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
Thèse N° : 88666

Rapporteurs avant soutenance :

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

Composition du Jury :

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

Abstract

Résumé en français

Acknowledgements

Contents

BLASTER	Blind and Sparse Technique for Echo Retrieval	10
RIR	Room Impulse Response	8
RTF	Room Transfer Function	19
TF	Time-Frequency	17
SSL	Sound Source Localization	21
RooGE	Room Geometry Estimation	17
AER	Acoustic Echo Retrieval	8
STFT	Short Time Fourier Transform	9
FFT	Fast Fourier Transform	
ReTF	Relative Transfer Function	20
TDOAs	Time Differences of Arrival	
AER	Acoustic Echo Retrieval	8
SSL	Sound Source Localization	21
RooGE	Room Geometry Estimation	17
DOA	Direction of Arrival	20
DOAs	Directions of Arrival	20
NMF	Nonnegative Matrix Factorization	18
CD	Continous Dictionary	10
DS	Delay-and-Sum	20
MVDR	Minimum-Variance-Distortionless-Response	20
LCMV	Linearly-Constrained-Minimum-Variance	20
MaxSINR	Maximum SINR	20
MaxSNR	Maximum SNR	20
NN	Neural Network	23
DNN	Deep Neural Network	21
GMM	Gaussian Mixture Model	23
GCC-PHAT	Generalized Cross Correlation with Phase Transform	23

Glossary:

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

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 number

Common indexing

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

Geometry

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

Signals

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

Acoustic

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

Mathematical Operation

$\star$	cross-correlation
$\circledast$	generalized cross-correlation
$*$	convolution

Examples

Acoustic Impulse Response for single source scenario:

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

Acoustic Transfer Function for single source scenario:

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

Time of Arrival between source and microphone

$$\tau_{ij} = \frac{\left\| \underline{\mathbf{x}}_i - \underline{\mathbf{s}}_j \right\|}{c_{\text{air}}} \tag{3}$$

PROLOGE

Part I

PROLOGUE

1

Overture

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

“Echoes shows the direction that we’re moving in.” —David Gilmour, about the making of “The Dark Side Of The Moon”

ECHOOOES ECHOOOES ECHOOOES

▸ 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 imaging sound, its constituents, and its generation. It is processed by our auditory system~~s~~ 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 leads us to conduct this process without any efforts, despite the presence of a huge level of background noise, for instance during a concert. This outstanding capability is not limited to humans and is common to all the creatures we are sharing the physical world.

Resources:

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

Nonetheless, we process *all* the information of the complex *acoustic scene* we are immersed into. In addition to the semantic content, a sound conveys also *temporal* and *spatial* information. For instance, the tickling of a metronome or clock provides units of time And when hearing someone shouting~~,~~ we unconsciously know where to turn our attention. Therefore, as for the content, this information is still carried by the sound. However, this information is determined by how sound *propagates* in t~~he~~ space and not in the source itself.

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

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

Some animals ~~are~~ evolved to “see” through echoes. ~~For instance, the~~ 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. As opposed to, in *passive* echo-location the ~~source~~ sound is not emitted, but rather only received. . “Locating

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. The French philosopher Denis Diderot in 18th century recorded ~~the~~ this incredible ability, which was labeled as “echo-location” only 300 years later by Donald Griffin.

it" means estimating its delay concerning to the direct sound. These delays are then processed as distances n the brain, in the same as 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 prey, see obstacles, and orientated in dark caves or ~~the~~ deep seas. However, the term "echo-location" here could be misleading as it may 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 change it in favor of *echo estimation*.

Remarkable examples of passive echo estimation in nature are not very known. Sand scorpions use the propagation of vibration 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 ~~to the~~ reverberation in their complex web[1]. They are not only able to localize the preys fast, but also identify them, and disambiguate from simple objects move by the wind or malicious visitors. In this case, instead of emitting sound, evolution taught them to uses complex structures (for scorpions their legs, for spiders webs) in order to feed and survive.

Echoes do not only serve for computing distances or localizing preys. For instance, they make speech more intelligible, provide music with "dimensionality"[Sacks 2014] and improve our sense of orientation and balancing. This phenomenon is material of studied in *room acoustics*, *pyschoacoustics* and *sound design*. In particular, the former study acoustic echoes for designing theatres, auditoriums, and meeting rooms, whose actual propose is to listen well.

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 turn, we will elaborate first the last two as they contextualize this thesis, immediately after, we will explain why and how echoes help.

1.1 Audio Signal Processing

Signal processing is the process of analyzing and modifying ~~a~~ *signals*, which are mathematical representations of quantities carrying information about a phenomenon. Then, *audio signal processing* represents the sound, such as music or speech, as signals and it involves applying various mathematical and algorithmic techniques to them. There are multiple reasons to do this, such as produce new signals with higher quality or and retrieve high-level information that the signal carries. To this end, complex systems are built which can be represented as a collection of simpler subsystems, with well-defined tasks, interacting with each other. In (audio) signal processing, these subsystems roughly fall into four categories: *representation*, *enhancement*, *estimation*, and *adaptive processing*. Many related problems can be then decomposed into blocks one or more of the following steps.

Audio is a more technical term, referring to sound coming from a recording, transmission, or electronic device. Acoustic, instead refer to the physical aspect of ~~the~~ sound. In this thesis, the two terms are used indistinctly.

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

- REPRESENTATION. The signal can be represented in many different ways, so that the *information* they contain becomes more suitable for specific tasks. It is generally implemented through change of *domain* or *feature*. In audio, the most famous representation is the Fourier basis, which changes the signal domain from time to frequencies.

