DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

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

Rapporteurs avant soutenance :

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

Composition du Jury :

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

Imprint

Echo-aware signal processing for audio scene analysis

Colophon

This thesis was typeset using LaTeX and the `memoir` documentclass. It is based on Aaron Turon's thesis *Understanding and expressing scalable concurrency*[1] (as re-implemented by Friedrich Wiemer[2]), itself a mixture of `classicthesis`[3] by André Miede and `tufte-latex`[4], based on Edward Tufte's *Beautiful Evidence*.

Graphics and plots are made with PGF/Ti*k*Z, Matplotlib, Mathcha and Excalidraw. Icons are downloaded from the `www.flaticon.com/` The bibliography was processed by Biblatex.

The body text is set to 10/14pt (long primer) on a 26pc measure. The margin text is set to 8/9pt (brevier) on a 12pc measure. `Linux Libertine` acts as both the text and display typeface.

[1] `https://people.mpi-sws.org/~turon/turon-thesis.pdf`

[2] `https://github.com/pfasante/phd_thesis`

[3] `https://bitbucket.org/amiede/classicthesis/`

[4] `https://github.com/Tufte-LaTeX/tufte-latex`

A nonna Leti,
"Lavati le man e va a lavorare!";

a nonna Carla,
"Uuuuuurrrrrcaaaaaaa, che brao ció!";

Empress Ropes[1],
like Morpheus, see you create absurd world;[2]
like Atlas, you hold up physical worlds;[3]
like Angeróna, you heal from sadness;[4]
like Diana, you points arrow and directions;[5]
like Urania, you help didascalic work;[6]

Unlike Narciso: reflections and echoes won't fool me,
unlike Echo: I answer to you.

Deviank, Ikitan, Maishi, Quolcat protect her.
— per Giorgia Cantisani

[1] *en.→it.→lat.*

[2] *and sweet mirages.*

[3] *and walls.*

[4] *and separate noises.*

[5] *the closest ref. to vectors and DoA.*

[6] *papers and thesis.*

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 the content and location of multiple sources of sound in an environnement. 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 available technologies solve audio scene analysis problems either ignoring how sound propagates in the environment or estimating it fully.

The central theme of this theses is acoustic echoes: the sound propagation elements bridging semantic and spatial information on sound sources. Indeed, as repetitions of a source signal, their semantic contributions can be aggregated to enhance this signal. Moreover, since they originate from an interaction with the environment, their paths 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 main questions arise: how to estimate acoustic echoes, and how to 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, in the first part, 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 differences of arrival of the first prominent echoes using physically-motivated regularizers. Furthermore, we present a novel, fully annotated dataset specifically designed for acoustic echo retrieval and echo-aware applications, paving the way for future echo-aware research.

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

► This thesis highlights the difficulty of exploiting acoustic echoes to improve indoor audio processing. As a first attempt to lay unified analytical and methodological foundations for these problems, it is hoped to serve as a starting point for promising new research in this field.

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, continuous dictionary.

Résumé en français

L'analyse de scène audio vise à récupérer des informations utiles à partir d'enregistrements microphoniques. La séparation et la localisation de sources sonores sont des exemples de ces problèmes, dans lesquels on s'intéresse à l'estimation du contenu et des positions de multiples sources de son dans un environnement. 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 de scènes sonores en tenant compte soit de la sémantique du son, soit des informations spatiales.

Le thème central de cette thèse est celui des échos acoustiques : les éléments de la propagation du son faisant le pont entre les informations sémantiques et spatiales des sources sonores. En effet, en tant que répétitions d'un signal source, leurs contributions sémantiques peuvent être aggrégées pour améliorer ce signal. De plus, comme ils sont issus d'une interaction avec l'environnement, leurs cheminements peuvent être retracés et utilisés pour estimer la géométrie de la scène sonore. Sur la base de ces observations, des méthodes récentes en traitement du signal audio tenant compte des échos ont été proposées. Deux questions principales se posent : comment estimer les échos acoustiques et comment exploiter leur connaissance ?

Ce travail de thèse vise à améliorer l'état de l'art actuel en traitement du signal audio dans des salles 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 ces fins, nous présentons tout d'abord deux approches : Une nouvelle méthode basée sur le cadre des dictionnaires continus qui ne nécessite pas de réglages de paramètres ni de données d'apprentissage, et un modèle d'apprentissage profond estimant le décalage temporel de l'arrivée des premiers échos à l'aide de régularisateurs physiquement motivés. En outre, nous présentons un nouvel ensemble de données entièrement annotées, spécialement conçu pour l'estimation d'échos acoustiques et les applications prenant en compte les échos, ouvrant la voie à de futures recherches dans ce nouveau domaine. La deuxième partie de cette thèse concerne l'extension de méthodes existantes d'analyse de scènes audio dans leur forme adaptée aux échos.

Le cadre de la NMF multicanale pour la séparation de sources audio, la méthode de localisation SRP-PHAT et la technique de formation de voies (beamforming) MVDR pour l'amélioration de la parole étendues à des versions "echo-aware". Ces applications montrent comment un simple modèle d'écho peut conduire à une amélioration des performances.

► Cette thèse souligne la difficulté d'exploiter les échos acoustiques pour améliorer le traitement audio à dans des salles. Elle constitue une première tentative de jeter des bases analytiques et méthodologiques unifiées pour résoudre ces problèmes et nous espérons qu'elle serve de point de départ à des de nouvelles recherches prometteuses dans ce domaine.

Mots-clés :
Echos acoustiques, récupération des échos acoustiques, estimation des réponses impulsionnelles d'une pièce ; analyse de scène sonore, acoustique des salles ; séparation de sources audio, estimation de la géométrie d'une pièce, filtrage spatial, localisation de sources sonores ; apprentissage profond, dictionnaires continus.

Résumé étendu en Français

► Ce résumé présente en français un aperçu des travaux abordés dans cette thèse. Le thème de l'analyse de scènes 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 et la localisation de sources sonores, où l'on s'intéresse à l'estimation de la parole 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 d'une pièce. Votre réaction typique sera probablement de tourner votre attention vers cette personne ou même d'aller vers elle. Cependant, pour les ordinateurs et les robots, ce problème de traitement du signal audio pour ces tâches reste un défi à relever.

Les sons transmettent des informations sémantiques (ce votre ami a dit), temporelles (quand il l'a dit) et spatiales (où ils l'ont dit). Nous pouvons modéliser ces contributions à l'aide de signaux décrivant le contenu du son, et à l'aide de la "réponse impulsionnelle" de la pièce, capturant la propagation du son dans l'espace. Certaines méthodes de traitement audio se concentrent sur ces premiers, ignorant ou décrivant grossièrement cette dernière en raison de la difficulté de l'estimer. Les réponses impulsionnelles de pièces intègrent tous les éléments de la propagation du son, tels que les échos, la réflexion diffuse et la réverbération.

Le thème central de cette thèse sont les échos acoustiques. 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 signal source, on peu améliorer le son cible en intégrant par rapport aux autres sources de bruit. Comme ces réflexions sont issues de l'interaction du signal source avec l'environnement, de part leurs temps d'arrivée, on peut remonter leur parcours et ainsi reconstruire la géométrie d'une scène sonore. Sur la base de ces observations, des méthodes de traitement du signal audio ont commencé à prendre en compte ces éléments de propagation du son dans l'analyse de scèness audio. Deux questions principales se posent : comment estimer les échos acoustiques, et comment exploiter leur connaissance ?

Ce travail de thèse vise à améliorer l'état de l'art actuel en traitement du signal audio d'intérieur de salles 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. De plus, il prolonge les méthodes classiques d'analyse de scènes audio des formes adaptées aux échos. Ces deux contributions sont développées dans les deux parties principales de la thèse, qui viennent après une introduction, comme résumé ci-dessous.

- Partie I, L'acoustique des salles rencontre traitement du signal
  Tout d'abord, nous donnons quelques définitions préliminaires en traitement du signal audio et énumérons quelques problèmes fondamentaux qui seront abordés tout au long de la thèse, à savoir l'estimation d'échos acoustiques, la séparation de sources audio, la localisation de sources sonores et l'estimation de la géométrie d'une pièce.

  - Le chapitre 2 construira un premier pont important : de l'acoustique au traitement du signal audio. Il définit d'abord le son, sa propagation dans l'environnement et puis les échos et leurs origines.

  - Dans le chapitre 3, nous passons de la physique au traitement du signal où les échos sont modélisés comme des éléments de filtres, appelés Réponses Impulsionnelles de la de Salle, opérant sur le signal source. Comme le traitement dans le domaine temporel natif est difficile, nous introduisons la représentation de Fourier, qui facilite à la fois l'exposition des méthodes et la mise en œuvre des algorithmes.

  Ce chapitre clôt la première partie introductive.

- Partie II - Estimation d'echos acoustiques
  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 définitions décrits dans la première partie, cette partie comprend d'abord un aperçu général des méthodes d'estimation d'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 une revue de l'état de l'art en estimation déchos acoustiques, c'est à dire, sur l'estimation de leurs propriétés. Après avoir présenté le problème, nous passons en revue la littérature. Afin de fournir un aperçu complet de l'estimation, certains ensembles de données et mesures d'évaluation fréquemment utilisés dans la littérature et ainsi que dans les chapitres suivant sont présentés. Les trois chapitres suivants présentent trois travaux que nous avons menés sur l'estimation des échos acoustiques.

  - Le chapitre 5 présente une nouvelle approche pour estimer les échos d'un enregistrement stéréophonique d'une source sonore inconnue telle que de la parole. Contrairement aux méthodes existantes, celle-ci s'appuie sur le cadre récent des dictionnaires continus et ne repose pas sur des réglages de paramètres. 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 l'état de l'art. 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 inférieurs, et la nature "off-grid" de l'approche donne généralement des erreurs d'estimation plus faibles. C'est prometteur,

puisque l'avantage pratique de connaître les temps d'arrivée de quelques échos par canal sera démontré dans la dernière partie de la thèse.

- Dans le chapitre 6, nous proposons des techniques d'apprentissage profond pour estimer les propriétés des échos acoustiques. À notre connaissance, il s'agit des premières méthodes de ce type pour ce problème, même si elles présentent des points communs avec des techniques d'apprentissage profond pour 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 du champs direct et des premiers échos proéminents ; puis l'exécution de cette estimation de manière plus robuste.

- Enfin, pour conclure cette deuxième partie, dans le chapitre 7, nous décrivons un ensemble de données que nous avons recueillies et spécifiquement conçu pour l'estimation des échos acoustiques. Cet ensemble de données comprend des mesures de réponses impulsionnelles multicanales, accompagnées des annotations des premiers échos et des positions 3D des microphones et des sources réelles et des images sous différentes configurations de murs dans une pièce rectangulaire. Ces données fournissent un nouvel outil pour l'évaluation comparative des méthodes récentes en traitement du signal audio "echo-aware" et des outils logiciels permettant d'accéder, de manipuler et de visualiser facilement les données.

▸ Partie III - Applications des Echos

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

- Un chapitre d'introduction (chapitre 8) rassemble les définitions standards et présente l'état de l'art actuel en traitement audio d'intérieur sous une même enseigne. Nous considérons quatre problèmes fondamentaux : la séparation de sources audio, la localisation de 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.

- Au chapitre 9, les échos sont utilisés pour améliorer les performances de méthodes classiques de séparation de sources audio. C'est 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 les "microphones-images", qui permet de mieux comprendre comment les algorithmes peuvent tirer parti de leur connaissance. Notre étude porte sur deux variantes de la séparation de sources par factorisation en matrices non-négatives multicanale : l'une utilise uniquement les amplitudes des fonctions de transfert et l'autre utilise également les

phases. Les résultats montrent que l'extension proposées bat l'originale en n'utilisant que quelques échos et que les échos permettent parfois la séparation dans des cas où elle était jugée inabordable.

- Le chapitre 10 aborde le problème de la localisation de sources audio dans le contexte de forts échos acoustiques. En utilisant le modèle de microphones-images présenté dans le chapitre précédent, nous montrons que ces contributions parasites peuvent être utilisées pour modifier la manière classique dont la localisation de sources 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 à la fois les angles d'azimut et d'élévation, tâche impossible en supposant une propagation idéale, comme le font les approches classiques. Ces résultats ont fait l'objet d'une publication pour une conférence internationale.

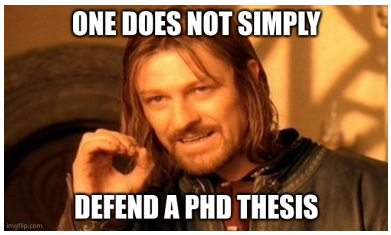
- Le chapitre 11 présente deux applications sensibles aux échos pouvant être utilisées sur l'ensemble de données `dEchorate`, présenté dans le chapitre 7. Nous illustrons l'utilisation de ces données en considérant deux problèmes possibles d'analyse de scènes audio : le filtrage spatial par échos et l'estimation de la géométrie d'une pièce. Afin de valider les données et de montrer leur potentiel, des algorithmes connus de l'état de l'art sont utilisés. Ainsi, pour chacune des applications, les méthodes envisagées sont contextualisées et résumées. Les résultats numériques confirment la valeur potentielle de cet ensemble de données pour la communauté du traitement du signal audio. L'ensemble des données et ces méthodes seront rendus publics pour que des contributeurs externes puissent les afin de pour développer des méthodes de traitement audio plus robustes.

- La dernière partie IV comprend le dernier chapitre (Chapitre 12), 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 un modèle dérivé de la physique de la propagation du son ou des modèles d'apprentissage profond sur simulateur acoustique. De plus, nous démontrons les avantages d'inclure la connaissance des échos acoustiques dans les méthodes de traitement du son. Pour ce qui est de l'évaluation sur données réelles, nous observons que des ensembles de données de référence disponibles librement 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 la recherche "echo-aware", couvrant une grande partie des applications en analyse de scènes audio.

Enfin, nous voulons souligner la difficulté de la tâche d'estimation et d'exploitation des échos acoustiques pour l'amélioration du traitement audio. Cette thèse ne constitue donc qu'une première tentative d'attaquer ces problèmes et pose

des bases analytiques sur la façon de les modéliser. Comme toutes premières investigations, beaucoup de choses peuvent être améliorées, et nous espérons que celle-ci pourra servir de point de départ à de futures recherches dans ce nouveau domaine prometteur.

Acknowledgements



"There is so much drama in the PhD". Excuse me, Snoop for crippling your words, but I could not find anything better. . . or maybe "one does not simply defend a phd thesis" (Excuse me, Boromir). This thesis work is just the emerged part of an iceberg, floating of an ocean of people (sorry for the cheesy metaphor). Without their technical and emotional support, this thesis would probably not exist. Fortunately, I was followed and accompanied by a committee of people, each entering the stage just at the right time.

Firstly, I wish to acknowledge the people giving me the opportunity to write this and are (and will be) *source* of inspiration.

- Nancy Bertin and Antoine Deleforge, PhD supervisors.
  Thank you for friendly mentoring me, teaching me how to address problems and be scientifically critical. You were always very patient and supportive during these years, despite my kaleidoscopic sabotage attempts. (and sorry for all the nuits blanches*, last-minutes rushing and all the 's' to correct).*
- Laurent Girin, Simon Doclo, Renaud Seguier, Fabio Antonacci, my defence jury.
  It was an honour (and a challenge) to have you on my thesis committee and for being the first ones to scrutinize these pages.
- Remi Gribonval and Srdan Kitić, CSID advisors.
  Your expertise and valuable comments helped to improve the quality of my work continuously.
- Clement Elvira, co-author.
  Thank you for all the good times in front of the terminal and the whiteboard, and showing me how to "continuously optimize" my work.
- Robin Scheibler and Ivan Dokmanić, co-authors.
  Thank you for inspiring and fruitful discussions and show me the echo-way.
- Sharon Gannot, Pinchas Tandeitnik, co-authors and advisors.
  Thank you for welcoming me in your lab in Israel and be an indispensable source of scientific and historical knowledge. Todah!
- Antoine Liutkus, former supervisor.
  Thank you for guiding me to the beauties of research and suggesting to apply to this PhD offer. Always be fresh!

This thesis is a coalescence of numerous interactions with friends in the city of Rennes, Nancy and Tel-Aviv. I would like to thank you all for keeping me entertained at, and outside of, work. I am indebted for you friendship, kindness and help:

Giorgia Cantisani, *magnifying glass under the emotional sun*;
Alexander Honorat, *an elegant beaver expert in organic market*;

Antoine Chatalic, *who will always be more open-source than you*;
Clèment Gaultier, *diamond of kindness, patience and bricolage*;
Clèment Elvira, *a living burning machine and aesthete*;
Adrien Llave, *bender of infrasound and distortions*;
Cassio Fraga Dantas, *master of explaining game rules. . . fast, but still safe*;
Laureline Perotin, *shaman of the great green lynx reminding us to have fun*;
Katharina Boudgoust, *painter of perspectives, often in pink.*;
Valentin Gilot *vaporized push-er of beats and style de malade*;
Corentin Louboutin, *Scalchi's doppelganger and crazy hiker*;
Jeremy Cohen, *the great ronin of the old Parafac order*;
Adreas Loukas, *admired master of graph, rock and good music*;
Jean-Joseph Martin, *only-do-it-at-home hacker and brew master*;
Lucas Franceschino, *the 'formal' knight of automatization*;
Claudio Rubattu, *composed researcher during daylight, bear-hugger at night).*
Leonardo Suriano, *container of humorous energy*;
and all the friend of the *Golden Age of Signal Processing* (see you soon at ICASSP).
I would like to thank the colleagues of the Panama team in Rennes and IRISA: Marin, Helena, Younes, Roilhi, Kilian, Hakim, Axel, Nathan, Mohammed, Corentin, Ewen, Romain, Stephanie, Frederic, Quentin (and Anaïs), Firas, Adrien and Mathias; the colleagues of the Multispeech team in Nancy: Nicolas, Nicholas, Sunnit, Manuel, Guillaume (Oh, Hi Mark!), Theo, Elodie and Thomas (See you in the slow-net.); and the colleagues Bar'Ilan university: David, David, Daniel, Gal, Nili, Shlomi. A special mention goes to the professors of the CSC team at the university of Padova that introduced me to the research field of audio signal processing.

Ammetto che fare un dottorato (e farlo all'estero) è una esperienza fantastica: si lavora[7], si fanno molte notti in bianco, si viaggia(va)[8] tanto, e ci sono sempre moltissimi buoni motivi per festeggiare. Questo entusiasmante periodo ha tuttavia comportato dei sacrifici. Il più grande è ovviamente quello di stare lontani (per fortuna solo fisicamente) dalla famiglia (e le sue bestie) e dagli amici. Quindi vogli qui esprimere la mia più grande gratitudine alla mia famiglia (Marta, Franco, Albi e Candy; Stefania, Claudio e nuMina) per avermi sempre supportato e sostenuto. Un pensiero dolce alle mie due nonne che sono recentemente andate a caccia di Cthulhu e Azatoth: chissà cosa avreste pensato delle eco per la source separation.
Last but not least, un ringraziamento enorme alle mie compagnie di amici: i Prana, i Mandela e gli Astrali. Per fortuna mi avete tenuto compagnia su (nota compagnia di messaggistica instantanea).
if I forgot to mention someone it was an honest mistake.

Giorgia, grazie per aver creduto in me e a spingermi nei momenti giusti. . . Oltre che a menzionarti in ogni parte di questa tesi, non so come fare per farti capire quanto ti sono grato (cf. dedication, itemize qui sopra e un po' qua e là). Ora è finita e possiamo andare al mare a St. Lunaire e non fare il bagno perchè è troppo l'acqua fredda.

Diego DI CARLO

[7] In Italia il dottorato non è considerato ancora un lavoro

[8] Maledetto COVID19

Contents

Glossary

From Latin glossarium *"collection of glosses", diminutive of* glossa *"obsolete or foreign word".*

AER Acoustic Echo Retrieval . . . 5

AIR Acoustic Impulse Response . . . 20

AOA Angle of Arrival . . . 129

ASR Automatic Speech Recognition . . . 59

ATF Acoustic Transfer Function . . . 20

AWGN Additive White Gaussian Noise . . . 32

BCE Blind Channel Estimation . . . 56

BEM Boundary Element Method . . . 21

BLASSO Beurling-LASSO . . . 71

BLASTER Blind And Sparse Technique for Echo Retrieval . . . 7

BSI Blind System Identification . . . 56

BSN Blind Sparse Nonnegative Channel Identification . . . 69

CASA Computational Auditory Scene Analysis . . . 43

CC Cross Correlation . . . 129

CD Continous Dictionary . . . 7

CNN Convolutional Neural Network . . . 78

DECHORATE Dataset dechorated by echoes . . . 60

DFT Discrete Fourier Transform . . . 34

DNN Deep Neural Network . . . 77

DOA Direction of Arrival . . . 55

DRR Direct-to-Reverberant Ratio . . . 27

DS Delay-and-Sum . . . 109

DTFT Discrete-Time Fourier Transform . . . 34

DWM Digital Waveguide Mesh . . . 21

EM Expectation Maximization . . . 56

ESPRIT Estimation of Signal Parameters via Rational Invariance Techniques 53

ESS Exponential Sine Sweep . . . 52

FDTD Finite-Difference-Time-Domain . . . 21

FEM Finite Element Method . . . 21

FFT Fast Fourier Transform . . . 40

FIR Finite Impulse Response . . . 56

FRI Finite Rate of Innovation . . . 59

FT Fourier Transform . . . 33

GA Geometrical (room) acoustics . . . 18

Notations

Linear Algebra

x, X	scalars
$\boldsymbol{x}, \mathbf{x}$	vectors
x_i	i-th entry of $\mathbf{x}$
$\mathbf{0}_I$	$I \times 1$ vector of zeros
$\mathbf{x}^\mathsf{T}$	transpose of the vector $\mathbf{x}$
$\mathbf{x}^\mathsf{H}$	conjugate-transpose (Hermitian) of the vector $\mathbf{x}$
$\mathrm{Re}[x]$	real part scalar (vector) x ($\mathbf{x}$)
$\mathrm{Im}[x]$	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 numbers

Common indexing

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

Geometry

$\underline{\mathbf{x}}_i$	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
r_j	distance of source j w. r. t. to a reference frame
θ_j	azimuth of source j w. r. t. to a reference frame
φ_j	elevation of source j w. r. t. to a reference frame
ϑ_j	angle of arrival of source j w. r. t. to a reference frame

Signals

$\tilde{x}(t)$	continuous time domain signal
$x[n]$	discrete time domain signal
$x_N[n]$	discrete and finite time domain signal
$\tilde{X}(f)$	continuous frequency domain
$X[k]$	discrete frequency domain
$X[k,l]$	discrete time-frequency domain

x_i	input signal recorded at microphone i
$\boldsymbol{x}$	$I \times 1$ multichannel input signal, i.e. $\boldsymbol{x} = [x_0, \ldots, x_{I-1}]$
$\boldsymbol{H}$	matrix of multichannel signal or filters, typically the mixing matrix ($I \times J$)
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 microphone 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 instant of the reflection r
c_{air}	speed of sound in air
T	temperature
H	relative humidity
p	sound pressure
h_{ij}	Room Impulse Response from source j to microphone i

Mathematical Operations

$\star$	continuous time convolution
$*$	discrete linear convolution
$\circledast$	discrete circular convolution

1

Prologue

"Echoes *shows the direction that we're moving in.*"

—David Gilmour,
making of "The Dark Side Of The Moon"

In a nutshell, this Ph. D. thesis is about acoustic Echoooes. 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 system 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 of 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[9] or clock provides units of time and when hearing someone shouting, we unconsciously know where to turn our attention. However, this latter information is determined by how sound *propagates* in space and not by the source itself.

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

Before reaching the ears, sound propagates in all directions and a portion of its energy arrives at us directly, and another 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. [10] 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.

[10] "*Écho. Citer ceux du Panthéon et du pont de Neuilly.*" — Gustave Flaubert, *Dictionnaire des idées reçues.*

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,

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*.

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].
How I use sonar to navigate the world

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 web[11]. 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.

[11] 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]

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*, *pyschoacoustics* 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 *signal*, 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 into categories, for instance, *representation*, *enhancement*, *estimation*, etc. Many related problems can be then decomposed into blocks, which may include one or more of the following steps.

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.

- 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.

  Enhancement and estimation can be conducted in *adaptive* fashion, namely with algorithms that are able to adapt themselves to the data. They typically process the data on the fly and are controlled by variable parameters resulting from previous estimation blocks. They usually rely on the 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 you are hearing right 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*[12]. 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.

[12] 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.*

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 1.1 lists some selected audio scene analysis problems that will be considered later in this thesis.

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

Problems	*From the recordings, can we...*
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 sound sources?
Microphone Calibration	... estimate the positions (and gains) 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	... estimate the type of source signal?
Speech Diarization	... who is speaking and when ?
Source Counting	... count the number of speakers?
Automatic Speech Recognition	... estimated 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.

- *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;

"*Everything is connected*"
—Douglas Adams, *Dirk Gently's Holistic Detective Agency*

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 The 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 audio[13]. However, it was not always the case. Many audio scene analysis methods make limiting assumptions on the sound propagation to derive efficient algorithms. One of the most common 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

[13] 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*.

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. 1993; Jan et al. 1995; Jan and Flanagan 1996] in source enhancement. However, only recently, these methods have regained interest in audio processing as manifested by the European project SCENIC [Annibale et al. 2011] and the UK research S^3A 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 reflections. However, prior to all of 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

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

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

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.

Chapter 3 We now move from physics to digital signal processing. At first, in § 3.1, this chapter formalizes fundamental concepts of audio signal

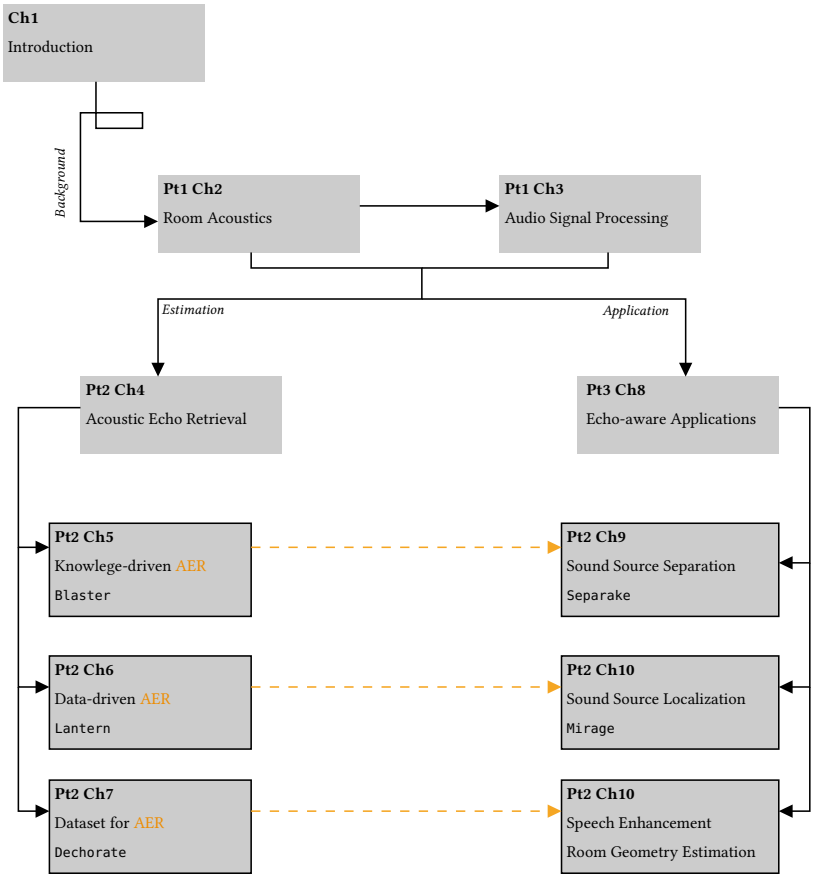


FIGURE 1.1: Schematic organization of the thesis, dependencies between chapters linked to author contributions.

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 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 details in Chapter 10. Finally, we will present a possible way to achieve a more robust estimation (§ 6.4) and discuss a possible way to scale this approach to multiple echoes (§ 6.5).

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 present a selection of algorithms and methodologies which we identified as potential beneficiaries of echo-aware additions. At first, we present a typical scenario that highlights some cardinal problems. Then, state-of-the-art approaches to address these problems are listed and commented in dedicated sections, highlighting the relationship with some acoustic propagation models, respectively. 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 Multichannel Audio Source Separation (MASS) methods. Existing methods typically 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 echo properties 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 based on the work of [Ozerov and Févotte 2010]: 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 In this chapter, we show that early-echo characteristics can, in fact, benefit SSL. To this end, we introduce the concept of Microphone Augmentation with Echoes (MIRAGE), based on the image microphone model. 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 anechoic propagation.

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 contributions

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: *Workin progess*
- **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: *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: *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: *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][14] . We emphasize the presence of clickable links by the logo. Code libraries are written in typewriter font, e. g. `dEchorate`.

[14] The colophon of the thesis reports more information on the template.

1.6.1 *Quick vademecum*

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;
- the following icon denotes the end of the chapter.

1.6.2 *The golden ratio of the thesis*

This thesis has been written 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 ACOUSTICS MEET SIGNAL PROCESSING

2

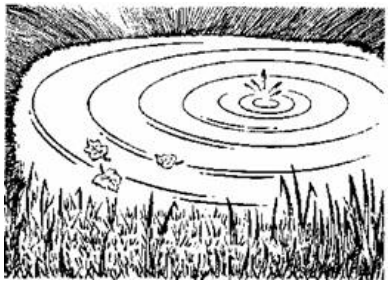
Elements of Room Acoustics

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

▸ 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 classical texts on room acoustics [Kuttruff 2016; Pierce 2019] and on partial differential equations [Duffy 2015].

2.1 Sound wave propagation



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

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*[15] 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:

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

where T is the air temperature [°C] and H is the relative air humidity [%].

[15] As 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.

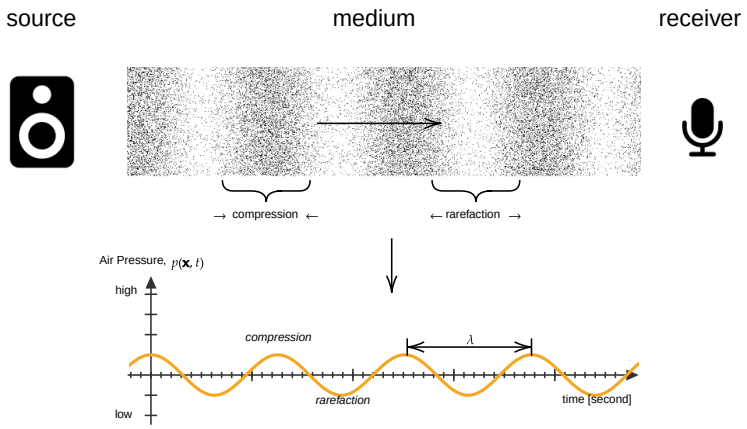


FIGURE 2.2: Illustration of the molecules under sound compression and rarefaction due to longitudinal sound wave and its waveform representation.

We can think of this process in the light of the classical *source-medium-receiver* model of communication theory: the *source* is anything that emits waves[16], the *medium* carries the waves from one point to another, and the *receiver* receives them.

[16] Example 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).

2.1.1 *The acoustic wave equation*

The acoustic wave equation is a second-order partial differential equation[17] which describes the evolution of acoustic pressure p as a function of the position $\underline{\mathbf{x}}$ and time t

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

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 $\left[\frac{\mathrm{m}}{\mathrm{s}}\right]$.
Bundary 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

[17] In 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

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

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

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 $s(\underline{\mathbf{x}}, t)$, the wave equation can be written

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

Thus, the corresponding Helmholtz's equation writes

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

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

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

The function $H(\underline{\mathbf{x}}, f \mid \underline{\mathbf{s}})$ 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

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].

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*[18] 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*.

[18] Impulse 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:

$$P(\underline{\mathbf{x}}, f) = \iiint_{\mathcal{V}_{\tilde{s}}} H(\underline{\mathbf{x}}, f \mid \underline{\mathbf{s}}) S(\underline{\mathbf{s}}, f)\, \mathrm{d}\underline{\mathbf{s}}, \tag{2.7}$$

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

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

$$H(\underline{\mathbf{x}}, f \mid \underline{\mathbf{s}}) = \frac{1}{4\pi \|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|} \mathrm{e}^{-\frac{\mathrm{i}2\pi f \|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|}{c}} \tag{2.8}$$

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

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

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 $q = \|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|$ from the source to the microphone.

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

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,

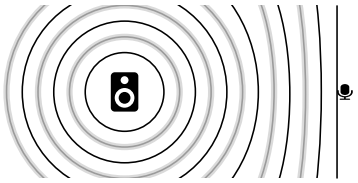


FIGURE 2.3: Visualization of the sound propagation. Since the sensor (i.e. a microphone) is drawn in the far field, the incoming waves can be approximated as plane waves.

$$\lambda = \frac{2\pi}{k} = \frac{c}{f} \quad [\mathrm{m}], \tag{2.10}$$

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 the receiver sensor can be considered in far-field if placed ∼69 cm away from the source.

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

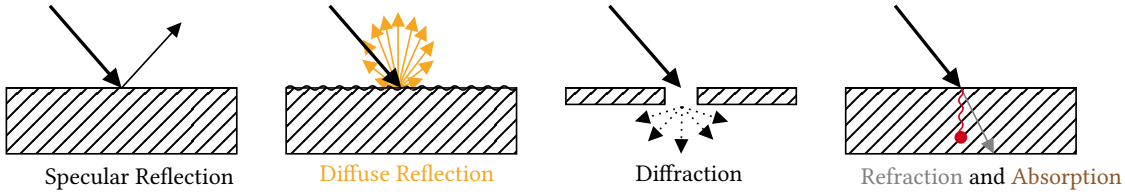


FIGURE 2.4: Different types of sound interaction with a surface.

