POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

A Deep Learning-based method for Multi-Zone Sound Field Synthesis

Laurea Magistrale in Music Engineering - Music and Acoustic Engineering

Author: Roberto Alessandri

Advisor: Prof. Fabio Antonacci

Co-advisor: Luca Comanducci

Academic year: 2021-2022

1. Introduction

Sound Field Control (SFS) has been an active research field in acoustics for several decades, with applications in immersive virtual/augmented reality, telepresence, gaming, noise cancellation, and personal sound zone generation. SFS methods use multiple loudspeakers (secondary sources) to synthesize a desired pressure field in a target region of space. Classical SFS approaches, such as Wave Field Synthesis (WFS) and Ambisonics, are based on analytic methods derived from the Helmholtz equation and assume large continuous distributions of loudspeakers. The other class of methods uses optimisation-based techniques, such as Pressure Matching (PM) [5] and Mode Matching (MM) [4], to minimise the error between the reproduced and desired sound fields. The Multi-zone Sound Field Synthesis (MZ-SFS) problem, which involves synthesising different sound fields inside multiple regions, has been addressed by techniques such as Acoustic Contrast Control (ACC) [2] and Amplitude Matching (AM) [1]. Recently, deep learning techniques have also been applied to sound field synthesis [3]. In this thesis's summary, we propose a Deep Learning-based Pressure Matching technique for the synthesis of Multi-Zones (MZ-DLPM) using a Uniform Linear Array (ULA). Our method estimates driving signals -i.e. the weights to be applied to each source to render the desired sound field - directly through a Convolutional Neural Network and optimizes the loss between the desired and estimated sound field. Through simulations, we compare the performance of the proposed technique with ACC, the original PM approach, and its AM variant. This summary is organised as follows. In *Section 2* the problem statement on MZ-DLPM is described. *Section 3* shows the proposed algorithm implementation, describing the network architecture and the procedure to obtain the optimal driving function. Experimental results are presented and discussed in *Section 4*. We provide simulation and evaluation results to validate our method, analysing how it performs when compared with the state-of-the-art methods. Finally, *Section 5* concludes this thesis's summary and proposes future works.

2. Problem Formulation

Let us consider L loudspeakers deployed in positions $\mathbf{r}'_l, l = 1, \cdots, L$ and M control points $\mathbf{r}_m, m = 1, \cdots, M$ used to measure the pressure in the area $\mathcal{A}$. And let's define as $\mathcal{A}_b$ and $\mathcal{A}_d$ the two regions inside the considered **free-field**

environment $\mathcal{Q}$ with high and low acoustic potential energy, respectively. The two zones are placed such as $\mathcal{A}_b \cup \mathcal{A}_d = \mathcal{A}$ and $\mathcal{A}_b \cap \mathcal{A}_d = \emptyset$, as shown in Figs. 3, 4. In the following description, $\mathbf{r}_{cp}$ to refer to all the $\mathbf{r}_m$ control points, while $\mathbf{r}_m$ represents a single control point. The sound field generated by an array of secondary sources in a set of control points at a given frequency can be expressed as

$$\mathbf{p}(\mathbf{r}_{cp}, \omega_k) = \sum_{l=1}^{L} \mathbf{G}(\mathbf{r}_{cp}|\mathbf{r'}_l, \omega_k)\mathbf{d}(\mathbf{r'}_l, \omega_k), \quad (1)$$

with $\mathbf{G}(\mathbf{r}_{cp}|\mathbf{r'}_l)$ denoting the acoustic transfer function from each loudspeaker to the control points and $\mathbf{d}(\mathbf{r'}_l, \omega_k)$ the driving signal containing the weights to be applied to the loudspeakers. We can now define the desired sound field values at the control points in the bright and dark zones $\mathcal{A}_b$ and $\mathcal{A}_d$ as

$$P^{des}(\mathbf{r}_m, \omega_k) = \begin{cases} \sum_{l=1}^{L} G(\mathbf{r}_m|\mathbf{r}'_l, \omega_k) D(\mathbf{r}'_l, \omega_k), m \in \mathcal{A}_b \\ 0, m \in \mathcal{A}_d. \end{cases} \quad (2)$$