- ENHANCEMENT. Measurements are affected by *noise*, which corrupts and hides the relevant information. Therefore, signal enhancement, namely, removing noise, is typically a necessary step. Examples of enhancement are removing background noise from a mobile phone recording or isolate 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, the frequency of a sound with respect to the background noise.

- ADAPTIVE PROCESSING. It deals with adaptive algorithms that are controlled by variable parameters resulting in previous estimation blocks. They usually rely on online optimization of objective function designed to meet specific requirements. Examples of these algorithmic are present in noise-canceling headphones or echo cancellation modules implemented in video conference call systems.

1.2 AUDIO SCENE ANALYSIS

Pay attention to what are you listening now: there might be music, someone talking to you, footsteps echoing in the other room, background noise due to cars, heating system, maybe rain or wind, the sound of your movement, and many others. Everything you hear now as well as its location in space is what is called the *audio scene*[2]. Therefore, ~~the~~ *audio scene analysis* is trivially the analysis of it. More specifically, the extraction and organization of all the information contained by the sound associated with an audio scene.

In audio signal processing, this process involves using algorithmic and mathematical tools to retrieve and organize such information. After recording the audio scene with microphones, complex systems, as described above, are used to access the information. Accessing different types of information at different levels of complexity leads to the definition of different *problems*. These problems focus on well-defined tasks ~~in the general audio scene analysis~~, and some are referred to with established names. Table 2.1 lists some selected audio scene analysis problems that will be considered later in this thesis.

Without going to philosophically, it is possible to re-cast these problems to some (simple) human interrogations ~~that~~ :

[2] 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.*

Thinking of the technologies behind Google Home and Amazon Alexa, one may wonder the ethical implication of audio scene analysis. During this thesis's work, these issues have resulted in discussions with colleagues and friends, but it will be discussed in another forum. Amazon Echo

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~~s~~ sources?
Microphone Calibration	... estimate the positions of the microphone position?
Room Geometry Estimation	... estimate the shape of the room?
Acoustic Echo Estimation	... estimate the echoes' properties?
Acoustic measurement	... estimate physical properties of the sound propagation?
Source Identification	... estimate~~d~~ the type of source signal?
Speech Diarization	... who is speaking and when ?
Source Counting	... count the number of speaker ?
Automatic Speech Recordings	... the content of the speech ?

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

- *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, by leveraging on the sound temporal information.

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

Our brain and the auditory system can instantly and effortlessly solve these problems, such that they may sound trivial tasks. However, they hide many difficult challenges when it comes to design efficient and robust algorithms. Moreover, most of these problems may exhibit strong inter-connections, and the solution of one of them depends on the solution of another. For instance, knowing when someone is speaking and its location in the room, sound source separation can be achieved more easily. It should not surprise and have strong parallelism with our everyday experience.

Finally, this is why echoes may help audio signal processing.

1.3 Echo-aware approach

As proven by natural behaviors, acoustic echoes are essential for human and animals for analyzing the audio scene: As repetition of a sound, they convey information about that sound. As characterized by temporal instant 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[3]. However, it was not always the case. Many audio scene analysis methods make strong assumptions of 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

[3] 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*.

nor reverberation is present in the audio scene. Even if this assumption can be seen as reasonable in some scenarios, it is easy to understand the underlying limitations when applied to real-world recordings. Furthermore, in some cases, they 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 their models, which we will refer to as *echo-aware methods*. One of the earliest examples in this direction are the works of Flanagan et al[Flanagan et al. 1993; Jan et al. 1995; Jan and Flanagan 1996] ~~for~~ in source enhancement. However, only recently, these methods have regained interest for audio processing as manifested by the European project SCENIC [Annibale et al. 2011] and the UK research S^3A project. In some recent studies, echoes are used boosts performances of typical audio scene analysis problems, e. g., speech enhancement [Dokmanić et al. 2015; Kowalczyk 2019], sound source localization [Ribeiro et al. 2010], and separation [Scheibler et al. 2018a; Leglaive et al. 2016; Remaggi et al. 2019], and room geometry estimation from sound [Remaggi et al. 2016; Dokmanić et al. 2013; Crocco et al. 2017],

All these methods show the importance and the benefits of modeling acoustic reflection; however, prior to all them is the Acoustic Echo Retrieval (AER). This step, which is typically given for granted in the above application, is extremely challenging, as shown throughout this entire 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 acoustic echoes and surpass~~ing~~ the limits of current approaches.
2. Extend previous classical methods for audio scene analysis by incorporating the knowledge of these elements of the sound propagation.

These two claims are elaborated in the two main part of the thesis which follow after an introductory one, as summarized below. However the parts are largely interconnected, as show in ~~the~~ Figure 1.1:

▶ Room Acoustic meets Signal Processing

?? This chapter will build a first important bridge: from acoustics to audio signal processing. It first defines sound and how it propagates in the environment ??, teasing out the fundamental concepts of this thesis: ~~the~~ echoes~~.~~ ?? and ~~the~~ Room Impulse Response (RIR) ??. By assuming some approximations, the RIR will be described in all its parts related to methods to compute them. Finally, in ??, how the human auditory system perceives reverberation will be reported.

