

Master Thesis

Gaspard Ulysse Fragnière

August 2022

1 How to L^AT_EX

How to make a reference to a paper: [6]

2 Draft

2.1 Grumiaux

[6] is a survey of several methods for sound source localization (SSL). Traditionally, this problem has been tackled using Signal Processing based methods. But in the recent years, methods based on deep learning have been developed and showed better results than traditional approaches. Those methods have been compiled in this paper. The survey is organized in the different following sections:

- **Section I:** Introduction
- **Section II:** Acoustic Environment and Sound Source Configuration
- **Section III:** Conventional SSL methods
- **Section IV:** Neural Network Architectures for SSL
- **Section V:** Input Features
- **Section VI:** Outputs strategies
- **Section VII:** Data
 1. Synthetic Data
 2. Real data
 3. Data augmentation techniques
- **Section VIII:** Learning Strategies
- **Section IX:** Conclusions and Perspectives

We are interested in the section about Synthetic Data and Data augmentation. Indeed those sections can be used as a starting point for building the state of the art. Its goal is to answer the following questions:

- Are there **existing methods** to generate virtually:
 - measured time data (single channel/multi-channel)?
 - measured source spectra (single channel/multi-channel)?
 - measured cross-spectral matrices in stationary environments (multi-channel only)?
- What **measurement scenarios** are used in the literature (time-stationary/non-stationary sources, number of microphones, temporal dimensions...)?
- What are the **existing setups** in multi-channel data generation with neural networks (conditioning variables, network architectures (convolutional, recurrent, Transformer, ...), generative algorithms (GAN/VAE), ...)

In [6], a classical method about data generation is introduced. The idea is the following: Simulate the Room Impulse Response (RIR) in order to simulate realistically room acoustics (e.g. reverberation). Then convolve dry audio signals with the RIR simulated. This can provide suited training data, since RIR for rooms of different size, different source position as well as different dry signals can be used for the training.

In [2], the datasets is created in the following way: a speaker with a visual marker is placed in front of camera and binaural microphone setup (dummy head). "The loud-speaker that emits fixed-length full-spectrum sounds is moved in front of the dummy-head/camera device and for each loud-speaker location, both the emitted sound and the image location of the visual marker are recorded. → not so useful

In [11], a GAN is used to simulate data. The GAN used in [11] is an implementation of [9]

[7] proposed a low-complexity model-based training data generation method that includes a deterministic model for the direct path and a statistical model for late reverberation. It has been demonstrated that the SSL neural network, trained using the data generated by this method, achieves comparable localization performance as the same architecture trained on a dataset generated by the usual ISM.

An investigation of several simulation methods was done by [4], with extensions of ISM, namely, ISM with directional sources, and ISM with a diffuse field due to scattering. [4] compared the simulation algorithms via the training of an MLP (in both regression and classification modes) and showed that ISM with scattering effects and directional sources leads to the best SSL performance.

[9] : We propose a learning-based method which uses Generative Adversarial Networks [12] to learn a stochastic mapping from perceptually-informed spectrograms into simple magnitude spectrograms.

Paper read and not useful:

- Deleforge 2013 [1]
-

2.2 Bianco

In Bianco a VAE is used to perform SSL. The idea is the following: based on VAEs to encode the phase of the relative transfer function (RTF) between two microphones to a latent parametric distribution. The resulting model estimates DOA and generates RTF phase.

VAEs learn from unlabeled data explicit latent codes for generating samples, and are inspiring examples of representation learning.

There is a link between DOA and RTF. Indeed, the RTF phase is encoded as a function of source azimuth (direction of arrival, DOA). Similarly as in [5], the goal of the NN (GAN or VAE) is to learn the distribution of a quantity that is a function of the DOA.

The experiments show, only **two labeled samples per DOA** permit the VAE-SSL to obtain better performance than SRP-PHAT (State of the art).

