David Burshtein, and Ehud Weinstein (2001). "Signal enhancement using beamforming and nonstationarity with applications to speech". In: *IEEE Transactions on Signal Processing* 49.8, pp. 1614–1626 (cit. on pp. 41, 56, 103, 133, 137).
- Gannot, Sharon, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov (2017). "A consolidated perspective on multimicrophone speech enhancement and source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.4, pp. 692–730 (cit. on pp. 100, 102, 136, 137).
- Garofolo, John S, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue (1993). "TIMIT acoustic-phonetic continuous speech corpus". In: *Linguistic data consortium* 10.5, p. 0 (cit. on p. 115).
- Gaultier, Clément, Saurabh Kataria, and Antoine Deleforge (2017). "VAST: The virtual acoustic space traveler dataset". In: *Lecture Notes in Computer Science*. Vol. 10169 LNCS, pp. 68–79. ISBN: 9783319535463. DOI: 10.1007/978-3-319-53547-0{_}7 (cit. on pp. 74, 75, 79, 80, 105, 106, 131).
- Genovese, Andrea F, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J Tashev (2019). "Blind room volume estimation from single-channel noisy speech". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 231–235 (cit. on pp. 57, 58).
- Gilloire, Andre and Martin Vetterli (1992). "Adaptive filtering in sub-bands with critical sampling: analysis, experiments, and application to acoustic echo cancellation". In: *IEEE transactions on signal processing* 40.ARTICLE, pp. 1862–1875 (cit. on p. 40).
- Gomez, Randy, Deborah Szapiro, Kerl Galindo, and Keisuke Nakamura (2018). "Haru: Hardware design of an experimental tabletop robot assistant". In: *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pp. 233–240 (cit. on p. 118).
- Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. MIT press Cambridge (cit. on pp. 74, 76).
- Griesinger, David (1997). "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces". In: *Acta Acustica united with Acustica* 83.4, pp. 721–731 (cit. on pp. 27, 28).
- Guillemain, Philippe and Richard Kronland-Martinet (1996). "Characterization of acoustic signals through continuous linear time-frequency representations". In: *Proceedings of the IEEE* 84.4, pp. 561–585 (cit. on p. 51).
- Habets, Emanuel AP (2006). "Room impulse response generator". In: *Technische Universiteit Eindhoven, Tech. Rep* 2.2.4, p. 1 (cit. on pp. 25, 156).
- Habets, Emanuël AP and Sharon Gannot (2007). "Generating sensor signals in isotropic noise fields". In: *The Journal of the Acoustical Society of America* 122.6, pp. 3464–3470 (cit. on p. 32).
- Hadad, Elior, Florian Heese, Peter Vary, and Sharon Gannot (2014). "Multichannel audio database in various acoustic environments". In: *2014 14th International Workshop on Acoustic Signal Enhancement, IWAENC 2014*, pp. 313–317. ISBN: 9781479968084. DOI: 10.1109/IWAENC.2014.6954309 (cit. on p. 86).

- Heinz, Renate (1993). "Binaural room simulation based on an image source model with addition of statistical methods to include the diffuse sound scattering of walls and to predict the reverberant tail". In: *Applied Acoustics* 38.2-4, pp. 145–159 (cit. on p. 23).
- Hershey, J. R., Z. Chen, J. Le Roux, and S. Watanabe (2016). "Deep clustering: Discriminative embeddings for segmentation and separation". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35 (cit. on p. 100).
- Huang, Yiteng and Jacob Benesty (2003). "A class of frequency-domain adaptive approaches to blind multichannel identification". In: *IEEE Transactions on signal processing* 51.1, pp. 11–24 (cit. on pp. 54, 59).
- Huang, Yiteng, Jacob Benesty, and Jingdong Chen (2005). "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment". In: *IEEE Transactions on Speech and Audio Processing* 13.5, pp. 882–895 (cit. on p. 108).
- Huggett, A St G (1953). "Human Senses". In: *British Medical Journal* 2.4851, p. 1417 (cit. on p. 2).
- Jager, Ingmar, Richard Heusdens, and Nikolay D Gaubitch (2016). "Room geometry estimation from acoustic echoes using graph-based echo labeling". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 52).
- Jan, E, Piergiorgio Svaizer, and James L Flanagan (1995). "Matched-filter processing of microphone array for spatial volume selectivity". In: *Proceedings of ISCAS'95-International Symposium on Circuits and Systems*. Vol. 2. IEEE, pp. 1460–1463 (cit. on p. 5).
- Jan, Ea-Ee and James Flanagan (1996). "Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors". In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 2. IEEE, pp. 917–920 (cit. on p. 5).
- Javed, Hamza A, Alastair H Moore, and Patrick A Naylor (2016). "Spherical microphone array acoustic rake receivers". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 111–115 (cit. on p. 133).
- Jensen, Jesper Rindom, Osama Saqib, and Sharon Gannot (2019). "An EM method for multichannel TOA and DOA estimation of acoustic echoes". In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 120–124 (cit. on pp. 53, 62, 149).