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.

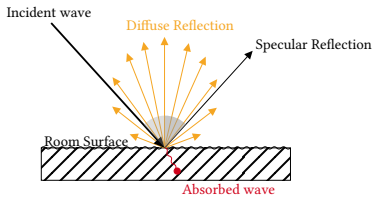


FIGURE 2.5: Specular and diffuse reflection.

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.

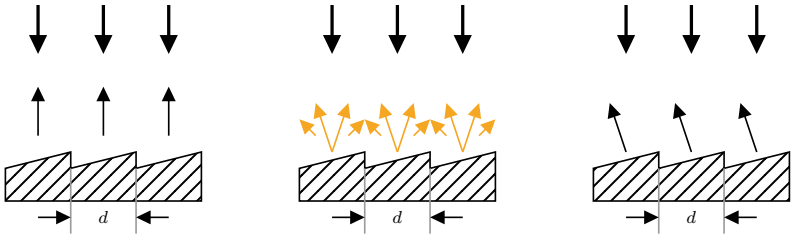
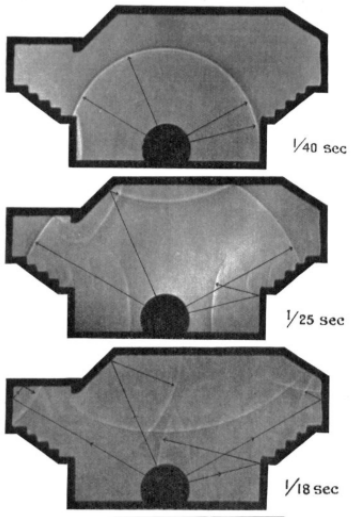


FIGURE 2.6: A reflector having irregularities on its surface with width d much smaller than the sound wavelength λ. Image courtesy of [Kuttruff 2016].

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.



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

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 the incident wave energy absorbed by the surface. Analytically, it is defined as [Kuttruff 2016; Pierce 2019]

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

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 wave[19] 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 frequencies is relaxed in favor of a frequency-independent scalar or at most a

[19] Since it is the square magnitude of the acoustic pressure, the phase information is lost.

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

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

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

▶ ECHOES ARE SPECULAR REFLECTIONS which distinguish themselves in terms of gain and 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 stories of Greece, Echo is a mountain nymph whose ability to speak was cursed: she was 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 though 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 thought 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.

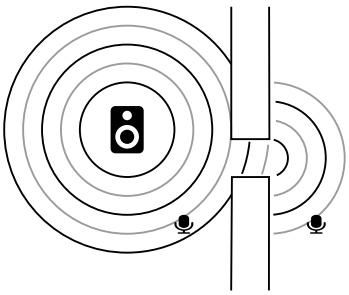


FIGURE 2.8: Schematic representation of sound diffraction. This effect allows to hear "behind walls".

2.3 ROOM ACOUSTICS AND ROOM IMPULSE RESPONSES

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,

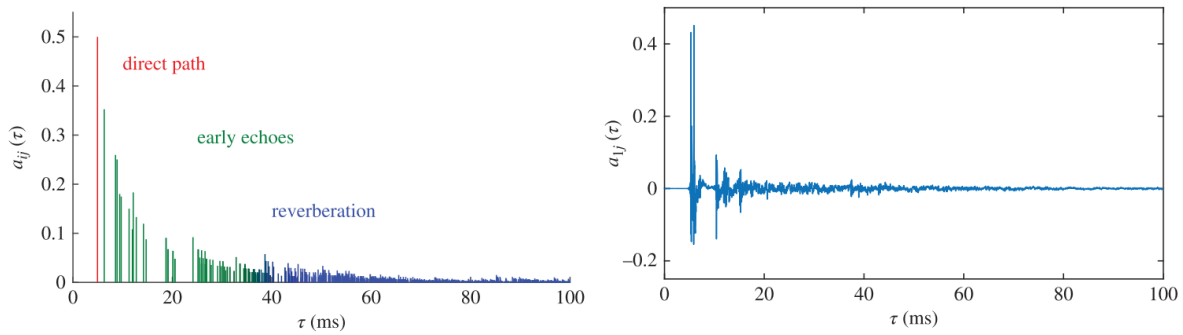


FIGURE 2.9: Schematic illustration of the shape of an RIR and the first 100 ms of a measured one. Image taken from [Vincent et al. 2018]

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)[20] 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.

[20] The 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 is 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:

$$h(t) = h^d(t) + h^e(t) + h^l(t), \tag{2.13}$$

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 in § 2.4.

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

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

where the symbol $\star$ is the continuous-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*

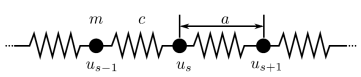
The documentation of the Wayverb acoustic simulator offers a complete overview of the State of the Art (SOTA) in acoustic simulator methods [Thomas 2017].

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.

▶ Wave-based Methods are iterative methods that divide the 3D bounded enclosure into a grid of interconnected nodes[21]. 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.

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



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.[22] 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

[22] For more details on this, the reader can refers to the *Courant–Friedrichs–Lewy conditions* and the documentation of the Wayverb .

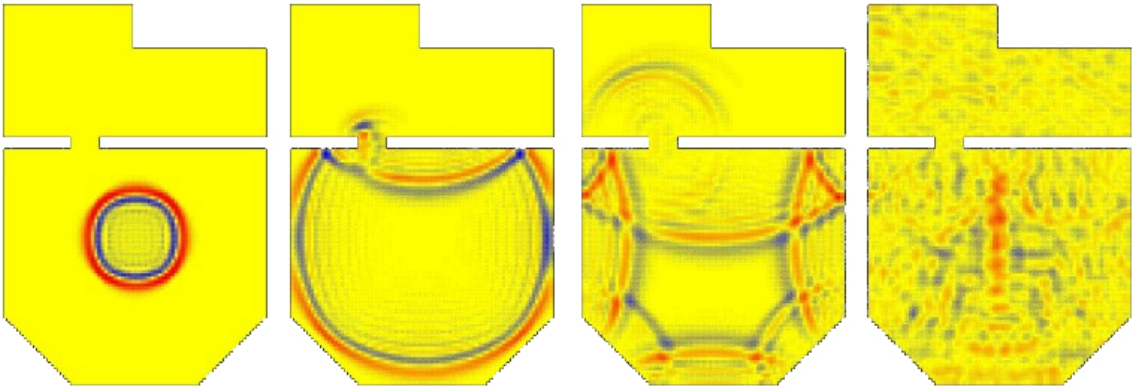


FIGURE 2.11: Simulation of Sound propagation at four consecutive timestamps using the DWM technique. A short, sharp, impulsive sound fired into the larger of two rooms causes a circular wavefront to spread out from the sound source. The wave is reflected from the walls and part of it passes through a 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*[23], 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.

[23] 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.

- 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.

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

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 descriptors[24] 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.

[24] such as the amount of reverberation or source-to-receiver distance.

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,

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.

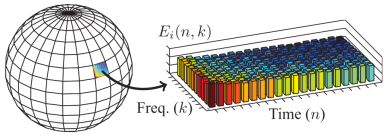


FIGURE 2.12: Directional-time-frequency energy map resulting form the diffuse rain algorithm [Schröder et al. 2007]. For each direction, that is receiver's spherical bin, a time-frequency histogram collects the energy of incoming rays. Image taken from [Schimmel et al. 2009]

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.

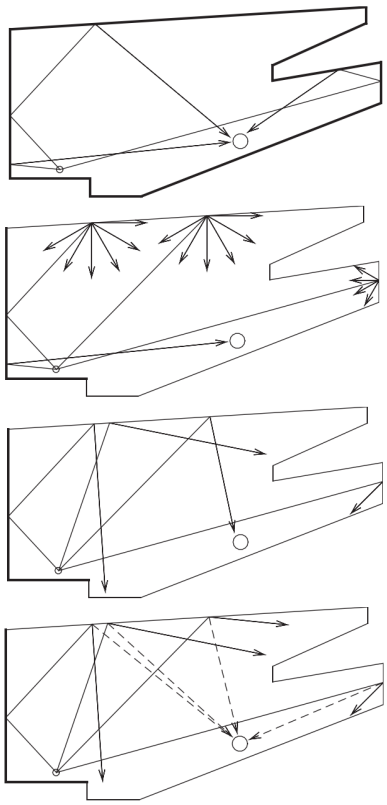


FIGURE 2.13: Visualization of ray-tracing method. From top to bottom: first the method will eventually find specular reflection; then diffuse reflections can be modeled either by splitting a ray into several new rays or a single random one. In the diffuse rain technique a shadow-ray is cast from each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

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

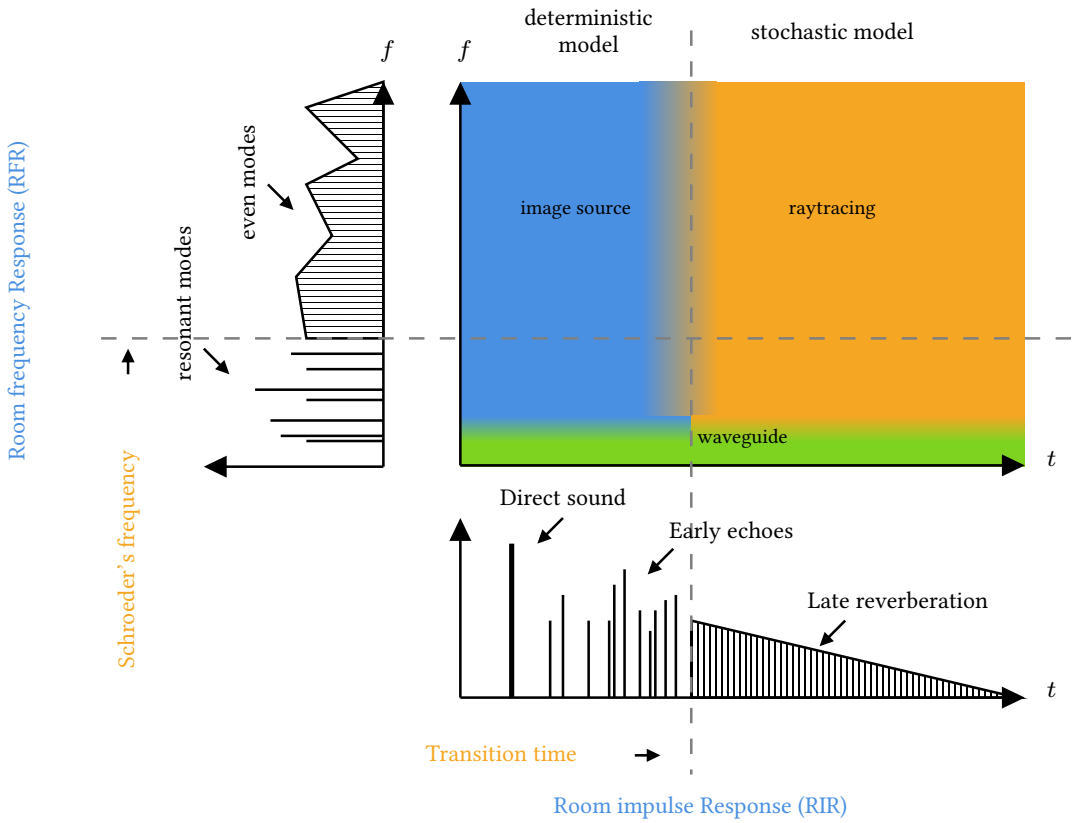


FIGURE 2.14: Time-Frequency regions of the RIR associated to the method that better simulate them. Image adapted from [Thomas 2017; Badeau 2019].

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 methods 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 know 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 of the Helmholtz's equation for cuboid enclosures with rigid boundaries.

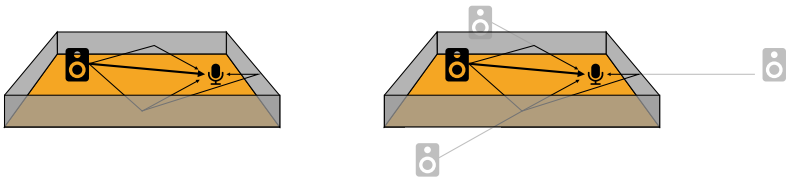


FIGURE 2.15: 3-D representation of the Image Source Method (ISM) and its propagation paths for selected echoes.

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.

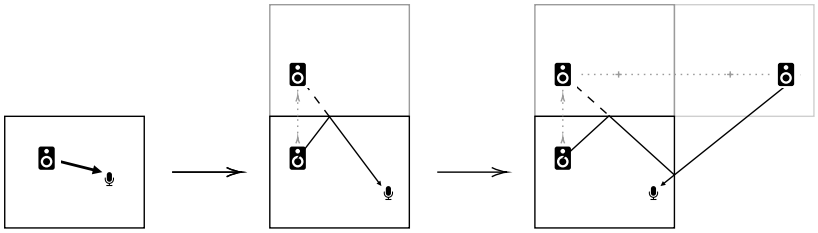


FIGURE 2.16: From left to right, path involving the direct path, one reflection obtained using first-order image, and two reflections obtained using two images. Image inspired from [Habets 2006].

As show 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 $\underline{\mathbf{s}}$ and the receiver at $\underline{\mathbf{x}}$ reads

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

where $\underline{\mathbf{s}}_r$ is the r-th image of the source and $\bar{\alpha}_r$ is the total frequency-independent[25]

[25] Which 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 attention 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), $\tau_r = \|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|/c$, or decreasing amplitudes, $\bar{\alpha}_r/(4\pi\|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|)$. 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. 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 encounters 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

$$H_{\text{ISM}}(\underline{\mathbf{x}}, f \mid \underline{\mathbf{s}}) = \sum_{r=0}^{R} \frac{\alpha_r(f)}{4\pi\|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|} \exp\left(-\mathrm{i}2\pi f \frac{\|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|}{c}\right), \tag{2.16}$$

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 (delayed) filters. This has a strong implication when modeling and estimating the RIRs as stream of Dirac impulses. Ideally they consists 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 differences, their energy may be aggregated together. This effect will be studied in detail is Chapter 5. Nevertheless, aggregating echoes is something that the human auditory system do 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 physical point of view. However in many applications it is important to correlate it to human perception. This will be important in order to define some acoustic parameter which will be used later in this thesis to characterize evaluation scenarios.

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 Chapter 10, 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 could be regarded as the moment when the diffuse tail cannot be distinguished from that of any other position or listener's orientation in the room [Lindau et al. 2012]. Analytically, it can be identified as the point dividing the early reflections from the late reverberation in a RIR. The previously cited work 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* ($\mathtt{RT}_{60}$), 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 $\mathtt{RT}_{60}$ 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*:

$$\mathtt{RT}_{60} \approx 0.161 \frac{V_{\mathrm{TOT}}}{\sum_l \alpha_l S_l} \quad [\mathrm{s}], \tag{2.17}$$

where V_{TOT} is the total volume of the room $[\mathrm{m}^3]$ and α_l and S_l are the absorption coefficient and the area $[\mathrm{m}^2]$ 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.

2.5 Conclusion

This chapter presented notions of room acoustics, 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.

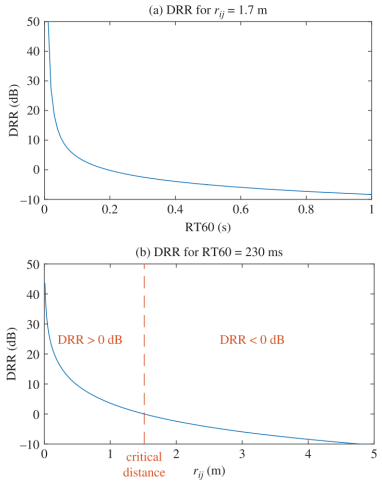


FIGURE 2.17: DRR as a function of the RT60 and the source distance rij based on Eyring's formula (Gustafsson et al., 2003). These curves assume that there is no obstacle between the source and the microphone, so that the direct path exists. The room dimensions are the same as in Figure 3.1.

3

Elements of Audio Signal Processing

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

▸ Synopsis We 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 books 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

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

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

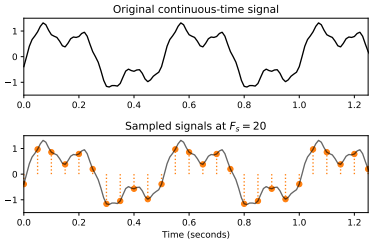


Figure 3.1: Continuous-time signal and its sampled version.

Today signals are typically processed, stored and analyzed by computers as *digital audio signals*. 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 $\tilde{\phi}_{\text{LP}}$ with frequency support in $]-F_s/2, F_s/2]$; [26] then it time-support is regularly discretized, $t = n/F_s$ for $n \in \mathbb{Z}$. This is expressed by

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

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.

Strickly 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.

[26] The ideal low-pass filter is $\tilde{\phi}_{\text{LP}}(t) = \text{sinc}(t) = \sin(\pi F_s t)/(\pi F_s t)$. The term sinc stands for *sinus cardinal* and was introduced in [Woodward and Davies 1952] in 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*"

Finally, at the end of the discretisation process, the $\tilde{x}(t)$ is represented as the finite time series or a vector,

$$x_N \in \mathbb{R}^N, \tag{3.3}$$

with entries $x_N[n]$ for $n = 0, \ldots, N-1$.
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 $44.1\,\mathrm{kHz}$ for music distribution on CDs and $8\,\mathrm{kHz}$ for first-generation speech communication. Now multiples of $8\,\mathrm{kHz}$ are typically used in audio processing: $(16, 48, 96, 128\,\mathrm{kHz})$.

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 describing 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 microphone 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 $\tilde{\boldsymbol{x}}(t) = [\tilde{x}_1(t), \ldots, \tilde{x}_I(t)]^\intercal \in \mathbb{R}^I$.

⇄ 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. Onmidirectional. 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 the models, it is not true in real scenario, where the physical properties of sources (resp. receivers) leads to frequency-dependent *directivity patterns*. 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 $i \in \{1, \ldots, I\}$. Let us assume that there are J sources indexed by $j \in \{1, \ldots, J\}$. Each microphone i and each source j have a well defined position in the space, $\underline{\mathbf{x}}_i$, $\underline{\mathbf{s}}_j$, respectively.

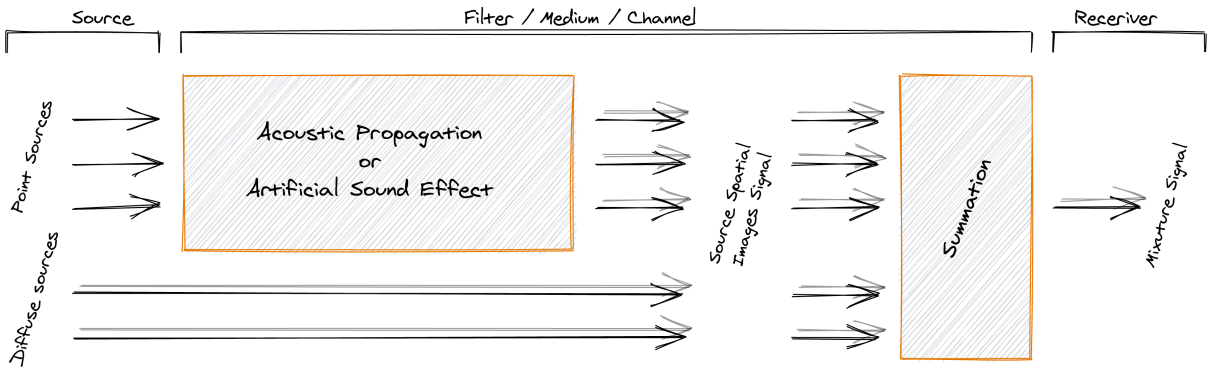


FIGURE 3.2: General mixing process, illustrated in the case of $J = 3$ sources, including three point sources and one diffuse source, and $I = 2$ channels.

The mixing process describes then the nature of the mixtures. In order to better formalized it, in source separation (e. g. in [Sturmel et al. 2012]) it is common to use an intermediate representation called *source spatial images*: $\tilde{c}_{ij}(t)$ describes the contribution of source j to microphone i. Consequently, the *mixture* $\tilde{x}_j$ is the combination of images associated to the source j. This is illustrated in Figure 3.2. 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. Convolutive 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 convolutive by nature and free-far-field natural recording are well approximated by anechoic mixtures.

instantaneous	$\tilde{c}_{ij} = \alpha_{ij}\tilde{s}(t)$
anechoic	$\tilde{c}_{ij} = \alpha_{ij}\tilde{s}_j(t - \tau_{ij})$
convolutive	$\tilde{c}_{ij} = (\tilde{h}_{ij} \star \tilde{s}_j)(t)$

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

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

of convolution between the RIRs and source signals:

$$\tilde{c}_{ij}(t) = \left(\tilde{h}_{ij} \star \tilde{s}_j\right)(t) \tag{3.4}$$

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

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

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:

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

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.

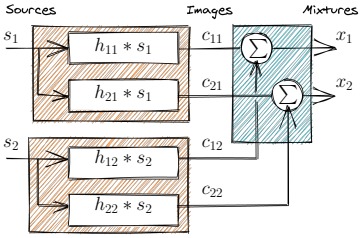


FIGURE 3.3: Graphical representation of the mixing model 3.5 for 2 sources and 2 microphones. The text is overlapping on purpose.

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 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:

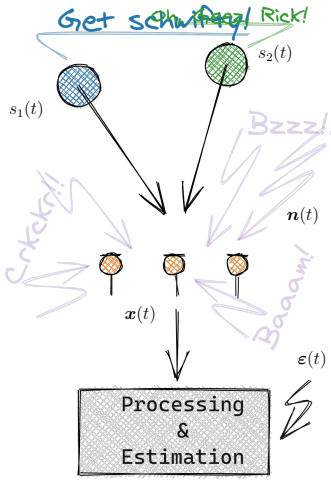


FIGURE 3.4: Graphical representation of the mixing model (3.5): $s_2(t)$ is the *interferer*, $\boldsymbol{n}(t)$ contributes to the *diffuse noise field*, and $\boldsymbol{\varepsilon}(t)$ model acquisition and modeling errors. *With proper prior knowledge, it can be understood.*

- 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 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.

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

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

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

$$\tilde{\boldsymbol{x}}(t) = \sum_{j=1}^{J} \tilde{\boldsymbol{c}}_j(t) + \tilde{\boldsymbol{n}}(t) \tag{3.8}$$

Note that the noise term $\tilde{\varepsilon}_{ij}$ of Eq. (3.7) does not correspond to the one of Eq. (3.6), but includes it. Thus, as not to weigh down the notation, we will use the same notation for both.

3.2 Signal model in the spectral domain

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.

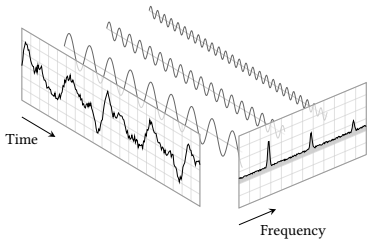


FIGURE 3.5: A signal resolved into its Fourier series: a linear combination of sines and cosines represented as peaks in the frequency domain. Adapted from the online tikz example.

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 (see Figure 3.5):

$$\tilde{X}(f) = (\mathcal{F}\,\tilde{x})(f) = \int_{-\infty}^{+\infty} \tilde{x}(t) \mathrm{e}^{-\mathrm{i}2\pi f t}\, \mathrm{d}t \quad \in \mathbb{C}, \tag{3.9}$$

where $f \in \mathbb{R}$ are the *natural frequency* in Hz and i is the imaginary unit. Apart from providing a space where an audio signal reveals its harmonic structure, the Fourier transform benefits from two fundamental properties: it is linear and it converts time-domain convolution into element-wise product. First, linearity allows to write Eq. (3.5) simply as:

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

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

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

As discussed in Chapter 2, 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

$$\tilde{H}_{ij}(f) = \sum_{r=0}^{R} \frac{\alpha_{ij}^{(r)}}{4\pi c \tau_{ij}^{(r)}} \mathrm{e}^{-\mathrm{i}2\pi f \tau_{ij}^{(r)}}. \tag{3.12}$$

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 the microphones capture and the related spectra are represented by finite- and discrete-time signals for which the properties (3.11) are valid only with some precautions.

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), $\mathcal{F}_{\mathrm{F_s}}$:

$$\tilde{X}_{F_s}(f) = (\mathcal{F}_{\mathrm{F_s}}\, x)(f) = \sum_{n=-\infty}^{+\infty} x[n]\mathrm{e}^{-\mathrm{i}2\pi f n/Fs}, \tag{3.13}$$

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. In case of uniformly spaced samples, 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

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

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) [27], $\mathbf{F}$:

$$X_F[k] = (\mathbf{F}\, x_N)[k] = \sum_{n=0}^{N-1} x_N[n]\mathrm{e}^{-\mathrm{i}2\pi k n/F}. \tag{3.15}$$

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 $\dddot{s}_N$ and $\dddot{h}_N$ of period N. Here $\dddot{\cdot}$ is used to denote periodicity. Let be S_F and H_F their DFT, respectively, then the convolution theorem for the DFT writes

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

where $\circledast$ denotes the *circular convolution*[28]. Note that now the results of the convolution, $\dddot{c}_N[n]$, is periodic with period N.

[27] This can be intrepreded as the projection onto the space spanned by a finite number of complex exponentials.

[28] 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 presents 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 denote with $\tilde{X}_{F_s}[k]$.

$$\tilde{X}(f) = \int_{-\infty}^{+\infty} \tilde{x}(t)\mathrm{e}^{-\mathrm{i}2\pi f t}\,\mathrm{d}t \quad \rightarrow \quad \tilde{X}_{F_s}(f) = \sum_{n=-\infty}^{\infty} x[n]\mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}}. \tag{3.17}$$

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 lower 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 porting 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,

$$\tilde{X}_{F_s}(f) = \sum_{n=-\infty}^{\infty} x[n]w[n]\mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}} \quad \leftrightarrow \quad \tilde{X}_N(f) = \sum_{n=0}^{N-1} x_N[n]\mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}}. \tag{3.18}$$

By the convolution theorem, we have that

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

where $\tilde{W}_{F_s}$ is the DTFT of the sampled window function $w[n]$.
Assuming the window function to be an ideal rectangular function, 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 approximation[29]

[29] When 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. (3.21), we finally obtain the DFT as in Eq. (3.15), that is

$$\tilde{X}_N(f_k) = \sum_{n=0}^{N-1} x_N[n] \mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}} \quad \leftrightarrow \quad X_F[k] = \sum_{n=0}^{N-1} x_N[n] \mathrm{e}^{-\mathrm{i}2\pi kn/F}. \tag{3.20}$$

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

$$f_k = \frac{k}{F} F_s. \tag{3.21}$$

Notice that F_s term disappeared in the right part of the equation above as it cancels out when using Eq. (3.21). 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 involve using a 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 domain is equivalent to a periodization in the other domain while truncating lead to convolving with a window function. Moreover, the chain of operation (sampling in time and truncation in time and sampling in frequencies) are valid in both way. 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 smooth the spectrum and in the other smooth the signal. All this relation and approximation that connects the FT to the DFT are well explained in this explanatory material Poisson Summation Formula, Revisited or in many classic books of digital signal processing, such as [Oppenheim 1987].

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

$$\boldsymbol{x}[n] = \sum_{j=1}^{J} \boldsymbol{c}_j[n] \quad \overset{\mathbf{F}}{\rightarrow} \quad \boldsymbol{X}[k] = \sum_{j=1}^{J} \boldsymbol{C}_j[k] \tag{3.22}$$

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

$$c_{ij}[n] = (h_{ij} * s)[n] \quad \overset{\mathbf{F}}{\rightarrow} \quad C_{ij}[k] = H_{ij}[k] S[k], \tag{3.23}$$

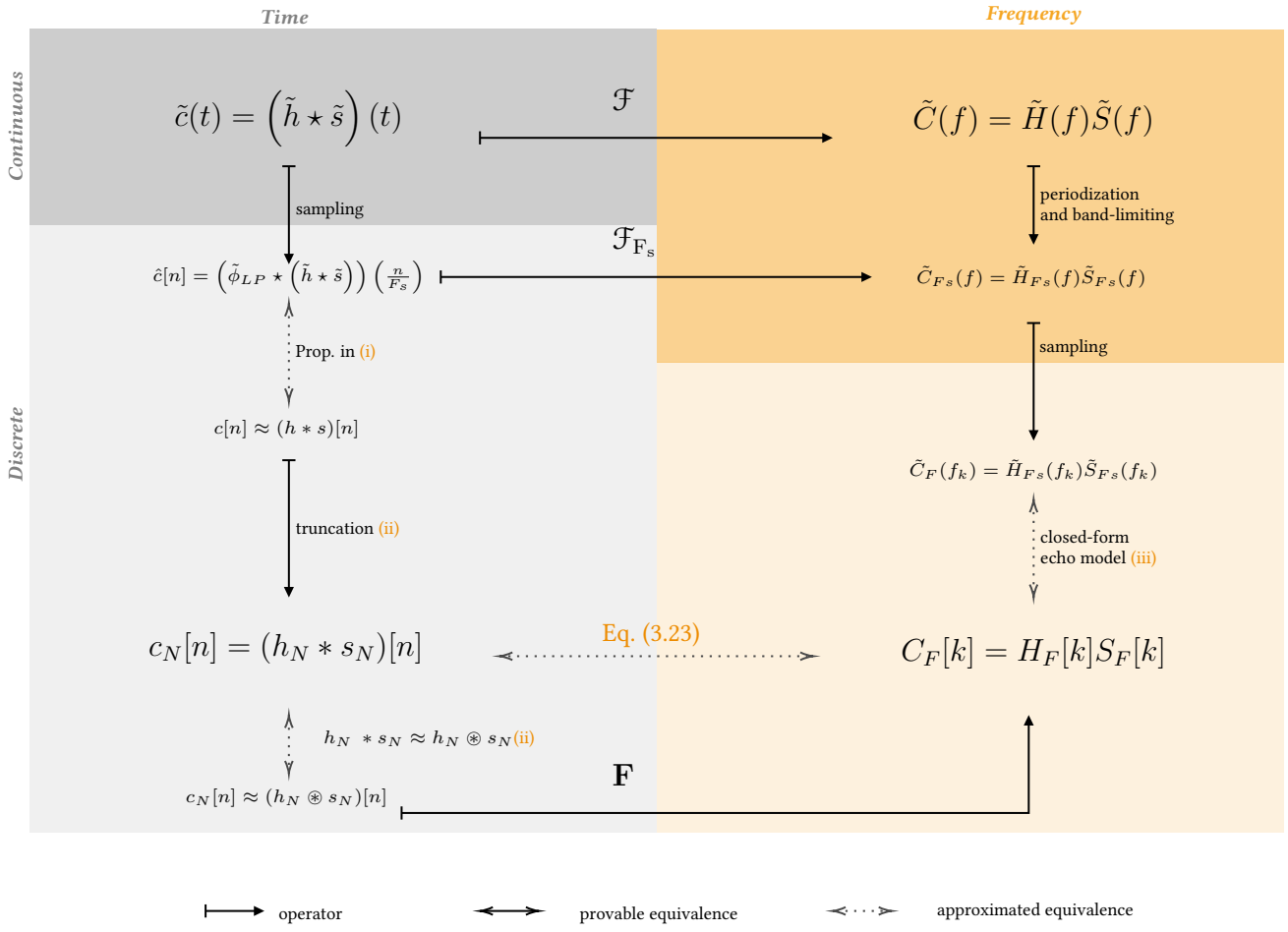
where $*$ is the finite-time linear convolution operator[30]. In general, this relation is wrong and the following paragraph will provide further insight on why is used in practice.

[30] 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 $n = 0, \cdots, D-L$.

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 approximation 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 by taking equally spaced frequencies, that is

$$H_{ij}[k] = \sum_{r=0}^{R} \frac{\alpha_{ij}^r}{4\pi c \tau_{ij}^r} \mathrm{e}^{-\mathrm{i}2\pi k F_s \tau_{ij}^r / F}. \tag{3.24}$$

▸ Although used in practice, the model (3.23) 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. (3.23) used in practice. In order,

(i) In [van denBoomgaard and van derWeij 2001, Proposition 2][31] the au-

[31] Even if this is a classic result and it is discussed is classical book, however the provided reference concisely discusses it.

thors 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 $h[n] \approx (\tilde{\phi}_{\text{LP}} * \tilde{h}^*)[n]$, 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-from DTFT (Eq. (3.24)) 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 $N \to \infty$ 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 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. In particular, the true peaks in the RIR do not necessarily correspond to the true echoes as shown in Figure 3.6. Tukuljac et al. made

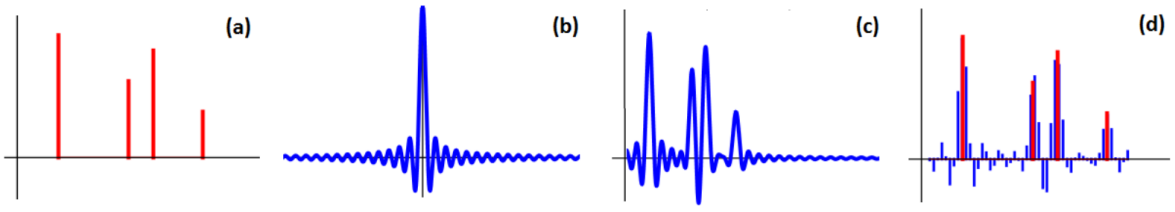


FIGURE 3.6: (a) Continuos-time stream of Diracs $\tilde{h}(t)$, (b) sinc kernel $\tilde{\phi}_{\text{LP}}(t)$, (c) smoothed stream of Diracs, (d) original stream of Dirac $\tilde{h}(t)$ (red) and its sampled (i. e., smoothed and discrete) version (blue). Image courtesy of [Tukuljac et al. 2018]

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 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 $L_{\texttt{win}}$ shorter than the entire signal length, $L_{\texttt{win}} \ll N$. 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

$$X[k,l] = \sum_{n=0}^{L_{\texttt{win}}-1} w[n]x[n+lL_{\texttt{hop}}]\mathrm{e}^{-\mathrm{i}2\pi kn/F} \quad \in \mathbb{C} \tag{3.25}$$

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

where $L_{\texttt{win}}$ is the window length and $L_{\texttt{hop}}$ 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 $x[n,l] = x[n+lL_{\texttt{hop}}]w[n]$.
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) power *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.