- **in:**
- **out:** Relative Transfer Function (RTF)

what is RTF: TODO: continue

2.3 Neekhara

[9] is concerned with finding a solution for Text to speech (TTS) problem. The claim is that using a GAN approach, they have been able to outperform by far naive approaches (user review) and being 100x faster than other DL approaches. More specifically this paper was concerned with creating a mapping from language to **perceptually informed spectrogram**. Indeed the difficulty of the problem lays in the fact that perceptually informed spectrograms are not invertible. Indeed a spectrogram is a compact representation of a signal where much of the information contained in a audio waveform has been lost. More specifically the problem at hand is phase estimation and magnitude estimation. Therefore a predictive model is required to fill the missing information and create natural sounding sound.

- **in:** (perceptually informed) spectrogram (or text on a more basic level)
- **out:** natural sounding audio waveform

2.4 NEURIPS

[8] also introduces a method for the TTS problem. Moreover in the introduction of the paper, there is a comparison of different method for text to speech (i.e. audio wave generation):

- **Pure signal processing approaches:** "The main issue with these pure signal processing methods is that the mapping from intermediate features to audio usually **introduces noticeable artifacts**"
- **Autoregressive NN based models:** An autoregressive model is a model that relies on past values to predict current ones. In this sense, an autoregressive model must be sequential. "These methods have produced state-of-the-art results in text-to-speech synthesis and other audio generation tasks. Unfortunately, inference with these models is inherently slow and inefficient because audio samples must be generated sequentially. Thus auto-regressive models are **usually not suited for real-time applications.**"
- **Non autoregressive models:** "While inference is fast on the GPU, the large size of the model makes it **impractical for applications with a constrained memory budget.**"
- **GAN for audio:** However their results show that adversarial loss alone is not sufficient for high quality waveform generation; it requires a KL-divergence based distillation objective as a critical component. To this date, making them work well in this domain has been challenging

In [8], a GAN (MelGAN) is introduced. The goal of this GAN is to perform audio waveform generation.

- **in:** mel-spectrogram
- **out:** audio waveform

2.5 EngelGanSynth:

[3] is concerned with demonstrating that GANs can in fact generate high-fidelity and locally-coherent naudio by modeling log magnitudes and instantaneous frequencies with sufficient frequency resolution in the spectral domain.

Using a GAN approach and the NSynth datasets (Dataset of standardized notes from instrument) to generate notes.

TODO: According to Adam, this paper is to generate spectrogram and not audio wave form. → check again

- **in:** instruments note from datasets NSynth
- **out:** instruments audio waveform.

2.6 gerstoft

[5] is concerned with training a GAN for generating audio based features.

In many array processing techniques (i.e. beamforming) first compute the Sample Covariance Matrix (SCM). The idea of the (Wasserstein) GAN here is to generate many SCM. More specifically, the goal of the GAN is to learn the **joint** probability distribution function of the **observable data** (array data) and the **target variable** (Wasserstein distance). The wasserstein distance is a metric to measure distance between probability distribution on a given metric space.

The idea is first to consider a model that describe relationship between **array data** and **DOA**. Using this model, another relationship between **CSM** and **array data** can be defined. Hence we have a mathematical relationship between **DOA** and **CSM**.

- **in:** True Sample Covariance Matrix with DOA
- **out:** Generator, i.e. probability distribution function

Idea: maybe it would be possible to generate for the pdf os SCM for several DOA. Moreover, a link between SCM and cross spectral matrix should be investigated.

2.7 Vera-Diaz

In ASL problem, the source's position can be estimated with at least three Time Difference of Arrival (TDoA) measurements with hyperbolic trilateration methods. Signals captured in everyday scenarios are contaminated with noise and multipath effects. Directly measuring TDoA in those cases is a difficult task that produces inaccurate localization results. [12] is concerned with finding a way to denoise those signals. Other methods uses GCC instead of TDoA. Such methods are more robust to noise and multipath effects, but not fully immune to it.