- Jia, Hongjian, Xiukun Li, Xiangxia Meng, and Yang Yang (2017). "Extraction of echo characteristics of underwater target based on cepstrum method". In: *Journal of Marine Science and Application* 16.2, pp. 216–224 (cit. on p. 51).
- Jin, Yuchen, Qiuyang Shen, Xuqing Wu, Jiefu Chen, Yueqin Huang, et al. (2020). "A Physics-Driven Deep-Learning Network for Solving Nonlinear Inverse Problems". In: *Petrophysics* 61.01, pp. 86–98 (cit. on pp. 85, 149).
- Kearney, Gavin, Marcin Gorzel, Henry Rice, and Frank Boland (2012). "Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields". In: *Acta Acustica united with Acustica* 98.1, pp. 61–71 (cit. on p. 27).
- Kelly, Ian J and Francis M Boland (2014). "Detecting arrivals in room impulse responses with dynamic time warping". In: *IEEE/ACM transactions on audio, speech, and language processing* 22.7, pp. 1139–1147 (cit. on pp. 50, 51).
- Kleeman, Lindsay and Roman Kuc (2016). "Sonar sensing". In: *Springer Handbook of Robotics*. Springer, pp. 753–782 (cit. on p. 150).
- Knapp, Charles and Glifford Carter (1976). "The generalized correlation method for estimation of time delay". In: *IEEE transactions on acoustics, speech, and signal processing* 24.4, pp. 320–327 (cit. on pp. 105, 121).
- Kodrasi, Ina and Simon Doclo (2017). "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods". In: *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, pp. 116–120 (cit. on pp. 41, 56, 133, 138).
- Kolarik, Andrew J, Silvia Cirstea, Shahina Pardhan, and Brian CJ Moore (2014). "A summary of research investigating echolocation abilities of blind and sighted humans". In: *Hearing research* 310, pp. 60–68 (cit. on p. 2).

- Koldovsky, Zbynek and Petr Tichavsky (2015). "Sparse reconstruction of incomplete relative transfer function: Discrete and continuous time domain". In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 394–398 (cit. on pp. 41, 56).
- Koldovský, Zbyněk, Jiří Málek, and Sharon Gannot (2015). "Spatial source subtraction based on incomplete measurements of relative transfer function". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.8, pp. 1335–1347 (cit. on p. 56).
- Korhonen, Teemu (2008). "Acoustic localization using reverberation with virtual microphones". In: *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Citeseer, pp. 211–223 (cit. on p. 119).
- Kowalczyk, Konrad (2019). "Raking early reflection signals for late reverberation and noise reduction". In: *The Journal of the Acoustical Society of America* 145.3, EL257–EL263. ISSN: 0001-4966. doi: [10.1121/1.5095535](https://doi.org/10.1121/1.5095535). URL: <http://dx.doi.org/10.1121/1.5095535> (cit. on pp. 5, 103, 133, 134, 137, 138, 149).
- Kowalczyk, Konrad, Emanuël AP Habets, Walter Kellermann, and Patrick A Naylor (2013). "Blind system identification using sparse learning for TDOA estimation of room reflections". In: *IEEE Signal Processing Letters* 20.7, pp. 653–656 (cit. on pp. 55, 65).
- Kreković, Miranda, Ivan Dokmanić, and Martin Vetterli (2016). "EchoSLAM: Simultaneous localization and mapping with acoustic echoes". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, pp. 11–15 (cit. on p. 119).
- Krokstad, Asbjørn, Staffan Strom, and Svein Sørsdal (1968). "Calculating the acoustical room response by the use of a ray tracing technique". In: *Journal of Sound and Vibration* 8.1, pp. 118–125 (cit. on p. 18).
- Kulowski, Andrzej (1985). "Algorithmic representation of the ray tracing technique". In: *Applied Acoustics* 18.6, pp. 449–469 (cit. on p. 22).
- Kuster, Martin (2008). "Reliability of estimating the room volume from a single room impulse response". In: *The Journal of the Acoustical Society of America* 124.2, pp. 982–993 (cit. on p. 50).
- (2012). "Objective sound field analysis based on the coherence estimated from two microphone signals". In: *The Journal of the Acoustical Society of America* 131.4, pp. 3284–3284 (cit. on p. 135).
- Kuttruff, Heinrich (2016a). *Room acoustics*. CRC Press (cit. on pp. 13, 16, 18, 50, 157).
- (2016b). *Room acoustics*. Crc Press (cit. on p. 135).
- Laufer, Bracha, Ronen Talmon, and Sharon Gannot (2013). "Relative transfer function modeling for supervised source localization". In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 1–4 (cit. on pp. 42, 75, 105, 106).
- Le Roux, Jonathan, John R Hershey, and Felix Weninger (2015). "Deep NMF for speech separation". In: *Proc. IEEE ICASSP*, pp. 66–70 (cit. on p. 101).