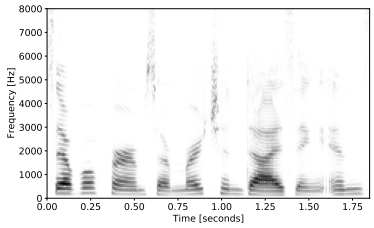


FIGURE 3.7: STFT spectrogram of an example speech signal. Higher energies are illustrated with darker colors.

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.

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 efficienciently 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.

3.2.5 *The final model*

The model (3.23) 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. To circumvent this issue, the *convolutional STFT*[32] for arbitrary window functions have been proposed [Gilloire and Vetterli 1992]. Although it leads to better approximation than the one of product, it is computationally and memory intensive. The exact effect of the time-domain convolution in the STFT would be modeled by cross-band filters [Avargel and Cohen 2007] performing a double convolution along time and frequency axes.[33] However, these models are not investigated in this thesis work.

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

[33] The interested reader can refer to [Girin et al. 2019] for a review on the topic applied to audio source separation

In this thesis, we will consider moderated reverberant scenario. This reflects in long room impulse responses, which would require to long analysis window. Instead, as we are interested in only the early echoes, we can consider filters shorter than typical window length used in speech processing (e. g. 64 ms). 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 $[l,k]$:

$$\boldsymbol{C}_j[l,k] \approx \boldsymbol{H}[k] S_j[l,k], \tag{3.26}$$

where the $\boldsymbol{H}_j[k] = [h_{1j}[k], \cdots, h_{Ij}[k]]^\intercal$ is the $I \times 1$ vector of the RTFs for source j. While this is a resonable approximation for short filters, it gets worsens as the filter length grows.

It is sometimes practical to concatenate all these vectors into an $I \times J$ matrix $\mathbf{H}[k] = [\boldsymbol{H}_1(f), \cdots, \boldsymbol{H}_J(f)]$ called *mixing matrix*.

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

$$\boldsymbol{X}[l,k] = \mathbf{H}[k]\boldsymbol{S}[l,k] + \boldsymbol{U}[l,k] \tag{3.27}$$

where $\boldsymbol{U}[l,k] = \boldsymbol{N}[l,k] + \varepsilon(l,k)$ 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. Besides, alternative models are common in the audio processing literature.

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):

$$\boldsymbol{D}_j[k] = \left[\frac{1}{4\pi q_{1j}} \mathrm{e}^{-\mathrm{i}2\pi f_k q_{1j}/c}, \cdots, \frac{1}{4\pi q_{Ij}} \mathrm{e}^{-\mathrm{i}2\pi f_k q_{Ij}/c}, \right] \tag{3.28}$$

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)[34] 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]

$$G_i[k] = \frac{H_i[k]}{H_1[k]}. \tag{3.29}$$

[34] In 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).

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

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

Notice that $\tilde{h}_i = \tilde{g}_i \star \tilde{h}_1$ 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 $\tilde{x}_i = \tilde{g}_i \star \tilde{x}_1$. 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 $g_i = h_i$;[35]
- 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 of the source signal. They can be used as *features* for the estimation of parameters of the acoustic propagation. Examples of this are many works in the Sound Source Localization domain (e. g., [Laufer et al. 2013; Deleforge et al. 2015a; Di Carlo et al. 2019].

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

Under the narrowband approximation, the ReTF can be approximated by the ratio of the DFTs of the channel signal $X_i[k]$ and $X_1[k]$.

$$G_i[k] \approx \frac{X_i[k]}{X_1[k]} \approx \frac{H_i[k]S[k]}{H_1[k]S[k]} = \frac{H_i[k]}{H_1[k]}. \tag{3.31}$$

It can be noticed that ideally, the ReTF removes the dependency from the source signal. However, this is true only in theory. In practice, the presence of external noise and the spectral sparsity of particular source signals (e. g., speech, and music) make the ratio diverging from the real ReTF. We will refer to the ReTFs computed with this vanilla approach as *instantaneous* ReTF. A way to improve the estimation of the ReTF is to perform the above operation in the STFT and then consider the average over multiple time frames. In this way, the non-stationarity and the spectral sparsity of the speech signals are alleviated. This approach is typically performed for computing the acoustic features known as *interchannel cues*, described below. More advanced estimators for the ReTF have been proposed in the literature and will be discussed in § 4.3.3. The reader can refer also to **??** for common and state of the art methods for ReTF estimation.

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

$$\boldsymbol{G}_j[k] = \frac{1}{H_{1j}[k]} \boldsymbol{H}_j[k]. \tag{3.32}$$

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

$$\boldsymbol{D}_j[k] = \left[1, \mathrm{e}^{-\mathrm{i}2\pi f_k (q_{2j} - q_{i'j})/c}, \ldots, \mathrm{e}^{-\mathrm{i}2\pi f_k (q_{Ij} - q_{i'j})/c}\right] \tag{3.33}$$

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 Chapter 10 as they are fundamental quantities for sound source localization.

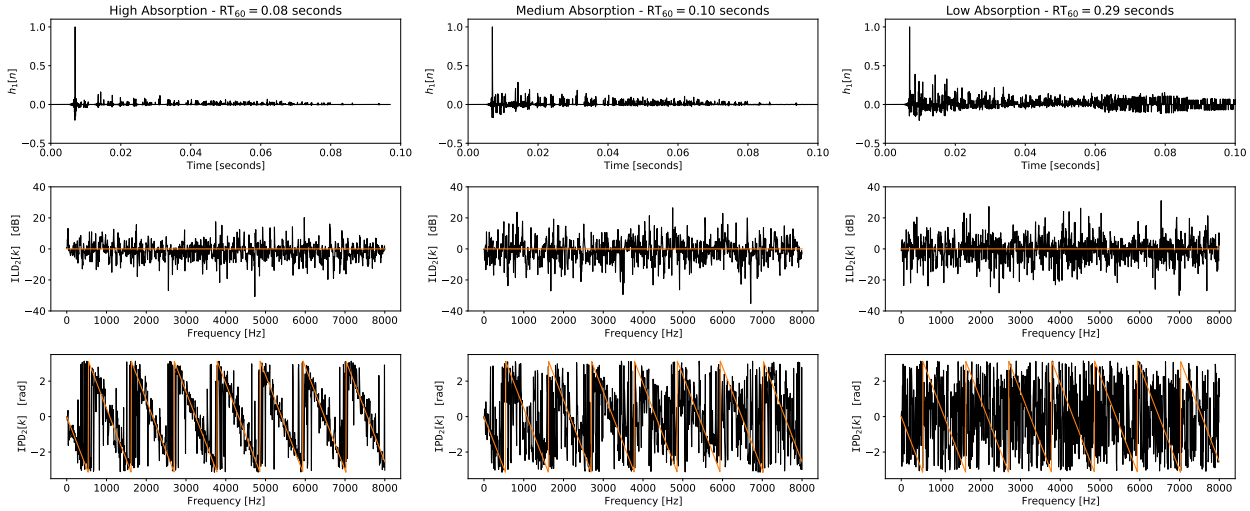


FIGURE 3.8: (Top) Synthetic RIR for different absorption conditions for the reference channel 1. (Below) ILD and IPD related the too channel 2 computed with respect to channel 1. Orange lines denote the theoretical far- and free-field ILD and IPD as defined by the relative steering vectors of Eq. (3.33). As the source is placed far from the microphones, the attenuation is similar on both the channels, whereas the delay is not.

► IN THE CONTEXT OF SPATIAL AUDITORY PERCEPTION [Bregman 1990] and Computational Auditory Scene Analysis (CASA) [Brown and Wang 2005], the ReTF is related to the *interchannel cues*[36]. In fact, the ReTF encodes the so-called Interchannel Level Difference (ILD) and the Interchannel Phase Difference (IPD)

$$\begin{aligned} \mathtt{ILD}_{ij}[k] &= 20\log_{10}|G[k]| \quad [\mathrm{dB}] \\ \mathtt{IPD}_{ij}[k] &= \angle G[k] \quad [\mathrm{rad}] \end{aligned} \tag{3.34}$$

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.

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

3.4 CONCLUSION

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 of 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 in Part III.

Part II

ACOUSTIC ECHO RETRIEVAL

4

Acoustic Echo Retrieval

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

▸ 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

As discussed in § 3.1.1, the continuous-time multi-channel signal model for one source signal $\tilde{s}(t)$ and I channels writes, for each channel i,

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

where $\tilde{h}_i(t)$ is the echo model for the RIR between the source and the i-th microphone. 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. This model can be generalized to frequency dependent attention, but it will be not addressed here.

The Acoustic Echo Retrieval (AER) problem consists in estimating the echoes' timings $\{\tau_i^{(r)}\}_{i,r}$ and attenuations (or gains) $\{\alpha_i^{(r)}\}_{i,r}$ from multichannel observations $\tilde{x}_i$ 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 this 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 generalization of *TOAs estimation*, or an instance of *acoustic channel estimation* (or *shaping*), *spike retrieval* and *onset detection* (see following § 4.3.1). 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 applications related to Sound Source Localization (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 we will make explicit the distinction between the two.
AER may be confused with the *acoustic echo cancellation* (AEC) problem in telecommunication and telephony which refers to the problem of estimating and suppressing feedbacks due to loudspeakers being too close to microphones.[37]

[37] AEC is a problem of very different nature, amounting to suppress from a mixture a known signal distorted by an unknown filter. This will not be further discussed in this thesis

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 broader category of *deconvolution problems* since a "clean" reference signal is used to *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 as a *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.

⇄ 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 sequential tasks: RIR estimation followed by echo extraction. The former can be seen as an instance of *channel estimation* (a. k. a. *system identification*) problem, 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 to enhance a few early reflections which are perceptually useful. *RIR-agnostic methods*, instead, try to overcome the challenging tasks of estimating the full acoustic channels and tuning peak-picking methods. They

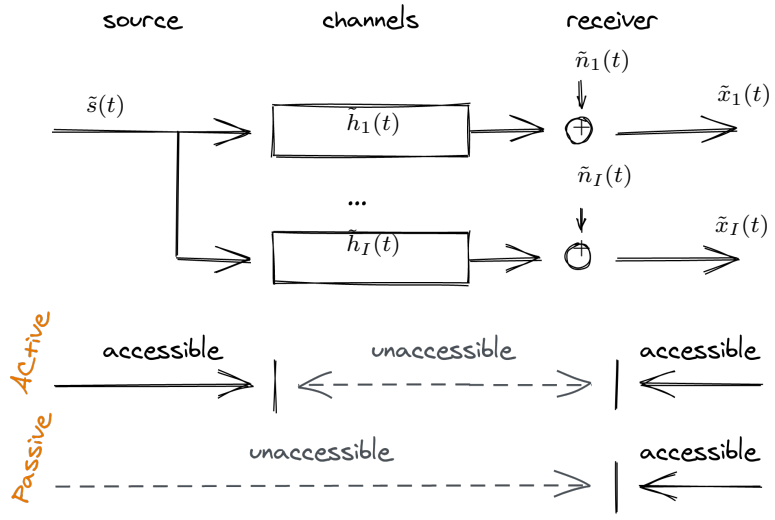


FIGURE 4.1: Schematic view of active vs. passive approaches.

attempt to estimate echo properties directly in the parameter space of echos' TOAs and amplitudes.

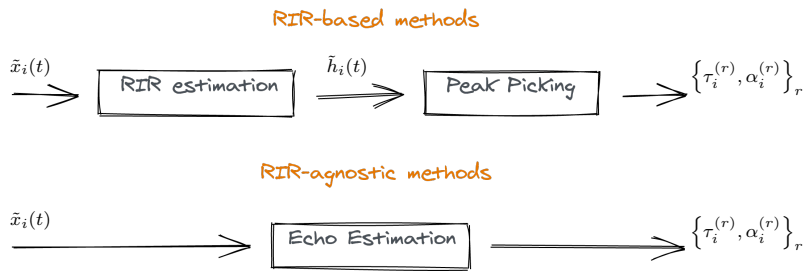


FIGURE 4.2: Schematics of RIR-based and RIR-agnostic approaches.

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 can be chosen, several methods were developed to measure real RIRs. Since the RIR corresponds to the room response to a perfect impulse, one can measure 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 terms of reproducibility and safety. Instead, modern computational techniques are used, involving the computation of a deconvolution or *cross-correlation*[38] 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

[38] Cross-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.

original MLS signal. This method was further improved in order to achieve better RIR estimation in [Aoshima 1981; Dunn and Hawksford 1993]. Unfortunately this technique introduces several artifacts which yield spurious peaks in 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 enjoys the following properties: the signal spans a user-defined frequency range; it compresses into a narrow impulse during autocorrelation[39]; and its Fourier inverse is available in closed form. The reader can find a review of the presented techniques applied to RIR measurements in [Szöke et al. 2019].
Sometimes the reference signal is known, but none of the above techniques can be used. For instance, when the source signal is narrowband or when the acoustic conditions are challenging. Therefore the RIR estimation problem needs to be addressed as a more general deconvolution problem, typically solved through optimization methods. This approach has been well studied in the literature and can be solved using standard Linear Least Squares Error methods with closed-form solutions [Krim and Viberg 1996; Glentis et al. 1999] However, in the case of signals 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, as it will discussed later.

[39] Ideally, if an infinite range of frequencies is used, then it compresses into the ideal Dirac function.

- 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 2016], $|\tilde{h}(t)|$, or the energy envelope[40] [Schroeder 1979] are used instead.
Provided a good estimation of the RIRs, the echoes' locations and amplitudes could be extracted manually by experts. However, even in an 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 multiples 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 in echo' timing estimation may lead to significant differences for 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 technique 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 algorithms on the RIR, followed by robust and manually-tuned peak

[40] 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

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

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

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 according to changes in the phase slope of the RIR spectrum. Other works apply onset detection techniques used in music information retrieval 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 reflections in the RIRs exhibit a similar shape to 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 a pattern (or atom) to be retrieved across the RIR. Unfortunately, in its pure form, this approach is unsuitable for real RIRs because of the frequency-dependent characteristic of the room absorption material. In order to improve the detection, in [Kelly and Boland 2014] this approach is extended to *Dynamic Time Warping* to account for the non-uniform compression, dilation and concurrency of echoes. Nevertheless, the idea of exploiting the direct path component to isolate the source-receiver coupling and thus identify first prominent reflections through deconvolution was used in [Annibale et al. 2012]. This technique is also referred to as *matching filter* or *direct-path compensation*.

Alternative 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 cross-wavelet transform based on previous works [Guillemain and Kronland-Martinet 1996; Loutridis 2005]. 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 spectrum.[41] One interesting property of the cepstrum is to separate the spectral envelop form the line-spectra components in different "quefrency" ranges (i. e., the final independent variable of the cepstrum domain). Therefore, it is used to detect periodicity in the spectrum of the observed signal, with many applications in speech processing as well as in hydraulic and mechanic applications. This approach seems promising since time-domain peaks are mapped to complex sinusoids in the frequency domain. However, this representation is highly sensible to external and sampling noise and the accuracy is limited by the approximation of the DFT operator.

[41] There are several concurrent definitions 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.

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 retrieved in closed-form. Although this method reaches the statistical optimality in the ML sense, the exact knowledge of the 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 the 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 allows 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.
  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 a potentially high number of candidates, which leads to intractable computational complexity when multiple reflections are considered. Moreover these methods require that all the distances between microphones 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 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 transform[42], they are able to disambiguate TOAs and reconstruct the reflectors'

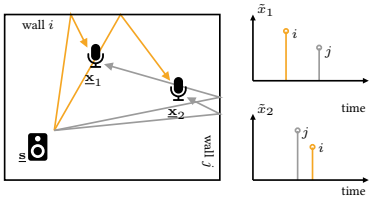


FIGURE 4.3: Illustration of different orders of arrival of wall reflections

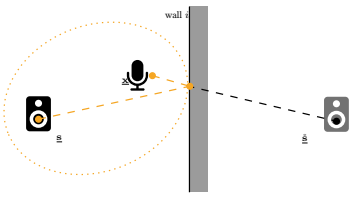


FIGURE 4.4: Illustration of the ellipsoid method: the length of the reflected path from the image source $\underline{\mathbf{s}}$ to the microphone at $\underline{\mathbf{x}}$ constrain the reflector line to be tangent to an ellipse.

[42] A 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 tranform is typically used as feature extractor to detect lines and edges in pictures.

positions and orientations. However, they all require a very specific acquisition setup and the optimized non-convex cost functions are sensible to local minima.

In general, none of the above methods come with specific strategies to cope with missing or spurious echo estimates. These may be 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 draws specific patterns which can be identified easily and robustly even in the presence of (a few) spurious and missing peaks. Moreover, this approach avoids the combinatorial search. However, it 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 expressed in closed-form and used in an 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 echo contributions can be extracted from the *cross-correlation* between the observed and the reference signals. Cross-correlation analysis is a mathematical tool for the identification of repeated patterns in a signal as a 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. Hence, 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 arrivals (elevation and azimuth), namely from TOAs to 2D-DOAs. The identification of strong reflections can then be 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 from 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.

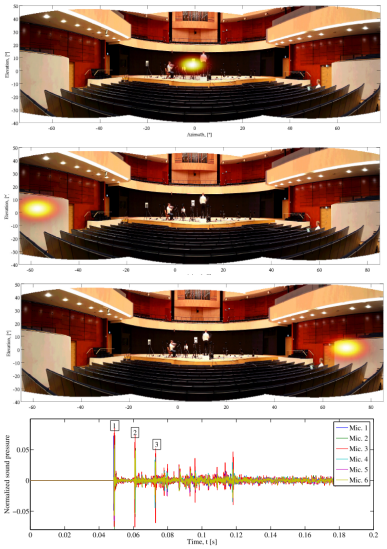


FIGURE 4.5: Visualization of the *audio camera*: The angular spectrum is overlapped to the corresponding images. Also shown are the impulse responses for 6 microphones. The numbered boxes indicate events shown in the audio camera. Images from [Tervo 2011].

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

4.3.3 *Passive and RIR-based method*

Passive approaches rely on external, unknown 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 *source-agnostic methods*. Although it makes the exposition clearer, there is a lot of overlap between these two methods nowadays. It could also be said that the accuracy of certain source-agnostic methods can be improved by the availability of source statistics.

- STATISTICAL METHODS exploit knowledge about the emitted signal. Since the nature of the source signal is often non-deterministic and it statistics can be modeled based on the signal category, e. g. speech or music. Two main approaches can be identified [Tong and Perreau 1998]:

  - *Second Order Moments approaches* derive closed-form solutions 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 MASS [Ozerov and Févotte 2010; Duong et al. 2010; Sawada et al. 2013; Leglaive et al. 2016; Leglaive et al. 2018; Scheibler et al. 2018c]. These methods are built on the well-studied framework of Multichannel Nonnegative

Matrix Factorization (NMF) [Ozerov and Févotte 2010; Sawada et al. 2013] which lends itself to account for various types of side information. Here the source signals are typically modeled as Gaussian distributions centred at zero and with unknown variances. In general, they aim at estimating both the acoustic channels and the source contribution. In particular, the work of [Duong et al. 2010] extends this framework to 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 of estimating 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 a sufficient level of details, indispensable for AER.

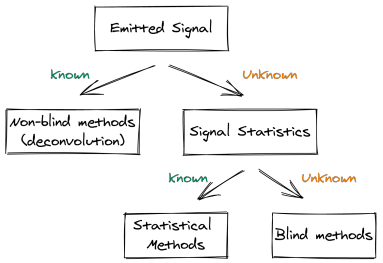


FIGURE 4.6: Classification of the State of the Art in channel estimation.

- SOURCE-AGNOSTIC 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 2016]. The formers are based on the key idea that the channel (or part of it) vector spans a one-dimensional subspace of noiseless observations. These methods have the attractive property that the channel estimates can often be obtained in closed-form by optimizing a quadratic cost function. However that they are typically computationally expensive and sensible to noise.
  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

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

by commutativity and associativity of the convolution operator. This identity 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 on 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 hyperparameters in [Lin et al. 2008]. Even if sparsity and non-negativity could be seen as a strong assumptions, works in speech enhancement [Ribeiro et al. 2010a; Dokmanić et al. 2015] 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 2016], 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.[43] 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.
An alternative approach is used in [Čmejla et al. 2019], where the ReIR estimation problem is treated as a special case of ReIR estimation. As mentioned in § 3.3.2, in the noiseless case, the ReIR identifies to the ReIR when the reference microphone is placed very close to the source. ReIR 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; Koldovskỳ et al. 2015; Koldovsky and Tichavsky 2015; Kodrasi and Doclo 2017]. By its definition, in the noiseless case, the ReIR 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 the 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].

[43] These approaches will be further formalized and detailed in Chapter 5.

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 an ideal low-pass kernel, 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 discussed in the previous chapter. Since these methods are fundamentally on-grid, the estimated 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 2016] reproduces the positions of the groundtruth peaks with remarkable precision.

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 methods 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, distinct 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 only consider 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 echo timing.

Recently, a fully blind, passive, off-grid and RIR-agnostic method was proposed by the 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*[44], [Condat and Hirabayashi 2013] and the theory of Finite Rate of Innovation (FRI) [Vetterli et al. 2002]. If the source signal is known, starting from the cross-relation identity, the AER problem translates into finding the annihilation filter for the RTFs, 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 a generalization of the Cadzow denoising framework [Condat and Hirabayashi 2015]. This method is shown to outperform conventional approaches by several orders of magnitude in precision in the 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.
In this direction, we contributed with two works: a knowledge-driven approach [Di Carlo et al. 2020], presented in Chapter 5, and a learning-based approach [Di Carlo et al. 2019], discussed in Chapter 6.

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

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 standard datasets for it 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 RIR corpora are freely available online, each of them meeting the demands of certain applications, usually SE and Automatic Speech Recognition (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, acoustic simulators typically based on the ISM are extensively used instead. While simulated datasets are more versatile, simple and quicker to obtain, they fail to fully capture the complexity and the richness of real acoustic environments. Due to these intrinsic issues, methods trained or validated on them may fail to generalize to real conditions, as will be shown in Chapter 7.

Database Name	Annotated			Number of				Key characteristics	Purpose
	Pos.	Echoes	Rooms	RIRs	Rooms	Mic×Pos.	Src		
[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	Circular dense array Circular placement of src.	RooGE SE†
[Remaggi et al. 2019]	✓	∼	✓	∼1.6k	4	48×2 +2×2	3-24	Circular 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
VoiceHome [Bertin et al. 2019]	✓	✗	✗	188	12	8×2	7-9	Various rooms, real homes	SE
`dEchorate` (Ch. 7)	✓	✓	✓	∼1.8k	1	30	6	Accurate annotation Different Echo-energy	**RooGE** **SE**

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. For instance in [Dokmanić et al. 2013] and in [Crocco et al. 2017] echoes are computed after the whole RooGE processing from some manually-selected peaks in the RIRs. The exact timing is not available. 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.

A good dataset for AER should include a variety of environments (room geometries and surface materials), of microphone placements (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 nearby 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 **could** be used in the context of echo-aware processing.

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 to 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:

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

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

- *The Normalized Projection Misaligment (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

$$\texttt{NPM}(\hat{h}, h) \stackrel{\text{def}}{=} 20\log_{10}\left(\frac{1}{\|h\|_2}\left\|h - \frac{h^\intercal\hat{h}}{\hat{h}^\intercal\hat{h}}\hat{h}\right\|_2\right) \quad \text{[dB]}, \quad (4.5)$$

  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 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]

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

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

In the context of RooGE, SSL and microphone calibration, echo 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 metrics considers only the 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 echo timings which are evaluated using regression/classification metrics of *information retrieval* and *machine learning*. Let be $\hat{\boldsymbol{\tau}} = \{\hat{\tau}_r\}_{r=0}^R$ and $\boldsymbol{\tau} = \{\tau_r\}_{r=0}^R$ the sets of estimated and reference echoes' TOAs. The following metrics are used:

- the *RMSE* is defined as

$$\texttt{RMSE}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) \overset{\text{def}}{=} \sqrt{\frac{1}{R+1}\sum_{r=0}^{R}|\hat{\tau}_r - \tau_r|^2} \quad \text{[seconds (or, samples)]}, \quad (4.7)$$

  This metric describes the mean error between estimated and reference echo 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.

Given 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:

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

Since these metrics rely on decision thresholds, their usage is not straightforward. In fact, in order to compare echoes, both estimated and reference echoes need to be labeled, pruned and matched first. 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 echo parameters 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 having taken hold in the recent years of signal processing:

⇄ 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 acoustic propagation, along with the audio. This knowledge is typically translated into mathematical models

which lead to solutions 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 in which it is used.
Regarding our contributions, Chapter 5 proposes a new knowledge-driven method for solving AER based on the theory of Continous Dictionarys (CDs).

⇄ DATA-DRIVEN METHODS, instead, are based on machine learning algorithms where information is automatically "learned from data" in *supervised* ways, therefore the knowledge of the physical laws is not need. Provided comprehensive and exhaustively annotated training datasets, such methods can learn functions 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. Alongside the great benefit of having black-box models that are able to learn on their own, this paradigm hides a few limitations.
First, these models rely on the information encoded in training data which are sometimes not representative enough of the real-world end-application. 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. It this case, prior knowledge is used to *generate* synthetic observations (i. e., from latent variables to the observations) and and the task of inverting the mapping is left to the learning method (i. e., from observation to latent variables). This leads to a second limitation of these approaches, that is *overfitting* to the data obeying the generative 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 have shown to achieve impressive performance. Therefore they are 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 my colleague Clement Elvira whose domain of expertise is in the Continous Dictionary (CD) framework, while my expertise lies in the audio and acoustic signal processing and modeling aspects. The section dedicated on the presentation of the CD framework applied to 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 highlighting the motivation behind it. Finally, this chapter recalls the main findings of the paper, while bringing additional insights in the existing models for AER.

" *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"

5.1 Introduction

Let us recall from Chapter 4 some knowledge-based methods addressing AER. Some existing 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 and an 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 2016]. 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 relies on a fixed grid of time samples 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 in the proposed approach is the recently proposed framework of Continous 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, the continuous-time signal recorded at microphone $i \in \{1, 2\}$ reads

$$\tilde{x}_i(t) = (\tilde{s} \star \tilde{h}_i^\star)(t) + \tilde{n}_i(t) \tag{5.1}$$

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

$$\tilde{h}_i^\star(t) = \sum_{r=0}^{R_i} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \tilde{\varepsilon}_i(t) \tag{5.2}$$

where R_i is the (unknown) number of echoes.
In the noiseless case, that is when $\tilde{n}_i = 0$ for $i \in \{1, 2\}$, we have the cross-relation identity

$$\tilde{x}_1 \star \tilde{h}_2^\star = \tilde{x}_2 \star \tilde{h}_1^\star \tag{5.3}$$

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 $\tilde{\phi}_{\text{LP}}$ 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

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

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 operate 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]:

$$\widehat{h}_1, \widehat{h}_2 = \underset{h_1, h_2}{\arg\min}\ \tfrac{1}{2}\|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2 + \lambda\|\mathbf{h}\|_1 \tag{5.5}$$

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 (built as shown in Figure 5.1) associated to a convolution where $2N$ and L respectively denote the microphone and filter signal lengths,
- $\mathbf{h} = [h_1^\intercal, h_2^\intercal]$ is the concatenation of the two vectorized discrete filters,
- 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).

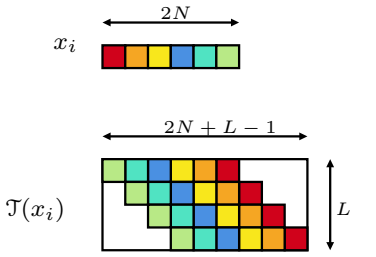


Figure 5.1: Graphical representation of the construction of $\mathcal{T}(x_i)$ from x_i

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:

$$\underset{u}{\arg\min}\ \tfrac{1}{2}\|v - Au\|_2^2 + \lambda\|u\|_1. \tag{5.6}$$

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), as follows

$$\widehat{h}_1, \widehat{h}_2 = \underset{h_1, h_2}{\arg\min}\ \mathcal{J}(h_1, h_2) + \mathcal{P}(h_1, h_2) \text{ s.t. } \mathcal{C}(h_1, h_2) \tag{5.7}$$

where $\mathcal{J}(h_1, h_2) = \frac{1}{2}\|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2$ 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 solutions and a constrained set. Thanks to this formulation, current state-of-the-art approaches can be summarized as in the Table 5.1.

Reference	$\mathcal{P}(h_1, h_2)$	$\mathcal{C}(h_1, h_2)$	Note
[Tong et al. 1994]	✗	$\|\mathbf{h}\|_2^2 = 1$	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]	$\lambda\|\mathbf{h}\|_1$	$\mathbf{p}^{(z)\mathsf{T}}\mathbf{h} = 1, \mathbf{h} \geq 0$	Iterative weighted ℓ_1 constraint.

TABLE 5.1: Some state-of-the-art penalties and constraints used in 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 2016], where the anchor constraint is replaced by an *iterative weighted* ℓ_1 equality constraint, i. e., such that at each iteration z, $\mathbf{p}^{(z)\mathsf{T}}\mathbf{h}^{(z)} = 1$.[45] 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 2016].

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

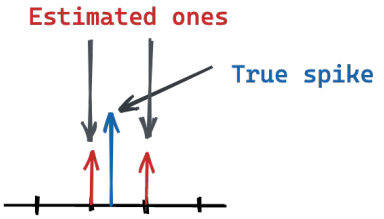


FIGURE 5.2: Schematics of the *bodyguard effect* affecting on-grid approaches. This is true only if h is non-negative.

▶ 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 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 inverse of F_s.
- *Bodyguard effect*. In addition to affecting the AER performance, on-grid methods may converge slowly to suboptimal solutions. In fact, as shown in Figure 5.2, instead of estimating the positive peaks at its true location, two smaller "bodyguard" peaks are estimated around it

instead. The estimation procedure may stop at this point returning two wrong peaks. Having smaller coefficients, these 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 need to be built, one for each pair of microphones. In addition, this leads to the risk that the optimization problem becomes ill-conditioned.

In the following section we will present a novel approach that aims at addressing the above limitations. It is based on a framework proposed for solving LASSO problems for continuous variables, hence the name CD. Before, let us show how 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

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

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

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

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

$$\tilde{\phi}_{\text{LP}} * \tilde{x}_1 * \tilde{h}_2^\star = \tilde{\phi}_{\text{LP}} * \tilde{x}_2 * \tilde{h}_1^\star \tag{5.9}$$

holds even during the introduced measurement process, hence

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

where $\mathcal{F}$ denotes the Fourier Transform (FT) described in § 3.2.
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 $\tilde{h}_i^\star$ 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, we consider the Discrete Fourier Transform (DFT) of $\tilde{x}_i$:

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

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

$$X_i[k] = \sum_{n=0}^{2N-1} x_i[n] \mathrm{e}^{-\mathrm{i}2\pi \frac{kn}{2N}} \tag{5.12}$$

is the DFT of the real vector $\tilde{x}_i$ as defined in Eq. (3.15) for positive frequencies only.

Let us define Δ_τ the following parametric vector of complex exponential

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

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

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

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

$$\Theta \overset{\text{def}}{=} [0, T] \times \{1, 2\}, \tag{5.15}$$

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

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

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 this type of 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) $\|\mu\|_{\mathrm{TV}}$ [46]. We now define the *linear* observation operator $\mathcal{A}\colon \mathcal{M}_+(\Theta) \to \mathbb{C}^{N+1}$, which is such that

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

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

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

$$\mathcal{A}\mu^\star = \mathbf{0}_{N+1}, \tag{5.18}$$

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

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

where $\bar{\mu}^\star$ is a (finite) discrete measure verifying $\bar{\mu}^\star(\{(0,1)\}) = 0$.
Denoting $y \overset{\text{def}}{=} -\mathcal{A}\delta_{(0,1)} \in \mathbb{C}^{N+1}$, the relation Eq. (5.18) becomes

$$\mathcal{A}\bar{\mu}^\star = y. \tag{5.20}$$

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 $\mu^\star$ by solving

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

Note that (5.21-$\mathcal{P}^0_{\mathrm{TV}}$) 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., $\|\mu\|_{\mathrm{TV}} = \sum_{r=0}^{R-1} |\alpha^{(r)}|$.