The contribution of this paper is a DNN named DeepGCC. This DNN takes as input a GCC (Generalized Cross Correlation) Matrix and estimates a Gaussian function. Then the SSL problem is solved by replacing the GCC typically used by DeepGCC (GCC) and then using a classical beamforming approach.

- **in:** GCC
- **out:** (x, y, z)

→ this is actually the inverse of what we want. Hence if we could reverse this network, maybe we could do something of the following:

$$(x, y, z) \rightarrow GCC \rightarrow CSM \quad (1)$$

Where the last step is done thanks to the relation between *GCC* and *CSM* via Fourier transform.

2.8 Hübner

This paper is concerned with dealing with a common issue when creating DL algorithm to solve the SSL problem, namely the complexity to gather data for training/validating. Indeed, the two current way to obtain data is by either generating them using simulation or by recording them in real life. Both methods require significant amount of resources (resp. time or storage)

TODO: Use those quotes to structure a bit the the state-of-the-art

quotes from Hübner: 2 "DOA estimation methods can be categorized into classical model-based methods and data-driven methods, which are prevalently implemented using deep neural networks (DNN)."

"Most of the DL methods include a feature extraction step rather than using the raw microphone signals. Popular features include:

- (i) the eigendecomposition of the spatial covariance matrix [14] (similar to MUSIC)
- (ii) generalized cross-correlation (GCC) based features [15–18]
- (iii) modal coherence [19]
- (iv) the Ambisonics intensity vectors [20]
- (v) phase and magnitude spectra [21] and
- (vi) phase spectra [11, 12]. Many of the features are phase-based as motivated by physical models and classical DOA estimators [9].

"

"One way to generate training data for DL-based DOA estimation is by recording sound emitted from a source (e.g., loudspeaker, human) in real acoustic environments [16, 17]. This approach is time-consuming and for high-quality datasets a precise ground truth position is essential, which requires expensive measurement equipment."

"Another popular method is the convolution of signals (e.g., speech) with room impulse responses (RIRs) that have either been recorded [14, 18] or simulated based on the source-image method [11, 12, 20, 21, 23]. The main drawbacks of these data generation methods are excessive time and storage consumption."

Conclusion: "We proposed a low complexity model-based training data generation method for phase-based DOA estimation. The proposed method models the microphone phases directly in the frequency domain to avoid computationally costly operations as present in state-of-the-art methods. The low computational complexity of the proposed method allows for online training data generation, which allows faster prototyping, and paves the way for **applications with a high data demand** such as moving sound sources simulation or large microphone arrays. An evaluation using measured RTFs yielded **comparable results** for phase-based DOA estimation when using **the proposed method and the computationally expensive source-image method for training data generation.**"

2.9 Difference between VAE and GAN

In the context of image generation:

"GANs generally produce better photo-realistic images but can be difficult to work with. Conversely, VAEs are easier to train but don't usually give the best results.

I recommend picking VAEs if you don't have a lot of time to experiment with GANs and photorealism isn't paramount.

There are exceptions such as Google's VQ-VAE 2 which can compete with GANs for image quality and realism. There is also VAE-GAN and VQ-VAE-GAN.

As a note, GANs and VAEs are not specifically for images and can be used for other data types/structures." (this is from a comment on stackstats -> not really usable as a source).

TODO: → read [10]

3 Goal

The goal of the project is to create a GAN to generate either:

- cross-spectral matrix
- the complex-valued sound pressure vector at the different microphone

with the corresponding labels (DoA).

In some way the goal of this project is to create a GAN to realistically add noise to either a cross spectral matrix or complex-valued signal sound pressure vector of an array of microphone. This noise pdf should be estimated instead of being modelled. Indeed a GAN approach is necessary to reduce the computation time and the storage required. Indeed we want to be able to generate randomly and in real time labeled data.