The goal of MZ-SFS systems is to minimise the squared error between the values of the desired pressure field $P^{des}(\mathbf{r}_m, \omega)$ and the estimated pressure field $P^{est}(\mathbf{r}_m, \omega)$ at the control points, and can be written (by omitting the arguments $\mathbf{r}'$ and ω_k) as

$$\min_{\mathbf{d}_{dlpm} \in \mathbb{C}^L} \sum_{m=1}^{M} |\mathbf{G}(\mathbf{r_m})\mathbf{d}_{dlpm} - \mathbf{p}^{des}(\mathbf{r}_m)|^2, \quad (3)$$

where $\mathbf{d}_{dlpm}$ is the output of our optimisation procedure, i.e. the DNN training. In classic DL methods the output of the network is directly compared with a predefined ground truth, by means of a loss function that is minimised. Since we don't have a ground truth set of driving signals, we use the output of our system - the estimated driving function $\mathbf{d}^{des}(\mathbf{r'}, \omega_k)$ - to compute through (1) our estimated pressure field $\mathbf{p}^{est}(\mathbf{r}_{cp}, \omega_k)$ and we apply the loss function to compare it with $\mathbf{p}^{des}(\mathbf{r}_{cp}, \omega_k)$. The input of the neural network model consists of a vector containing the concatenation of the real and imaginary parts of the desired bright zone, i.e.

$$\tilde{\mathbf{p}}_b^{des}(\mathbf{r}_{cp}, \omega_k) = \begin{bmatrix} \Re(\mathbf{p}_b^{des}(\mathbf{r}_{cp}, \omega_k) \\ \Im(\mathbf{p}_b^{des}(\mathbf{r}_{cp}, \omega_k) \end{bmatrix}, \quad (4)$$

where $\mathbf{p}_b^{des}(\mathbf{r}_{cp}, \omega_k)$ being the desired pressure field in the bright zone at the control points,

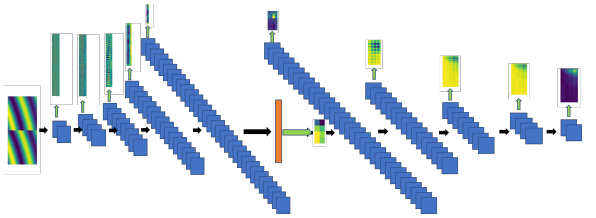

Figure 1: Schematic representation of the Neural Network. For simplicity we represent only the layers with stride 2 × 2, the reshape layer and their outputs.

and $\Re(\cdot)$ and $\Im(\cdot)$ representing the real and imaginary parts of a complex number, respectively. We omit the dark zone because it's an area where all values are equal, hence it would not add any discriminative information from the learning purpose. Terming $\mathcal{U}$ a series of nested functions that represent our Neural Network, defined as

$$\mathcal{U}(\cdot) = \bigcirc_{i=1}^{I} f_i = f_I \circ \cdots \circ f_1, \quad (5)$$

we can express the solution of our system as

$$\mathbf{d}_{dlpm} = \mathcal{U}(\tilde{\mathbf{p}}_b^{des}). \quad (6)$$

3. Proposed Method

In this section we will present the model proposed for multi-zone synthesis. A first part will be dedicated to the depiction of the network architecture. The second part of this section will describe the proposed training procedure.

3.1. Neural Network Architecture

In the proposed method we adopt a Neural Network that follows the basis of an encoder-decoder structure. The structure is shown in Fig. 1.

The Encoder is structured as follows:

- The first layer takes as input the vector $\tilde{\mathbf{p}}_b^{des}(r, \omega) \in \mathbb{R}^{2M \times K}$;
- 10 convolutional layers having (*i*) 32, (*ii*) 32, (*iii*) 64, (*iv*) 64, (*v*) 128, (*vi*) 128, (*vii*) 256, (*viii*) 256, (*ix*) 512, (*x*) 512 filters, respectively;
- The output of the last layer is a flattened monodimensional vector;
- odd layers' kernels are regularised with the *L2* regularisation.

The bottle-neck layer, is a dense layer composed by $(2L/32)(K/32)$ neurons. Its output is regularised through the *Elastic Net* regularisation.

To be fed to the decoder, the bottle-neck layer is followed by a reshaping layer.
The structure of the Decoder is the following:

- The first layer takes as input a tensor of shape $(2L/32) \times (K/32) \times 1$;
- 10 de-convolutional layers having (xi) 512, (xii) 512, $(xiii)$ 256, (xiv) 256, (xv) 128, (xvi) 128, $(xvii)$ 64, $(xviii)$ 64, (xix) 32, (xx) 32 filters, respectively;
- The output layer, composed by 1 filter, returns a tensor of shape $(2L) \times (K) \times 1$;

Finally, common characteristics throughout the neural network are:

- All layers have a kernel size of 3×3;
- All layers, escluding the output layer of the decoder, have Parametric ReLU (PReLU) as activation function;
- Odd layers have a stride of 2×2, while even layers and the output layer of the Decoder have a stride of 1×1;
- Every layers' input is zero-padded evenly in the left/right and up/down parts.