- Lebarbenchon, Romain, Ewen Camberlein, **Di Carlo, Diego**, Clément Gaultier, Antoine Deleforge, and Nancy Bertin (2018a). "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge". In: *Proc. of LOCATA Challenge Workshop-a satellite event of IWAENC* (cit. on p. 9).
- Lebarbenchon, Romain, Ewen Camberlein, Diego Di Carlo, Clément Gaultier, Antoine Deleforge, and Nancy Bertin (2018b). "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge". In: *Proc. of LOCATA Challenge Workshop-a satellite event of IWAENC* (cit. on pp. 105, 148).
- Lee, Daniel D. and H. Sebastian Seung (2001). "Algorithms for Non-negative Matrix Factorization". In: *Advances in Neural Information Processing Systems* 13. Ed. by T. K. Leen, T. G. Dietterich, and V. Tresp. MIT Press, pp. 556–562 (cit. on p. 111).
- Leglaive, Simon, Roland Badeau, and Gaël Richard (2015). "Multichannel audio source separation with probabilistic reverberation modeling". In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 1–5 (cit. on pp. 101, 108, 149).
- (2016). "Multichannel audio source separation with probabilistic reverberation priors". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.12, pp. 2453–2465 (cit. on pp. 5, 54, 55, 108).

- Leglaive, Simon, Roland Badeau, and Gaël Richard (2018). "Student's t source and mixing models for multichannel audio source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.6, pp. 1154–1168 (cit. on pp. 54, 55).
- Li, Bo, Tara N Sainath, Ron J Weiss, Kevin W Wilson, and Michiel Bacchiani (2016a). "Neural network adaptive beamforming for robust multichannel speech recognition". In: *Google Research* (cit. on p. 103).
- Li, Xiaofei, Laurent Girin, Radu Horaud, and Sharon Gannot (2016b). "Estimation of the direct-path relative transfer function for supervised sound-source localization". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.11, pp. 2171–2186 (cit. on p. 105).
- Li, Xiaofei, Laurent Girin, and Radu Horaud (2019). "Expectation-maximisation for speech source separation using convolutive transfer function". In: *CAAI Transactions on Intelligence Technology* 4.1, pp. 47–53 (cit. on p. 101).
- Lin, Yuanqing and Daniel D Lee (2006). "Bayesian regularization and nonnegative deconvolution for room impulse response estimation". In: *IEEE Transactions on Signal Processing* 54.3, pp. 839–847 (cit. on p. 50).
- Lin, Yuanqing, Jingdong Chen, Youngmoo Kim, and Daniel D Lee (2007). "Blind sparse-nonnegative (BSN) channel identification for acoustic time-difference-of-arrival estimation". In: *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 106–109 (cit. on pp. 55, 64, 65, 70).
- (2008). "Blind channel identification for speech dereverberation using l1-norm sparse learning". In: *Advances in Neural Information Processing Systems*, pp. 921–928 (cit. on pp. 55, 65, 66, 73).
- Lindau, Alexander, Linda Kosanke, and Stefan Weinzierl (2012). "Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses". In: *Journal of the Audio Engineering Society* 60.11, pp. 887–898 (cit. on p. 28).
- Looney, David and Nikolay D Gaubitch (2020). "Joint Estimation Of Acoustic Parameters From Single-Microphone Speech Observations". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 431–435 (cit. on p. 149).
- Loutridis, Spyros J (2005). "Decomposition of impulse responses using complex wavelets". In: *Journal of the Audio Engineering Society* 53.9, pp. 796–811 (cit. on p. 51).
- Luo, Yi and Nima Mesgarani (2019). "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8, pp. 1256–1266 (cit. on p. 101).
- Makino, Shoji (2018). *Audio Source Separation*. Vol. 433. Springer (cit. on pp. 100, 107).
- Markovich, Shmulik, Sharon Gannot, and Israel Cohen (2009). "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.6, pp. 1071–1086 (cit. on pp. 41, 103).
- Morgan, Dennis R, Jacob Benesty, and M Mohan Sondhi (1998). "On the evaluation of estimated impulse responses". In: *IEEE Signal processing letters* 5.7, pp. 174–176 (cit. on p. 59).
- Müller, Meinard (2015). *Fundamentals of Music Processing*. Springer Verlag. ISBN: 978-3-319-21944-8 (cit. on pp. 39, 113).
- Nabian, Mohammad Amin and Hadi Meidani (2020). "Physics-driven regularization of deep neural networks for enhanced engineering design and analysis". In: *Journal of Computing and Information Science in Engineering* 20.1 (cit. on pp. 85, 131, 149).
- Nadisic, Nicolas, Arnaud Vandaele, Nicolas Gillis, and Jeremy E Cohen (2020). "Exact Sparse Nonnegative Least Squares". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5395–5399 (cit. on p. 149).