4 Important words

- SSL: Sound source localization
- alternatively SSL: Semi-Supervised Learning
- ASL: Acoustics Source Localization
- Source Spectra
- GCC: Generalized Cross Correlation
- CSM: Cross Spectral Matrix (note that there is a connection between the CSM and the GCC features via the Fourier transform)
- SCM: Sample Covariance Matrix
- CB map: Conventional Beamforming map (also known as Acoustic Power Map???)

5 State-of-the-art

The problem of Sound Source Localization (SSL) is typically dealt with with two approaches: Model-based and Data-driven. The issue that we try to adress is that for data driven approaches, significant quantities of well structures data are required. In the litterature, two main ways of obtaining data have been observed:

- Record sound emitted with a loudspeaker or human voice in a real acoustic enviromnent. The issue with such methods is that it can be very tedious te record in different environment. Moreover, recording sufficiently data is very time consuming. Finally, to build a high quality data sets, expensive equipment is required to have an accurate groundtruth. **TODO: add reference**
- Simulate a Room Impulse Response (RIR) in order to recreate realistic room acoustics (e.g. reverberation). Then convolve dry audio signals with the RIR simulated. This can provide suited training data, since RIR for rooms of different size, different source position as well as different dry signals can be used for the training. The issue with such a method is the important amount of time and data required. **TODO: add reference**

Therefore we would like to find another to generate datas. But before tackling this problem, it is important to note that DL-based approach do not use raw data (direct recording of microphone input) but instead features extracted from the raw data. This is crucial because it means that recording or simulating raw microphone data is no longer necessary, if good quality features could be generated directly. The features that are typically used in DL approaches are:

1. Generalized Cross-Correlation (GCC) based features **TODO: add reference+ check**
2. Eigendecomposition of the Spatial Covariance Matrix **TODO: add reference+ check**
3. Modal coherence **TODO: add reference+ check**
4. Ambisonics intensity vectors **TODO: add reference+ check**
5. Phase and Magnitude Spectra **TODO: add reference+ check**
6. Phase Spectra **TODO: add reference+ check**

References

- [1] Antoine Deleforge, Florence Forbes, and Radu Horaud. Variational em for binaural sound-source separation and localization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 76–80. IEEE, 2013.
- [2] Antoine Deleforge, Radu Horaud, Yoav Y Schechner, and Laurent Girin. Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):718–731, 2015.
- [3] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.
- [4] Femke B Gelderblom, Yi Liu, Johannes Kvam, and Tor Andre Myrvoll. Synthetic data for dnn-based doa estimation of indoor speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4390–4394. IEEE, 2021.
- [5] Peter Gerstoft, Herbert Groll, and Christoph F Mecklenbräuker. Parametric bootstrapping of array data with a generative adversarial network. In *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE, 2020.
- [6] Pierre-Amaury Grumiaux, Srđan Kitić, Laurent Girin, and Alexandre Guérin. A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, 152(1):107–151, 2022.
- [7] Fabian Hübner, Wolfgang Mack, and Emanuël AP Habets. Efficient training data generation for phase-based doa estimation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 456–460. IEEE, 2021.
- [8] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [9] Paarth Neekhara, Chris Donahue, Miller Puckette, Shlomo Dubnov, and Julian McAuley. Expediting tts synthesis with adversarial vocoding. *arXiv preprint arXiv:1904.07944*, 2019.
- [10] Yuanyuan Sun, Lele Xu, Lili Guo, Ye Li, and Yongming Wang. A comparison study of vae and gan for software fault prediction. In Sheng Wen, Albert Zomaya, and Laurence T. Yang, editors, *Algorithms and Architectures for Parallel Processing*, pages 82–96, Cham, 2020. Springer International Publishing.
- [11] Elizabeth Vargas, James R Hopgood, Keith Brown, and Kartic Subr. On improved training of cnn for acoustic source localisation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:720–732, 2021.
- [12] Juan Manuel Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. Acoustic source localization with deep generalized cross correlations. *Signal Processing*, 187:108169, 2021.