3.2. Procedure

In the following, we'll term as $\mathcal{S}$, the set of virtual sources placed outside the listening environment $\mathcal{Q}$. Since we are considering a free-field environment we can use transfer functions to compute the pressure field $\mathbf{p}^{des}_{b,cp}(\mathbf{r}_{cp})$ emitted by each virtual source $\mathbf{r}_s \in \mathcal{S}$, i.e.

$$\mathbf{p}^{des}_{b,cp}(\mathbf{r}_{cp}, \omega_k) \approx \mathbf{G}(\mathbf{r}_{cp}|\mathbf{r'}_s, \omega), \ s \in \mathcal{S}. \quad (7)$$

Since we are considering only control points, eq. (6) can hence be reformulated as

$$\mathbf{d}_{dlpm} = \mathcal{U}(\tilde{\mathbf{p}}^{des}_{b,cp}). \quad (8)$$

We reorganise the output of the training procedure $\mathbf{d}_{dlpm} \in \mathbb{R}^{2L \times K \times 1}$ in a complex formulation as

$$\mathbf{d}^{C}_{dlpm,l} = \mathbf{d}_{dlpm,l} + j\mathbf{d}_{dlpm,L+l}, \ l = 1, \dots, L \quad (9)$$

being $\mathbf{d}^{C}_{dlpm}$ the complex reformulation of our estimated driving signal, and j the imaginary unit.
With this complex formulation we use (1) to compute our estimated pressure field at the control points as

$$\mathbf{p}^{est}_{b,cp}(\mathbf{r}_{cp}) = \sum_{l=1}^{L} \mathbf{G}(\mathbf{r}_{b,cp}|\mathbf{r}'_l)\mathbf{d}^{C}_{dlpm}(\mathbf{r}'_l), \quad (10)$$

$$\mathbf{p}^{est}_{d,cp}(\mathbf{r}_{cp}) = \sum_{l=1}^{L} \mathbf{G}(\mathbf{r}_{d,cp}|\mathbf{r}'_l)\mathbf{d}^{C}_{dlpm}(\mathbf{r}'_l). \quad (11)$$

We apply the Mean Absolute Error (MAE) to different components of the pressure fields to build our loss function, defined as

$$\begin{aligned}\mathcal{L}_{MAE}(\mathbf{p}^{des}_{cp}, \mathbf{p}^{est}_{cp}) = (\lambda_{abs}(||\mathbf{p}^{des}_{b,cp}| - |\mathbf{p}^{est}_{b,cp}||)+ \\ (|\angle\mathbf{p}^{des}_{b,cp} - \angle\mathbf{p}^{est}_{b,cp}|))+, \\ +\lambda_d(\lambda_{abs}(||\mathbf{p}^{des}_{d,cp}| - |\mathbf{p}^{est}_{d,cp}||)\end{aligned} \quad (12)$$

where the loss is calculated over the whole batch, and λ_{abs} and λ_{dark} are two weights empirically estimated. Note that since our goal is to correctly reproduce the bright zone and only to attenuate the dark zone, we completely discarded the phase of the dark zone. A schematic representation of the training procedure is shown in Fig. 2.

4. Results

In this section we show simulation results aiming to demonstrate the effectiveness of the proposed technique. We start by describing the metrics used to evaluate our system, and the setup built to run our simulations. We then present the obtained results along with discussions and interpretations.

4.1. Evaluation Metrics

In order to evaluate the accuracy of the reconstructed soundfield, we computed the Mean Squared Error (MSE) for the amplitude distribution between the ground-truth and estimated pressure fields as

$$MSE_{abs} = \frac{\sum_{n=1}^{N}(|\mathbf{p}_{des}(\mathbf{r}_n, \omega)| - |\mathbf{p}_{est}(\mathbf{r}_n, \omega)|)^2}{N}, \quad (13)$$

where $|\cdot|$ represent the absolute value operator and N is the number of evaluation points.
Following [3] we also compute the Structural Similarity Index Measure (SSIM), which is a measure usually applied in image processing problems, that quantifies how much two images are similar, being 1 the value for identical images. In the considered scenario, it measures the accuracy of the reproduced wavefronts. Finally, we evaluate the difference between the acoustic potential energy in the bright and dark zones with the acoustic contrast (AC). Using the apex * to denote complex conjugate, AC is defined as

$$AC = \frac{e_b(n)}{e_d(n)} = \frac{\sum_{n=1}^{N} \mathbf{p}^*_b(n)\mathbf{p}_b(n)}{\sum_{n=1}^{N} \mathbf{p}^*_d(n)\mathbf{p}_d(n)}. \quad (14)$$