- Nakashima, Hiromichi, Mitsuru Kawamoto, and Toshiharu Mukai (2010). "A localization method for multiple sound sources by using coherence function". In: *2010 18th European Signal Processing Conference*. IEEE, pp. 130–134 (cit. on p. 119).

- Naylor, Patrick A, Anastasis Kounoudes, Jon Gudnason, and Mike Brookes (2006). "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1, pp. 34–43 (cit. on p. 50).
- Naylor, Patrick and Eds Gaubitch (2010). *Dereverberation, Speech*. Berlin, Germany: Springer-Verlag (cit. on p. 135).
- Neely, Stephen T and Jont B Allen (1979). "Invertibility of a room impulse response". In: *The Journal of the Acoustical Society of America* 66.1, pp. 165–169 (cit. on p. 102).
- Nesta, Francesco and Maurizio Omologo (2012). "Convulsive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 222–230 (cit. on p. 101).
- Nguyen, Quan, Laurent Girin, Gérard Bailly, Frédéric Elisei, and Duc-Canh Nguyen (2018). "Autonomous sensorimotor learning for sound source localization by a humanoid robot". In: (cit. on pp. 75, 82, 105, 106).
- Nugraha, Aditya Arie, Antoine Liutkus, and Emmanuel Vincent (2016). "Multichannel audio source separation with deep neural networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.9, pp. 1652–1664 (cit. on p. 100).
- O'Donovan, Adam E, Ramani Duraiswami, and Dmitry N Zotkin (2010). "Automatic matched filter recovery via the audio camera". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 2826–2829 (cit. on pp. 54, 57).
- O'Donovan, Adam, Ramani Duraiswami, and Dmitry Zotkin (2008). "Imaging concert hall acoustics using visual and audio cameras". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 5284–5287 (cit. on pp. 54, 57).
- Oppenheim, Alan V (1987). *Signals and Systems: An Introduction to Analog and Digital Signal Processing*. MIT Center for Advanced Engineering Study (cit. on p. 39).
- Ozerov, Alexey and Cédric Févotte (2009). "Multichannel nonnegative matrix factorization in convulsive mixtures for audio source separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3, pp. 550–563 (cit. on pp. 54, 148).
- (2010). "Multichannel nonnegative matrix factorization in convulsive mixtures for audio source separation". In: *IEEE Trans. Audio, Speech, Language Process.* 18.3, pp. 550–563 (cit. on pp. 100, 101, 111–115).
- Pan, Hanjie, Thierry Blu, and Martin Vetterli (2016). "Towards generalized FRI sampling with an application to source resolution in radioastronomy". In: *IEEE Transactions on Signal Processing* 65.4, pp. 821–835 (cit. on p. 150).
- Parhizkar, Reza, Ivan Dokmanić, and Martin Vetterli (2014). "Single-channel indoor microphone localization". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1434–1438 (cit. on p. 52).
- Park, Yongsung, Woojae Seong, and Youngmin Choo (2017). "Compressive time delay estimation off the grid". In: *The Journal of the Acoustical Society of America* 141.6, EL585–EL591 (cit. on p. 62).
- Paul, Douglas B and Janet M Baker (1992). "The design for the Wall Street Journal-based CSR corpus". In: *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, pp. 357–362 (cit. on p. 89).
- Pavlović, Milan, Dragan M Ristić, Irini Reljin, and Miomir Mijić (2016). "Multifractal analysis of visualized room impulse response for detecting early reflections". In: *The Journal of the Acoustical Society of America* 139.5, EL113–EL117 (cit. on p. 51).
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830 (cit. on p. 115).
- Peled, Yotam and Boaz Rafaely (2013). "Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field". In: *IEEE transactions on audio, speech, and language processing* 21.12, pp. 2532–2540 (cit. on pp. 51, 103, 133).

- Perotin, Lauréline, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin (2018). "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector". In: *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, pp. 241–245 (cit. on pp. 75, 105, 106).
- Pierce, Allan D (2019). *Acoustics: an introduction to its physical principles and applications*. Springer (cit. on pp. 13, 18).
- Plinge, Axel, Florian Jacob, Reinhold Haeb-Umbach, and Gernot A. Fink (2016). "Acoustic microphone geometry calibration". In: *IEEE Signal Processing Magazine* July, pp. 14–28 (cit. on p. 90).
- Purwins, Hendrik, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath (2019). "Deep learning for audio signal processing". In: *IEEE Journal of Selected Topics in Signal Processing* 13.2, pp. 206–219 (cit. on p. 74).
- Qi, Yuanlei, Feiran Yang, Ming Wu, and Jun Yang (2019). "A Broadband Kalman Filtering Approach to Blind Multichannel Identification". In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 102.6, pp. 788–795 (cit. on p. 55).
- Raffel, Colin, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel (2014). "mir_eval: A transparent implementation of common MIR metrics". In: *Proc. ISMIR* (cit. on p. 116).
- Rao, Chengping, Hao Sun, and Yang Liu (2020). "Physics-informed deep learning for incompressible laminar flows". In: *arXiv preprint arXiv:2002.10558* (cit. on pp. 85, 149).