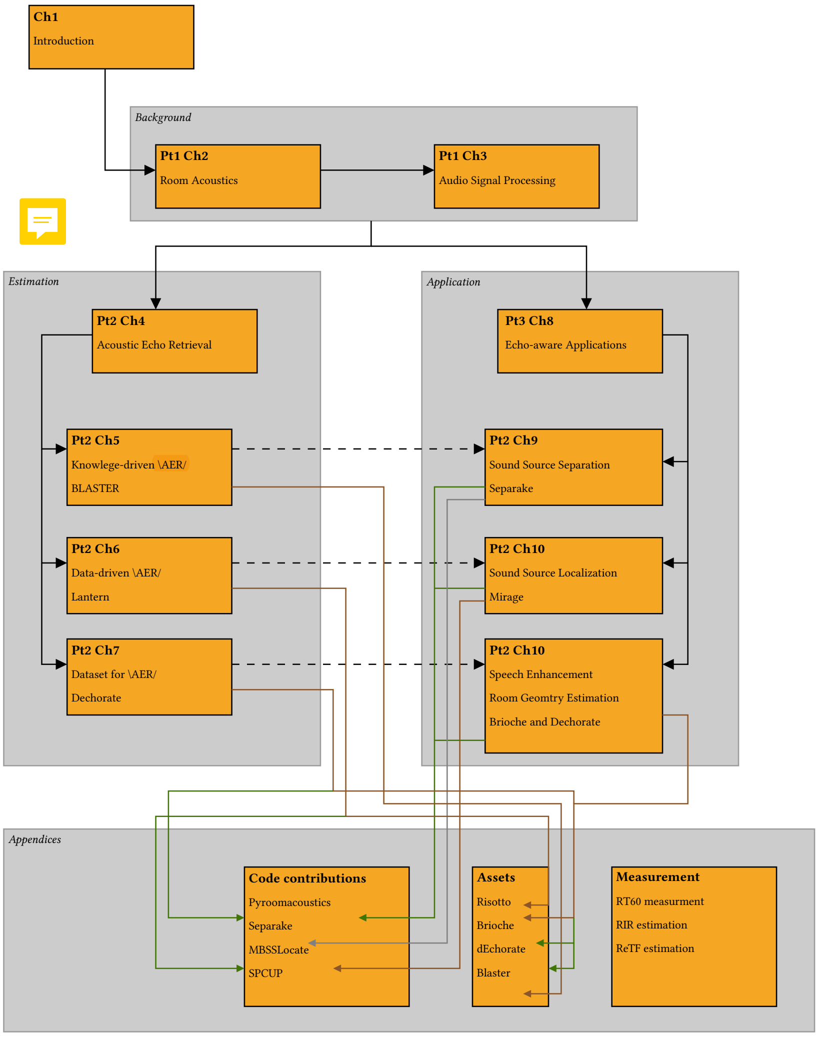

FIGURE 1.1: Schematic rganization of the thesis, dependecies between chapters linked to author contributions and producted assets.

?? Let us now move from physics to digital signal processing. At first in ??, this chapter formalizes fundamental concepts of audio signal processing such as signal, mixtures, and noise in the time domain. In ??, we will present the signal representation that we will use throughout the entire thesis: the Short Time Fourier Transform (STFT). Finally, ~~after assuming the narrowband approximation~~, in ??, some essential models for the Room Impulse Response (RIR) are described.

▸ ACOUSTIC ECHOES ESTIMATION

This part focuses on how to estimate~~d~~ echoes for the only observation of microphone recordings.

?? This chapter aims to provide the reader with knowledge of the state-of-the-art of Acoustic Echo Retrieval (AER). After presenting the AER problem in ??, the chapter is divided into three main sections: ?? defines the categories of methods thank to which the literature can be clustered and analyzed in detail later in ??. Finally, in ?? some datasets and evaluation metrics for AER are presented.

?? This chapter

?? This chapter proposes a novel approach for *off-grid* AER from a stereophonic recording of an unknown sound source such as speech. In contrast with existing methods, the proposed approach, named Blind and Sparse Technique for Echo Retrieval (BLASTER). ~~It is~~ built 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.

▶ Echo-aware Audio Scene Analysis

Chapter 2 In this chapter, we will present algorithms and methodologies for audio scene analysis in the context of signal processing. At first, in section § 2.1, we present a typical scenario for defining some cardinal problems. Therefore in section § 2.2, state-of-the-art approaches to address these problems are listed and commented, highlighting the relationship with some acoustic propagation models. The content presented here serves as a basis for a deeper investigation conducted in each of the following chapters.

?? In this chapter, echoes are used for boosting the performance of classical Audio Source Separation methods. At first, we describe existing methods that either ignore the acoustic propagation or attempt to estimate it fully. Instead, these works investigate whether sound separation can benefit from the knowledge of early acoustic echoes derived from the known locations of a few *image microphones*. The improvements are shown for two variants of a method based on non-negative matrix factorization: one that uses only magnitudes of the transfer functions and uses the phases. 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.

?? This chapter

?? 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 (**??**) and room geometry estimation (**??**). This investigation is conducted using state-of-the-art algorithms described and contextualized in the corresponding sections. In the final section (**??**), the main results are summarized, and future perspectives will be presented.

▶ Finally, the dissertation concludes with **??**, which summarizes the contributions and raises several additional research questions.

