Master Thesis

Gaspard Ulysse Fragnière

August 2022

1 State-of-the-art

The problem of Acoustical Source Localization (ASL) is an important problem. It was many applications suach as smart assistant (e.g. Google Home, Alexa, ...), industrial applications, **TODO: add more?**. Traditionnaly this problem is tackled with methods based on the physics of sound propagation (e.g. TDoA, beamforming) or with statistcal inference (e.g.Sparse Bayesian Learning).

The recent success of Deep Learning (DL) based method in other field of research (e.g. Computer Vision) led to believe that Deep Neural Networks (DNN) based approaches could provide state-of-the-art result in solving the ASL problem. [3], [10], [12], [13], [17] and [20] propose state-of-the-arts DL-based methods for Source Characterization. It is important to note that in [3], [12], [13], [20], the Cross Power Spectra (CPS) is used as features for source characterization. The CPS is a direct representation of the signals received in the array of microphone.

A common issue faced while implementing deep learning based methods is that 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. In the literature, [8] and [6] have such an approach.

- 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 prosition 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 storage required. E.g. [4], [16] and [1] created their datasets in this way.

Therefore we would like to find another to generate datas, using Deep Learning approach. It is important to note that DL-based approach do not necessarily use raw data (direct recording of microphone input) but instead features extracted from the raw data. This is crucial because it means that recording, simulating or generating raw microphone data is no longer necessary, if good quality features could be generated directly.

[14], [11], [5] use Generative Adversial Network (GAN) to generate realistic audio waveform. [14] and [11] specifically focus on the generation of audio waveform conditioned on a spectogram

(cGAN). On the other hand, [5] design a GAN to generate realistic audio waveform of single music notes played by an instrument. It is important to note that the GAN designed by [14] is the one implemented in [18].

[2] proposes an approach to generate another acoustic feature: the phase of the relative transfer function (RTF) between two microphones. In this paper a Variational Auto Encoder (VAE) is designed to simultaneously generate phases of RTF and classifying them by their Direction of Arrival (DoA).

[7] use a GAN to generate Sample Cross Spectra Matrices (named as Sample Covariance Matrices) for a given DoA. In their approach, the GAN is trained with data only coming from one DoA, making it unable to generate sample for different DoA.

In [9] introduce a low complexity model-based method for generating samples of microphones phases. This method proposed is not based on DL. Indeed, it is based on a statistical noise model, a deterministic direct-path model for the point source, and a statistical model. The claim of this paper is that the low complexity of the proposed model makes it suited for online training data generation.

[19] introduce a CNN for denoising (i.e. removing the effects of reverberation and multipath effects) of of the Generalization Cross Correlation (GCC) matrix of an array of microphone. This is interesting, since such an network could maybe be inverted to noise GCC matrix and hence make them realistic.

[15] introduce a GAN approach to generate artificial Acoustical Impulse Response (AIR). An AIR is high-dimensional, consisting of thousands of coefficients. AIRs are used typically in the problem of classification of acoustic environment.

References

[1] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466. IEEE, 2018.

[2] Michael J Bianco, Sharon Gannot, and Peter Gerstoft. Semi-supervised source localization with deep generative modeling. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

[3] Paolo Castellini, Nicola Giulietti, Nicola Falcionelli, Aldo Franco Dragoni, and Paolo Chiariotti. A neural network based microphone array approach to grid-less noise source localization. *Applied Acoustics*, 177:107947, 2021.

[4] Soumitro Chakrabarty and Emanuël AP Habets. 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)*, pages 136–140. IEEE, 2017.

[5] 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.

[6] Eric L Ferguson, Stefan B Williams, and Craig T Jin. Sound source localization in a multipath environment using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2386–2390. IEEE, 2018.

[7] 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.

[8] Weipeng He, Petr Motlicek, and Jean-Marc Odobez. Deep neural networks for multiple speaker detection and localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 74–79. IEEE, 2018.

[9] 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.

[10] Adam Kujawski, Gert Herold, and Ennes Sarradj. A deep learning method for grid-free localization and quantification of sound sources. *The Journal of the Acoustical Society of America*, 146(3):EL225–EL231, 2019.

[11] 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.

[12] Soo Young Lee, Jiho Chang, and Seungchul Lee. Deep learning-based method for multiple sound source localization with high resolution and accuracy. *Mechanical Systems and Signal Processing*, 161:107959, 2021.

[13] Wei Ma and Xun Liu. Phased microphone array for sound source localization with deep learning. *Aerospace Systems*, 2(2):71–81, 2019.

[14] Paarth Neekhara, Chris Donahue, Miller Puckette, Shlomo Dubnov, and Julian McAuley. Expediting tts synthesis with adversarial vocoding. *arXiv preprint arXiv:1904.07944*, 2019.

[15] Constantinos Papayiannis, Christine Evers, and Patrick A Naylor. Data augmentation of room classifiers using generative adversarial networks. *arXiv preprint arXiv:1901.03257*, 2019.

[16] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin. Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 241–245. IEEE, 2018.

[17] Wagner Gonçalves Pinto, Michaël Bauerheim, and Hélène Parisot-Dupuis. Deconvoluting acoustic beamforming maps with a deep neural network. 2021.

[18] 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.

[19] Juan Manuel Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. Acoustic source localization with deep generalized cross correlations. *Signal Processing*, 187:108169, 2021.

[20] Pengwei Xu, Elias JG Arcondoulis, and Yu Liu. Deep neural network models for acoustic source localization. In *Berlin Beamforming Conference*, 2021.