- Rascon, Caleb and Ivan Meza (2017). "Localization of sound sources in robotics: A review". In: *Robotics and Autonomous Systems* 96, pp. 184–210 (cit. on pp. 104, 150).
- Remaggi, Luca, Philip JB Jackson, Philip Coleman, and Wenwu Wang (2016). "Acoustic reflector localization: novel image source reversion and direct localization methods". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.2, pp. 296–309 (cit. on pp. 5, 50, 58, 141).
- Remaggi, Luca, Philip JB Jackson, and Wenwu Wang (2019). "Modeling the Comb Filter Effect and Interaural Coherence for Binaural Source Separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12, pp. 2263–2277 (cit. on p. 5).
- Ribeiro, Flávio, Demba Ba, Cha Zhang, and Dinei Florêncio (2010a). "Turning enemies into friends: Using reflections to improve sound source localization". In: *2010 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 731–736 (cit. on pp. 5, 47, 55, 104, 119).
- Ribeiro, Flávio, Cha Zhang, Dinei A Florêncio, and Demba Elimane Ba (2010b). "Using reverberation to improve range and elevation discrimination for small array sound source localization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, pp. 1781–1792 (cit. on p. 119).
- Rickard, Scott (2007). "The DUET blind source separation algorithm". In: *Blind Speech Separation*, pp. 217–241 (cit. on pp. 100, 101).
- Ristić, Dragan M, Milan Pavlović, Dragana Šumarac Pavlović, and Irini Reljin (2013). "Detection of early reflections using multifractals". In: *The Journal of the Acoustical Society of America* 133.4, EL235–EL241 (cit. on p. 51).
- Rix, Antony W, John G Beerends, Michael P Hollier, and Andries P Hekstra (2001). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE, pp. 749–752 (cit. on p. 139).
- Rotili, Rudy, Claudio De Simone, Alessandro Perelli, Simone Cifani, and Stefano Squartini (2010). "Joint multichannel blind speech separation and dereverberation: A real-time algorithmic implementation". In: *International Conference on Intelligent Computing*. Springer, pp. 85–93 (cit. on p. 108).
- Roy, Robert, Arogyaswami Paulraj, and Thomas Kailath (1986). "ESPRIT-A subspace rotation approach to estimation of parameters of cisoids in noise". In: *IEEE transactions on acoustics, speech, and signal processing* 34.5, pp. 1340–1342 (cit. on p. 51).
- Rudin, Walter (1987). "Real and complex analysis (mcgraw-hill international editions: Mathematics series)". In: (cit. on p. 68).

- Rui, Yong and Dinei Florencio (2004). "Time delay estimation in the presence of correlated noise and reverberation". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. IEEE, pp. ii–133 (cit. on p. 119).
- Sacks, Oliver (2014). *Musicophilia*. Adelphi Edizioni spa (cit. on p. 2).
- Sainath, Tara N, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al. (2017). "Multichannel signal processing with deep neural networks for automatic speech recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.5, pp. 965–979 (cit. on pp. 75, 103, 104).
- Salvati, Daniele, Carlo Drioli, and Gian Luca Foresti (2016). "Sound source and microphone localization from acoustic impulse responses". In: *IEEE Signal Processing Letters* 23.10, pp. 1459–1463 (cit. on pp. 52, 119).
- (2018). "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 2.2, pp. 103–116 (cit. on pp. 75, 104, 129).
- Santos, João F and Tiago H Falk (2014). "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.12, pp. 2197–2206 (cit. on p. 139).
- Saqib, Usama, Sharon Gannot, and Jesper Rindom Jensen (2020). "Estimation of acoustic echoes using expectation–maximization methods". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2020.1, pp. 1–15 (cit. on p. 53).
- Sato, Haruo, Michael C Fehler, and Takuto Maeda (2012). *Seismic wave propagation and scattering in the heterogeneous earth*. Springer Science & Business Media (cit. on p. 150).
- Sato, Yoichi (1975). "A method of self-recovering equalization for multilevel amplitude-modulation systems". In: *IEEE Transactions on communications* 23.6, pp. 679–682 (cit. on p. 54).
- Savioja, Lauri and U Peter Svensson (2015). "Overview of geometrical room acoustic modeling techniques". In: *The Journal of the Acoustical Society of America* 138.2, pp. 708–730 (cit. on pp. 18, 22, 23, 28).
- Sawada, Hiroshi, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda (2013). "Multichannel extensions of non-negative matrix factorization with complex-valued data". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.5, pp. 971–982 (cit. on p. 100).
- Scheibler, Robin, Ivan Dokmanić, and Martin Vetterli (2015). "Raking echoes in the time domain". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 554–558 (cit. on pp. 103, 133).
- Scheibler, Robin, Eric Bezzam, and Ivan Dokmanić (2018a). "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms". In: *Proc. IEEE ICASSP*. accepted. Calgary, CA (cit. on p. 114).