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-$\mathcal{P}^0_{\mathrm{TV}}$) is relaxed, leading to the so-called Beurling-LASSO (BLASSO) problem

$$\widehat{\mu} = \underset{\mu \in \mathcal{M}_+(\Theta)}{\arg\min} \ \tfrac{1}{2}\|y - \mathcal{A}\mu\|_2^2 + \lambda \|\mu\|_{\mathrm{TV}} \quad \text{s.t.} \quad \mu(\{(0,1)\}) = 0. \tag{5.22-$\mathcal{P}^\lambda_{\mathrm{TV}}$}$$

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 the problem in 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, optimizing over the infinite-dimensional space of measures is not possible, and hence non-convex optimization with respect to the measures *parameters* is performed instead.

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.

$$\arg\min_{u} \tfrac{1}{2}\|v - Au\|_2^2 + \lambda\|u\|_1 \text{ s.t. } u \geq 0$$

$$\downarrow$$

$$\arg\min_{u} \tfrac{1}{2}\|y - \mathcal{A}\mu\|_2^2 + \lambda\|\mu\|_{\text{TV}} \text{ s.t. } \mu \in \mathcal{M}_+(\Theta)$$

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. In the BLASSO, this is translated assuming the channels being spare non-negative measures, i. e., $\mu = \sum_r \alpha^{(r)}\delta(t - \tau^{(r)})$.
- *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 of 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 [Di Carlo et al. 2020, Supplementary Material] and in § 12.2. In a nutshell, the algorithm iterates over the following steps until a condition on the cost function is met.

1. *Anchor constraint.* At first the anchor constraint in added arbitrarily on one of the two filters. This is used to initialize the two filters.

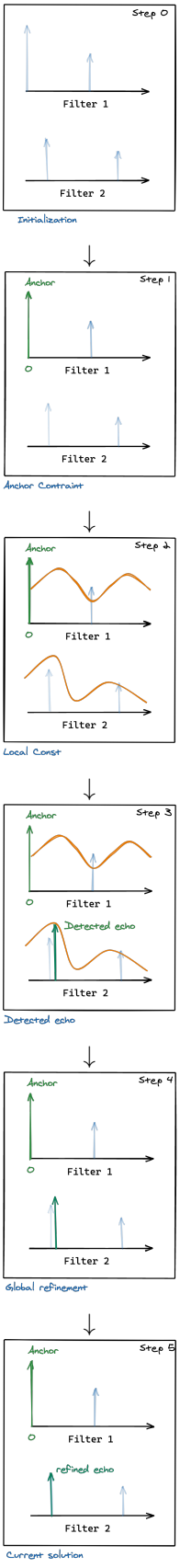


FIGURE 5.3: Illustration of the sliding Frank-Wolfe algorithm proposed in [Denoyelle et al. 2019] in BLASTER.

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 a previously found solution is 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. Otherwise, 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, rely on of the regularization parameter λ 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 $\lambda_{\max}$ 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 $\lambda_{\max}$ is upper bounded by $\max_{\theta\in\Theta}|y^{\intercal}\mathcal{A}\delta_{\theta}|$. 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 2016].[47] 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 levels of noise (SNR) and reverberation time ($\texttt{RT}_{60}$).

[47] Reference implementations for IL1C and BSN were kindly provided by the authors of [Crocco and Del Bue 2016].

The quality of AER estimation is assessed in terms of precision[48] in percentage as in the literature of onset detection [Böck et al. 2012] and the RMSE in

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

samples. Both metrics evaluate only the *matched* peaks, where a *match* is defined as being within a small window $\tau_{\max}$ 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]	$\mathtt{RT}_{60}$ [s]
$\mathcal{D}^{\text{(valid)}}$	broadband noise	⚄	⚄
$\mathcal{D}^{\mathtt{SNR}}$	broadband noise, speech	⚃	400 ms
$\mathcal{D}^{\mathtt{RT}_{60}}$	broadband noise, speech	20 dB	⚃

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 law.

$\mathcal{D}^{\text{(valid)}}$ is used for tuning the hyperparameter λ and the peak-picking parameters for IL1C and BSN using $\mathtt{RT}_{60}$ and SNR randomly drawn from $\mathcal{U}[0, 1]$ (sec) and $\mathcal{U}[0, 20]$ (dB) respectively; $\mathcal{D}^{\mathtt{SNR}}$ features SNR value uniformly sampled in $[0, 6, 14, 20, \infty]$ while the $\mathtt{RT}_{60}$ is kept fixed to 400 ms; akin the $\mathcal{D}^{\mathtt{RT}_{60}}$ is built sampling $\mathtt{RT}_{60}$ 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 $\mathcal{D}^{\mathtt{SNR}}$ and $\mathcal{D}^{\mathtt{RT}_{60}}$ speech utterances from the TIMIT dataset [Garofolo et al. 1993] are also included. The signal duration is kept fixed to 1 s with sampling frequency $F_s = 16$ kHz. For a given $\mathtt{RT}_{60}$ value and room with random dimensions, a unique absorption coefficient is assigned to all surfaces based on Sabine's formula (Eq. (2.17)). Then, the two microphones and the source are randomly positioned inside the room. The parameters of the 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 ($\texttt{RT}_{60} = 200$ ms, $\texttt{SNR} = 20$ dB), up to 68% of echoes can be retrieved by BLASTER with errors lower than a quarter of sample in that case. This is promising since the practical advantage of knowing the timings of two echoes per channel has been demonstrated in [Di Carlo et al. 2019] (See Chapter 10), and in [Scheibler et al. 2018c] (See Chapter 9).

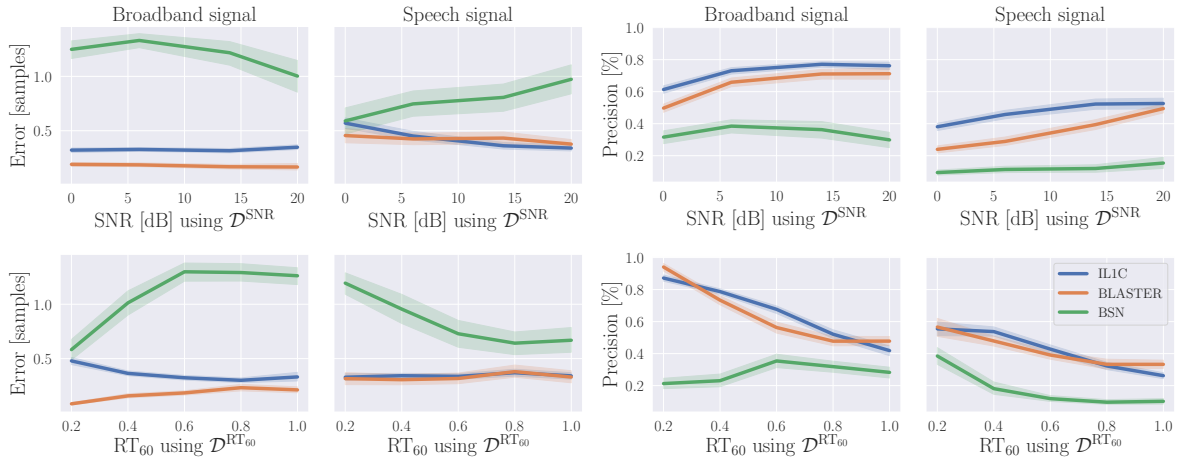


FIGURE 5.4: Mean error (left) and precision (right) versus SNR level (top) and $\texttt{RT}_{60}$ level (bottom) using broadband and speech signals for the task of recovering $R = 7$ echoes. A threshold of $\tau_{\text{max}} = 2$ samples is used to compute the precision. Error bands denotes 95% confidence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

	Precision [%]									
	R = 2 echoes					R = 7 echoes				
τ_{max}	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, $\texttt{RT}_{60} = 200$ ms and $\texttt{SNR} = 20$ dB.

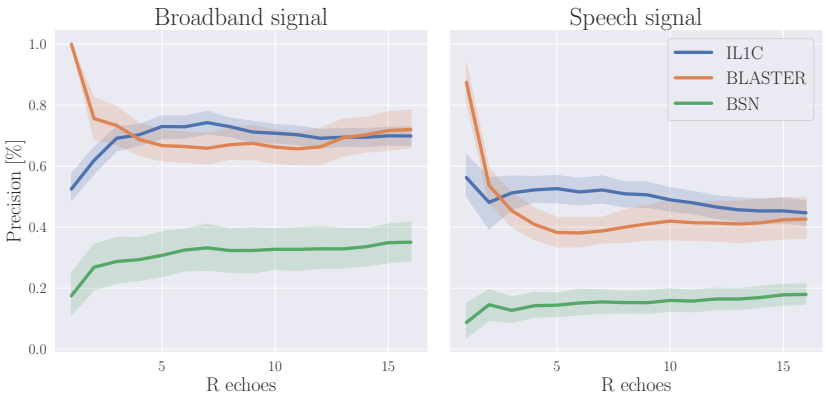


FIGURE 5.5: Precision versus number of echoes R to be retrieved for broadband (left) and speech (right) signals with RT_{60} = 400 ms and SNR = 20 dB. Error bands denotes 95% confidence intervals.

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. The 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 classical methods for BCE, discussed in the previous chapter. Despite an heavy mathematical formulation, it can seen as as the continuous extension of a 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 Multi-channel 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 to deal 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;
- 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

"I'm here 'cause I am stuck
Somebody jump in my computer server
And take the information out"
—Venetian Snares, *Gentleman*

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"

▶ Synopsis In this chapter, we 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 details in Chapter 10. Finally, we will present a possible way to achieve a more robust estimation (§ 6.4) and discuss a possible way to scale this approach to multiple echoes (§ 6.5).

Deep learning notions come from readings 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 towards audio signal processing. In the remaining sections, we will present part of the previously published work [Di Carlo et al. 2019] and ongoing 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 as a follow-up on 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 per channel. The generalization to multiple echoes will only be discussed as future work.

6.1 Introduction

Recently, Deep Neural Networks (DNNs) have 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, DNNs have been widely used in many domains. They belong to the class of *supervised* and *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 led to a considerable improvement in performance, in particular in audio and music source separation, speech enhancement, and source localization (see related sections in Chapter 8). Moreover, with respect to traditional machine learning methods, the use of

deep learning models presents several advantages. They are flexible and adaptable across tasks; for instance, Convolutional Neural Network (CNN)-based models from Computer Vision can be adapted to audio source separation (e. g., [Xiao et al. 2016; Sainath et al. 2017; Perotin et al. 2018]), and to source localization (e. g., [Chakrabarty and Habets 2017; Vesperini et al. 2018; Nguyen et al. 2018; Salvati et al. 2018]). Furthermore, DNNs reduce — even sometimes remove — the step of designing hand-crafted features, by including feature learning as part of the learning process.
Other supervised learning methods have been proposed in the audio processing literature, such as GMM-based frameworks [Deleforge et al. 2015a; Gaultier et al. 2017]. This framework allows to build partially informed, invertible regression models based on probabilistic prior knowledge and are competing with DNN models in case of a small training dataset. However, they require stronger assumptions on the nature of the mapping between input-output pairs, such as (local-)linearity or independence between variables.

At first, the Gaussian Locally-Linear Mapping (GLLiM) framework based on Gaussian Mixture Model (GMM) was considered to address the AER problem. Then, we oriented our attention towards DNN models, which gave 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 observations. 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.

The use of supervised learning models (not only the DNNs), presents some disadvantages. One of the most obvious 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 (RT_{60}) and Signal-to-Noise-Ratio (SNR) levels;
- a precise (up to computer precision) annotation of the data is directly available,
  e. g. spatial or temporal information on the audio scene events;
- and, 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. For instance, authors of [Deleforge et al. 2015b] showed, in a SSL context, that training on real data collected in one part of a room and testing in another part of the same room does not generalize well. Interestingly, the works in [Deleforge et al. 2015a; Nguyen et al. 2018] propose an automatic approach for collecting and annotating real data automatically as a pre-calibration step. However, this approach is possible only with specific technologies and equipments (in both works, 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.

6.2 Deep learning models

6.2.1 *Multi-layer perceptrons*

Multi-layer Perceptrons (MLPs) are simple and basic modules of DNNs and are also known as *fully-connected* or *dense layers*. They consist of a sequence of layers, each of which defined by an affine vector transformation followed by a an entry-wise non-linear 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 an input to a representation space where problems (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 inputs. Although scaling can be compensated by data normalization, they are often not suitable to capture temporal- and frequency-variations of audio signals. 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 constitute one of the key features 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. This means that, the activation functions make the model capable of accounting for more complex relationships between the input and the output than an affine transformation. Typical examples of these functions are the hyperbolic tangent, the Sigmoid function, and the Rectified Linear Unit (ReLU).

6.2.2 *Convolutional neural networks*

The CNNs consist 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 properties, thanks to the convolution operation;
- reduced model complexity, since the same kernel can be used at different input location;
- the ability to detect local structures at a different levels of abstraction in the network.

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

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

where

- $\mathbf{X}[:,:,i] \in \mathbb{R}^{F \times T \times I}$ and $\mathbf{Y}[:,:,k] \in \mathbb{R}^{A \times B \times K}$ are input and output tensors, respectively.
- $i \in [0, I]$ denotes the channel dimensions;[49]
- $k \in [0, K]$ denotes the channel dimension in the output;
- the tensors $\mathbf{W}$ and $\mathbf{b}$ denotes the weight and bias tensors, respectively;
- f is an activation function;
- $*$ denotes discrete convolution;

[49] In the deep learning community, input dimensions are often referred to as channels Based on the application, such channels *do not necessarily* correspond to the channels of microphone recordings.

A CNN can be designed with either 1D, 2D or 3D convolutions (e. g., along time, frequency and channel dimension), or a combination of them. In general, in audio, 1D convolution layers are used to process the time-domain or frequency-domain input, whereas 2D convolutional layers are used for time-frequency representations, such as spectrograms. The output of a convolutional layer consists of a collection of convolved versions of the input signal, which are typically referred to as *feature maps*. 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 maps. 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. Typical pooling operators are *max-pooling* and *mean-pooling*, which samples non-overlapping patches by keeping the largest 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 MLPs suffer from, and allow to the extract of spatial features. In contrast, MLPs offer a simple and generic mappings from high-dimensional spaces to lower ones, suitable for classification and regression problems. Therefore, hybrid architectures are now the standard for deep learning models.

6.2.4 *Learning and optimization*

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

the predicted values $\hat{\mathbf{Y}}$ and the desired ones $\mathbf{Y}$. The optimization process then iteratively updates the parameters $\boldsymbol{\theta}$ so the loss function is minimized. In the case of vanilla gradient descent this writes:

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

where η is referred to as the *learning rate* and controls the length of the *gradient step* at each iteration. Due to the composite structure of the DNNs, the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta}}(\hat{\mathbf{Y}}, \mathbf{Y})$ is computed via the chain rule, using the *back propagation* algorithm. Most gradient descent variants used today adapt the learning rate along iterations, such as the popular *Adam optimizer* [Kingma and Ba 2014].
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 improve convergence and to solve computational and memory load issues occurring when using training sets. It approximates the gradient at each step on a mini-bach 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 conditions on the loss function (or other metrics), for instance, when it is not improving after certain amount of iterations. This criteria is typically computed on a dedicated data set known as *validation* set to avoid overfitting;
- *batch normalization* consists of scaling the data by estimating their statistics during training, which usually leads to better performance and faster convergence;
- *dropout* is a regularization technique that reduces overfitting by randomly omitting some network units during training.

In this chapter, we will propose a framework based on the following 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 DNNs to estimate the timings of the first and strongest echo in each channel from stereophonic recordings. This setup is related to "table-top" scenarios, which are commonly encountered in typical home smart audio devices and it is motivated by an echo-aware Sound Source Localization (SSL) application that will be discussed in Chapter 10. To this end, we consider the following simple setup: two microphones placed close to a surface, as illustrated in Figure 6.1. The reason why we consider only one microphone pair is that this approach can be generalized to microphone arrays using data-fusion-like techniques. In fact, by considering all the pairs, it is possible to aggregate their estimation in a second stage.[50] In the reminder of this chapter, we will discuss how to achieve robust first-echo estimation.

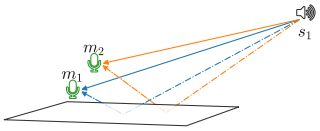


Figure 6.1: Typical setup with one sound source recorded by two microphones. The illustration shows the direct sound path (blue lines) and the resulting first-order echoes (orange lines).

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

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),

$$H_i[k] = \sum_{r=0}^{R=1} \alpha_i^{(r)} \, \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_i^{(r)}} \; + \; \varepsilon_i[k], \tag{6.1}$$

where $\tau_i^{(r)}$ and $\alpha_i^{(r)}$ are the echoes' times of arrival and amplitudes, respectively. Here, $f_k = kF_s/F$ denotes k-th of F frequency bins and the error term $\varepsilon_i[k]$ collects later echoes, the reverberation tail, diffusion, and noise.
As we consider only the first echo per channel, we have $R = 1$. Given the close-surface scenario, the first echo is the earliest and strongest.

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

$$\begin{cases} \texttt{ILD}[k] = & 20\log_{10}\left|\frac{1}{T}\sum_{l=1}^{T}\frac{X_2[k,l]}{X_1[k,l]}\right|, \\ \overline{\texttt{IPD}}[k] = & \frac{1}{T}\sum_{l=1}^{T}\frac{X_2[k,l]/|X_2[k,l]|}{X_1[k,l]/|X_1[k,l]|}, \end{cases} \tag{6.2}$$

where $X_i[k,l]$ is the STFT of the i-th microphone. In practice we do not consider the IPD in radiants, but rather its complex exponential, denoted as $\overline{\texttt{IPD}}$.
More precisely, the input of the network is

$$\xi = \left[\texttt{ILD}, \mathrm{Re}\{\overline{\texttt{IPD}}\}, \mathrm{Im}\{\overline{\texttt{IPD}}\}\right],$$

namely, the concatenation of the above features in all frequencies into a single vector.. Here $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote real and imaginary part operators, respectively. Finally, the entries of ξ corresponding to $\overline{\texttt{IPD}}[0]$ are discarded because always equal 1.

Initial investigations showed that learning the echoes amplitudes yielded large estimation errors, probably due to the complexity of the task. Thus, we considered echo timings with respect to the direct-path TOA in a reference microphone for the following three main reasons: first, the origin of time cannot be known in the blind setting; second, motivated by an application to passive SSL discussed in Chapter 10, Time Differences of Arrival (TDOA) can be easily converted to angles. Then, the timings of the first echoes in the two

channels can be fully parameterized by the following three quantities:

$$\begin{cases} \tau_{\texttt{TDOA}} &= \frac{1}{c}\|\underline{\mathbf{x}}_2 - \underline{\mathbf{s}}\| - \frac{1}{c}\|\underline{\mathbf{x}}_1 - \underline{\mathbf{s}}\| = \tau_2^{(0)} - \tau_1^{(0)} \quad [\text{s}], \\ \tau_{\texttt{iTDOA}} &= \frac{1}{c}\|\mathring{\underline{\mathbf{x}}}_2 - \underline{\mathbf{s}}\| - \frac{1}{c}\|\mathring{\underline{\mathbf{x}}}_1 - \underline{\mathbf{s}}\| = \tau_2^{(1)} - \tau_1^{(1)} \quad [\text{s}], \\ \tau_{\texttt{TDOE}_1} &= \frac{1}{c}\|\mathring{\underline{\mathbf{x}}}_1 - \underline{\mathbf{s}}\| - \frac{1}{c}\|\underline{\mathbf{x}}_1 - \underline{\mathbf{s}}\| = \tau_1^{(1)} - \tau_1^{(0)} \quad [\text{s}], \\ \tau_{\texttt{TDOE}_2} &= \frac{1}{c}\|\mathring{\underline{\mathbf{x}}}_2 - \underline{\mathbf{s}}\| - \frac{1}{c}\|\underline{\mathbf{x}}_2 - \underline{\mathbf{s}}\| = \tau_2^{(1)} - \tau_2^{(0)} \quad [\text{s}], \end{cases} \tag{6.3}$$

where $\mathring{\underline{\mathbf{x}}}_i$ denotes the position of the image of the microphone at position $\underline{\mathbf{x}}_i$ with respect to the reflector. Note that, since $\tau_{\texttt{TDOE},2} = \tau_{\texttt{iTDOA}} + \tau_{\texttt{TDOE},1} - \tau_{\texttt{TDOA}}$, only the first three will be considered. These three quantities are directly connected to the early parts of the RIRs, as illustrated in Figure 6.2. Let $V = [\tau_{\texttt{TDOA}}, \tau_{\texttt{iTDOA}}, \tau_{\texttt{TDOE}_1}] \in \mathbb{R}^3$ be the vector of parameters of interest. Since it will be useful later, let define $\mathcal{V} = \{\texttt{TDOA}, \texttt{iTDOA}, \texttt{TDOE}_1\}$ the set of three **indexes** denoted by TDOA, Image TDOA (iTDOA), Time Difference of Echoes (TDOE).

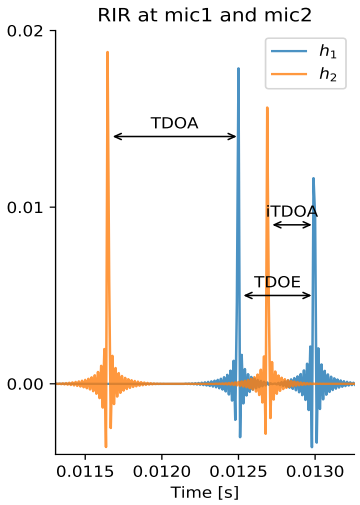


FIGURE 6.2: Superposition of two synthetic RIRs and visualization of time differences of arrival between direct paths (TDOA), first echoes (iTDOA) and direct path and first echo (TDOE).

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

While, given the early part of the RTF, the V is uniquely defined, the opposite is not. Therefore, in general, different values of V can yield the same ReTF and hence the same ILD/IPD features. In particular, this happens when $\tau_1^{(1)} - \tau_1^{(0)} = \tau_2^{(1)} - \tau_2^{(0)}$ for $r = \{0, 1\}$. That is, TDOEs are equal, or equivalently, the TDOA is equal to the iTDOA. Geometrically, this correspond to the case where real and image source have the same angle of arrival, which occur in very specific source / array / reflector configurations. Since this degenerate case is fundamentally unsolvable in the blind setting, we preventively pruned, all the entries with $|\tau_{\texttt{TDOA}} - \tau_{\texttt{iTDOA}}| < 10^{-6}$ from all the dataset. In particular, they are removed from the training set so that the training is not affected by these degenerate cases. Moreover, these are degenerate situations are known to be non-solvable and are not included in the evaluation to not bias the metrics.

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,

$$\mathcal{L}_\theta(V, \hat{V}) = \frac{1}{|\mathcal{V}|} \sum_{\xi \in \Xi} \sum_{v \in \mathcal{V}} |\tau_{v,\xi} - \hat{\tau}_{v,\xi}|^2 \tag{6.4}$$

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

The normalized Root Mean Squared Error (nRMSE)[51]is taken as validation metric to assess the quality of the estimation $\hat{V}$. Here we consider the vectors $\boldsymbol{\tau}_v, \hat{\boldsymbol{\tau}}_v \in \mathbb{R}^N$ containing the reference and estimated value for all the N

[51] The nRMSE takes values between 0 (perfect fit) and ∞ (bad fit). If it is equal to 1, then the prediction is no better than a constant.

samples in the set. The nRMSE is define as

$$\texttt{nRMSE}(\hat{\boldsymbol{\tau}}_v, \boldsymbol{\tau}_v) = \frac{\|\hat{\boldsymbol{\tau}}_v - \boldsymbol{\tau}_v\|_2}{\left\|\boldsymbol{\tau}_v - \texttt{avg}_\xi(\boldsymbol{\tau}_v)\right\|_2}, \tag{6.5}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm and $\texttt{avg}_\xi(\cdot)$ is the average operation over the sample ξ in the set. The network is manually tuned on a validation set to find the best combination of number of hidden layers, layer sizes and p_{do}.

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 § 2.3.2. Room lengths and widths are uniformly drawn at random in $[3, 9]$ m, and heights in $[2, 4]$ m. Random source and microphone positions are used, respecting the close-surface scenario.[52] In particular, the microphones are at most 30 cm from the close-surface, placed 10 cm from each other. The (frequency-independent) absorption coefficients of the close-surface is sampled uniformly in $]0, 0.5[$, while the ones of the other walls' are sampled in $]0.5, 1[$. The same realistic diffusion profile [Gaultier et al. 2017] is used for all surfaces. Around $90,000$ RIRs are generated this way, with a RT_{60} 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 points with 50% overlap. Finally the features are computed as in (6.2), that is taking the average over the entire duration of the sequence and yielding a vector of size $D = 1534$ for each observation.
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 speech utterances of various lengths (from 1 s to 6 s), randomly selected from the TIMIT corpus [Garofolo et al. 1993], normalized with respected to their energy.

[52] A rejection-sampling strategy was used to approximate uniform distributions.

▶ 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 the TDOA between two microphone recordings. It is known to achieve good performance in the case of broadband source signals. However, it was shown that as soon as the acoustic conditions become challenging due to strong early echoes, high reverberation level, and noise, the method performances decrease [Chen et al. 2006].

TDOA estimation errors using the proposed approach and GCC-PHAT are presented in Table 6.1. From these results, one can see that training a MLP to estimate TDOAs brings similar performances as GCC-PHAT in terms of

	Input	nRMSE TDOA	iTDOA	TDOE
MIRAGE-MPL	wn	0.18	0.28	0.25
MIRAGE-MPL	wn+n	0.68	0.69	0.89
MIRAGE-MPL	sp	0.31	0.34	0.56
MIRAGE-MPL	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 Errors (nRMSEs) for TDOA estimation using the MLP architecture.

nRMSE for clean white noise and speech signals. However, the estimation of iTDOA and TDOE seems to be a more challenging task for such a simple MLP model. For both methods, the performance are generally worse on speech signals than on white noise signals, which is due to the spectral sparsity of speech signals. However, in the noiseless scenario, it is interesting to see that even a simple architecture trained on ReTF-based features and broadband data is somewhat able to generalize to speech. Unfortunately, when some external noise is added, the performances of the methods degrade in all the cases. This is a well-known and expected behavior for GCC-PHAT, and for the MLP it suggests that noise should also be considered in the training phase, as already investigated for instance in [Deleforge et al. 2015a; Chakrabarty and Habets 2017]. Nevertheless, our results confirm the possibility of retrieving the strongest echoes from only two-microphone recordings, in the absence of noise.

6.4 Robustly learning the first echo

The above study was presented in our published work [Di Carlo et al. 2019]. To overcome the limitations of of our simple MLP-based method, we introduced the following three modifications:

- use a more elaborate architecture inspired by a state-of-the-art SSL method;
- include corrupting noise in the training data;
- use a robust loss function which is able to return a level of confidence on the estimations.

Motivated by their success, we consider a CNN architecture (see § 6.2.2). To this end, we inspire from the work of [Chakrabarty and Habets 2017; Nguyen et al. 2018] in SSL. As shown in Figure 6.3, the architecture consists of two convolutional modules made of a one-dimensional convolutional layer (Conv1D) followed by max-pooling along 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 in the correct dimension, we re-arranged the input so that the each of the features $\{\texttt{ILD}, \mathrm{Re}\{\texttt{IPD}\}, \mathrm{Im}\{\texttt{IPD}\}\}$ is considered as a channel for the Conv1D. After each layer, a dropout probability $p_{\mathrm{do}} = 0.3$ is applied to prevent overfitting.

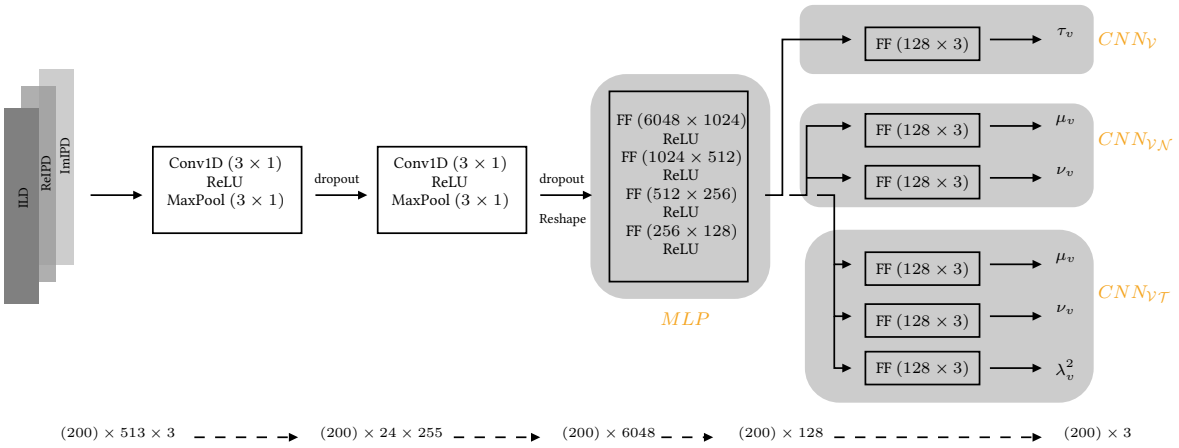


FIGURE 6.3: Architecture of the proposed deep neural network. Input dimensions for each stage are reported. The first dimension is the batch size $|\Xi| = 200$

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 mentioned earlier, our final goal is to generalize this approach to microphone arrays 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. 2019], we proposed to use predicted values and the errors on the validation set as means and variances to build a Gaussian priors (see Chapter 10 for details on this). Instead, here, we explicitly modify the CNN model to estimated these parameters. This idea recalls the Mixture Density Networks (MDNs) as presented in in [Bishop 1994], where the output of a neural network parametrizes a Gaussian mixture model. It is also the foundation of Deep Generative Networks (or generative DNNs) for which the two most striking examples are Variational Auto-Encoders (VAEs) [Kingma and Welling 2014; Rezende et al. 2014], Generative Adversarial Networks (GANs) [Goodfellow et al. 2014]. The rationale behind this design choice is to allow the learning model to assess its prediction quality. In order to achieve this, both the loss functions and the network outputs need to be modified accordingly.

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

The general form of the Gaussian probability density function is

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

$$p(\tau_v \mid \xi;\boldsymbol{\theta}) \sim \mathcal{N}\big(\tau_v; \mu_v(\xi;\boldsymbol{\theta}), \sigma_v^2(\xi;\boldsymbol{\theta})\big). \tag{6.6}$$

From now on, we omit the dependency on $(\xi;\boldsymbol{\theta})$ for the sake of clarity. Finally, the corresponding training loss function is the negative the negative log-posterior probability over the batch Ξ, that is,

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\theta}}^{\mathcal{N}}(V,\hat{V}) &= -\log \prod_{\xi\in\Xi}\prod_{v\in\mathcal{V}} p(\tau_v|\xi;\boldsymbol{\theta}) \\ &\overset{c}{=} \frac{1}{|\mathcal{V}|}\sum_{\xi\in\Xi}\sum_{v\in\mathcal{V}} \log\sigma_v^2 + \frac{|\tau_v-\mu_v|^2}{\sigma_v^2}, \end{aligned} \tag{6.7}$$

where $\overset{c}{=}$ denotes equivalence up to a constant.
Hence, for a given input ξ, the network attempts to find mean and variance parameters maximizing the *a posteriori* probability of TDOAs according to the considered Gaussian model. The means are then used as TDOA estimates, while variances can be seen as a measure of uncertainty from the model. Hereafter we denote by $\hat{V}_{\mathcal{N}}$ the set of 6 network outputs.
This approach can then be generalized to other this of distributions. For instance, we can consider a Student's t-distribution, which is more robust to outliers, due to its heavy-tailed nature. Using this distribution has already been proposed in the context of binaural SSL, e. g., in [Zohny and Chambers 2014; Deleforge and Forbes 2016].

▶ THE STUDENT'S T-BASED LOSS FUNCTION can be derived similarly as the Gaussian case assuming a t-distribution posterior on the prediction. The corresponding negative log-posterior probability over the batch Ξ now writes

$$\mathcal{L}_{\boldsymbol{\theta}}^{\mathcal{T}}(V,\hat{V}) = \sum_{\xi\in\Xi}\sum_{v\in\mathcal{V}} \frac{1}{2}\log(\nu_v\pi_v) + \frac{1}{2}\log(\lambda_v^2) - \log\Gamma\left(\frac{\nu_v+1}{2}\right) + \log\Gamma\left(\frac{\nu_v}{2}\right) + \frac{\nu_v+1}{2}\log\left(1+\frac{|\mu_v-\tau_v)|^2}{\nu_v\lambda_v^2}\right) \tag{6.8}$$

where Γ is the Gamma function. Similarly to the Gaussian case, for a given input, the network returns the 3 parameters of 3 t-distributions (μ_v,ν_v,λ_v), one for each index $v\in\mathcal{V}$. The set of 9 network outputs is denoted by $\hat{V}_{\mathcal{T}}$.