1.5 List of Contribution

This dissertation draws heavily on the ~~earlier~~ work and writing in 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: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 156–160
- Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019). "Mirage: 2d source localization using microphone pair augmentation with echoes". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 775–779
- Deleforge, Antoine, Diego Di Carlo, 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, Diego Di Carlo, Clément Gaultier, Antoine Deleforge, and Nancy Bertin (2018). "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge". In: *arXiv preprint arXiv:1812.05901*
- Scheibler, Robin, Diego Di Carlo, Antoine Deleforge, and Ivan Dokmanić (2018b). "Separake: Source separation with a little help from echoes". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901

1.6 Don't Panic!

The reader will have already noticed that a large margin is left free on each manuscript page. We will use it to insert personal comments, historical notes, additional insights, and figures and tables to complete each subject. This graphic template is inspired by the work of Tufte and Graves-Morris[Tufte and Graves-Morris 1983][4] . We emphasize the presence of clickable links by the ↗ logo and code library, which are written in typewriter font, e. g. `dEchorate`. A list of design choice follows:

1.6.1 *Quick vademecum*

for the readers:

- Bibliographic references are denoted as [Kuttruff 2016].

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

- Figures, Tables, and other floating objects and equations are numbered within the chapter number.
- Equations are referred as ??
- ↗ denotes clickable external link
- orange is used for clickable internal link, such as § 1.1 and acronyms FFT.
- grey is used for clickable internal 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 reader attention towards ~~the~~ important points.
- ⇄ indicate the presence of definition by dichotomy.
- The end of the chapter is shown with a logo signature.

1.6.2 *The golden ratio of the thesis*

This thesis has been written following personal stylistic rules:

- At most three levels of sub-headings: section, subsection, and Tufte's *new-thought*.
- The usage of dichotomies is emphasized.
- Each paragraph is introduced briefly at the end of the previous one.
- no indentation, but well-separated text blocks.

Part II

ECHO-AWARE APPLICATION

2

Problems in Audio Scene Analysis

▸ Synopsis In this chapter, we will present algorithms and methodologies for audio scene analysis in the context of signal processing. At first, in section § 2.1, we present a typical scenario for defining some cardinal problems. Therefore in section § 2.2, state-of-the-art approaches to address these problems are listed and commented, highlighting the relationship with some acoustic propagation models. The content presented here serves as a basis for a deeper investigation conducted in each of the following chapters.

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

The material presented here results from the personal elaboration 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].

2.1 Audio Scene Analysis Problems

As mentioned in the first chapter, the audio scene analysis aims to parcel all the 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.

2.1.1 *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 it is a indoor scenario, all the elements of reverberation (in particular echoes) are presents. Diffuse background noise 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 j sources and i 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 the relative chapter, the signal recorded at the i-th microphones reads

$$x_i[n] = \sum_{j=1}^{J} \big(h_{ij}(\underline{\mathbf{x}}_i|\underline{\mathbf{s}}_j) * s_j\big)[n] + n_i[n], \tag{2.1}$$

or alternatively, using the source spatial image signals,

$$\begin{aligned} x_i[n] &= \sum_{j=1}^{J} c_{ij}[n] + n_i[n] \\ c_{ij}[n] &= \big(h_{ij}(\underline{\mathbf{x}}_i|\underline{\mathbf{s}}_j) * s_j\big)[n] \end{aligned} , \tag{2.2}$$

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 **????**, we can decouple the information of indoor microphone natural recordings into two orthogonal contributions: the RIRs (thus the mixing matrix) accounting for only the sound propagation, and the source signals that depend only its content.

2.1.2 *Problems 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.

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

Table 2.1: List of audio scene analysis problems considered in this thesis accompanied by their mathematical description.

As introduced in depending on the application, these problems can be said either *informed* or *blind* and the related scenario *active* or *passive*. These two dichotomies emphasize the amount of prior knowledge available for solving them. As opposed to the active scenario, where the source signal is known, transmitted, and available, the passive one considers only the microphone

measurements. For instance, when addressing the active echo estimation problem or RIR measurement, the exact time of emission of the source signal is known, as well as the source signal itself.
The second dichotomy refers to the possibility of exploiting prior knowledge to facilitate the solution of the problem. This information may derive from annotations, meta-data that accompany the application. 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* and *strongly guided* problems: For the former, 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) etc. For the latter, specific information about the mixing process and the speakers' identity can be used.

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

2.2 Literature overview

Here we present the general overview of the literature related to the problems considered in this thesis: multichannel audio source separation, and 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 of the related chapters . Since Room Geometry Estimation (RooGE) is manly based on echo estimation and labeling, its discussion is reported in ????.

2.2.1 *on Multichannel Sound Source Separation*

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

In the context of the multichannel speech recordings, some of the most successful and popular methods used nowadays include spatial filtering, Time-Frequency (TF) masking, and end-to-end regression. In this thesis, we deliberately distinguish between the spatial filtering, which will be discussed in the following subsection, and TF masking.

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

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. Finally, by inverting it, the resulting STFTs estimates of the source signal are obtained. One of the most popular masking rules is adaptive Wiener filtering. For each time-frequency bin, the STFTs of the estimated source spatial images of the j-th source at the i microphone, writes

$$\hat{C}_{ij} = W_{\texttt{Wiener}} X_i = \frac{|C_{ij}|^2}{\sum_{j=0}^{J} |C_{ij}|^2} X_i \tag{2.3}$$

where the fraction compute the TF mask $W_{\texttt{Wiener}}$.
In order to be computed, the Wiener Filter requires the knowledge of all the spatial source images sources, or equivalently, the mixing filters and the source signals. Therefore, this approach has been generalized in several ways to account for both these unknowns. As opposed to spatial filtering that operates considering the mixing filters, the source signals are indispensable to weigh each of the TF bins.