- (2018b). "Pyroomacoustics: A python package for audio room simulation and array processing algorithms". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 351–355 (cit. on p. 71).
- Scheibler, Robin, Diego Di Carlos, Antoine Deleforge, and Ivan Dokmanic (2018c). "Separake: Source Separation with a Little Help from Echoes". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2018-April, pp. 6897–6901. ISBN: 9781538646588. doi: [10.1109/ICASSP.2018.8461345](https://doi.org/10.1109/ICASSP.2018.8461345). URL: <http://arxiv.org/abs/1711.06805> (cit. on p. 5).
- Scheibler, Robin, **Di Carlo, Diego**, Antoine Deleforge, and Ivan Dokmanić (2018d). "Separake: Source separation with a little help from echoes". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901 (cit. on p. 9).
- Scheibler, Robin, Diego Di Carlo, Antoine Deleforge, and Ivan Dokmanić (2018e). "Separake: Source separation with a little help from echoes". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901 (cit. on pp. 54, 72, 107, 108, 110, 111, 114, 115).
- Scheuing, Jan and Bin Yang (2006). "Disambiguation of TDOA estimates in multi-path multi-source environments (DATEMM)". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 4. IEEE, pp. IV–IV (cit. on p. 52).

- Schimmel, Steven M, Martin F Muller, and Norbert Dillier (2009). "A fast and accurate "shoebox" room acoustics simulator". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 241–244 (cit. on pp. 23, 80, 125).
- Schmidt, Mikkel N and Rasmus K Olsson (2006). "Single-channel speech separation using sparse non-negative matrix factorization". In: *Ninth International Conference on Spoken Language Processing* (cit. on pp. 101, 113).
- Schröder, Dirk, Philipp Dross, and Michael Vorländer (2007). "A fast reverberation estimator for virtual environments". In: *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society (cit. on p. 23).
- Schroeder, Manfred R (1979). "Integrated-impulse method measuring sound decay without using impulses". In: *The Journal of the Acoustical Society of America* 66.2, pp. 497–500 (cit. on pp. 49, 50).
- Schwartz, Ofer, Sharon Gannot, and Emanuël AP Habets (2014). "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.2, pp. 240–251 (cit. on pp. 103, 137).
- (2016). "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm". In: *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1123–1127 (cit. on pp. 103, 138).
- Smaragdis, Paris, Madhusudana Shashanka, and Bhiksha Raj (2009). "A sparse non-parametric approach for single channel separation of known sounds". In: *Advances in neural information processing systems*, pp. 1705–1713 (cit. on pp. 101, 113).
- Sokol, Joshua (2017). "The thoughts of a spiderweb". In: *Obtenido de: https://www.quantamagazine.org/the-thoughts-of-a-spiderweb-20170523* (cit. on p. 2).
- Stoica, Petre and Kenneth C Sharman (1990). "Maximum likelihood methods for direction-of-arrival estimation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.7, pp. 1132–1143 (cit. on p. 105).
- Stöter, Fabian-Robert, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji (2019). "Open-unmix-a reference implementation for music source separation". In: (cit. on p. 100).
- Sturmel, Nicolas, Antoine Liutkus, Jonathan Pinel, Laurent Girin, Sylvain Marchand, Gaël Richard, Roland Badeau, and Laurent Daudet (2012). "Linear mixing models for active listening of music productions in realistic studio conditions". In: *Proceedings of the Audio Engineering Society Convention*. 8594. IEEE (cit. on p. 31).
- Sun, D L and G J Mysore (2013). "Universal speech models for speaker independent single channel source separation". In: *IEEE ICASSP*, pp. 141–145 (cit. on pp. 111, 114, 115).
- Svaizer, Piergiorgio, Alessio Brutti, and Maurizio Omologo (2011). "Use of reflected wavefronts for acoustic source localization with a line array". In: *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*. IEEE, pp. 165–169 (cit. on p. 119).
- Szöke, Igor, Miroslav Skácel, Ladislav Mošner, Jakub Palísek, and Jan Honza Černocký (2019). "Building and evaluation of a real room impulse response dataset". In: *IEEE Journal of Selected Topics in Signal Processing* 13.4, pp. 863–876 (cit. on pp. 50, 57, 58).
- Taghizadeh, Mohammad Javad, Afsaneh Asaei, Saeid Haghigatshoar, Philip N Garner, and Herve Bourlard (2015). "Spatial sound localization via multipath euclidean distance matrix recovery". In: *IEEE Journal of Selected Topics in Signal Processing* 9.5, pp. 802–814 (cit. on p. 119).
- Takao, Kazuaki, M Fujita, and T Nishi (1976). "An adaptive antenna array under directional constraint". In: *IEEE Transactions on Antennas and Propagation* 24.5, pp. 662–669 (cit. on p. 103).