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

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

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

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 perturb the observed microphone signals with Additive White Gaussian Noise (AWGN), leading to SNR values uniformly distributed in $[0,30]$ dB. Meanwhile, the test set includes observations featuring SNR levels of 0, 10, or 20 dB. The protocol to generate the observations and the feature extraction step are the same (§ 6.3.3), but the investigation is restricted to broadband source signals (white noise) only. Informal test on speech data yielded to 3.3 times larger errors respect to noise signals and close to random. One of the main reason is due to the missing frequencies in speech spectrum. Such spectral holes lead values close to zero at both the numerator and denominator of the observation used to calculate ILD and IPD (see § 3.3.2). Alternative technique are currently investigates to overcome this limitation, e. g., robust ReTF estimator or DNN handling missing data. Finally, the network is manually tuned on a validation set in order to find the best combination of number of hidden layers and their sizes, which led to the model in Figure 6.3.
Figure 6.4 shows the performances of the proposed models $\texttt{CNN}_{\mathcal{V}}$, $\texttt{CNN}_{\mathcal{V}_{\mathcal{N}}}$ and $\texttt{CNN}_{\mathcal{V}_{\mathcal{T}}}$ with respect to $\texttt{MLP}_{\mathcal{V}}$ and the baseline $\texttt{GCC}-\texttt{PHAT}$ for the task of estimating TDOA, iTDOA and TDOE. Again we report the performances in terms of the nRMSE computed on test samples. It can be seen that the proposed modifications (CNN architecture, robust loss function, and considering noise in training) yield considerable improvement in performances, compared to

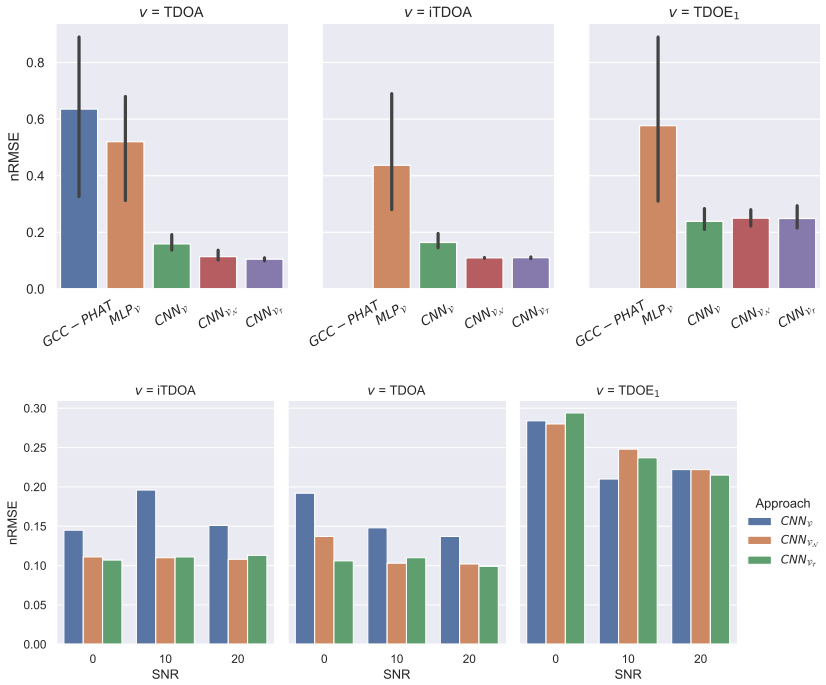


FIGURE 6.4: Normalize root mean squared error (the lower, the better) for TDOA estimation using the proposed architectures on white noise source signals.

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

the vanilla and baseline methods To give a sense to the values, a TDOA's nRMSE lower than 0.2 corresponds to an angular RMSE lower that 4 degree (more details in Chapter 10), resulting in a excellent estimation. In particular, the new methods are more robust to noise, as indicated by error bars. While both $\texttt{CNN}_{\mathcal{V}_\mathcal{N}}$ and $\texttt{CNN}_{\mathcal{V}_\mathcal{T}}$ outperform the $\texttt{CNN}_\mathcal{V}$ for TDOA and iTDOA estimation, there is no significant[53] difference between the three of them for TDOE estimation.

[53] Significance assessed with a t-test for $p < 0.05$.

Deeper insights about this can be gathered from Figure 6.5, where performances are reported for the different SNR levels considered in the test. These results suggest that using a simple MSE-based loss function leads to errors more sensitive to external noise. Moreover, they shows that TDOE estimation is a more difficult task than TDOA estimation, yielding bigger estimation errors as well as a smaller difference between the three methods. The former task is probably made more challenging. In fact, the ranges that the TDOEs span are much larger values which highly depends not only on the source and microphone relative position, but also on the surface position and orientation. In general, there does not seem to be a significant difference between using a Gaussian distribution or a Student's t-distribution at the network's output in our experiments. 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 at using deep learning models to learn the complex mapping from stereophonic recordings to echoes' timings. As part of a first investigation, we considered a simple, yet common scenario where two microphones are placed close to a reflective surface and attend a to a single sound source. The problem is then restricted to estimating the the

relative delays between the direct path and the first echo timings within the two channels. Moreover, this simplification is motivated by further application to echo-aware SSL, which will be proposed and tested in Chapter 10.
We studied two standard DNN architectures, which are trained in a *virtually supervised* fashion on data generated by an acoustic simulator. Finally, we discuss three ways to improving such performances, i. e., training on noisy data, using a CNN-based architecture, and using robust loss functions.
The investigation so far was limited to a microphone pair. Taking inspiration from data-fusion approaches, in Chapter 10, we will propose to aggregate the contributions of real and image microphone pairs to preform 2D SSL with only two microphones.

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

- The first one would be a generalization to more reflections, which is far from being trivial. Informal investigations spanned different training paradigms, such as, *curriculum learning* [Bengio et al. 2009], *teacher-student learning* [Tu et al. 2019], where the network is iterative trained to estimate more and more echoes. Another idea would be to use a collection of DNNs, each pre-trained for single echo estimation.

- Due to severe drops in performance observed using speech signals, we could consider more robust audio features (e. g., ReTF using state-of-the-art ReTF estimators) or DNN architectures designed to handle missing entries in the input feature. This latter direction was investigated in [Deleforge et al. 2015a] to make a GMM-based model able to generalize to speech data. However the best author knowledge, it is still an open problem which could be connected to the so-called (audio) "inpainting" inverse problem [Bilen et al. 2015].

- Finally, another interesting direction would be to use physics-driven neural networks as explored in [Karpatne et al. 2017; Nabian and Meidani 2020; Rao et al. 2020; Jin et al. 2020], In these work physically-motivated regularizers or entire loss-functions obeying physical laws are used to ensure the correctness of the solution or restrict the search space. It can be done by adding simple formulas, such as the temperature-density relation for fluids as in [Karpatne et al. 2017], or entire PDEs as in the other cited works.

7

dEchorate: Datasets for Acoustic Echo Estimation

"Once you've been alone
And everybody knows
Work begins alone"

—Strapping Young Lad, *Monument*

Keywords: Room impulse response, early reflection, acoustic echoes, audio database, microphone arrays.
Resources:

- Code repository
- Dataset data

▶ 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 Chapter 11 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 in 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. Conversely, datasets for RooGE typically features design choices which are not suitable for many SE application, involving the positioning of microphone(s) and source(s). `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

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

different reverberation levels sensed by uniform linear arrays (ULAs); and `MIRaGE` [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 m $\times$ 6 m $\times$ 2.4 m. The 6 facets of the room (walls, ceiling, floor) are covered by acoustic panels allowing controllable reverberation time ($\texttt{RT}_{60}$). 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 $[4, 5, 7.5, 10]$ cm[54]. Each array is steered towards a different vertical edge of the room for calibration and reproducibility purposes.

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

Loudspeakers	(directional, direct) 4× Avantone Pro Mixcube (directional, indirect) 2× Avantone Pro Mixcube (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). Theses 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

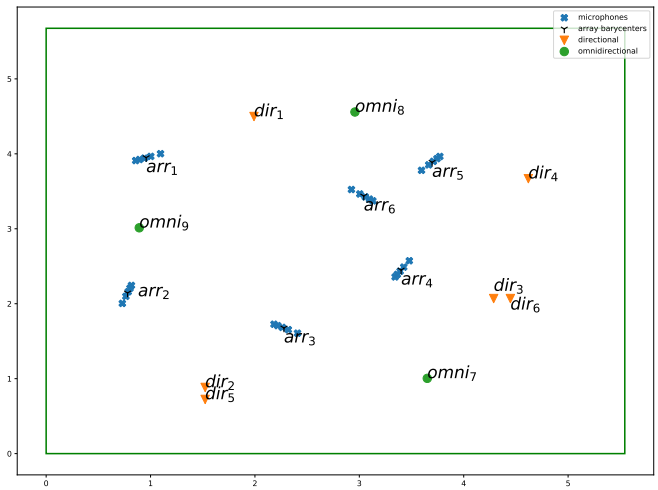


FIGURE 7.2: Illustration of the recording setup - top view.

(a) Overall setup

(b) Microphone array

(c) Revolving panels

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
one-hot	000000	✗	✗	✗	✗	✗	✗
	010000	✗	✓	✗	✗	✗	✗
	001000	✗	✗	✓	✗	✗	✗
	...			...			
	000001	✗	✗	✗	✗	✗	✓
incremental	011000	✗	✓	✓	✗	✗	✗
	011100	✗	✓	✓	✓	✗	✗
	...			...			
	011111	✗	✓	✓	✓	✓	✓

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

and "popping" effects, the signals were linearly faded in and out over 0.2 seconds with a Tuckey taper window[55]. 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.

[55] The code to generate the reference signals and to process them is available togheter 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] at 48 kHz: 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 off-grid 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 position 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.

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

(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 GUI[56] (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 2016; 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 Chapter 11.

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

[56] This GUI is available in the dataset package.

	Metrics	bIPS	dMDS	dcMDS
Geom.	Max.	0	6.1	1.07
	Avg.±Std.	0	1.8 ± 1.4	0.39 ± 0.2
Signal	Max.	5.86	1.20	1.86
	Avg.±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%

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.

in average) in both the signal *and* geometric spaces with 98.1% of matching first order echoes 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 echograms[57], that is $\mathbf{H}_{l,n} = \bar{\eta}_n(l) = |h_n(l)| / \max |h_n(l)|$, where $l = 0, \ldots, L-1$ 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.

[57] The echogram is defined either as the absolute value or as the squared value of the RIR.

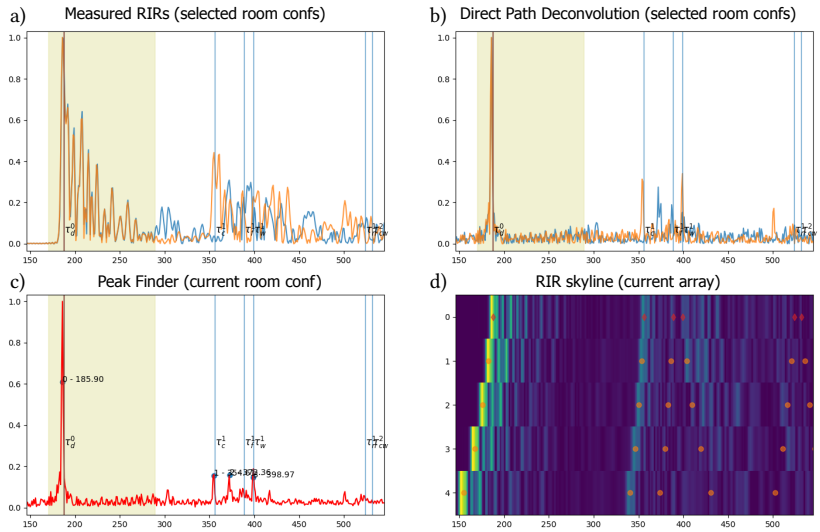


FIGURE 7.5: Detail of the GUI used to manually annotate the RIRs. For a given source and microphone, a) and b) shows 2 RIRs for 2 different room walls configuration (blue and orange) before and after the direct path deconvolution respectively. c) shows the results of the peak finder for one of the deconvolved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

- DIRECT PATH DECONVOLUTION/EQUALIZATION was used to compensate the frequency response of the source loudspeaker and microphone [Antonacci et al.

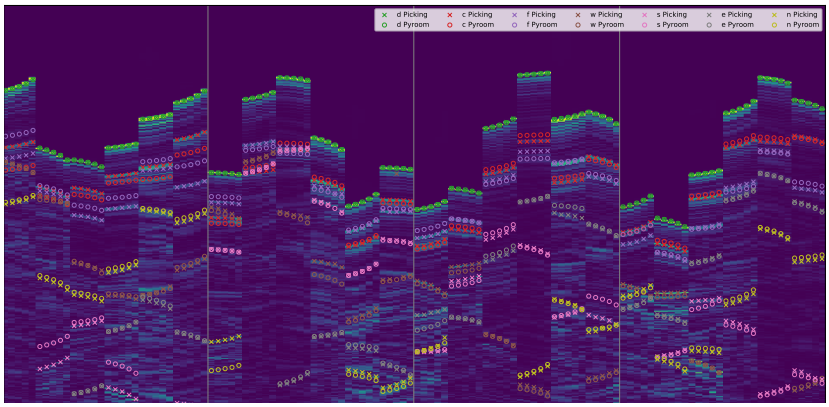


FIGURE 7.6: RIR `Skyline` annotated with observed peaks (×) together with their geometrically-expected position (○) computed with the `Pyroomacoustic` simulator. As specified in the legend, different colors are used to indicate the room facets responsible for the reflection: direct path (`d`), ceiling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

2012; Eaton et al. 2016]. In particular, the direct path of the RIR was manually isolated and used 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 FINDER[58] was used on equalized echograms $\bar{\eta}_n(l)$ to provide an initial guess on the peak positions.

[58] 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 2$^{\text{nd}}$ 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
1	1	chirp	False	...	2020-01-22__22-50-36.wav
1	1	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 study. 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.
Future works will explore different directions. By such as 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. Moreover, the dataset could be updated by including more robust annotations derived from more advanced algorithms for calibration and AER.

Part III

ECHO-AWARE APPLICATION

8

Audio Scene Analysis meets Signal Processing

▶ Synopsis In this chapter, we present a selection of algorithms and methodologies which we identified as potential beneficiaries of echo-aware additions. At first, we present a typical scenario that highlights some cardinal problems. Then, state-of-the-art approaches to address these problems are listed and commented in dedicated sections, highlighting the relationship with some acoustic propagation models, respectively. 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 introduction gathers the common knowledge shared across the following chapters. Here we make a strong transition: we assume the echo properties are known a priori, so that our focus is only 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 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 in Figure 8.1. As an indoor scenario, all the elements of reverberation (in particular echoes) are present. Diffuse background noise

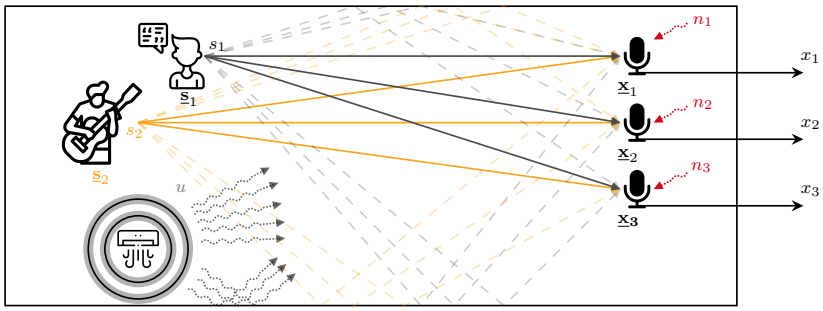


FIGURE 8.1: Illustration of an indoor audio scene recorded by three microphones. The microphone signals capture the reverberant mixture of several sound sources such as speech, music (guitar) and background diffuse noise (an air conditioner).

is present as 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 source and microphone as well-defined points with coordinate $\underline{\mathbf{s}}$ and $\underline{\mathbf{x}}$, 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

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

or alternatively, using the source spatial image signals,

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

Note that the filter $h_{ij}(\underline{\mathbf{x}}_i|\underline{\mathbf{s}}_j)$ denotes the RIR where we intentionally highlight the dependencies on geometry, namely, accounting for the whole sound propagation for the source position $\underline{\mathbf{s}}_j$ to the microphone position $\underline{\mathbf{x}}_i$. In fact, as discussed throughout Part I, we can decouple the information of indoor microphone natural recordings into two complementary sub-problems: the estimation of the RIRs (thus the mixing matrix) that account for only the sound propagation, and of the source signals accounting only the semantic content. The former problem can be seen as an instance a more general the Blind Channel Estimation (BCE) problem where the (acoustic) channel is the RIR (see the review in Chapter 4).

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

Audio scene analysis problems	*from the mixtures* $\{x_i\}_i$*, can we estimate...*	Chapter
Audio Source Separation	the source signals $\{s_j\}_j$, or the spatial images $\{c_{ij}\}_{ij}$?	Chapter 9
Spatial filtering	the source signals $\{s_j\}_j$, knowing the filters $\{h_{ij}(\underline{\mathbf{x}}_i\|\underline{\mathbf{s}}_j)\}_{ij}$?	Chapter 11 § 11.1
Sound Source Localization	the source positions $\{\underline{\mathbf{s}}_j\}_j$?	Chapter 10
Room Geometry Estimation	the shape of the room?	Chapter 11 § 11.2

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

exact time of emission of the source signal is known, as well as the source 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 Audio Source Separation

Multichannel Audio Source Separation (MASS) 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 processing. 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. Both share many methods, which are accordingly modified, taking into account scenarios and applications. In the context of multichannel speech recordings, some of the most successful and popular methods used nowadays include [Vincent et al. 2011]:

Many publications have addressed this issue, including books [Vincent et al. 2018; Makino 2018], and overview papers [Cardoso 1998; O'grady et al. 2005; Gannot et al. 2017; Girin et al. 2019].

- spatial filtering, which uses spatial information only;
- TF masking, which uses spectral information only;
- Multichannel Wiener Filter (MWF) which combines the above two;
- and end-to-end regression which use powerful end-to-end (i. e., from waveforms to waveform) learning models (e. g., DNN).

In this thesis, we deliberately distinguish between 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 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

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

where X_i is the STFT of the microphone channel. Here the dependency on the TF bin is omitted for clarity.
In order to be computed, the Wiener filter requires estimating source spatial images statistics, in particular an estimate of the power in each TF bin (and discarding the phase). In these thesis we stress the difference between source separation and spatial filtering. In the former, source signals and mixing filters statistics are needed to weigh each of the TF bins of the microphone channel's STFT. As opposed to, in the latter problem, the "mask" of one of the sources, 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, e. g. in the case of MWF [Ozerov and Févotte 2010]. 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 model. 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 Multichannel NMF [Ozerov and Févotte 2010] is one of the most successful framework for source separation using MWF, that is, combining spatial filtering and TF masking. It uses the NMF model for modeling the source Power Spectral Density (PSD) within a convolutive mixing model (i. e., spatial model, as discussed in § 3.2.5). It then deploys optimization-based framework for estimating both the mixing matrix and an estimate of the source PSD based on NMF. Once these PSD and filter parameters are estimated, separation can be achieved with MWF. One of the main advantage of this approach is that it allows to easily incorporated prior knowledge on the problem. Thanks to the narrowband approximation (Eq. (3.26)), estimation of spatial and semantic content are two complementary sub-problems that can be solved independently. 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, ReTFs, etc.

Here, we will focus on multichannel source separation based on Nonnegative Matrix Factorization (NMF). NMF refers to a set of technique to model spectra of complex sounds by a sum of basic components. Modeling sound structure is beneficial for source separation, since it make separation possible in many challenging scenario. 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.

However, the work in [Luo and Mesgarani 2019] showed that even with oracle TF masks, 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; Défossez et al. 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 prove 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 is hard to inject prior knowledge. Instead, Multichannel NMF-based frameworks accounts for this freedom.

"Classic" (pre-deep learning) MASS spans almost 25 years (1990-2015) and hundreds of MASS techniques have been proposed.[59] With extreme simplification,

[59] The reader can refer to the review papers [Gannot et al. 2017; Girin et al. 2019] for more details.)

one can be broadly grouped such techniques according to how they model sound propagation of the mixing process:

- (*no propagation*) those that simply ignore it, e. g. [Kokkinis et al. 2011; Le Roux et al. 2015; Di Carlo et al. 2017];
- (*free field propagation*) those that assume a single anechoic path, e. g. [Rickard 2007; Nesta and Omologo 2012; Higuchi et al. 2014] ;
- (*reverberant propagation*) those that model the sound propagation path entirely, e. g. [Ozerov and Févotte 2010; Duong et al. 2010; Li et al. 2019a];
- (*reverberant propagation*) and those that attempt to separately estimate the contribution of the early echoes and the contribution of the late tail, e. g. [Leglaive 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 source over multiple TF bins, for which the assignation is harder. When computing the TF masking operation, these bins may introduce strong artifacts. In the second case, the algorithm need to estimated 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 Chapter 9, where a new method for speech source separation based on the Multichannel NMF framework and echoes is described.

8.3 OVERVIEW ON SPATIAL FILTERING

For a comprehensive review on spatial filtering methods, the reader can refer to the book [Van Trees 2004].

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 a vast literature including several reference books dedicated to the topic. For more details in this direction, the reader can refer to, e.g., the book [Van Trees 2004]. In audio, this topic has been extensively reviewed in the context of speech enhancement in a recent publication [Gannot et al. 2017][60].

[60] The content of this work has been extended in the book [Vincent et al. 2018].

In spatial filtering, the RIRs (and related models, e. g., RTFs, steering vectors or ReTFs) 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].

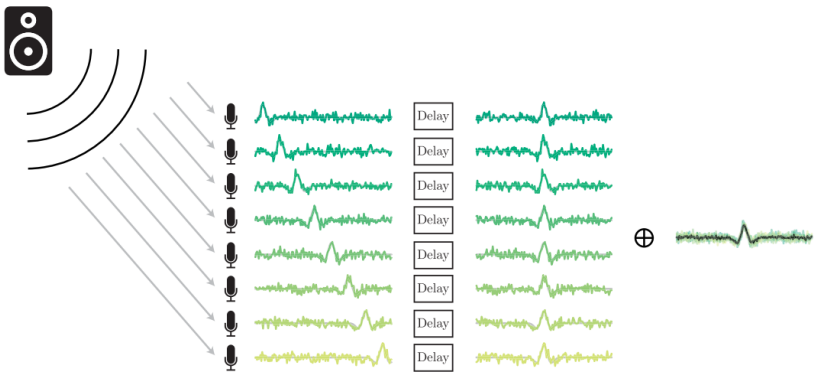


FIGURE 8.2: Illustration of the DS Beamforming in the time-domain applied to a pulse source signal. The delays of the recorded signals at each microphone are compensated. By summing the shifted signals, the source signal is enhanced. Image courtesy of Robin Scheibler, author of [Scheibler et al. 2015].

8.3.1 *Beamforming*

Beamforming is one of the most famous techniques used in spatial filtering, and the two term are almost equivalent. One of the most famous beamformer is the Delay-and-Sum (DS) beamformer. 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 TF 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 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 receiver* [Flanagan et al. 1993; Jan et al. 1995; Dokmanić et al. 2015; Peled and Rafaely 2013; Scheibler et al. 2015; Kowalczyk 2019];
- (*reverberant propagation*) methods based on BCE considering the RIR as the acoustic channel (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, DNNs have been proposed for solving this task, either to estimate the beamformer filter [Li et al. 2016a; Xiao et al. 2016; Sainath et al. 2017] or in an end-to-end task [Ernst et al. 2018]. Moreover, DNNs have been used to estimate some of the parameters, such as the DOAs [Salvati et al. 2018; Chazan et al. 2019] or speech activity detection for 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

The reader can find more details is SSL in the recent review articles [Rascon and Meza 2017; Argentieri et al. 2015] as well as in [Vincent et al. 2018, Chapter 4].

Most of Sound Source Localization (SSL) methods consist in determining the distance and direction 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 2016]. 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-Direction of Arrival (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 is 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:

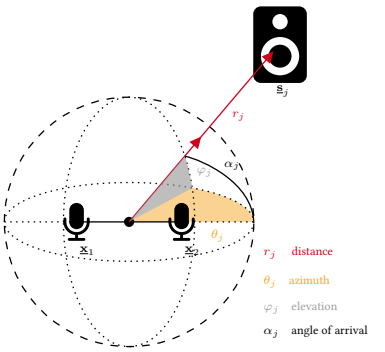


FIGURE 8.3: Geometrical illustration of the Sound Source Localization, showing the azimuth θ_j, the elevation φ_j and angle of arrival α_j for the source j at $\underline{\mathbf{s}}_j$ and two microphones at $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2$ respectively.

- 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 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.

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.

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 (TDOAs) 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 [Knapp and Carter 1976; Blandin et al. 2012; 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 [DiBiase et al. 2001], 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]. Lot of recent work has been dedicated to make SSL methods more robust to noise and reverberation even in challenging scenario of multiple moving speakers, such as [Li et al. 2019b; Kitic and Guérin 2018]. The interested reader is invited to see the results of the recent LOCATA challenge [Evers et al. 2020].

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. 2015a; 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. 2015a; Gaultier et al. 2017], GMMs-based models were used to learn the mapping from features derived from the ReTF of pair of microphones. The work in [Nguyen et al. 2018] compare GMM- and DNN-based approach for SSL for the interesting case of a robot gathering training examples. In [Vesperini et al. 2018], the author proposes to use DNN 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

"*If you dont pay attention, your learing methods risk to* learn a model from model"
—Zybnek Koldowsky

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 simplistic, with the consequence that the models may not be able to generalize to real-world conditions.

- REVERBERATION AND, IN PARTICULAR, ACOUSTIC ECHOES are typically treated as nuisance 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.

* 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 using only two microphones,
  - and in Chapter 11, we will discuss some applications of the `dEchorate` dataset (Chapter 7) for some 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 Multichannel Audio Source Separation (MASS) methods. Existing methods typically 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 echo properties 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 based on the work of [Ozerov and Févotte 2010]: 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.

"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"

The material presented in this chapter results from a collaboration with Robin Scheibler and Ivan Dokmanić and was published in [Scheibler et al. 2018c]. 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 NMF method accounting for echoes and using pre-trained dictionaries.

9.1 Literature review in echo-aware Multichannel Audio Source Separation

The scientific literature on MASS is vast, still active, and spans decades. For a brief introduction, the reader can refer to § 8.2 and to recent books [Vincent et al. 2018; Makino 2018]. 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.

Here, we will focus on Nonnegative Matrix Factorization (NMF)-based MASS, which allows to easily incorporate side information on both the sources and the acoustic propagation. This is a key feature, as opposed to the end-to-end

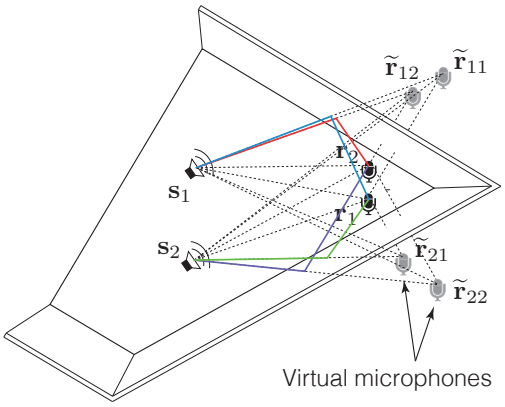


FIGURE 9.1: Typical setup with two speakers recorded by two microphones. The illustration shows the virtual microphone model (grey microphones) with direct sound path (dotted lines) and resulting first-order echoes (colored lines).
Image taken from our paper [Scheibler et al. 2018c].

Deep Neural Network (DNN)-model. These latter models are very popular nowadays and are able to reach impressive performances in the general case. However, when it comes include side information or adapt to specific tasks, it is far from trivial.

To the best author knowledge, in the literature only few works can be found that incorporate the knowledge of echoes into sound source separation. 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 $\mathtt{RT}_{60}$.
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 [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 [Bergamo et al. 2004; Korhonen 2008]. 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*[61] previously proposed in [Dokmanić et al. 2015] and is thus dubbed `Separake`. This model allows for a simple and intuitive way to structure the mixing matrix so that less parameters need to be estimated.

[61] Rake receivers are a particular type of beamformers accounting for the effect of multipath propagation (see § 11.1)

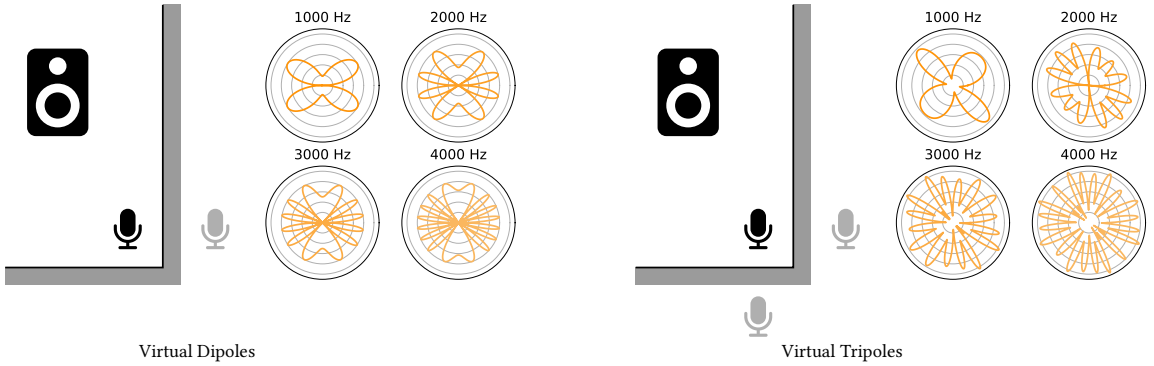


FIGURE 9.2: The frequency-dependent directivity pattern for a virtual array built with a real and one (left) and two (right) image microphones. Image taken from the presentation of the paper [Scheibler et al. 2018c].

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. 2015]. 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. Note that the signal virtual microphones are not accessible directly, as the observations will remain the mixture of convolutive sound at the real microphones. However, such formulation rather offer a useful interpretation and a way to model incoming reverberant signals.

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

$$\tilde{H}_{ij}(f) = \sum_{r=0}^{R} \alpha_{ij}^{(r)} \mathrm{e}^{-\mathrm{i}2\pi f \tau_{ij}^{(r)}}. \tag{9.1}$$

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 of a reference microphone to zero. 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 all these 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

$$X_i[k,l] = \sum_{j=1}^{J} H_{ij}[k] S_j[k,l] + N_i[k,l] \tag{9.2}$$

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,

$$\boldsymbol{X}[k,l] = \mathbf{H}[k]\boldsymbol{S}[k,l] + \boldsymbol{N}[k,l], \tag{9.3}$$

where $\boldsymbol{X}[k,l], \boldsymbol{N}[k,l] \in \mathbb{C}^{I\times 1}$, $\boldsymbol{S}[k,l] \in \mathbb{C}^{J\times 1}$ and $\mathbf{H}[k,l] \in \mathbb{C}^{I\times J}$.

Let the squared magnitude of the spectrogram of the j-th source be $\mathbf{P}_j = \left[|S_j|^2\right]_{kl} \in \mathbb{R}^{F\times T}$. As depicted in Figure 9.3, the spectrogram can be modeled as the product of 2 non-negative matrices:

$$\mathbf{P}_j = \mathbf{D}_j \mathbf{Z}_j, \tag{9.4}$$

where $\mathbf{D}_j$ is the non-negative *dictionary* whose columns are called *atoms* and can be interpreted as interpreted as spectral templates of the source, while the latent variables $\mathbf{Z}_j$, called *activations*, indicating when and how this templates are activated.

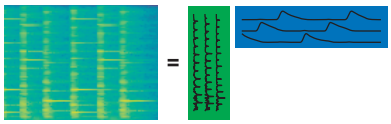


FIGURE 9.3: Spectrogram of a sound source signal decomposed into dictionary and activation. Image taken from the slides accompanying the paper of this work [Scheibler et al. 2018c].

9.3 MULTICHANNEL AUDIO SOURCE SEPARATION BY NMF

NMF-based Multichannel Audio Source Separation (MASS) can then be cast as an inference problem in which we minimize the error between the observed $\boldsymbol{X}$ 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. As introduced in § 8.2, the NMF-based MASS consists in two steps: the estimation of source's Power Spectral Density (PSD) via NMF and filters and the actual separation via Wiener Filter.

Here 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. The reader is invite to refer 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 appear in [Ozerov and Févotte 2010]. The observed multi-channel squared magnitude spectra are denoted $\mathbf{V}_i = \left[|X_i[k,l]|^2\right]_{kl}$ and their non-negative factorizations

$$\widehat{\mathbf{V}}_i = \sum_{j=1}^{J} \mathrm{diag}(\mathbf{Q}_{ij})\mathbf{D}_j\mathbf{Z}_j, \quad i = 1, \ldots, I \tag{9.5}$$

where $\mathbf{Q}_{ij} = \left[|H_{ij}[k]|^2\right]_k$ is the vector of squared magnitudes of the approximate RTF between microphone i and source j.

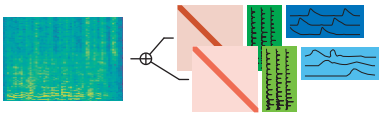


FIGURE 9.4: Schematics of the signal model used for MU-NMF. Image taken from the slides accompanying the paper of this work [Scheibler et al. 2018c].