One of the most successful framework to the Gaussian Model based on Multichannel Nonnegative Matrix Factorization [Ozerov and Févotte 2010; Sawada et al. 2013]. It combines the Nonnegative Matrix Factorization (NMF) and narrowband spatial model (discussed in **??**) and deploys optimization-based framework for estimating both the mixing matrix and the sources. This approach will be further discussed **??**. On of the main advantage of this approach is that allows to easily incorporated prior knowledge on the problems. In fact, thanks to the NMF formulation, information about sources can easily incorporated, even learned a priori [Schmidt and Olsson 2006; Smaragdis et al. 2009]. In addition, thanks to the narrowband approximation, filter and source content are decoupled, allowing the user to define proper model for the RIRs, or ReTF, can be implemented as well.

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 the same task [Hershey et al. 2016; Wang et al. 2018]. Alternatively, deep learning techniques are 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 EM-NMF source separation framework of [Ozerov and Févotte 2010].

However, it has been shown that even with oracle TF [Luo and Mesgarani 2019], the estimation is still affected by artifacts. This limitation affects all the approaches operating in the TF domain. To overcome this, end-to-deep deep learning models [Luo and Mesgarani 2019; Tzinis et al. 2020], which now hold the record in source separation. These models work directly in the time domain: both input and output are time-domain waveforms. Despite the separation qualities, all deep learning methods rely on trained black-box models for which is hard to inject prior knowledge. Instead, Multichannel NMF-based frameworks provide accounts for this option.

▶ MULTICHANNEL NMF SOURCE SEPARATION METHODS can be grouped according to how they model sound propagation of the mixing process:

- those that simply ignore it [Le Roux et al. 2015];
- (*free field propagation*) those that assume a single anechoic path [Rickard 2007; Nesta and Omologo 2012] ;
- (*reverberant propagation*) those that model the Room Transfer Functions (RTFs) entirely [Ozerov and Févotte 2010; Duong et al. 2010; Li et al. 2019];
- (*reverberant propagation*) and those that attempt to separately estimate the contribution of the early echoes and the contribution of the late tail [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 compting 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 **??**, where a new method for speech source separation based on the Multichannel NMF framework and echoes is described.

2.2.2 *on Spatial Filtering*

For a comprehensive review on spatial filtering methods, the reader can refers 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 vast research field that interested the signal processing and telecommunication communities since several decades. It produces an enormous literature as well as well-affirmed book, which will not be covered in this thesis In audio, this topic has been recently review in the context of speech enhancement in recent publication [Gannot et al. 2017][5] As opposed to Audio Source Separation, whose techniques cover both signal- and multi-channel recordings, Spatial Filtering explicitly exploits the microphones' different spatial distribution. Nevertheless, the two problems are intertwined, and some techniques can be used reciprocally.

In spatial filtering, the RIRs (and related models, e. g., RTFs, steering vectors or ReTF) play a central role. Intuitively, giving the mixing model in Eq. (2.1), 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 for the following two reasons: First, it is due to a fundamental trade-off

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

between denoising and filtering given by the number of microphones available. Second, the inversion of the RIRs is not straightforward.[6].

- BEAMFORMING is one of the most famous techniques used in spatial filtering. 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 [Frost 1972; Van Veen and Buckley 1988]. Thus, the target source signal is enhanced, while noise, interferences, and reverberation being suppressed. **??** illustrate this ideas. This idea has been extended to Frequency and Time-Frequency processing. 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.

- MANY BEAMFORMERS CRITERIA have been proposed. Among all, some of the most famous are the Delay-and-Sum (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 **??**. 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 Direction of Arrival (DOA) [Takao et al. 1976; Applebaum and Chapman 1976; Cox et al. 1987; Van Veen and Buckley 1988];

  - (*multipath propagation*) methods based on rake rake receiver[**Jan1995matched**; Flanagan et al. 1993; Dokmanić et al. 2015; Peled and Rafaely 2013; Scheibler et al. 2015; Kowalczyk 2019];

  - (*reverberant propagation*) methods based on full acoustic channel estimation (See **??**);

  - (*reverberant propagation*) methods based on Directions of Arrival (DOAs) and the statistical modeling of the diffuse sound field, [Thiergart and Habets 2013; Schwartz et al. 2014];

  - (*reverberant propagation*) methods based Relative Transfer Function (ReTF) [Gannot et al. 2001; Doclo and Moonen 2002; Cohen 2004; Markovich et al. 2009];

[6] The work in [Neely and Allen 1979] discusses the issues of inverting RIRs. Several techniques were investigated to overcome this problem, which is also known as Room Response Equalization [Cecchi et al. 2018]

- (*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, Deep Neural Network (DNN) have been proposed for solving this task, either to estimate the beamformer filter [**li2016neural directly**; Xiao et al. 2016; Sainath et al. 2017] or in an end2end task [Ernst et al. 2018] Moreover, DNN has been used to estimate some of parameters, such as the DOAs [Salvati et al. 2018; Chazan et al. 2019], ReTF estimation [Chazan et al. 2018].

▶ Early echoes, in the literature thus far, are neither considered nor modeled as noise terms. 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 **??** together with their implementation.