- Tammen, Marvin, Ina Kodrasi, and Simon Doclo (2018). "Iterative Alternating Least-Squares Approach to Jointly Estimate the RETFs and the Diffuse PSD". In: *Speech Communication; 13th ITG-Symposium*. VDE, pp. 1–5 (cit. on pp. 41, 59, 135, 138).
- Tervo, Sakari (2011). "Localization and tracing of early acoustic reflections". PhD thesis (cit. on pp. 54, 141).

- Tervo, Sakari and Archontis Politis (2015). "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.10, pp. 1539–1551 (cit. on p. 54).
- Tervo, Sakari, Teemu Korhonen, and Tapani Lokki (2011). "Estimation of reflections from impulse responses". In: *Building Acoustics* 18.1-2, pp. 159–173 (cit. on p. 53).
- Thiergart, Oliver and Emanuël AP Habets (2013). "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 659–663 (cit. on p. 103).
- Thiergart, Oliver, Maja Taseska, and Emanuël AP Habets (2014). "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.12, pp. 2182–2196 (cit. on p. 137).
- Thomas, Matthew Reuben (2017). "Wayverb: A Graphical Tool for Hybrid Room Acoustics Simulation". PhD thesis. University of Huddersfield (cit. on pp. 21, 24).
- Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288 (cit. on p. 64).
- Tong, Lang and Sylvie Perreau (1998). "Multichannel blind identification: From subspace to maximum likelihood methods". In: *Proceedings of the IEEE* 86.10, pp. 1951–1968 (cit. on pp. 54, 55).
- Tong, Lang, Guanghan Xu, and Thomas Kailath (1994). "Blind identification and equalization based on second-order statistics: A time domain approach". In: *IEEE Transactions on Information Theory* 40.2, pp. 340–349 (cit. on pp. 55, 65).
- Tropp, Joel A and Stephen J Wright (2010). "Computational methods for sparse solution of linear inverse problems". In: *Proceedings of the IEEE* 98.6, pp. 948–958 (cit. on p. 149).
- Tu, Yan-Hui, Jun Du, and Chin-Hui Lee (2019). "Speech Enhancement Based on Teacher-Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12, pp. 2080–2091 (cit. on p. 85).
- Tufte, Edward R and Peter R Graves-Morris (1983). *The visual display of quantitative information*. Vol. 2. 9. Graphics press Cheshire, CT (cit. on p. 10).
- Tukuljac, Helena Peic (2020). "Sparse and Parametric Modeling with Applications to Acoustics and Audio". PhD thesis. École polytechnique fédérale de Lausanne (cit. on pp. 57, 73).
- Tukuljac, Helena Peic, Antoine Deleforge, and Rémi Gribonval (2018). "MULAN: a blind and off-grid method for multichannel echo retrieval". In: *Advances in Neural Information Processing Systems*, pp. 2182–2192 (cit. on pp. 37, 38, 50, 56, 57, 63).
- Tuzlukov, Vyacheslav (2018). *Signal processing noise*. CRC Press (cit. on p. 32).
- Tzinis, Efthymios, Zhepei Wang, and Paris Smaragdis (2020). "Sudo rm-rf: Efficient Networks for Universal Audio Source Separation". In: *arXiv preprint arXiv:2007.06833* (cit. on p. 101).
- Usher, John (2010). "An improved method to determine the onset timings of reflections in an acoustic impulse response". In: *The Journal of the Acoustical Society of America* 127.4, EL172–EL177 (cit. on pp. 50, 51).
- Välimäki, Vesa, Julian Parker, Lauri Savioja, Julius O Smith, and Jonathan Abel (2016). "More than 50 years of artificial reverberation". In: *Audio engineering society conference: 60th international conference: dreams (dereverberation and reverberation of audio, music, and speech)*. Audio Engineering Society (cit. on pp. 22, 27).
- Van Trees, Harry L (2004). *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons (cit. on pp. 102, 133).
- Van Veen, Barry D and Kevin M Buckley (1988). "Beamforming: A versatile approach to spatial filtering". In: *IEEE assp magazine* 5.2, pp. 4–24 (cit. on pp. 102, 103).
- Vanderveen, Michaela C, Constantinos B Papadias, and Arogyaswami Paulraj (1997). "Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array". In: *IEEE Communications letters* 1.1, pp. 12–14 (cit. on p. 53).

- Varzandeh, Reza, Maja Taseska, and Emanuël AP Habets (2017). "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation". In: *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, pp. 11–15 (cit. on p. 59).
- Venkateswaran, Sriram and Upamanyu Madhow (2012). "Localizing multiple events using times of arrival: a parallelized, hierarchical approach to the association problem". In: *IEEE Transactions on Signal Processing* 60.10, pp. 5464–5477 (cit. on p. 52).
- Verhaevert, Jo, Emmanuel Van Lil, and Antoine Van de Capelle (2004). "Direction of arrival (DOA) parameter estimation with the SAGE algorithm". In: *Signal Processing* 84.3, pp. 619–629 (cit. on p. 53).