- THE MU RULES MINIMIZE the *divergence*[62] 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

$$\mathcal{C}_{\text{MU}}(\Theta_{\text{MU}}) = \sum_{jkl} \mathcal{D}_{\text{IS}}(\mathbf{V}_i[k,l]|\widehat{\mathbf{V}}_i[k,l]) + \gamma \sum_j \|\mathbf{Z}_j\|_1, \tag{9.6}$$

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, we obtain:

$$\mathbf{Q}_{ij} \leftarrow \mathbf{Q}_{ij} \odot \frac{\left[\widehat{\mathbf{V}}_j^{-2} \odot \mathbf{V}_j \odot (\mathbf{Z}_j\mathbf{D}_j)\right]\mathbf{1}_{1\times T}}{\left[\widehat{\mathbf{V}}_j^{-1} \odot (\mathbf{Z}_j\mathbf{D}_j)\right]\mathbf{1}_{1\times T}} \tag{9.7}$$

$$\mathbf{Z}_j \leftarrow \mathbf{Z}_j \odot \frac{\sum_i (\mathrm{diag}(\mathbf{Q}_{ij})\mathbf{D}_j)^\intercal \left(\mathbf{V}_j \odot \widehat{\mathbf{V}}_j^{-2}\right)}{\sum_i (\mathrm{diag}(\mathbf{Q}_{ij})\mathbf{D}_j)^\intercal \widehat{\mathbf{V}}_j^{-1} + \gamma}, \tag{9.8}$$

$$\mathbf{D}_j \leftarrow \mathbf{D}_j \odot \frac{\sum_i \mathrm{diag}(\mathbf{Q}_{ij})^\intercal \left(\mathbf{V}_j \odot \widehat{\mathbf{V}}_j^{-2}\right)\mathbf{Z}_j{}^\intercal}{\sum_i \mathrm{diag}(\mathbf{Q}_{ij})^\intercal \widehat{\mathbf{V}}_j^{-1}\mathbf{Z}_j{}^\intercal}, \tag{9.9}$$

where multiplication $\odot$, power, and division are element-wise and $\mathbf{1}_{1\times T}$ is a vector of T ones.

9.3.2 *NMF using Expectation Maximization (EM-NMF)*

Unlike the MU algorithm that independently maximizes the log-likelihood of mixture spectral magnitudes over individual channels, the EM-NMF maximizes

[62] The divergence is an asimmetrical "distance" measure. The Itakura–Saito diverge was proposed to account for perceptum (dis)similarity between spectra. Other type of divergence are used in Audio Source Separation, such as the Kullback-Leibler divergence.

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

$$S_j[k,l] \sim \mathcal{N}_c(0, (\mathbf{D}_j \mathbf{Z}_j)_{kl}), \tag{9.10}$$

and the magnitude spectrum $\mathbf{P}_j$ of (9.4) can be understood as the variance of source j. This underlying statistical model for the source signal can be used (with little changes) in the MU-NMF approach. However, the MU-based approach can be applied deterministically, without any statistical assumption on the latent variables.
Under this model, and assuming uncorrelated noise, the microphone signals also follow a complex Gaussian distribution with covariance matrix

$$\mathbf{\Sigma}_{\boldsymbol{X}}[k,l] = \mathbf{H}[k]\mathbf{\Sigma}_{\boldsymbol{S}}[k,l]\mathbf{H}^{\mathsf{H}}[k] + \mathbf{\Sigma}_{\boldsymbol{N}}[k,l], \tag{9.11}$$

where $\mathbf{\Sigma}_{\boldsymbol{S}}$ and $\mathbf{\Sigma}_{\boldsymbol{N}}$ 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,

$$\mathcal{C}_{\text{EM}}(\Theta_{\text{EM}}) = \sum_{kl} \operatorname{trace}\left(\boldsymbol{X}[k,l]\boldsymbol{X}[k,l]^{\mathsf{H}}\mathbf{\Sigma}_{\boldsymbol{X}}^{-1}[k,l]\right) + \log\det \mathbf{\Sigma}_{\boldsymbol{X}}[k,l]. \tag{9.12}$$

where the $\Theta_{\text{EM}} = \left\{\mathbf{H}, \{\mathbf{D}_j, \mathbf{Z}_j\}_j, \mathbf{\Sigma}_{\boldsymbol{N}}\right\}$ 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 likelihood[63] 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.

[63] The complete data log-likelihood includes both observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{S}$

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 is 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 algorithm[64]. 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*.

[64] 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 Audio Source Separation

To evaluate the usefulness of echoes in source separation, we modified the 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,

$$\begin{aligned} H_{ij}[k] &= \sum_{r=0}^{R} \alpha_{ij}^{(r)} \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_{ij}^{(r)}}, \\ \mathbf{H}[k] &= [H_{ij}[k]]_{ij}, \end{aligned} \tag{9.13}$$

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 far-field anechoic regime) leads to a constant and equal $\mathbf{Q}_{ij}$ for all j and i. A consequence is that, if we also assume a unique, universal dictionary, namely, $\mathbf{D} = \mathbf{D}_j \ \forall j$, then MU-NMF framework can be simplified.
Indeed, (9.5) becomes the same for all i,

$$\widehat{\mathbf{V}}_i = \sum_j \mathbf{D}\mathbf{Z}_j = \mathbf{D}\sum_j \mathbf{Z}_j,$$

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 the results.

9.5.1 *Setup*

An array of three microphones arranged on the corners of an equilateral triangle with edge length $0.3\,\mathrm{m}$ is placed in the corner of a 3D room with 7 walls. We select 40 sources at random locations at a distance ranging from $2.5\,\mathrm{m}$ to $4\,\mathrm{m}$ from the microphone array. Pairs of sources are chosen so that they are at least $1\,\mathrm{m}$ 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.

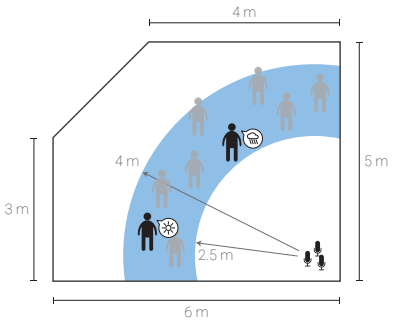


FIGURE 9.5: The simulated scenario. Image taken from the paper of this work [Scheibler et al. 2018c]

			Number of echoes R						
	anechoic	learn	0	1	2	3	4	5	6
$\gamma =$	10	10^{-1}	10	10^{-3}	0	0	0	0	0

TABLE 9.1: Value of the regularization parameter γ used with the universal dictionary.

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.[65] 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.

[65] *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.*

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, thus, without echoes in the mixture.
2. *learn*: the RTFs are learned from the data together with 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 [Garofolo et al. 1993] using the NMF solver in `scikit-learn` Python package [Pedregosa et al. 2011]. Figure 9.6 illustrates the two different dictionaries used in this work.

- 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 $\mathbf{D} = [\mathbf{D}_1^{\mathtt{M}} \cdots \mathbf{D}_{25}^{\mathtt{M}}\, \mathbf{D}_1^{\mathtt{F}} \cdots \mathbf{D}_{25}^{\mathtt{F}}]$. 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 $\mathbf{D} \in \mathbb{R}^{1025 \times 500}$.

(a) Universal dictionary (b) Speaker-specific dictionary

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. 2018c].

- 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 implementation[66] 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:

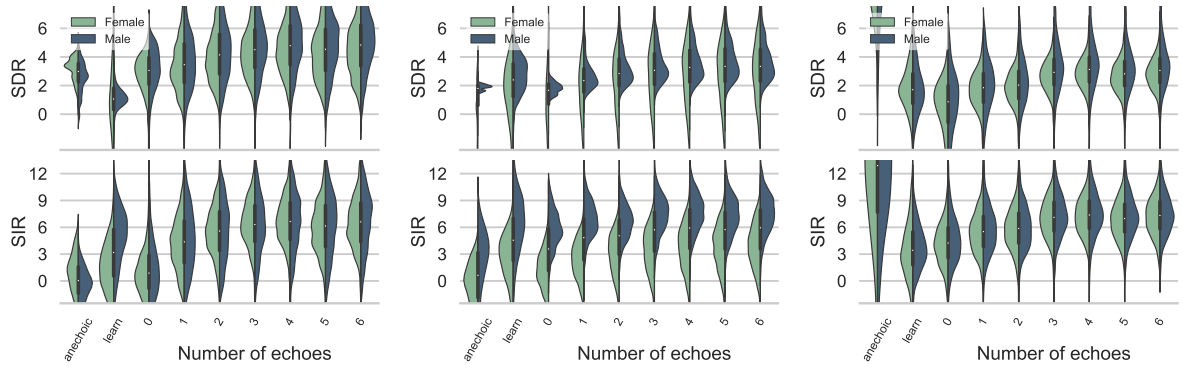
[66] Multichannel nonnegative matrix factorization toolbox (in Matlab)

- 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};
- separation is achieved via MWF as described in [Ozerov and Févotte 2010].
- the code was further modified to deal with fixed dictionary and channel model matrices, which are normalized in order to avoid indeterminacy issues [Ozerov and Févotte 2010].

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 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].



(a) MU-NMF, Universal dictionary (b) MU-NMF, Speaker-specific dictionary (c) EM-NMF, Speaker-specific dictionary

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.

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.

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).

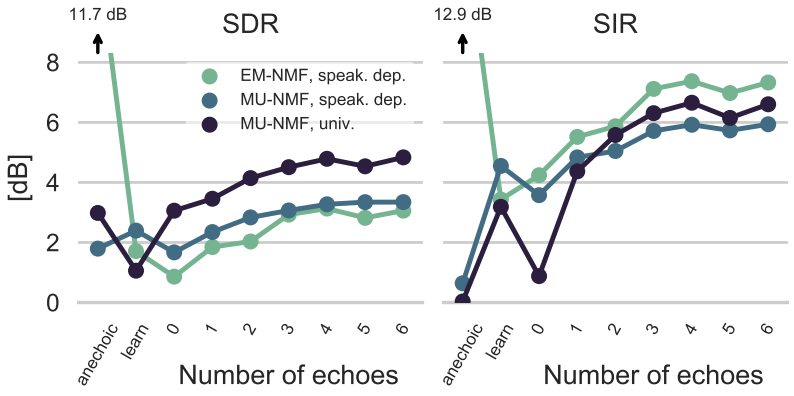


FIGURE 9.8: Summary of the median SDR and SIR for the different algorithms evaluated.

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 a 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 shows 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 In this chapter, we show that early-echo characteristics can, in fact, benefit SSL. To this end, we introduce the concept of Microphone Augmentation with Echoes (MIRAGE), based on the image microphone model. 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 anechoic propagation.

Together with Chapter 6, this chapter describes methods and results published in [Di Carlo et al. 2019], which considers only stereophonic recordings. In this sense, this chapter provides an application for LANTERN, the learning-based echo estimation method presented in the above mentioned chapter.

"Scanning all around [...] Ancient blocks of sound Mirage"

—Meat Puppets, *Mirage*

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"

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 reverberation. It is typical to observe that Direction of Arrival (DOA) estimation degrades with increasing acoustic reflections [Chen et al. 2006]. For these reasons, most sound source localization methods regard reverberation, and in particular acoustic echoes, as a nuisance. Some prior works, e. g., [Rui and Florencio 2004; Chen et al. 2006; Zhang et al. 2007], modeled the reverberation as a noise term. However, the commonly used dereverberation methods do not reduce strong early echoes. Alternatively, approaches in [Weinstein et al. 1994; Taghizadeh et al. 2015; Salvati et al. 2016] attempt to solve SSL by estimating the full RIRs (or ReTFs [Li et al. 2016b]), which is known to be a very difficult task.

Instead, echo-aware sound source localization methods take another direction: they exploit the closed-form relation between echo timings and the audio scene geometry as expressed in the Image Source Method (ISM). Early works such as [Korhonen 2008; Ribeiro et al. 2010a; Ribeiro et al. 2010b; Svaizer et al. 2011] use knowledge form the room geometry to estimate the position of the sound source with respect to the array. 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 used ad-hoc reflectors as artificial *pinnae* and a simple reflection model. In the work [Kreković et al. 2016], the authors address the problem of Acoustic Source Localization and Mapping (SLAM)[67] using echoes. The authors of [An et al. 2018] leverage on cameras, depth sensors, and laser sensors to identify reflectors and build a corresponding acoustic model that is used later for localizing sources. Finally, in a very recent work, the well-known MUSIC framework is modified to account for an echo model in the spherical harmonic representation [Birnie et al. 2020]

[67] Acoustic SLAM enables the estimation of a moving robot's position in relation to a number of external acoustic sources [Evers and Naylor 2018].

All the above mentioned echo-aware methods are explicitly knowledge-driven, namely, they use closed-form solutions based on physics, acoustics, and signal processing models. The benefit of manipulating "simple" models is however paid by strong assumptions, simplifications and a need for hand-crafted features. In order to overcome these limitations, data-driven methods have been proposed to address SSL (see § 8.4). This approach leverages training datasets to implicitly learn the mapping from audio features to the source position. Such data can be obtained from real recordings or using physics-based simulators. These methods were showed to overcome some limitations of physics-based model, especially when some assumptions are violated. However, they are typically trained for specific applications and use-cases (e. g., specific array geometries, acoustic conditions, etc.) and fail whenever test conditions strongly mismatch training conditions.

10.2 Proposed approach

In this work, we propose to combine the best of the two worlds:

- to use a data-driven model to estimate sound propagation parameters;
- to use a knowledge-driven model to map such parameters to the source's DOA.

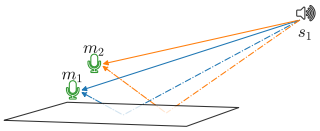


Figure 10.1: Typical setup with one sound source recorded by two microphones. The illustration shows the direct sound path (blue lines) and the resulting first-order echoes (orange lines).

Let us first introduced a simple yet common scenario: two microphones, one source, and a nearby reflective surface, as illustrated in Figure 10.1. This may occur when the sensors are placed on a table or next to a wall. Striking examples of this scenario are 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? To answer this question, we propose to use the class of deep learning models LANTERN, presented in Chapter 6. These models are trained on a close-surface dataset to estimate the first early echoes per channel 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

Microphone Augmentation with Echoes (MIRAGE) framework. It exploits echo times of arrival by expressing them as TDOAs in a *virtual 4-microphone array* formed by the true microphone pair and its image with respect to the reflective surface. This model is based on the *image microphones* model (See § 9.1). Note that answering this question is an *impossible* task in free- and far-field stereophonic conditions, due to the well known spatial ambiguity of a single TDOA, referred to as the *cone of confusion* [Bregman 1990]., that is a cone-shaped region where a source provide the same TDOA and (see Figure 10.2). For instance, this makes it impossible for humans and device with only 2 microphone to determine the elevation and the front-back position of a sound source using binaural cues alone.

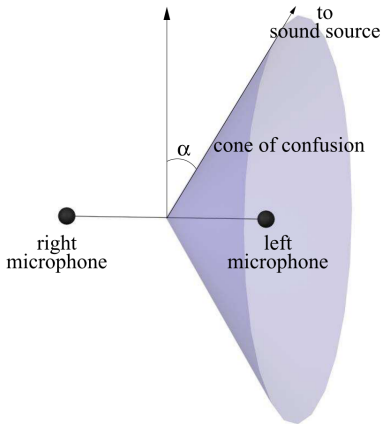


FIGURE 10.2: Illustration of the cone of confusion.

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 ($J = 1$). Recalling the signal model presented in Eq. (4.1), the relationship between the signal x_i recorded by the i-th sensor placed at fixed position $\underline{\mathbf{x}}_i$ and the signal s emitted by the source at fixed position $\underline{\mathbf{s}}$ writes

$$\begin{aligned} \tilde{x}_i(t) &= (\tilde{h}_i \star \tilde{s})(t) + \tilde{n}_i(t) \\ \tilde{h}_i(t) &= \sum_{r=0}^{R} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \tilde{\varepsilon}_i(t), \end{aligned} \tag{10.1}$$

where $\tilde{n}_i$ denotes possible measurement noise and $\tilde{\varepsilon}_i$ collects later echoes, the reverberation tail, diffusion, and undermodeling noise. In this work, we will consider only the first strongest echo, therefore $R = 1$. Note that for $r = 0$ denotes the direct propagation path, where $\tau_i^{(0)}$ is the direct path's time of arrival from the source to the i-th microphone, and $\alpha_i^{(0)}$ captures both the sound wave attenuation and air absorption effects. In the remainder of this chapter, 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 far-field assumption, traditional SSL methods use the Time Difference of Arrival (TDOA),

$$\tau_{\texttt{TDOA}} \overset{\text{def}}{=} \tau_2^{(0)} - \tau_1^{(0)} \quad \text{[second]},$$

as a proxy for the estimation of the Angle of Arrival (AOA), ϑ, since:

$$\vartheta \approx \arccos(c\, \tau_{\texttt{TDOA}} \,/\, d) \quad \text{[rad]}, \tag{10.2}$$

where c is the speed of sound and d is the inter-microphone distance, considered small compared to the source's distance. This relation is illustrated in Figure 10.3.

Then, SSL reduces to estimating the TDOA, which can be done by a Cross Correlation (CC)-based methods, e. g., the Generalized Cross Correlation with Phase Transform (GCC-PHAT) method [Knapp and Carter 1976; Blandin et al.

The "generalized" cross-correlation methods add weighting functions (e. g. the Phase Transform (PHAT), or the smoothed coherence transform (SCOT)) to the frequency-domain CC. Their purpose is to improve the estimation of the time delay depending on specific characteristics of the signal and noise. See [Chen et al. 2006] for an overview.

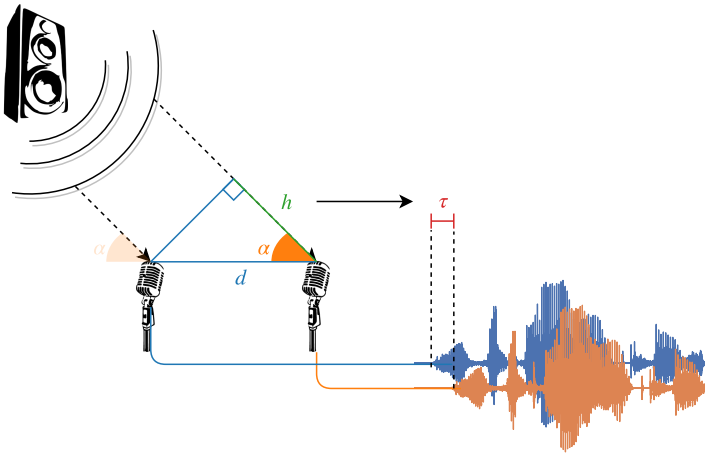


FIGURE 10.3: Illustration of the relation between DOA and TDOA with ones source and two microphones. Knowing the distance d between the two microphones, simple trigonometry yields the AOA ϑ according to Eq. (10.2).

2012]. Given the STFT X_1 and X_2 of the two microphones signals, the CC and GCC-PHAT *angular spectra* are defined as:

$$\Psi_{\texttt{CC}}(\tau) = \sum_{k,l} X_1[k,l] X_2^*[k,l] \mathrm{e}^{-\mathrm{i}2\pi f_k \tau}, \tag{10.3}$$

$$\Psi_{\texttt{PHAT}}(\tau) = \sum_{k,l} \frac{X_1[k,l] X_2^*[k,l]}{|X_1[k,l] X_2^*[k,l]|} \mathrm{e}^{-\mathrm{i}2\pi f_k \tau}, \tag{10.4}$$

where $*$ denotes the complex conjugate, $[k,l]$ indexes a TF bin of the STFT and f_k is the k-th natural frequency in the STFT.
The weighting function $1/|X_1[k,l]X_2^*[k,l]|$ is called Phase Transform (PHAT). This function was introduced to suppress the source's autocorrelation component from the angular spectrum.
Then, the TDOA estimate is given by

$$\hat{\tau}_{\texttt{TDOA}} = \arg\max_{\tau} \Psi(\tau),$$

with Ψ being either $\Psi_{\texttt{CC}}(\tau)$ or $\Psi_{\texttt{PHAT}}(\tau)$. Note that these functions can also be expressed directly as a function of the AOA using (10.2), hence the term *angular spectrum*. Despite the theoretical limitation of CC-based methods presented in [Chen et al. 2006], they are known to work well in practice under moderate reverberation and noise. 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 microphone array is compact and not linear[68], 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 method returns the source's DOA, namely azimuth-elevation pair (θ, ϕ), by estimating TDOAs in all microphone pairs. In order to achieve

[68] In the case where the microphones are complanar, the angle can be estimated up to an "up-down" ambigity.

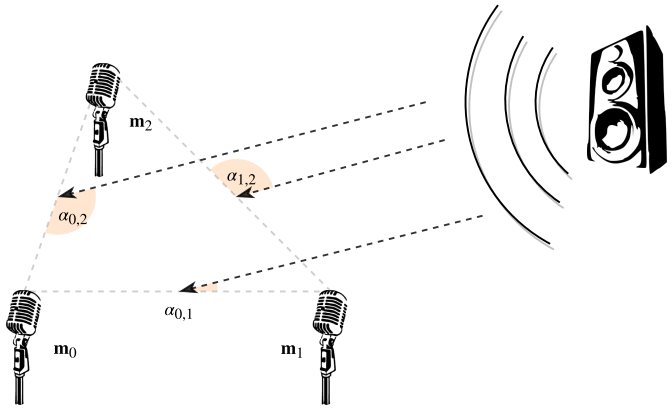


FIGURE 10.4: Illustration of the different DOAs at each microphone pairs attending one sound source. Knowing the position of the microphones, the angle with respect to a reference frame can be deduced in closed-form.

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 a reference frame in the array, typically centered at its barycenter. This function will exhibit a local maximum in the direction of the active source.
The algorithm can be decomposed into in the following steps:

See `MBSSLocate` for a free MATLAB implementation and comprehensive documentation of this algorithm.

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 AOAs (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 $\Psi_{\text{SRP}}(\theta, \phi)$;
5. the DOA(s) maximizing Ψ_{SRP} is (are) used as estimate(s).

10.4 MICROPHONE ARRAY AUGMENTATION WITH ECHOES

We now introduce the proposed concept of Microphone Augmentation with Echoes (MIRAGE). In the well-known Image Source Method (ISM) all the 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 to ISM, namely, the image-microphone model (see § 9.1). As illustrated in 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

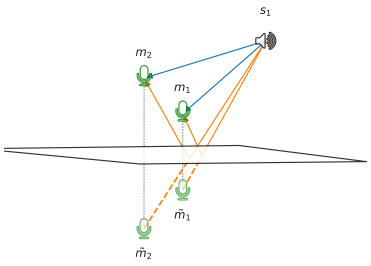


FIGURE 10.5: Illustration of the images $\mathring{m}_1$ and $\mathring{m}_2$ of microphones m_1 and m_2 in the presence of a reflective surface and a source. Blue lines correspond to direct paths, orange lines correspond to echo paths.

one defined in § 6.3, that is,

$$\begin{aligned}
\tau_{\texttt{TDOA}} &= \tfrac{1}{c}\|\underline{\mathbf{x}}_2 - \underline{\mathbf{s}}\| - \tfrac{1}{c}\|\underline{\mathbf{x}}_1 - \underline{\mathbf{s}}\| = \tau_2^{(0)} - \tau_1^{(0)}, \\
\tau_{\texttt{iTDOA}} &= \tfrac{1}{c}\|\mathring{\underline{\mathbf{x}}}_2 - \underline{\mathbf{s}}\| - \tfrac{1}{c}\|\mathring{\underline{\mathbf{x}}}_1 - \underline{\mathbf{s}}\| = \tau_2^{(1)} - \tau_1^{(1)}, \\
\tau_{\texttt{TDOE}_1} &= \tfrac{1}{c}\|\mathring{\underline{\mathbf{x}}}_1 - \underline{\mathbf{s}}\| - \tfrac{1}{c}\|\underline{\mathbf{x}}_1 - \underline{\mathbf{s}}\| = \tau_1^{(1)} - \tau_1^{(0)} \\
\tau_{\texttt{TDOE}_2} &= \tfrac{1}{c}\|\mathring{\underline{\mathbf{x}}}_2 - \underline{\mathbf{s}}\| - \tfrac{1}{c}\|\underline{\mathbf{x}}_2 - \underline{\mathbf{s}}\| = \tau_2^{(1)} - \tau_2^{(0)}
\end{aligned} \tag{10.5}$$

where $\mathring{\underline{\mathbf{x}}}_i$ denotes the position of the image of $\underline{\mathbf{x}}_i$ with respect to the reflector. It is easy to see that $\tau_{\texttt{TDOE}_2} = \tau_{\texttt{TDOE}_1} + \tau_{\texttt{iTDOA}} - \tau_{\texttt{TDOA}}$. Therefore, we will consider only the set $V = [\tau_{\texttt{TDOA}}, \tau_{\texttt{iTDOA}}, \tau_{\texttt{TDOE}_1}] \in \mathbb{R}^3$. These three quantities are directly connected to RIRs, as illustrated in Figure 10.6.

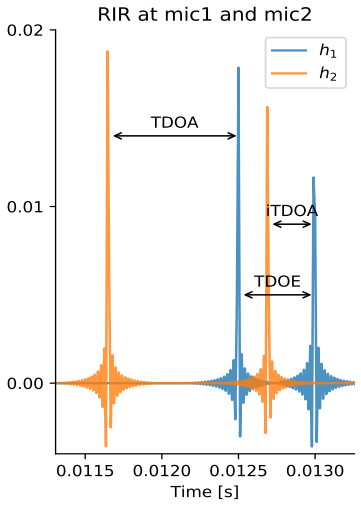


FIGURE 10.6: Superposition of two RIRs and visualization of time difference of arrival between direct paths (TDOA), first echoes (iTDOA) and direct path and first echo (TDOE).

To learn the parameters V, we consider the LANTERN data-driven approaches presented in Chapter 6. It consists in a class of DNN architectures trained to perform multi-target regression from the input audio features Interchannel Level Difference (ILD) and Interchannel Phase Difference (IPD) to the parameters $V \in \mathbb{R}^3$. In fact, those models were proposed exactly to address the task of estimating the first and strongest echo per channel in a close-surface scenario. Note that here the network provides directly a (unique) set of TDOA estimates, not "continuous" CC values.

Following the SRP-PHAT-like 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), the TDOA estimates in V can in principle be used to estimate the 2D direction of arrival of the source if provide as continues CC values. To this end, the 3 real-valued estimates in V are interpolated to create *virtual* local angular spectra (see § 10.3.2). In the first investigation of [Di Carlo et al. 2019], we proposed to use Gaussians functions centered at those estimates with variances equal to the prediction errors made by the DNN on the validation set. Later, we modified the networks in order to estimated both these means and the variances, as explained in § 6.4.

10.4.1 *Experimental Results*

In this section we will report the experimental results for the proposed approach. At first, we will consider the continuation of the investigation started in § 6.3, namely, the analysis is limited to stereophonic recordings in a single-source close-surface scenario.

Here we compare the MIRAGE approach using the LANTERN MLP (MIRAGE-MLP) learning model. Note that in § 6.4 we demonstrated that MLP is outperformed by more recent DNN architecture. However, here we are reporting the results obtained in the work [Di Carlo et al. 2019], antecedent to the experimental results reported in § 6.4. A free and open-source Matlab implementation of SRP-PHAT[69] is used to aggregate local angular spectra obtained from the model output.[70] External noises in the recordings are simulated by perturbing the observation with AWGN noise at 10 dB SNR (wn+n, sp+n, respectively). In this stereophonic experiment, the baseline method GCC-PHAT is not considered: it can access only the true microphone signals, thus, it is unable to perform 2D-SSL.

[69] MBSSLocate

[70] For SRP-PHAT we used a sphere sampling with 0.5° resolution and coordinates $\theta \in [-179, 180]$ and $\phi \in [0, 90]$ are used for the DOA search.

DOA		ACCURACY $< 10°$		ACCURACY $< 20°$	
	Input	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)

TABLE 10.1: Mean angular errors in degree (with accuracies (%)) for 2D SSL (azimuth and elevation) with 10° and 20° tolerance.

The dataset, the implementation, and the training procedure are the same as in § 6.3.3. In particular, the test set features both white noise (wn), and speech (sp) signals convolved with RIRs generated by the acoustic simulator [Schimmel et al. 2009]. Moreover, some observations feature background AWGN at 10 dB SNR level (wn+n, sp+,n respectively).

▶ NUMERICAL RESULTS

Table 10.1 reports the performance of the full MIRAGE-MLP 2D-SSL pipeline. Within a tolerance of 20°, the MIRAGE-MLP model allows estimation of 2D-DOA with an accuracy of 79% for azimuth and 88% for elevation for a source emitting white noise in noiseless conditions. It is interesting to notice that our model trained and validated with white noise sources somewhat generalizes to speech sources. In fact, for noiseless speech signals the performances are just slightly lower than the corresponding broadband case. As already observed in § 6.3, external noise severely degrades performances.

With this work, we demonstrated how a simple echo model could allow 2D SSL with only two microphones, using simulated data. Motivated by these results, we put lot of effort in collecting real-world data in order to test the whole MIRAGE pipeline, including LANTERN. One direction was offered by a collaboration with Randy Gomez from the Honda Research Institute. The idea was to test the proposed method on the microphone array of the autonomous robot platform called *HARU* [Ackerman 2018; Gomez et al. 2018]. We collected real-world data, but unfortunately due to several problems identified in the resulting dataset, these recordings were discarded.

10.5 CONCLUSION

This chapter demonstrated how a simple echo model could boost an SSL algorithm in strongly echoic conditions when microphones are placed close to a reflector. We proposed to use a successful algorithm for multichannel SSL on the virtual array created by image microphones. In order to leverage such an array, the echoes' parameters need to be estimated. To this end, we used 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, the task is still very challenging in the presence of noise 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 approaches to SSL. The learning approach could still be significantly improved by considering other acoustic features (such as advance ReTF methods), other architectures, and

other improvements. For instance, handling the missing frequencies of the speech while training on a broadband signal such as in [Gaultier et al. 2017], using physics-driven regularizers such as in [Nabian and Meidani 2020] and making the learning phase independent of the array geometry. Secondly, the approach needs to be tested on real-world data, for which the `dEchorate` could be used as a valuable testing dataset in the future.

11

Echo-aware Applications of `dEchorate`

"Echoes of the same old psalm
Sing with me
In unity
Join in with me"
—Leprous, *Mirage*

Keywords: Early reflection, Speech Enhancement, Beamforming, Room Geometry Estimation, Reflector Localization.
Resources:

- `dEchorate`
- `Risotto`
- `Brioche`

▶ 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 Tandeitnik at the Bar'Ilan University, Israel. The algorithms presented here are straightforward extension of the ones 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.

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*.

- 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, ReTFs) 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.[71] Since then, ReIRs have been incorporated in powerful beamforming algorithms, used for both dereverberation and noise reduction (e. g. [Schwartz et al. 2014; Kodrasi and Doclo 2017]).

[71] Note that, as opposed to channel estimation, estimating the ReTFs is a non-blind problem (See § 3.3)

- 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 [Flanagan et al. 1993; Jan et al. 1995]. 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 as 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. 2015] (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.

11.1.2 *Background in spatial filtering*

Given the narrowband STFT signal model presented is § 3.2.5, the signals captured by I microphones listening to a single sound source ($J = 1$) in a noisy reverberant room reads:

$$\boldsymbol{X}[k,l] = \boldsymbol{H}[k]S[k,l] + \boldsymbol{N}[k,l], \tag{11.1}$$

where $\boldsymbol{X}[k,l], \boldsymbol{H}[k], \boldsymbol{N}[k,n] \in \mathbb{C}^I$ 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 $\boldsymbol{H}[k]$ 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,

$$\boldsymbol{H} = \left[H_i^{\mathtt{dp}} + H_i^{\mathtt{ee}} + H_i^{\mathtt{lr}}\right]_i \tag{11.2}$$

where the summands correspond to the direct-path (`dp`), early echoes (`ee`) and late reverberation (`lr`), respectively. It is sometimes common to consider the entire *early contributions*, $\boldsymbol{H}^{\mathtt{ec}}$ comprising both direct and echo paths. We can now model the such part of the RIR associated to the i-th channel according the echo model, that is,

$$H_i^{\mathtt{ec}} = H_i^{\mathtt{dp}} + H_i^{\mathtt{ee}} = \sum_{r=0}^{R} \alpha_i^{(r)} \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_i^{(r)}}, \tag{11.3}$$

where $r = 0$ is the index of the direct propagation. Given such a decomposition of the RIRs, the vector $\boldsymbol{X}$ can be expressed accordingly as:

$$\boldsymbol{X} = \boldsymbol{X}^{\mathtt{ec}} + \boldsymbol{X}^{\mathtt{lr}} + \boldsymbol{N}. \tag{11.4}$$

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. 2015; 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 [Bradley et al. 2003]. Based on this, we define as the signal of interest the following

$$\boldsymbol{X}^{\mathtt{ec}} = \boldsymbol{X}^{\mathtt{dp}} + \boldsymbol{X}^{\mathtt{ee}} = \boldsymbol{H}^{\mathtt{ec}} S. \tag{11.5}$$

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, $\boldsymbol{\Sigma}_x = \mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^{\mathsf{H}}\right\} \in \mathbb{C}^{I\times I}$, as

$$\boldsymbol{\Sigma}_x = \boldsymbol{H}^{\mathrm{ec}} \boldsymbol{\Sigma}_s \boldsymbol{H}^{\mathrm{ec}\mathsf{H}} + \boldsymbol{\Sigma}_s^{\mathtt{lr}} + \boldsymbol{\Sigma}_n, \tag{11.6}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator. Here $\boldsymbol{\Sigma}_x^{\mathtt{lr}}$ and $\boldsymbol{\Sigma}_n$ denote the PSD matrices of the late reverberation and noise, respectively. We will describe each term in turn.

- THE SOURCE covariance matrix $\mathbf{\Sigma}_s$ is assumed here to be diagonal, since we assume that all the spatial information is expressed by the filters $\mathbf{H}^{\text{ec}}$ and the source signals are independent to each other, that is,

$$\mathbf{\Sigma}_s = \sigma_s^2 \mathbf{I}, \tag{11.7}$$

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 2016].

$$\mathbf{\Sigma}_x^{\text{lr}} = \sigma_{\text{lr}}^2 \mathbf{\Gamma}, \tag{11.8}$$

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.[72] 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 term $\boldsymbol{N}$ is assumed to be stationary over the recording. Therefore its covariance matrix can be 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 exploiting speech non-stationarity [Gannot et al. 2001].