2.2.3 *on Sound Source Localization*

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

Sound Source Localization (SSL) consists in determining the position of sources from microphone recordings 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. 2010; 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 coordinates, most of the existing methods aim at estimating 2-dimensional DOA, namely the angles for on the unit sphere with the center in a reference point. This reference point is usually the center of the microphone array. This angles are called *azimuth* and *elevation* as shown is **??**. Second, they assume far-field scenarios. The main reasons for adopting such simplifications are the followings: First, estimating the distance is known to be a much more challenging task than estimating the DOAs [Vesa 2009]. Second, the task is decoupled from the more ambitious on room geometry estimation. Third, the far-field scenario is a reasonable assumption when using a compact array recording distant talking speech. Finally, in far-field settings, sometimes the only DOAs are sufficient to achieve reasonable speech enhancement performances [Gannot et al. 2017].

Despite these approximations, the SSL problem still challenges today's computational methods, particularly in the presence of reverberation or interfering sources. Popular approaches for this task consists in two 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.

- Knowledge-based approaches rely on a physic model for sound propagation [Knapp and Carter 1976; Stoica and Sharman 1990; DiBiase et al. 2001; Dmochowski et al. 2007; Lebarbenchon et al. 2018] These models rely on closed-form mapping from the sound's direct path Time Differences of Arrival at the microphone pair and the source's azimuth angle in this pair. If multiple microphone pairs are available and form a non-linear array, their TDOAs can be aggregated to obtain 2D directions of arrival [DiBiase et al. 2001]. Furthermore, the main difference between these approaches 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 [Dmochowski et al. 2007], generalized-cross-correlation [Knapp and Carter 1976; DiBiase et al. 2001; Lebarbenchon et al. 2018], blind system identification [Chen et al. 2006], maximum likelihood [Stoica and Sharman 1990; Laufer et al. 2013], direct-path ReTF [Li et al. 2016b]. The main limitations of these approaches result in the approximation considered in the models. In particular, common to all of them is to assumption sound propagation being free-field. Thus, they intensely suffer in environments it is violated, e. g., in the presence of strong acoustic echoes and reverberation as discussed as shown in [Chen et al. 2006].

- 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 [Laufer et al. 2013; Deleforge et al. 2015; Vesperini et al. 2018; Chakrabarty and Habets 2017; Adavanne et al. 2018; Perotin et al. 2018; Gaultier et al. 2017] (to cite a few examples). Such data can be obtained from annotated real recordings [Deleforge et al. 2015; Nguyen et al. 2018] or using physics-based 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, the data-driven literature can broadly dichotomize into two approaches: 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; Adavanne et al. 2018], where the task is performed with DNNs models. In the latter, learning models are used as a substitute for either feature extraction or the mapping. For instance, in [Laufer et al. 2013; Deleforge et al. 2015; Gaultier et al. 2017; Nguyen et al. 2018], Gaussian Mixture Models (GMMs)-based models ware used to learning the mapping from features derived from the ReTF of pair of microphones. In [Vesperini et al. 2018], the author proposes to use Neural Network (NN) models to estimate source location using features computed through Generalized Cross Correlation with Phase Transform (GCC-PHAT). Despite the considerable benefit of data-driven approaches in learning complex functions, their main limitation lies in the training data. First, these data are typically tuned for specific microphone arrays and fail whenever test conditions strongly mismatch training conditions. Moreover, due to the cumbersome task of collecting building annotated datasets that cover as many possible scenarios as possible, physics-based simulators are used. Therefore, as they "learn a model from model" which, in turn, rely on assumptions, they may not be able to generalize to real-world conditions.

▸ To CONCLUDE most of the methods developed for SSL, and in particular DOAs estimation, including the above listed, regard reverberation and, in particular, acoustic echoes as a nuisance. The recent DNN based supervised learning approaches have proven to succeed in the presence of harsh acoustic conditions. However, they are based on black-box, where knowledge about sound propagation is not trivial to inject. Based on these limitations, we propose to combines the best of the two worlds: using DNN to estimate echoes **??** and use well-understood knowledge-based method to map echoes to source DOAs **??**.

2.3 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.

Bibliography

Adavanne, Sharath, Archontis Politis, and Tuomas Virtanen (2018). “Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1462–1466 (cit. on pp. 22, 23).

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. 8).

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. 20).

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. 21).

Bregman, Albert S (1990). “Auditory scene analysis”. In: *McAdams and Bigand, editors, Thinking in Sound*, pp. 10–36 (cit. on p. 6).

Capon, Jack (1969). “High-resolution frequency-wavenumber spectrum analysis”. In: *Proceedings of the IEEE* 57.8, pp. 1408–1418 (cit. on p. 20).

Cecchi, Stefania, Alberto Carini, and Sascha Spors (2018). “Room response equalization—A review”. In: *Applied Sciences* 8.1, p. 16 (cit. on p. 20).

Chakrabarty, Soumitro and Emanuël AP Habets (2017). “Broadband DOA estimation using convolutional neural networks trained with noise signals”. In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 136–140 (cit. on pp. 22, 23).

Chazan, Shlomo E, Jacob Goldberger, and Sharon Gannot (2018). “DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6712–6716 (cit. on p. 21).

Chazan, Shlomo E, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot (2019). “Multi-microphone speaker separation based on deep DOA estimation”. In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1–5 (cit. on p. 21).

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 p. 22).

Cherry, Colin (1953). “Cocktail party problem”. In: *Journal of the Acoustical Society of America* 25, pp. 975–979 (cit. on p. 17).

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. 20).

Cox, Henry, Robertm Zeskind, and Markm Owen (1987). “Robust adaptive beamforming”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35.10, pp. 1365–1376 (cit. on p. 20).

Crocco, Marco and Alessio Del Bue (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 p. 21).

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 p. 8).