- Vesa, Sampo (2009). "Binaural sound source distance learning in rooms". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.8, pp. 1498–1507 (cit. on p. 104).
- Vesa, Sampo and Tapio Lokki (2010). "Segmentation and analysis of early reflections from a binaural room impulse response". In: *Helsinki University of Technology: Technical Report TKK-ME-RI, TKK Reports in Media Technology* (cit. on p. 51).
- Vesperini, Fabio, Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, and Francesco Piazza (2018). "Localizing speakers in multiple rooms by using deep neural networks". In: *Computer Speech & Language* 49, pp. 83–106 (cit. on pp. 75, 105, 106).
- Vincent, Emmanuel, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian P Rosca (2007). "First stereo audio source separation evaluation campaign: data, algorithms and results". In: *International Conference on Independent Component Analysis and Signal Separation*. Springer, pp. 552–559 (cit. on pp. 114, 116).
- Vincent, Emmanuel, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot (2014). "From blind to guided audio source separation: How models and side information can improve the separation of sound". In: *IEEE Signal Processing Magazine* 31.3, pp. 107–115 (cit. on p. 99).
- Vincent, Emmanuel, Tuomas Virtanen, and Sharon Gannot (2018). *Audio source separation and speech enhancement*. John Wiley & Sons (cit. on pp. 20, 29, 39, 78, 97, 100, 102, 104, 107).
- Wallach, Hans, Edwin B Newman, and Mark R Rosenzweig (1973). "The precedence effect in sound localization (tutorial reprint)". In: *Journal of the audio engineering society* 21.10, pp. 817–826 (cit. on p. 27).
- Wang, Zhong-Qiu, Jonathan Le Roux, and John R Hershey (2018). "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 100).
- Weinstein, Ehud, Alan V Oppenheim, Meir Feder, and John R Buck (1994). "Iterative and sequential algorithms for multisensor signal enhancement". In: *IEEE Transactions on Signal Processing* 42.4, pp. 846–859 (cit. on p. 119).
- Woodward, Philip M and Ian L Davies (1952). "Information theory and inverse probability in telecommunication". In: *Proceedings of the IEE-Part III: Radio and Communication Engineering* 99.58, pp. 37–44 (cit. on p. 29).
- Xiao, Xiong, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu (2016). "Deep beamforming networks for multi-channel speech recognition". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5745–5749 (cit. on pp. 75, 103, 104).
- Xu, Guanghan, Hui Liu, Lang Tong, and Thomas Kailath (1995). "A least-squares approach to blind channel identification". In: *IEEE Transactions on signal processing* 43.12, pp. 2982–2993 (cit. on pp. 54, 55, 62).
- Yu, Meng, Wenye Ma, Jack Xin, and Stanley Osher (2011). "Multi-Channel $l_{\{1\}}$ Regularized Convex Speech Enhancement Model and Fast Computation by the Split Bregman Method". In: *IEEE transactions on audio, speech, and language processing* 20.2, pp. 661–675 (cit. on p. 65).
- Zahorik, Pavel (2002). "Direct-to-reverberant energy ratio sensitivity". In: *The Journal of the Acoustical Society of America* 112.5, pp. 2110–2117 (cit. on p. 28).
- Zannini, Cecilia Maria, Albenzio Cirillo, Raffaele Parisi, and Aurelio Uncini (2010). "Improved TDOA disambiguation techniques for sound source localization in reverberant environments". In: *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, pp. 2666–2669 (cit. on p. 52).

- Zhang, Cha, Zhengyou Zhang, and Dinei Florêncio (2007). "Maximum likelihood sound source localization for multiple directional microphones". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 1. IEEE, pp. I-125 (cit. on p. 119).
- Zohny, Zeinab and Jonathon Chambers (2014). "Modelling interaural level and phase cues with Student's t-distribution for robust clustering in MESSL". In: *2014 19th International Conference on Digital Signal Processing*. IEEE, pp. 59–62 (cit. on p. 83).
- Di Carlo, Diego**, Antoine Deleforge, and Nancy Bertin (2019). "Mirage: 2D source localization using microphone pair augmentation with echoes". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 775–779 (cit. on p. 9).
- Di Carlo, Diego**, Clement Elvira, Antoine Deleforge, Nancy Bertin, and Rémi Gribonval (2020). "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 156–160 (cit. on p. 9).
- Di Carlo, Diego**, Pinchas Tandeitnik, Sharon Gannot, Antoine Deleforge, and Nancy Bertin (2021). "dEchorate: a calibrated Room Impulse Response database for acoustic echo retrieval". In: *Workin progesess* (cit. on p. 9).
- van den Boomgaard, Rein and Rik van der Weij (2001). "Gaussian convolutions numerical approximations based on interpolation". In: *Scale-Space and Morphology in Computer Vision: Third International Conference, Scale-Space 2001 Vancouver, Canada, July 7–8, 2001 Proceedings* 3. Springer, pp. 205–214 (cit. on p. 37).

