08/10/2020

Problems in Audio Scene Analysis

→ not very catchy / explicit
We don't know what to expect

▸ 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.

a selection of ...
which we will need later / which we identified as potential beneficiaries of echo-aware additions ...

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.

OVERALL this § is too vague, be more specific / show the motivation

should come sooner!

much better

there is something to redispatch so that important message is in the 1st §!

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].

synthesis? (elaboration on? to be checked)

2.1 Audio Scene Analysis Problems

?? extract?

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.

(not necessarily "all")
SSL ∈ ASA for instance

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~~ an indoor scenario, all the elements of

reverberation (in particular echoes) are present. 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 mainly based on echo estimation and labeling, its discussion is reported in ????.

2.2.1 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 1. 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

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

where the fraction compute the TF mask W_{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 Multi-channel 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 ??. One of the main advantage of this approach is that allows to easily incorporate prior knowledge on the problem. 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~~ can be 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

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

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

[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].

[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]

- 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];

- (*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

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' positions 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 following: 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

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].

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 physical 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.

quite good

unclear

- 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 ??.

the

sounds weird in a SOTA chapter (or, the counterpart is missing in source sep. ?)

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.

Maybe there is an outline pb.
2.2 is overweight. Maybe I would split

2.2. On Multichannel Source Sep.
2.3. On Spatial Filt.
2.4. On Source Loc.
2.5. Conclusion.

The chapter doesn't work very well... maybe too many things? some synthesis parts are very good, but some others drown into details, and the overall result doesn't seem very straightforward.
Let's see what Antoine think, but I'd recommend some cuts.