Deleforge, Antoine, Florence Forbes, and Radu Horaud (2015). "Acoustic space learning for sound-source separation and localization on binaural manifolds". In: *International journal of neural systems* 25.01, p. 1440003 (cit. on pp. 22, 23).

Deleforge, Antoine, Diego Di Carlo, 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. 11).

Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019). "Mirage: 2d source localization using microphone pair augmentation with echoes". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 775–779 (cit. on p. 11).

Di Carlo, Diego, Clement Elvira, Antoine Deleforge, Nancy Bertin, and Rémi Gribonval (2020). "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 156–160 (cit. on p. 11).

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. 11).

DiBiase, Joseph H, Harvey F Silverman, and Michael S Brandstein (2001). "Robust localization in reverberant rooms". In: *Microphone Arrays.* Springer, pp. 157–180 (cit. on p. 22).

Dmochowski, Jacek P, Jacob Benesty, and Sofiene Affes (2007). "Broadband MUSIC: Opportunities and challenges for multiple source localization". In: *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.* IEEE, pp. 18–21 (cit. on p. 22).

Doclo, Simon and Marc Moonen (2002). "GSVD-based optimal filtering for single and multimicrophone speech enhancement". In: *IEEE Transactions on signal processing* 50.9, pp. 2230–2244 (cit. on p. 20).

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 p. 8).

Dokmanić, Ivan, Robin Scheibler, and Martin Vetterli (2015). "Raking the Cocktail Party". In: *IEEE Journal on Selected Topics in Signal Processing* 9.5, pp. 825–836. ISSN: 19324553. DOI: 10.1109/JSTSP.2015.2415761 (cit. on pp. 8, 20).

Duong, Ngoc QK, Emmanuel Vincent, and Rémi Gribonval (2010). "Under-determined reverberant audio source separation using a full-rank spatial covariance model". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, pp. 1830–1840 (cit. on p. 19).

Ernst, Ori, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger (2018). "Speech dereverberation using fully convolutional networks". In: *2018 26th European Signal Processing Conference (EUSIPCO).* IEEE, pp. 390–394 (cit. on p. 21).

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. 8, 20).

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. 20).

Gannot, Sharon, David Burshtein, and Ehud Weinstein (2001). "Signal enhancement using beamforming and nonstationarity with applications to speech". In: *IEEE Transactions on Signal Processing* 49.8, pp. 1614–1626 (cit. on p. 20).

Gannot, Sharon, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov (2017). "A consolidated perspective on multimicrophone speech enhancement and source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.4, pp. 692–730 (cit. on pp. 19, 22).

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. 22, 23).

Hershey, J. R., Z. Chen, J. Le Roux, and S. Watanabe (2016). "Deep clustering: Discriminative embeddings for segmentation and separation". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35 (cit. on p. 18).

Jan, E, Piergiorgio Svaizer, and James L Flanagan (1995). “Matched-filter processing of microphone array for spatial volume selectivity”. In: *Proceedings of ISCAS'95-International Symposium on Circuits and Systems*. Vol. 2. IEEE, pp. 1460–1463 (cit. on p. 8).

Jan, Ea-Ee and James Flanagan (1996). “Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors”. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 2. IEEE, pp. 917–920 (cit. on p. 8).

Knapp, Charles and Glifford Carter (1976). “The generalized correlation method for estimation of time delay”. In: *IEEE transactions on acoustics, speech, and signal processing* 24.4, pp. 320–327 (cit. on p. 22).

Kowalczyk, Konrad (2019). “Raking early reflection signals for late reverberation and noise reduction”. In: *The Journal of the Acoustical Society of America* 145.3, EL257–EL263. ISSN: 0001-4966. DOI: `10.1121/1.5095535`. URL: `http://dx.doi.org/10.1121/1.5095535` (cit. on pp. 8, 20).

Kuttruff, Heinrich (2016). *Room acoustics*. CRC Press (cit. on p. 11).

Laufer, Bracha, Ronen Talmon, and Sharon Gannot (2013). “Relative transfer function modeling for supervised source localization”. In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 1–4 (cit. on pp. 22, 23).

Le Roux, Jonathan, John R Hershey, and Felix Weninger (2015). “Deep NMF for speech separation”. In: *Proc. IEEE ICASSP*, pp. 66–70 (cit. on p. 19).

Lebarbenchon, Romain, Ewen Camberlein, Diego Di Carlo, Clément Gaultier, Antoine Deleforge, and Nancy Bertin (2018). “Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge”. In: *arXiv preprint arXiv:1812.05901* (cit. on pp. 11, 22).

Leglaive, Simon, Roland Badeau, and Gaël Richard (2015). “Multichannel audio source separation with probabilistic reverberation modeling”. In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 1–5 (cit. on p. 19).

— (2016). “Multichannel audio source separation with probabilistic reverberation priors”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.12, pp. 2453–2465 (cit. on p. 8).

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. 21).

Li, Xiaofei, Laurent Girin, Radu Horaud, and Sharon Gannot (2016b). “Estimation of the direct-path relative transfer function for supervised sound-source localization”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.11, pp. 2171–2186 (cit. on p. 22).

Li, Xiaofei, Laurent Girin, and Radu Horaud (2019). “Expectation-maximisation for speech source separation using convolutive transfer function”. In: *CAAI Transactions on Intelligence Technology* 4.1, pp. 47–53 (cit. on p. 19).

Luo, Yi and Nima Mesgarani (2019). “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation”. In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8, pp. 1256–1266 (cit. on p. 18).