[72] Given the distance $d_{ii'}$ between to microphone i and i', the interchannel coherence function in the continuous-frequency domain writes

$$\tilde{\gamma}_{ii'}[k] = \frac{\sin(2\pi f_k d_{ii'}/c)}{2\pi f_k d_{ii'}/c}.$$

Then, the matrix $\mathbf{\Gamma} = \left[\tilde{\gamma}_{ii'}\right]$ is built 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 $Y \in \mathbb{C}$:

$$Y = \boldsymbol{W}^{\mathsf{H}} \boldsymbol{X} = \boldsymbol{W}^{\mathsf{H}} \boldsymbol{H} S + \boldsymbol{W}^{\mathsf{H}} \boldsymbol{N},$$

where the vector $\boldsymbol{W} \in \mathbb{C}^I$ 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. (3.33), that is,

$$\boldsymbol{D}' = \left[1, \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_i^{(0)}/c}, \ldots, \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_I^{(0)}/c}\right] \tag{11.9}$$

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

$$\boldsymbol{W}_{\text{dp}} = \frac{\boldsymbol{D}'}{\|\boldsymbol{D}'\|}, \tag{11.10}$$

where $\|\cdot\|$ denotes the Euclidean norm.

▸ THE MVDR beamformer optimizes the following criterion[73]

$$\boldsymbol{W}_{\texttt{MVDR}} = \arg\min_{\boldsymbol{W}} \boldsymbol{W}^{\mathsf{H}}\boldsymbol{\Sigma}_n\boldsymbol{W} \text{ s.t. } \boldsymbol{W}^{\mathsf{H}}\boldsymbol{H} = 1. \tag{11.11}$$

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 $\boldsymbol{\Sigma}_n$.
The Least-Square minimization through the Lagrangian multiplier method yields the following closed-form optimal solutions

$$\boldsymbol{W}_{\texttt{MVDR}} = \frac{\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}}{\boldsymbol{H}^{\mathsf{H}}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}}. \tag{11.12}$$

Note that these techniques do not require any reference signal, only the knowledge of the source's filter $\boldsymbol{H}$ 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 $\boldsymbol{\Sigma}_n$ and $\boldsymbol{H}$, respectively.

[73] The MVDR design is equivalent to the Minimum Power Distortionless Response (MPDR) beamformer which minimize the following

$$\boldsymbol{W}_{\texttt{MPDR}} = \arg\min_{\boldsymbol{W}} \boldsymbol{W}^{\mathsf{H}}\boldsymbol{\Sigma}_x\boldsymbol{W} \text{ s.t. } \boldsymbol{W}^{\mathsf{H}}\boldsymbol{H} = 1.$$

However, it exhibit higher sensitivity to misalignment errors than the MVDR beamformer [Gannot et al. 2017, Section V.A].

11.1.4 *Noise, steering vectors, rake filters, and relative transfer functions*

The vectors $\boldsymbol{H}$ 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 built on the knowledge of the TOAs of the source signal and their integration in the MVDR criterion is straightforward: the $\boldsymbol{H}$ simply identifies with the corresponding relative steering vector $\boldsymbol{D}$, that is,

$$\boldsymbol{H}_{\texttt{DP}}[k] = \boldsymbol{D}[k] = \left[\mathrm{e}^{-\mathrm{i}2\pi f_k \tau_i^{(r)}}\right]_i. \tag{11.13}$$

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 DOA[74], thus, the steering vectors can be easily computed. However, DOAs need to be estimate aside, using Sound Source Localization (SSL) methods (See § 8.4).

[74] 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 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:

$$\boldsymbol{H}_{\texttt{RAKE}}[k] = \left[\sum_{r=0}^{R} \alpha_i^{(r)} \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_i^{(r)}}\right]_i. \tag{11.14}$$

As before, both the echoes' delays and amplitudes are considered relatively to the reference microphone.

- 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 H_i, its ReTFs with respect to the first channel is given by

$$\boldsymbol{H}_{\texttt{ReTF}}[k] = \frac{\boldsymbol{H}[k]}{H_1[k]} \tag{11.15}$$

  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 $\boldsymbol{\Sigma}_n$. 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 $\boldsymbol{\Sigma}_n$ and using it in Eq. (11.12).[75] For instance,

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_{s_q} + \boldsymbol{\Sigma}_n, \tag{11.16}$$

  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 diarization[76] tools. If these excerpts are long enough, the whole covariance matrix $\boldsymbol{\Sigma}_n$, 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 $\boldsymbol{\Sigma}_s^{\texttt{lr}}$ 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,

$$\boldsymbol{\Sigma}_{\texttt{noise-late}} = \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_s^{\texttt{lr}} = \boldsymbol{\Sigma}_u + \sigma^2_{\texttt{lr}}\boldsymbol{\Gamma}. \tag{11.17}$$

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, $\boldsymbol{\Sigma}_u$, 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);

[75] 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 contraint is satisfied [Gannot et al. 2017].

[76] 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	$\boldsymbol{W} = \boldsymbol{H}_{\texttt{DP}} / \|\boldsymbol{H}_{\texttt{DP}}\|$
MVDR-DP	$\min\ \boldsymbol{W}^{\mathsf{H}}\boldsymbol{\Sigma}_u\boldsymbol{W}$ s.t. $\boldsymbol{W}^{\mathsf{H}}\boldsymbol{H}_{\texttt{DP}} = 1$	$\boldsymbol{W} = (\boldsymbol{H}_{\texttt{DP}}{}^{\mathsf{H}}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{H}_{\texttt{DP}})^{-1}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{H}_{\texttt{DP}}$
MVDR-ReTF	$\min\ \boldsymbol{W}^{\mathsf{H}}\boldsymbol{\Sigma}_u\boldsymbol{W}$ s.t. $\boldsymbol{W}^{\mathsf{H}}\boldsymbol{H}_{\texttt{ReTF}} = 1$	$\boldsymbol{W} = (\boldsymbol{H}_{\texttt{ReTF}}{}^{\mathsf{H}}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{H}_{\texttt{ReTF}})^{-1}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{H}_{\texttt{ReTF}}$
MVDR-Rake*	$\min\ \boldsymbol{W}^{\mathsf{H}}\boldsymbol{\Sigma}_u\boldsymbol{W}$ s.t. $\boldsymbol{W}^{\mathsf{H}}\boldsymbol{H}_{\texttt{RAKE}} = 1$	$\boldsymbol{W} = (\boldsymbol{H}_{\texttt{RAKE}}{}^{\mathsf{H}}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{H}_{\texttt{RAKE}})^{-1}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{H}_{\texttt{RAKE}}$
MVDR-DP-Late	$\min\ \boldsymbol{W}^{\mathsf{H}}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})\boldsymbol{W}$ s.t. $\boldsymbol{W}^{\mathsf{H}}\boldsymbol{H}_{\texttt{DP}} = 1$	$\boldsymbol{W} = (\boldsymbol{H}_{\texttt{DP}}{}^{\mathsf{H}}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})^{-1}\boldsymbol{H}_{\texttt{DP}})^{-1}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})^{-1}\boldsymbol{H}_{\texttt{DP}}$
MVDR-ReTF-Late	$\min\ \boldsymbol{W}^{\mathsf{H}}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})\boldsymbol{W}$ s.t. $\boldsymbol{W}^{\mathsf{H}}\boldsymbol{H}_{\texttt{ReTF}} = 1$	$\boldsymbol{W} = (\boldsymbol{H}_{\texttt{ReTF}}{}^{\mathsf{H}}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})^{-1}\boldsymbol{H}_{\texttt{ReTF}})^{-1}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})^{-1}\boldsymbol{H}_{\texttt{ReTF}}$
MVDR-Rake-Late*	$\min\ \boldsymbol{W}^{\mathsf{H}}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})\boldsymbol{W}$ s.t. $\boldsymbol{W}^{\mathsf{H}}\boldsymbol{H}_{\texttt{RAKE}} = 1$	$\boldsymbol{W} = (\boldsymbol{H}_{\texttt{RAKE}}{}^{\mathsf{H}}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})^{-1}\boldsymbol{H}_{\texttt{RAKE}})^{-1}(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{\texttt{lr}})^{-1}\boldsymbol{H}_{\texttt{RAKE}}$

TABLE 11.1: Summary of the considered beamformers. (*) denotes echo-aware beamformers.

- the ReTFs, $\boldsymbol{H}_{\texttt{ReTF}}$, is estimated from the observed signal using the Generalized Eigenvalue Decomposition (GEVD) method described in [Doclo and Moonen 2003];
- $\boldsymbol{H}_{\texttt{DP}}$, $\boldsymbol{H}_{\texttt{RAKE}}$ are computed using known relative delays and amplitudes available in the DECHORATE dataset;
- the late reverberation power, $\sigma^2_{\texttt{lr}}$, is estimated in closed-form knowing the filters $\boldsymbol{H}$ 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 ($\texttt{SNR} \in \{0, 10, 20\}$). Using the `dEchorate` data, we considered the room configuration `011111` ($\texttt{RT}_{60} \approx 600$ ms) and all the possible combinations of target/array's positions. Both real and corresponding synthetic RIRs are used, which are then convolved with anechoic utterances from the WSJ corpus and corrupted by recorded diffuse noise. The correspondence between real and synthetic RIRs is a property of the `dEchorate` dataset and is described in Chapter 7. 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 [Dokmanić et al. 2015; Kodrasi and Doclo 2017; 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, $\boldsymbol{X}^{\texttt{ec}}$ in Eq. (11.5). On one hand, this choice numerically penalizes both direct-path-based and ReTF-based beamformers which aim at extract the direct-path signal only and the reverberant signal at the reference microphone, respectively. On the other hand, considering only the direct path or the full reverberant signal would be equally unfair for the other beamformers. Moreover, choosing the early contribution as target signal is perceptually motivated by the fact that

early reflection contributes to speech intelligibility [Bradley et al. 2003]. 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 $\texttt{SNRR}_\texttt{i}$ at the reference microphone and the $\texttt{SNRR}_\texttt{o}$ at the filter output. Denoting with X_1 the signal at the reference microphone and $X_1^{\texttt{ec}}$ the target speech at the same microphone (See Eq. (11.5)), these quantities are defined as follows:

$$\begin{aligned} \texttt{SNRR}_\texttt{i} &= 10\log_{10}\left(\frac{\sigma^2_{X_1^{\texttt{ec}}}}{\sigma^2_{X_1}-\sigma^2_{X_1^{\texttt{ec}}}}\right) \quad [\text{dB}] \\ \texttt{SNRR}_\texttt{o} &= 10\log_{10}\left(\frac{\sigma^2_{\boldsymbol{W}^{\texttt{H}}\boldsymbol{X}^{\texttt{ec}}}}{\sigma^2_{\boldsymbol{W}^{\texttt{H}}\boldsymbol{X}}-\sigma^2_{\boldsymbol{W}^{\texttt{H}}\boldsymbol{X}^{\texttt{ec}}}}\right) \quad [\text{dB}] \\ \Delta\texttt{SNRR} &= \texttt{SNRR}_\texttt{o} - \texttt{SNRR}_\texttt{i} \quad [\text{dB}] \end{aligned} \tag{11.18}$$

- 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].[77] 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

[77] 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).

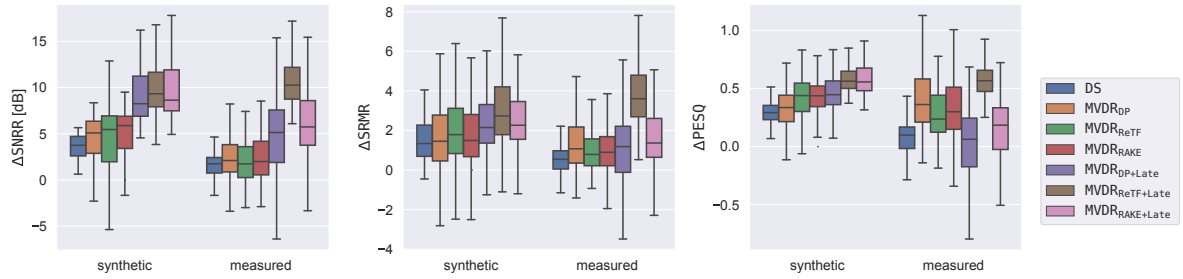


FIGURE 11.1: Comparison of echo-aware beamforming for the room configuration `011111` ($RT_{60} \approx 600$ ms) on measured and synthetic data for all combinations of source-array positions in the `dEchorate` dataset.

Mean Opinion Score (MOS)[78] is inferred using a linear regression model. An implementation of this metrics is available at the `speechmetrics` Python library.

[78] The MOS is a measure used in the domain of Quality of Experience and telecommunications engineering, representing overall quality of a stimulus or system. MOS were originally obtained from perceptual tests involving human subjects.

▶ 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 that more information is used, the better performances are. ReTFs- 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, ReTFs 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 ReTFs, 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 ReTFs can capture more information which, in turn, yields to sightly better enhancement. Nevertheless, the PESQ metrics suggest that for this ideal scenario both echo-aware (Rake) and echo-agnostic (ReTFs) 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 ΔPESQ metric, as it also accounts for artifacts. Here, the echo-agnostic beamformer considering late reverberation $\texttt{MVDR}_{\texttt{ReTF+Late}}$ outperforms the other methods, maintaining the trend exhibited on simulated data. In general, it looks like the $\texttt{MVDR}_{\texttt{Rake+Late}}$ has more variance than $\texttt{MVDR}_{\texttt{ReTF+Late}}$, 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 in 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] (see below), 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).

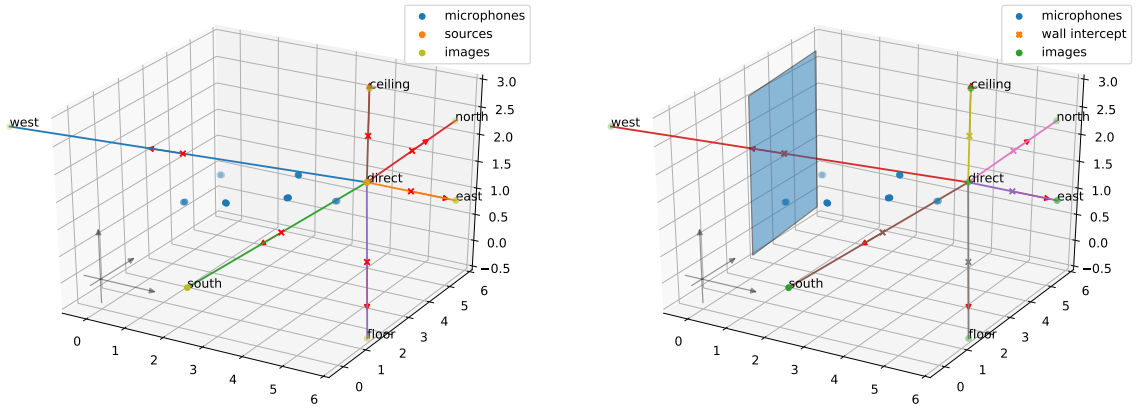
More advance techniques have been proposed in the literature of reflector localization for different setups and scenarios. A comprehensive review can be found in [Remaggi et al. 2016; Crocco et al. 2017].
Nonetheless, when the echoes' TOAs and their labeling are known for 4 spatially-separated non-coplanar 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 (on 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).



(a) Source images estimation

(b) Reflector estimation

FIGURE 11.2: Images source estimation (right) and corresponding reflector estimation (left) for one of the sound sources in the DECHORATE dataset.

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.

source id	1		2		3		4	
wall	DE	AE	DE	AE	DE	AE	DE	AE
west	0.74	8.99°	4.59	8.32°	5.89	5.75°	**0.05**	2.40°
east	**0.81**	0.08°	0.9	0.50°	*69.51*	*55.70°*	0.31	0.21°
south	3.94	*16.08°*	**0.18**	1.77°	*14.37*	*18.55°*	0.82	1.65°
north	1.34	0.76°	1.40	8.94°	**0.63**	0.17°	2.08	1.38°
floor	**5.19**	1.76°	7.27	2.66°	7.11	2.02°	5.22	1.90°
ceiling	1.16	0.28°	0.67	0.76°	**0.24**	1.16°	0.48	0.26°

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.

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 and 2.5° errors. Small errors are due to concurrency of multiple factors, such as tiny offsets in the annotation and the ideal shoebox approximation[79]. 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 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 misclassified. 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.

[79] In the real recording room, some gaps were present between revolving panels in the walls

11.3 Conclusions

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 know 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 ReTFs-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 ReTFs, 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 `dEchorate`data 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. 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 $\texttt{RT}_{60}$, 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, it could be possible to develop techniques to convert one to the other, using style transfer and domain adaptation techniques.

Part IV

EPILOGUE

12

Echo-aware Reflective Reflection

"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*

▶ 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 the case of passive stereophonic recordings.

B. We revisited some fundamental *audio scene analysis problems* and methodologies under an echo-aware perspective.

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 directions above. The pursuit of these two goals took the form of the following methods and outcomes.

- A Knowledge-driven Method for AER dubbed BLASTER. This approach enables direct and *off-grid* estimation of the echoes' properties in stereophonic passive recordings. In particular, the "knowledge" we used is the echo model for the early part of the RIRs. Due to its off-grid nature, the method overcomes some theoretical limitation of on-grid approaches. Although it is currently not outperforming the state-of-the-art when retrieving more than 2 echoes per channel, this investigation is motivated by theoretical guarantees.

- 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 a virtually-supervised fashion. The proposed model combines recent deep learning methodologies to reduce estimation error with a loss functions able to measure uncertainty. Having a measure for confidence allows to aggregate multiple observations of the same scenario in a data-fusion-like approach.

- An Echo-aware Dataset designed for both AER and echo-aware applications, dubbed DECHORATE. These annotated data should fill a gap between existing datasets and it is designed for validating future echo-aware research. The dataset is accompanied by software utilities to easily access, manipulate, and visualize the data and baseline methods

in echo-related tasks. Moreover, by evaluating state-of-the-art methods on these data and comparing their results with corresponding RIRs generated by acoustic simulators, we showed that current methodologies did not always generalize well to real acoustic conditions.

- An ECHO-AWARE AUDIO SOURCE SEPARATION METHOD, dubbed SEPARAKE. It is based on the popular Multichannel NMF framework, which allows simple yet effective integration of the echoes' properties. Assuming their knowledge, we can reformulate such a framework in terms of image microphones and virtual arrays. Results show how this leads to enough spatial diversity to get a performance boost over the vanilla version of two classical NMF-based algorithms.

- An ECHO-AWARE SOUND SOURCE LOCALIZATION METHOD, dubbed MIRAGE. By regarding echoes as image microphones, this method allows for source's azimuth and elevation estimation from 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 method allows for simple extensions to multi-channel recordings as long as the geometry of the array is available.

- The following LIBRARIES for echo-aware processing:

  - `dEchorate` — code for DECHORATE, Room Impulse Response estimation and annotation. dEchorate
    Joint work with Pinchas Tandeitnik.
  - `Risotto` — a collection of state-of-the-art methods for estimating Relative Impulse Responses. Risotto
    Joint work with Z. Koldovský, S. Markovich-Golan and S. Gannot.
  - `Brioche` — a collection of state-of-the-art beamforming techniques including, but not limited to, echo-aware approaches. Brioche
    Joint work with S. Gannot.
  - `Blaster` — code for BLASTER, its results and related state-of-the-art methods. Blaster
    Joint work with C. Elvira.
  - `Separake` — code for SEPARAKE including an Python implementation of the Matlab toolbox Multichannel NMF [Ozerov and Févotte 2010] for audio source separation. Separake
    Joint work with R. Scheibler.
  - `pyMBSSLocate` — Python implementation of the Matlab toolbox `MBSSLocate` [Lebarbenchon et al. 2018b] for sound source localization. pyMBSSLocate
    Joint work with of R. Lebarbenchon and E. Camberlein.

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 towards echo-aware audio signal processing. Nevertheless, there are many issues and many potential areas of improvement in the methods presented here. Besides, it is clear that we have only scratched the surface of many problems related to echo 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**

  This thesis' work highlighted the difficulty of estimating echoes accurately in arbitrary scenarios. Using prior information, the problem can be substantially relaxed, for instance, knowing the microphone array's geometry. The impact of the microphone array (and antenna) geometry has been well studied in various branches of signal processing, such as telecommunication. Therefore, the knowledge of array geometry could 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 searched for. Even if this approach was used for the considered echo-aware applications, it was not for the proposed AER methods, which are currently restricted to the stereophonic case. Recent works in [Jensen et al. 2019] (resp. in [Kowalczyk 2019]) demonstrated the benefit of using this kind of steering vectors in AER (resp. echo-aware methods).

  Besides this, we also noticed that a critical parameter for AER is the *number of early echoes* within a limited time window. Therefore the *echo counting* problem appears to be a reasonable direction to investigate. Estimating early reflections knowing their number can be recast as a *sparse coding* problem. Unfortunately, problems in this class are not necessarily easier to solve, but it was shown that 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 instead of the simple magnitude and phase of the instantaneous ReTF. On the same direction, one could directly enhance the recorded sound with respect to noise and late reverberation. The most promising direction in this regard is to use recent results on relative early transfer functions estimation [Schwartz et al. 2016; Kodrasi and Doclo 2017; Tammen et al. 2018]. Relative early transfer functions are transfer functions accounting for the early part of the ReTF only and are used for speech enhancement and dereverberation.

  Alternatively, statistical models for reverberation that are parametrized by the RT_{60} or the Direct-to-Reverberant Ratio (DRR) can be used. Particularly appealing are the works of [Leglaive et al. 2015] and [Badeau 2019], which define a framework to deal with early reflections within a statistical model. This exploratory direction is also motivated by recent results in estimating these parameters blindly from 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 models have the potential to solve issues due to the strong approximation used in the image source model, such as flat walls, frequency-independent absorption coefficients and specular-only reflections.

Learning-based approaches and especially Deep Neural Networks (DNNs) are potent tools. Unfortunately, their potential was not fully exploited in this thesis work, but many exciting directions can be pursued. Of particular interest are the recent developments in physics-driven neural networks [Nabian and Meidani 2020; Rao et al. 2020; Jin et al. 2020]. In these work physically-motivated regularizers or entire loss-functions obeying physical laws are used to ensure the correctness of the solution or restrict the search space. It can be done by adding simple physical relations, such as the temperature-density relation for fluids as in [Karpatne et al. 2017], or entire PDEs as in the other cited works.

- Using echoes

In this thesis, we presented a few selected applications of acoustic echoes to audio scene analysis problems. This could be extended to other related problems such as binaural hearing, source tracking and acoustic measurements.

Humans have two ears and complex auditory systems that provide a natural filtering of incoming sounds. Moreover, reflections from the 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 settings in 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], and to build perceptual evaluation metrics.

The MIRAGE framework presented in Chapter 10 can be extended in the long term to sound source localization and tracking, as well as to acoustic feedback cancellation (better know as acoustic echo cancellation) or to microphone array self-calibration. This has a great potential for industrial applications, as illustrated by vocal assistant devices such as Google Home or Amazon Echo. These smart speakers could benefit from echo-aware processing, not only in recording sounds but also in sound production. For instance, echo processing could provide dynamic *sweet-spots* that adapt 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, the frameworks proposed in the thesis could be expanded beyond the field of indoor audio signal processing. Other research fields putting relevant focus on reflections include 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 processing on real data. A natural step is to use these data for the applications discussed in this thesis and particularly for AER, which was the dataset's initial goal. Moreover, new lines of research using 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 of a new type of learning-based acoustic simulators.

- CROSSING THE DIRECTIONS
  Ultimately, the two parts of estimation and application in this dissertation should be combined together. So far we only showed how from audio features it was possible to estimate echoes and how from echoes it was possible to estimate audio scene analysis information, e. g. source content and location. These problems have an innate *uroboric* nature: where, what, when and how are connected — the knowledge of one helps the estimation of the others, in a 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

Resources:

- Code
- Open-access paper with supplementary material

▶ Synopsis This appendix briefly describes the Sliding Franck-Wolfe Algorithm used to solve the Non-negative BLASSO presented is Chapter 5. The material reported here are extracted from the supplementary material accompanying the work in [Di Carlo et al. 2020]. The main author of this work is Clement Elvira, to whom I extend my greatest thanks, and I will report it for sake of completeness.

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 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]

$$\sup_{\theta \in \Theta} \lambda^{-1} \left| \left\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^{(t)} \right\rangle \right| \leq 1. \tag{12.1}$$

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 (12.1) 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 (12.1).
We have shown earlier that SFW stops as soon as an iterate $\mu^{(t)}$ satisfies (12.1) at some iteration t. Since the estimated coefficients $\left\{c_r^{(t)}\right\}_{r=1}^{R}$ are (non-

Algorithm 1: Sliding Frank-Wolfe algorithm for solving (5.22-$\mathcal{P}^{\lambda}_{\mathrm{TV}}$).

Input: Observation operator $\mathcal{A}$, positive scalar λ, precision ε

Output: Channels represented as a measure $\widehat{\mu}$

// Initialization

$\mathbf{y} \leftarrow -\mathcal{A}\delta_{(0,1)}$ // observation vector

$\mu^{(0)} = 0_{\mathcal{M}}$ // estimated filters

$\mathcal{E}^{(0)} = \emptyset$ // estimated echoes

$x_{\max} = (2\lambda)^{-1}\|\mathbf{y}\|_2^2$;

// Starting algorithm

repeat

  $t \leftarrow t+1$ // Iteration index

  // 1. Add new element to the support

  Find $\theta^{\mathsf{new}} \in \arg\max_{\theta\in\Theta} \mathrm{Re}\big(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^{(t-1)}\rangle\big)$;

  $\eta^{(t)} \leftarrow \lambda^{-1}\,\mathrm{Re}\big(\langle \mathcal{A}\delta_{\theta^{\mathsf{new}}}, \mathbf{y} - \mathcal{A}\mu^{(t-1)}\rangle\big)$;

  if $\eta^{(t)} \leq 1+\varepsilon$ **then**

    Stop and return $\widehat{\mu} = \mu^{(t-1)}$ is a solution ;

  end

  $\mathcal{E}^{(t-½)} \leftarrow \mathcal{E}^{(t-½)} \cup \{\theta^{\mathsf{new}}\}$;

  $R^{(t)} \leftarrow \mathsf{card}(\mathcal{E}^{(t-½)})$ // Number of detected echoes

  // 2. Lasso update of the coefficients

  $$\mathbf{c}^{(t-½)} \leftarrow \operatorname*{arg\,min}_{\mathbf{c}\in\mathbf{R}_+^{R^{(t)}}} \frac{1}{2}\Big\|\mathbf{y} - \sum_{\theta\in\mathcal{E}^{(t-½)}} c_\theta \mathcal{A}\delta_\theta\Big\|_2^2 + \lambda\|\mathbf{c}\|_1$$

  approximated using a proximal gradient algorithm ;

  // 3. Joint update for a given number of spikes

  $$\mathcal{E}^{(t)}, \mathbf{c}^{(t)} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta^{R^{(t)}}, \mathbf{c}\in[0,x_{\max}]^{R^{(t)}}} \frac{1}{2}\Big\|\mathbf{y} - \sum_{r=1}^{R^{(t)}} \mathbf{c}_r \mathcal{A}\delta_{\theta_r}\Big\|_2^2 + \lambda\|\mathbf{c}\|_1$$

  approximated using a non-convex solver initialized with $(\mathcal{E}^{(t-½)}, \mathbf{c}^{(t-½)})$;

  // 4. Eventually remove zero amplitude Dirac masses

  $\mathcal{E}^{(t)} \leftarrow \Big\{\theta_r^{(t)} \in \mathcal{E}^{(t)} \mid \mathbf{c}_r^{(t)} \neq 0\Big\}$;

  $\mathbf{c}^{(t)} \leftarrow \Big\{\mathbf{c}_r^{(t)} \mid \mathbf{c}_r^{(t)} \neq 0\Big\}$;

  $\mu^{(t)} \leftarrow \sum_{r=1}^{\mathsf{card}(\mathcal{E}^{(t)})} \mathbf{c}_r^{(t)} \delta_{\theta_r^{(t)}}$;

until *until convergence*;

negative) scalars, (12.1) can be rewritten as

$$\sup_{\theta\in\Theta} \lambda^{-1} \operatorname{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^\star\rangle) \leq 1. \tag{12.2}$$

In particular, using the real part in the implementation allows to remove the imaginary part that may appear due to the imprecision.

- PRECISION OF THE STOPPING CRITERION.
  Unfortunately, condition (12.1) cannot be met due to the machine precision, *i.e.*, the solution of (5.22-$\mathcal{P}^\lambda_{\mathrm{TV}}$) is computed up to some prescribed accuracy. In this paper, we say that the algorithm stops as soon as

$$\sup_{\theta\in\Theta} \lambda^{-1} \operatorname{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^\star\rangle) \leq 1 + \varepsilon \tag{12.3}$$

  where ε is a positive scalar set to $\varepsilon = 10^{-3}$.

- FINDING NEW PARAMETERS (LINE 7).
  The new parameter is found by solving

$$\arg\max_{\theta\in\Theta} \operatorname{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\widehat{\mu}\rangle). \tag{12.4}$$

  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 library[80].

[80] 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

$$\arg\min_{\boldsymbol{\theta}\in\Theta^R, \mathbf{c}\in\mathbf{R}^R} \tfrac{1}{2}\Big\|\mathbf{y} - \sum_{r=1}^R \mathbf{c}_r \mathcal{A}\delta_{\theta_r}\Big\|_2^2 + \lambda\|\mathbf{c}\|_1 \tag{12.5}$$

  is equivalent to the solution of

$$\arg\min_{\boldsymbol{\theta}\in\Theta^R, \mathbf{c}\in[0,x_{\max}]^R} \tfrac{1}{2}\Big\|\mathbf{y} - \sum_{r=1}^R \mathbf{c}_r \mathcal{A}\delta_{\theta_r}\Big\|_2^2 + \lambda\|\mathbf{c}\|_1 \tag{12.6}$$

  where

$$x_{\max} = \frac{1}{2\lambda}\|\mathbf{y}\|_2^2. \tag{12.7}$$

  Indeed, let us denote $\boldsymbol{\theta}^\star, \mathbf{c}^\star$ the minimizers of (12.5). For any $\boldsymbol{\theta} \in \Theta^R$, the couple $\boldsymbol{\theta}, \mathbf{0}_R$ is admissible for (12.5) so we have by definition

$$\tfrac{1}{2}\Big\|\mathbf{y} - \sum_{r=1}^R \mathbf{c}^\star_r \mathcal{A}\delta_{\theta^\star_r}\Big\|_2^2 + \lambda\|\mathbf{c}^\star\|_1 \leq \tfrac{1}{2}\|\mathbf{y}\|_2^2. \tag{12.8}$$

  Hence

$$0 \leq c^\star_r \leq \|\mathbf{c}^\star\|_1 \leq \tfrac{1}{2\lambda}\|\mathbf{y}\|_2^2 \triangleq x_{\max}. \tag{12.9}$$

  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 80.

(Incomplete) Recommended Listenings

An incomplete selection of the most played music that fueled this Ph.D. work (according to Deezer).

First Ph.D. Year

- Gojira
  – *L'Enfant Savage*
- Xploding Plastix
  – *Amateur Girlfriends Go Proskirt Agents*
- Clark
  – *Totems Flare*
- Aufgang
  – *Aufgang*
- The Knife
  – *Silent Shout*
- General Elektriks
  – *To Be a Stranger*
- Twenty One Pilot
  – *Blurryface*
- There's A Light
  – *A Long Lost Silence*
- Corpo-Mente
  – *Corpo-Mente*
- Igorrr
  – *Sinusoid Savage*
- John Frusciante
  – *A Sphere In The Heart Of Silence*
- Sister Iodine
  – *Flame Desastre*
- Metronomy
  – *The English Riviera*

Second Ph.D. Year

- Beyond Creation
  – *The Aura*
- Venetian Snares
  – *Traditional Synthesizer Music*
- Plaid
  – *The Digging Remedy*
- Hannes Grossmann
  – *The Crypts of Sleep*
- Thylacine
  – *Transsiberian*
- Baroness
  – *Purple*
- Princess Nokia
  – *A Girl Cried Red*
- Ruby My Dear
  – *Brame*
- Devin Townsend
  – *Empath*
- Tatran
  – *Shvat*
- Sunn O)))
  – *White2*
- The Gaslamp Killer
  – *Instrumentalepathy*
- R̈ic̈nn
  – *L̈ian*
- Öxxö XööX
  – *Nämïdää*
- Rosalía
  – *El Mal Querer*

Third Ph.D. Year

- Author & Punisher
  – *Beastland*
- Polo & Pan
  – *Caravelle*
- The Claypool Lennon Delirium
  – *South of Reality*
- Dark Funeral
  – *Diabolis Interium*
- Zeal & Ardor
  – *Devil Is Fine*
- Ruby My Dear
  – *Stranger in Paradise*
- Tool
  – *Fear Inoculum*
- Giobia
  – *The Magnifier*
- Edoardo Maggiolo
  – *Mappe.*
- Coldplay
  – *X&Y*
- Childish Gambino
  – *"Awaken, My Love!"*

Thesis redaction

- Death
  – *Individual Thought Patterns*
- The Faceless
  – *In Becoming a Ghost*
- Foehn Trio
  – *Highlines*
- Mandela
  – *The Sound of Grass*
- Moderat
  – *Modelar*
- Rings of Saturn
  – *Ultu Ulla*
- Obscura
  – *Cosmogenesis*
- Mort Garson
  – *Mother Earth's Plantasia*
- Caroline Lavelle
  – *Spirit*
- Wintergatan
  – *Wintergatan*
- Three Trapped Tigers
  – *Numbers: 1-13*
- Caroline Lavelle
  – *Spirit*

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. 57).