Makino, Shoji (2018). *Audio Source Separation*. Vol. 433. Springer (cit. on p. 17).

Markovich, Shmulik, Sharon Gannot, and Israel Cohen (2009). “Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.6, pp. 1071–1086 (cit. on p. 20).

Neely, Stephen T and Jont B Allen (1979). “Invertibility of a room impulse response”. In: *The Journal of the Acoustical Society of America* 66.1, pp. 165–169 (cit. on p. 20).

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. 19).

Nguyen, Quan, Laurent Girin, Gérard Bailly, Frédéric Elisei, and Duc-Canh Nguyen (2018). “Autonomous sensorimotor learning for sound source localization by a humanoid robot”. In: (cit. on pp. 22, 23).

Nugraha, Aditya Arie, Antoine Liutkus, and Emmanuel Vincent (2016). “Multichannel audio source separation with deep neural networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.9, pp. 1652–1664 (cit. on p. 18).

Ozerov, Alexey and Cédric Févotte (2010). “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation”. In: *IEEE Trans. Audio, Speech, Language Process.* 18.3, pp. 550–563 (cit. on pp. 18, 19).

Peled, Yotam and Boaz Rafaely (2013). “Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field”. In: *IEEE transactions on audio, speech, and language processing* 21.12, pp. 2532–2540 (cit. on p. 20).

Perotin, Lauréline, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin (2018). “CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector”. In: *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, pp. 241–245 (cit. on p. 22).

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 p. 21).

Remaggi, Luca, Philip JB Jackson, Philip Coleman, and Wenwu Wang (2016). “Acoustic reflector localization: novel image source reversion and direct localization methods”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.2, pp. 296–309 (cit. on p. 8).

Remaggi, Luca, Philip JB Jackson, and Wenwu Wang (2019). “Modeling the Comb Filter Effect and Interaural Coherence for Binaural Source Separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12, pp. 2263–2277 (cit. on p. 8).

Ribeiro, Flávio, Demba Ba, Cha Zhang, and Dinei Florêncio (2010). “Turning enemies into friends: Using reflections to improve sound source localization”. In: *2010 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 731–736 (cit. on pp. 8, 21).

Rickard, Scott (2007). “The DUET blind source separation algorithm”. In: *Blind Speech Separation*, pp. 217–241 (cit. on pp. 18, 19).

Sacks, Oliver (2014). *Musicofilia*. Adelphi Edizioni spa (cit. on p. 5).

Sainath, Tara N, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al. (2017). “Multichannel signal processing with deep neural networks for automatic speech recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.5, pp. 965–979 (cit. on p. 21).

Salvati, Daniele, Carlo Drioli, and Gian Luca Foresti (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 p. 21).

Sawada, Hiroshi, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda (2013). “Multichannel extensions of non-negative matrix factorization with complex-valued data”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.5, pp. 971–982 (cit. on p. 18).

Scheibler, Robin, Ivan Dokmanić, and Martin Vetterli (2015). “Raking echoes in the time domain”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 554–558 (cit. on p. 20).

Scheibler, Robin, Diego Di Carlos, Antoine Deleforge, and Ivan Dokmanic (2018a). “Separake: Source Separation with a Little Help from Echoes”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2018-April, pp. 6897–6901. ISBN: 9781538646588. DOI: 10.1109/ICASSP.2018.8461345. URL: http://arxiv.org/abs/1711.06805 (cit. on p. 8).

Scheibler, Robin, Diego Di Carlo, Antoine Deleforge, and Ivan Dokmanić (2018b). “Separake: Source separation with a little help from echoes”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901 (cit. on p. 11).

Schmidt, Mikkel N and Rasmus K Olsson (2006). “Single-channel speech separation using sparse non-negative matrix factorization”. In: *Ninth International Conference on Spoken Language Processing* (cit. on p. 18).

Schwartz, Ofer, Sharon Gannot, and Emanuël AP Habets (2014). “Multi-microphone speech dereverberation and noise reduction using relative early transfer functions”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.2, pp. 240–251 (cit. on p. 20).

— (2016). “Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm”. In: *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1123–1127 (cit. on p. 21).

Smaragdis, Paris, Madhusudana Shashanka, and Bhiksha Raj (2009). “A sparse non-parametric approach for single channel separation of known sounds”. In: *Advances in neural information processing systems*, pp. 1705–1713 (cit. on p. 18).

Sokol, Joshua (2017). “The thoughts of a spiderweb”. In: *Obtenido de: https://www. quantamagazine. org/the-thoughts-of-a-spiderweb-20170523* (cit. on p. 5).

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. 22).

Stöter, Fabian-Robert, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji (2019). “Open-unmix-a reference implementation for music source separation”. In: (cit. on p. 18).

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. 20).

Thiergart, Oliver and Emanuël AP Habets (2013). “An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 659–663 (cit. on p. 20).

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. 11).

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. 18).

Van Trees, Harry L (2004). *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons (cit. on p. 19).

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. 20).

Vesa, Sampo (2009). “Binaural sound source distance learning in rooms”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.8, pp. 1498–1507 (cit. on p. 21).

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. 22, 23).

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. 17).

Vincent, Emmanuel, Tuomas Virtanen, and Sharon Gannot (2018). *Audio source separation and speech enhancement*. John Wiley & Sons (cit. on pp. 15, 17, 19, 21).

Wang, Zhong-Qiu, Jonathan Le Roux, and John R Hershey (2018). “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 18).

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 p. 21).