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. 154).

Ackerman, Evan (2018). "HARU: An Experimental Social Robot From Honda Research". In: *IEEE Specturm* (cit. on p. 133).

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 p. 112).

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. 61).

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. 57, 68).

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 pp. 55, 56).

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 (JASA)* 65.4, pp. 943–950 (cit. on pp. 23–25, 74, 116).

An, Inkyu, Myungbae Son, Dinesh Manocha, and Sung-eui Yoon (2018). "Reflection-aware sound source localization". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 66–73 (cit. on p. 128).

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. 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: *International Symposium on Communications, Control and Signal Processing*. IEEE, pp. 1–4 (cit. on pp. 52, 53, 145).

Antonacci, Fabio, Augusto Sarti, and Stefano Tubaro (2010). "Geometric reconstruction of the environment from its response to multiple acoustic emissions". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2822–2825 (cit. on p. 54).

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 (TASLP)* 20.10, pp. 2683–2695 (cit. on pp. 54, 57, 96).

Aoshima, Nobuharu (1981). "Computer-generated pulse signal applied for sound measurement". In: *The Journal of the Acoustical Society of America (JASA)* 69.5, pp. 1484–1488 (cit. on p. 52).

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. 109).

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. 110).

Asaei, Afsaneh, Mohammad Golbabaee, Herve Bourlard, and Volkan Cevher (2014). "Structured sparsity models for reverberant speech separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 22.3, pp. 620–633 (cit. on p. 116).

Avargel, Yekutiel and Israel Cohen (2007). "System identification in the short-time Fourier transform domain with crossband filtering". In: *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)* 15.4, pp. 1305–1319 (cit. on p. 40).

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. 96).

Badeau, Roland (2019). "Common mathematical framework for stochastic reverberation models". In: *The Journal of the Acoustical Society of America (JASA)* 145.4, pp. 2733–2745 (cit. on pp. 22–24, 153).

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. 154).

Beck, Amir, Petre Stoica, and Jian Li (2008). "Exact and approximate solutions of source localization problems". In: *IEEE Transactions on Signal Processing (TSP)* 56.5, pp. 1770–1778. ISSN: 1053587X. DOI: 10.1109/TSP.2007.909342 (cit. on pp. 95, 144).

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. 52, 53).

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. 89).

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. 116).

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. 60).

Bilen, Çağdaş, Alexey Ozerov, and Patrick Pérez (2015). "Joint audio inpainting and source separation". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 251–258 (cit. on p. 89).

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 (TASLP)* 28, pp. 279–293 (cit. on p. 128).

Bishop, Christopher M (1994). "Mixture density networks". In: (cit. on p. 86).

— (2006). *Pattern recognition and machine learning*. springer (cit. on p. 77).

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. 111, 129, 130).

Böck, Sebastian, Florian Krebs, and Markus Schedl (2012). "Evaluating the Online Capabilities of Onset Detection Methods." In: *International Symposium on Music Information Retrieval (ISMIR)*, pp. 49–54 (cit. on pp. 62, 73).

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 (TSP)* 64.6, pp. 1405–1419 (cit. on p. 153).

Bradley, John S, Hiroshi Sato, and Michel Picard (2003). "On the importance of early reflections for speech in rooms". In: *The Journal of the Acoustical Society of America (JASA)* 113.6, pp. 3233–3244 (cit. on pp. 137, 142).

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. 72, 157).

Bregman, Albert S (1990). "Auditory scene analysis". In: *McAdams and Bigand, editors, Thinking in Sound*, pp. 10–36 (cit. on pp. 3, 43, 129).

Brown, Guy J and DeLiang Wang (2005). "Separation of speech by computational auditory scene analysis". In: *Speech Enhancement*. Springer, pp. 371–402 (cit. on p. 43).

Bryan, Nicholas J (2020). "Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation". In: *SIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 154).

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. 66, 71).

Capon, Jack (1969). "High-resolution frequency-wavenumber spectrum analysis". In: *Proceedings of the IEEE* 57.8, pp. 1408–1418 (cit. on p. 109).

Cardoso, J-F (1998). "Blind signal separation: statistical principles". In: *Proceedings of the IEEE* 86.10, pp. 2009–2025 (cit. on p. 106).

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. 138).

Cecchi, Stefania, Alberto Carini, and Sascha Spors (2018). "Room response equalization—A review". In: *Applied Sciences* 8.1, p. 16 (cit. on p. 108).

Chakrabarty, Soumitro and Emanuël AP Habets (2017). "Broadband DOA estimation using convolutional neural networks trained with noise signals". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 136–140 (cit. on pp. 78, 85, 112).

Chazan, Shlomo E, Jacob Goldberger, and Sharon Gannot (2018). "DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6712–6716 (cit. on p. 110).

Chazan, Shlomo E, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot (2019). "Multi-microphone speaker separation based on deep DOA estimation". In: *European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1–5 (cit. on p. 110).

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. 84, 112, 127, 129, 130).

Cheng, Tian, Matthias Mauch, Emmanouil Benetos, Simon Dixon, et al. (2016). "An attack/decay model for piano transcription". In: International Symposium on Music Information Retrieval (ISMIR) (cit. on pp. 52, 53).

Cherry, Colin (1953). "Cocktail party problem". In: *Journal of the Acoustical Society of America (JASA)* 25, pp. 975–979 (cit. on p. 105).

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 (TSP)* 59.5, pp. 2182–2195 (cit. on p. 58).

Čmejla, Jaroslav, Tomáš Kounovskỳ, Sharon Gannot, Zbyněk Koldovskỳ, 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. 58, 60, 92).

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. 110).

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. 53, 59).

— (2015). "Cadzow denoising upgraded: A new projection method for the recovery of Dirac pulses from noisy linear measurements". In: (cit. on pp. 49, 59).

Cox, Henry, Robertm Zeskind, and Markm Owen (1987). "Robust adaptive beamforming". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)* 35.10, pp. 1365–1376 (cit. on p. 109).

Crocco, 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. 57, 68, 74, 76).

— (2016). "Estimation of TDOA for room reflections by iterative weighted l 1 constraint". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3201–3205 (cit. on pp. 57, 58, 65, 68, 73, 95, 111).

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. 58, 144).

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. 54–56).

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, 49, 52–54, 57, 58, 60, 144, 145).

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. 71).

Défossez, Alexandre, Nicolas Usunier, Léon Bottou, and Francis Bach (2019). "Music Source Separation in the Waveform Domain". In: *arXiv preprint arXiv:1911.13254* (cit. on p. 107).

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. 52, 53).

— (2008b). "Finding the onset of a room impulse response: Straightforward?" In: *The Journal of the Acoustical Society of America (JASA)* 124.4, EL248–EL254 (cit. on pp. 52, 58, 97).

Deleforge, Antoine and Florence Forbes (2016). "Rectified binaural ratio: A complex T-distributed feature for robust sound localization". In: *24th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1257–1261 (cit. on p. 87).

Deleforge, Antoine, Florence Forbes, and Radu Horaud (2015a). "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, 78, 82, 85, 89, 112).

Deleforge, Antoine, Radu Horaud, Yoav Y Schechner, and Laurent Girin (2015b). "Co-localization of audio sources in images using binaural features and locally-linear regression". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 23.4, pp. 718–731 (cit. on p. 78).

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. 66, 72, 157).

Di Carlo, Diego, Ken Déguernel, and Antoine Liutkus (2017). "Gaussian framework for interference reduction in live recordings". In: *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society (cit. on p. 108).

Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019). "Mirage: 2D source localization using microphone pair augmentation with echoes". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 775–779. URL: https://hal.archives-ouvertes.fr/hal-02160940v1 (cit. on pp. 42, 59, 75, 77, 85, 86, 127, 132).

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: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 156–160. URL: https://hal.archives-ouvertes.fr/hal-02469901 (cit. on pp. 59, 65, 72, 157).

DiBiase, Joseph, Harvey Silverman, and Michael Brandstein (2001). "Robust localization in reverberant rooms". In: *Microphone Arrays*. Springer, pp. 157–180 (cit. on pp. 56, 58, 81, 111, 112, 130).

Dmochowski, Jacek P, Jacob Benesty, and Sofiene Affes (2007). "Broadband MUSIC: Opportunities and challenges for multiple source localization". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 18–21 (cit. on p. 111).

Doclo, Simon and Marc Moonen (2002). "GSVD-based optimal filtering for single and multimicrophone speech enhancement". In: *IEEE Transactions on Signal Processing (TSP)* 50.9, pp. 2230–2244 (cit. on pp. 42, 110).

— (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. 141).

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, 54, 60, 144).

Dokmanić, Ivan, Robin Scheibler, and Martin Vetterli (2015). "Raking the cocktail party". In: *IEEE journal of selected topics in signal processing* 9.5, pp. 825–836 (cit. on pp. 5, 57, 110, 116, 117, 136, 137, 141).

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 (cit. on pp. 95, 144).

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. 52).

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 (TASLP)* 18.7, pp. 1830–1840 (cit. on pp. 56, 57, 108).

Eaton, James, Nikolay D Gaubitch, Alastair H Moore, and Patrick A Naylor (2015). "The ACE challenge—Corpus description and performance evaluation". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 1–5 (cit. on p. 142).

Eaton, James, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor (2016). "Estimation of Room Acoustic Parameters: The ACE Challenge". In: *IEEE/ACM Transactions on Audio Speech and Language Processing (TASLP)* 24, pp. 1681–1693. ISSN: 23299290. DOI: 10.1109/TASLP.2016.2577502 (cit. on p. 97).

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. 54, 55).

Ernst, Ori, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger (2018). "Speech dereverberation using fully convolutional networks". In: *European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 390–394 (cit. on p. 110).

Evers, Christine and Patrick A Naylor (2018). "Acoustic slam". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 26.9, pp. 1484–1498 (cit. on p. 128).

Evers, Christine, Heinrich Loellmann, Heinrich Mellmann, Alexander Schmidt, Hendrik Barfuss, Patrick A Naylor, and Walter Kellermann (2020). "The LOCATA challenge: Acoustic source localization and tracking". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* (cit. on p. 112).

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 (TASLP)* 18.7, pp. 1766–1774 (cit. on p. 142).

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. 52).

— (2007). "Advancements in impulse response measurements by sine sweeps". In: *Audio Engineering Society Convention 122*. Audio Engineering Society (cit. on pp. 52, 94).

Ferguson, Eric L, Stefan B Williams, and Craig T Jin (2019). "Improved multipath time delay estimation using cepstrum subtraction". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 551–555 (cit. on p. 53).

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. 119).

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: *European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 161–165 (cit. on p. 54).

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, 110, 136).

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. 109).

Gannot, Sharon, David Burshtein, and Ehud Weinstein (2001). "Signal enhancement using beamforming and nonstationarity with applications to speech". In: *IEEE Transactions on Signal Processing (TSP)* 49.8, pp. 1614–1626 (cit. on pp. 41, 42, 58, 110, 136, 138, 140).

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 (TASLP)* 25.4, pp. 692–730 (cit. on pp. 106–108, 139, 140).

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 pp. 74, 84, 122).

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. 77, 78, 82, 84, 112, 134).

Genovese, Andrea F, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J Tashev (2019). "Blind room volume estimation from single-channel noisy speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 231–235 (cit. on pp. 59, 60).

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 (TSP)* 40.ARTICLE, pp. 1862–1875 (cit. on p. 40).

Girin, Laurent, Sharon Gannot, and Xiaofei Li (2019). "Audio source separation into the wild". In: *Multimodal Behavior Analysis in the Wild*. Elsevier, pp. 53–78 (cit. on pp. 40, 106, 107).

Glentis, G-O, Kostas Berberidis, and Sergios Theodoridis (1999). "Efficient least squares adaptive algorithms for FIR transversal filtering". In: *IEEE signal processing magazine* 16.4, pp. 13–41 (cit. on p. 52).

Gomez, Randy, Deborah Szapiro, Kerl Galindo, and Keisuke Nakamura (2018). "HARU: Hardware design of an experimental tabletop robot assistant". In: *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 233–240 (cit. on p. 133).

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative adversarial nets". In: *Advances in neural information processing systems (NIPS)* 27, pp. 2672–2680 (cit. on p. 86).

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. MIT press Cambridge (cit. on pp. 77, 79).

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. 53).

Habets, Emanuël AP and Sharon Gannot (2007). "Generating sensor signals in isotropic noise fields". In: *The Journal of the Acoustical Society of America (JASA)* 122.6, pp. 3464–3470 (cit. on p. 32).

Habets, Emanuel (2006). *Room impulse response generator*. Tech. rep. (cit. on p. 25).

Hadad, Elior, Florian Heese, Peter Vary, and Sharon Gannot (2014). "Multichannel audio database in various acoustic environments". In: *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 313–317. ISBN: 9781479968084. DOI: 10.1109/IWAENC.2014.6954309 (cit. on p. 91).

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: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35 (cit. on p. 107).

Higuchi, Takuya, Norihiro Takamune, Tomohiko Nakamura, and Hirokazu Kameoka (2014). "Underdetermined blind separation and tracking of moving sources based ONDOA-HMM". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3191–3195 (cit. on p. 108).

Huang, Yiteng and Jacob Benesty (2003). "A class of frequency-domain adaptive approaches to blind multichannel identification". In: *IEEE Transactions on Signal Processing (TSP)* 51.1, pp. 11–24 (cit. on pp. 56, 61).

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. 54).

Jan, E, Piergiorgio Svaizer, and James L Flanagan (1995). "Matched-filter processing of microphone array for spatial volume selectivity". In: *International Symposium on Circuits and Systems (ISCAS)*. Vol. 2. IEEE, pp. 1460–1463 (cit. on pp. 5, 110, 136).

Jan, Ea-Ee and James Flanagan (1996). "Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 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: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 111–115 (cit. on p. 136).

Jensen, Jesper Rindom, Usama Saqib, and Sharon Gannot (2019). "An EM method for multichannel TOA and DOA estimation of acoustic echoes". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 120–124 (cit. on pp. 55, 65, 153).

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. 53).

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. 89, 154).

Karpatne, Anuj, William Watkins, Jordan Read, and Vipin Kumar (2017). "Physics-guided neural networks (pgnn): An application in lake temperature modeling". In: *arXiv preprint arXiv:1710.11431* (cit. on pp. 89, 154).

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 (TASLP)* 22.7, pp. 1139–1147 (cit. on pp. 52, 53).

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (cit. on p. 81).

Kingma, Diederik and Max Welling (2014). "Auto-encoding variational Bayes." In: *International Conference on Learning Representations (ICLR)* (cit. on p. 86).

Kitic, Srdan and Alexandre Guérin (2018). "TRAMP: Tracking by a realtime Ambisonic-based particle filter". In: *Proc. of LOCATA Challenge Workshop - a satellite event of IWAENC* (cit. on p. 112).

Kleeman, Lindsay and Roman Kuc (2016). "Sonar sensing". In: *Springer Handbook of Robotics*. Springer, pp. 753–782 (cit. on p. 154).

Knapp, Charles and Glifford Carter (1976). "The generalized correlation method for estimation of time delay". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)* 24.4, pp. 320–327 (cit. on pp. 111, 129).

Kodrasi, Ina and Simon Doclo (2017). "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods". In: *Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, pp. 116–120 (cit. on pp. 42, 58, 136, 141, 153).

Kokkinis, Elias K, Joshua D Reiss, and John Mourjopoulos (2011). "A Wiener filter approach to microphone leakage reduction in close-microphone applications". In: *IEEE transactions on audio, speech, and language processing* 20.3, pp. 767–779 (cit. on p. 108).

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. 42, 58).

Koldovskỳ, 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 (TASLP)* 23.8, pp. 1335–1347 (cit. on p. 58).

Korhonen, Teemu (2008). "Acoustic localization using reverberation with virtual microphones". In: *International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Citeseer, pp. 211–223 (cit. on pp. 116, 127).

Kowalczyk, Konrad (2019). "Raking early reflection signals for late reverberation and noise reduction". In: *The Journal of the Acoustical Society of America (JASA)* 145.3, EL257–EL263. ISSN: 0001-4966. DOI: 10.1121/1.5095535. URL: http://dx.doi.org/10.1121/1.5095535 (cit. on pp. 5, 110, 136, 137, 140, 141, 153).

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. 57, 68).

Kreković, Miranda, Ivan Dokmanić, and Martin Vetterli (2016). "EchoSLAM: Simultaneous localization and mapping with acoustic echoes". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 11–15 (cit. on p. 128).

Krim, Hamid and Mats Viberg (1996). "Two decades of array signal processing research: the parametric approach". In: *IEEE signal processing magazine* 13.4, pp. 67–94 (cit. on p. 52).

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 (JASA)* 124.2, pp. 982–993 (cit. on pp. 52, 53).

— (2012). "Objective sound field analysis based on the coherence estimated from two microphone signals". In: *The Journal of the Acoustical Society of America (JASA)* 131.4, pp. 3284–3284 (cit. on p. 138).

Kuttruff, Heinrich (2016). *Room acoustics*. CRC Press (cit. on pp. 13, 16, 18, 52, 138).

Laufer, Bracha, Ronen Talmon, and Sharon Gannot (2013). "Relative transfer function modeling for supervised source localization". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 1–4 (cit. on pp. 42, 112).

Le Roux, Jonathan, John R Hershey, and Felix Weninger (2015). "Deep NMF for speech separation". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 66–70 (cit. on p. 108).

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*. URL: https://arxiv.org/abs/1812.05901 (cit. on pp. 111, 152).

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. 119).

Leglaive, Simon, Roland Badeau, and Gaël Richard (2015). "Multichannel audio source separation with probabilistic reverberation modeling". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 1–5 (cit. on pp. 108, 116, 153).

— (2016). "Multichannel audio source separation with probabilistic reverberation priors". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 24.12, pp. 2453–2465 (cit. on pp. 5, 56, 57, 116).

— (2018). "Student's t source and mixing models for multichannel audio source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 26.6, pp. 1154–1168 (cit. on pp. 56, 57).

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. 110).

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 (TASLP)* 24.11, pp. 2171–2186 (cit. on pp. 112, 127).

Li, Xiaofei, Laurent Girin, and Radu Horaud (2019a). "Expectation-maximisation for speech source separation using convolutive transfer function". In: *CAAI Transactions on Intelligence Technology* 4.1, pp. 47–53 (cit. on p. 108).

Li, Xiaofei, Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud (2019b). "Online localization and tracking of multiple moving speakers in reverberant environments". In: *IEEE Journal of Selected Topics in Signal Processing* 13.1, pp. 88–103 (cit. on p. 112).

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. 57, 67, 68, 73).

— (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. 57, 68, 69, 76).

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: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 431–435 (cit. on p. 153).

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. 53).

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 (TASLP)* 27.8, pp. 1256–1266 (cit. on p. 107).

Makino, Shoji (2018). *Audio Source Separation*. Vol. 433. Springer (cit. on pp. 106, 115).

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 (TASLP)* 17.6, pp. 1071–1086 (cit. on pp. 42, 110).

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. 61).

Müller, Meinard (2015). *Fundamentals of Music Processing*. Springer Verlag. ISBN: 978-3-319-21944-8 (cit. on pp. 39, 120).

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. 89, 134, 154).

Nadisic, Nicolas, Arnaud Vandaele, Nicolas Gillis, and Jeremy E Cohen (2020). "Exact Sparse Nonnegative Least Squares". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5395–5399 (cit. on p. 153).

Nakashima, Hiromichi, Mitsuru Kawamoto, and Toshiharu Mukai (2010). "A localization method for multiple sound sources by using coherence function". In: *European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 130–134 (cit. on p. 127).

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 (TASLP)* 15.1, pp. 34–43 (cit. on p. 53).

Naylor, Patrick and Eds Gaubitch (2010). *Dereverberation, Speech*. Berlin, Germany: Springer-Verlag (cit. on p. 138).

Neely, Stephen T and Jont B Allen (1979). "Invertibility of a room impulse response". In: *The Journal of the Acoustical Society of America (JASA)* 66.1, pp. 165–169 (cit. on p. 108).

Nesta, Francesco and Maurizio Omologo (2012). "Convolutive 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. 108).

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: *Workshop on Crossmodal Learning for Intelligent Robotics in conjunction with IEEE/RSJ IROS* (cit. on pp. 78, 85, 112).

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 (TASLP)* 24.9, pp. 1652–1664 (cit. on p. 107).

O'Donovan, Adam E, Ramani Duraiswami, and Dmitry N Zotkin (2010). "Automatic matched filter recovery via the audio camera". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2826–2829 (cit. on pp. 56, 58).

O'Donovan, Adam, Ramani Duraiswami, and Dmitry Zotkin (2008). "Imaging concert hall acoustics using visual and audio cameras". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5284–5287 (cit. on pp. 56, 58).

O'grady, Paul D, Barak A Pearlmutter, and Scott T Rickard (2005). "Survey of sparse and non-sparse methods in source separation". In: *International Journal of Imaging Systems and Technology* 15.1, pp. 18–33 (cit. on p. 106).

Oppenheim, Alan V (1987). *Signals and Systems: An Introduction to Analog and Digital Signal Processing*. MIT Center for Advanced Engineering Study (cit. on pp. 36, 39).

Ozerov, Alexey and Cédric Févotte (2010). "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation". In: *IEEE Transaction in Audio, Speech and Language Processing (TASLP)* 18.3, pp. 550–563 (cit. on pp. 7, 56, 57, 106–108, 115, 118–123, 152).

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 (TSP)* 65.4, pp. 821–835 (cit. on p. 154).

Parhizkar, Reza, Ivan Dokmanić, and Martin Vetterli (2014). "Single-channel indoor microphone localization". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1434–1438 (cit. on p. 54).

Park, Yongsung, Woojae Seong, and Youngmin Choo (2017). "Compressive time delay estimation off the grid". In: *The Journal of the Acoustical Society of America (JASA)* 141.6, EL585–EL591 (cit. on p. 65).

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. 94).

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 (JASA)* 139.5, EL113–EL117 (cit. on p. 53).

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. 122).

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 (TALSP)* 21.12, pp. 2532–2540 (cit. on pp. 110, 136).

Perotin, Lauréline, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin (2018). "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector". In: *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, pp. 241–245 (cit. on pp. 78, 112).

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. 95).

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. 77).

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. 58).

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: *International Symposium on Music Information Retrieval (ISMIR)* (cit. on p. 123).

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. 89, 154).

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. 110, 154).

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 (TASLP)* 25.2, pp. 296–309 (cit. on pp. 5, 52, 53, 60, 144).

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 (TASLP)* 27.12, pp. 2263–2277 (cit. on pp. 5, 60).

Rezende, Jimenez Danilo, Shakir Mohamed, and Daan Wierstra (2014). "Stochastic backpropagation and approximate inference in deep generative models". In: *International Conference on Machine Learning (ICML)* (cit. on p. 86).

Ribeiro, Flávio, Demba Ba, Cha Zhang, and Dinei Florêncio (2010a). "Turning enemies into friends: Using reflections to improve sound source localization". In: *IEEE International Conference on Multimedia and Expo*. IEEE, pp. 731–736 (cit. on pp. 5, 49, 57, 111, 127).

Ribeiro, Flavio, 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 (TASLP)* 18.7, pp. 1781–1792 (cit. on p. 127).

Rickard, Scott (2007). "The DUET blind source separation algorithm". In: *Blind Speech Separation*, pp. 217–241 (cit. on pp. 107, 108).

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 (JASA)* 133.4, EL235–EL241 (cit. on p. 53).

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. 142).

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 (TASSP)* 34.5, pp. 1340–1342 (cit. on p. 53).

Rudin, Walter (1987). "Real and complex analysis (mcgraw-hill international editions: Mathematics series)". In: (cit. on p. 71).

Rui, Yong and Dinei Florencio (2004). "Time delay estimation in the presence of correlated noise and reverberation". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 2. IEEE, pp. ii–133 (cit. on p. 127).

Sacks, Oliver (2014). *Musicofilia*. 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 (TASLP)* 25.5, pp. 965–979 (cit. on pp. 78, 110).

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. 54, 127).

— (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. 78, 110).

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 (TASLP)* 22.12, pp. 2197–2206 (cit. on p. 142).

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. 55).

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. 154).

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. 56).

Savioja, Lauri and U Peter Svensson (2015). "Overview of geometrical room acoustic modeling techniques". In: *The Journal of the Acoustical Society of America (JASA)* 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 (TASLP)* 21.5, pp. 971–982 (cit. on pp. 56, 57, 107).

Scheibler, Robin, Ivan Dokmanić, and Martin Vetterli (2015). "Raking echoes in the time domain". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 554–558 (cit. on pp. 109, 110, 136).

Scheibler, Robin, Eric Bezzam, and Ivan Dokmanić (2018a). "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. accepted. Calgary, CA (cit. on p. 122).

— (2018b). "Pyroomacoustics: A python package for audio room simulation and array processing algorithms". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 351–355 (cit. on p. 74).

Scheibler, Robin, Diego Di Carlo, Antoine Deleforge, and Ivan Dokmanić (2018c). "Separake: Source separation with a little help from echoes". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901. URL: https://arxiv.org/abs/1711.06805 (cit. on pp. 5, 56, 75, 115–119, 121, 123).

Scheibler, Robin, **Di Carlo, Diego**, Antoine Deleforge, and Ivan Dokmanić (2018d). "Separake: Source separation with a little help from echoes". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901 (cit. on p. 9).

Scheuing, Jan and Bin Yang (2006). "Disambiguation of TDOA estimates in multi-path multi-source environments (DATEMM)". In: *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Vol. 4. IEEE, pp. IV–IV (cit. on p. 54).

Schimmel, Steven M, Martin F Muller, and Norbert Dillier (2009). "A fast and accurate "shoebox" room acoustics simulator". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 241–244 (cit. on pp. 23, 84, 133).

Schmidt, Mikkel N and Rasmus K Olsson (2006). "Single-channel speech separation using sparse non-negative matrix factorization". In: *International Conference on Spoken Language Processing* (cit. on pp. 107, 120).

Schröder, Dirk, Philipp Dross, and Michael Vorländer (2007). "A fast reverberation estimator for virtual environments". In: *Audio Engineering Society 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 (JASA)* 66.2, pp. 497–500 (cit. on pp. 51, 52).

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 (TASLP)* 23.2, pp. 240–251 (cit. on pp. 110, 136, 140).

— (2016). "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm". In: *European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1123–1127 (cit. on pp. 110, 141, 153).

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 (NIPS)*, pp. 1705–1713 (cit. on pp. 107, 120).

Sokol, Joshua (2017). "The thoughts of a spiderweb". In: *Quanta Magazine* (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. 112).

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).

Stöter, Fabian-Robert, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji (2019). "Open-Unmix - A Reference Implementation for Music Source Separation". In: *Journal of Open Source Software* 4.41, p. 1667. DOI: 10.21105/joss.01667. URL: https://doi.org/10.21105/joss.01667 (cit. on p. 107).

Sun, D L and G J Mysore (2013). "Universal speech models for speaker independent single channel source separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 141–145 (cit. on pp. 119, 122).

Svaizer, Piergiorgio, Alessio Brutti, and Maurizio Omologo (2011). "Use of reflected wavefronts for acoustic source localization with a line array". In: *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*. IEEE, pp. 165–169 (cit. on p. 127).

Szöke, Igor, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Honza Černockỳ (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. 52, 59, 60).

Taghizadeh, Mohammad Javad, Afsaneh Asaei, Saeid Haghighatshoar, 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. 127).

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. 109).

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. 42, 61, 138, 141, 153).

Tervo, Sakari (2011). "Localization and tracing of early acoustic reflections". PhD thesis (cit. on pp. 56, 144).

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 (TASLP)* 23.10, pp. 1539–1551 (cit. on p. 56).

Tervo, Sakari, Teemu Korhonen, and Tapio Lokki (2011). "Estimation of reflections from impulse responses". In: *Building Acoustics* 18.1-2, pp. 159–173 (cit. on p. 56).

Thiergart, Oliver and Emanuël AP Habets (2013). "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 659–663 (cit. on p. 110).

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 (TASLP)* 22.12, pp. 2182–2196 (cit. on p. 140).

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. 67).

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. 56, 57).

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. 57, 68).

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 (TASLP)* 27.12, pp. 2080–2091 (cit. on p. 89).

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. 59, 76).

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, 52, 59, 66).

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. 107).

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 (JASA)* 127.4, EL172–EL177 (cit. on pp. 52, 53).

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: 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. 108, 136).

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 p. 109).

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. 55).

Varzandeh, Reza, Maja Taseska, and Emanuël AP Habets (2017). "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation". In: *Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, pp. 11–15 (cit. on p. 61).

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 (TSP)* 60.10, pp. 5464–5477 (cit. on p. 54).

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. 55).

Vesa, Sampo (2009). "Binaural sound source distance learning in rooms". In: *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)* 17.8, pp. 1498–1507 (cit. on p. 111).

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. 53).

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. 78, 112).

Vetterli, Martin, Pina Marziliano, and Thierry Blu (2002). "Sampling signals with finite rate of innovation". In: *IEEE Transactions on Signal Processing (TSP)* 50.6, pp. 1417–1428 (cit. on p. 59).

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. 122, 123).

Vincent, Emmanuel, Maria G Jafari, Samer A Abdallah, Mark D Plumbley, and Mike E Davies (2011). "Probabilistic modeling paradigms for audio source separation". In: *Machine Audition: Principles, Algorithms and Systems*. IGI global, pp. 162–185 (cit. on p. 106).

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. 105).

Vincent, Emmanuel, Tuomas Virtanen, and Sharon Gannot (2018). *Audio source separation and speech enhancement*. John Wiley & Sons (cit. on pp. 20, 29, 39, 40, 103, 106, 108, 110, 115).

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: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 107).

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 (TSP)* 42.4, pp. 846–859 (cit. on p. 127).

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. 78, 110).

Xu, Guanghan, Hui Liu, Lang Tong, and Thomas Kailath (1995). "A least-squares approach to blind channel identification". In: *IEEE Transactions on Signal Processing (TSP)* 43.12, pp. 2982–2993 (cit. on pp. 56, 57, 65).

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 (TALSP)* 20.2, pp. 661–675 (cit. on p. 68).

Zahorik, Pavel (2002). "Direct-to-reverberant energy ratio sensitivity". In: *The Journal of the Acoustical Society of America (JASA)* 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: *IEEE International Symposium on Circuits and Systems*. IEEE, pp. 2666–2669 (cit. on p. 54).

Zhang, Cha, Zhengyou Zhang, and Dinei Florêncio (2007). "Maximum likelihood sound source localization for multiple directional microphones". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. IEEE, pp. I–125 (cit. on p. 127).

Zohny, Zeinab and Jonathon Chambers (2014). "Modelling interaural level and phase cues with Student's t-distribution for robust clustering in MESSL". In: *19th International Conference on Digital Signal Processing*. IEEE, pp. 59–62 (cit. on p. 87).

Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019). "Mirage: 2D source localization using microphone pair augmentation with echoes". In: *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: *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 progess* (cit. on p. 9).

van denBoomgaard, Rein and Rik van derWeij (2001). "Gaussian convolutions numerical approximations based on interpolation". In: *Scale-Space and Morphology in Computer Vision*. Springer, pp. 205–214 (cit. on p. 37).

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : *Traitement du signal avec des echos acoustiques pour l'analyse des scènes audio*

Mot clés : Traitement des signaux audio, échos acoustiques, estimation des canaux aveugles, séparation de sources sonores, localisation de sources sonores, estimation de la géométrie d'une salle

Résumé : La plupart des méthodes de traitement du signal audio considèrent la réverbération et en particulier les échos acoustiques comme une nuisance. Cependant, ceux-ci transmettent des informations spatiales et sémantiques importantes sur les sources sonores et des méthodes essayant de les prendre en compte ont donc récemment émergé.. Dans ce travail, nous nous concentrons sur deux directions. Tout d'abord, nous étudions la manière d'estimer les échos acoustiques à l'aveugle à partir d'enregistrements microphoniques. Deux approches sont proposées, l'une s'appuyant sur le cadre des dictionnaires continus, l'autre sur des techniques récentes d'apprentissage profond. Ensuite, nous nous concentrons sur l'extension de méthodes existantes d'analyse de scènes audio à leurs formes sensibles à l'écho. Le cadre NMF multicanal pour la séparation de sources audio, la méthode de localisation SRP-PHAT et le formateur de voies MVDR pour l'amélioration de la parole sont tous étendus pour prendre en compte les échos. Ces applications montrent comment un simple modèle d'écho peut conduire à une amélioration des performances.

Title: *Echo-aware signal processing for audio scene analysis*

Keywords: Audio Signal Processing, Acoustic Echoes, Blind Channel Estimation, Sound Source Separation, Sound Source Localization, Room Geometry Estimation

Abstract: Most of audio signal processing methods regard reverberation and in particular acoustic echoes as a nuisance. However, they convey important spatial and semantic information about sound sources and, based on this, recent echo-aware methods have been proposed. In this work we focus on two directions. First, we study the how to estimate acoustic echoes blindly from microphone recordings. Two approaches are proposed, one leveraging on continuous dictionaries, one using recent deep learning techniques. Then, we focus on extending existing methods in audio scene analysis to their echo-aware forms. The Multichannel NMF framework for audio source separation, the SRP-PHAT localization method, and the MVDR beamformer for speech enhancement are all extended to their echo-aware versions.