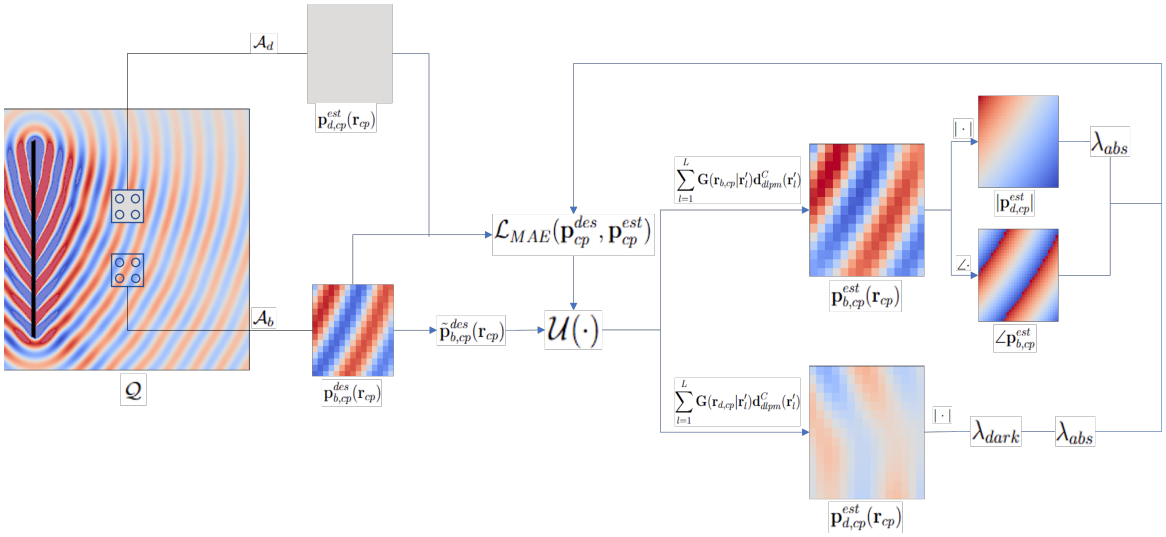

Figure 2: Schematic representation of the training procedure.

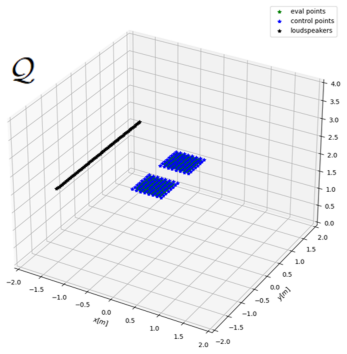

Figure 3: Experimental Setting in 3D environment

4.2. Simulation Setup

As shown in Fig 3 our environment is a free-field cubic room of dimensions $[-2m, 2m] \times [-2m, 2m] \times [0m, 4m]$, with the position $\mathbf{r}_0 = [0m, 0m, 2m].T$ being its centre and origin. Two square target regions $\mathcal{A}_b$ and $\mathcal{A}_d$ defined as in eq. 2 are placed for the generation of the two zones with high-and-low acoustic potential energy. The bright evaluation zone is centered at $[0.0m, 0.5m, 2m].T$, while the dark evaluation zone is centered at $[0.0m, -0.5m, 2m].T$. Both zones have a side of 0.5 m. We refer to *evaluation points* to define the points used for the evaluation of our system, i.e. the ones over which the metrics are calculated, while we refer to *control points* to define the points used for training our system, precisely the ones over which the loss is minimised. We'll use the subscripts *eval* and *cp* to refer to evaluation and control points, respectively. Evaluation points compose a dense distribution by means of a spacing of $\delta_{eval} \approx 0.02$ m, while control points are more sparse due to a spacing of $\delta_{cp} \approx 0.05$ m. Furthermore the side of the control zones is of 0.6 m. To sum up, each evaluation zone is composed by 512 evenly distributed points in an area of 0.25 m^2, while each control zone is composed by 128 evenly distributed points in an area of 0.36 m^2. Finally, we use a ULA of $L = 64$ secondary sources, linearly distributed in the range $-1.5m \times [-1.5m, 1.5m] \times 2m$. Since closed-cabinet loudspeakers behave similarly to point sources we can model our secondary sources using the Green's Function

$$G(\mathbf{r}|\mathbf{r}', \omega) = \frac{e^{-j(\frac{\omega}{c})\|\mathbf{r}-\mathbf{r}'\|}}{4\pi\|\mathbf{r}-\mathbf{r}'\|}, \tag{15}$$

where $\mathbf{r}$ will correspond to $\mathbf{r}_{cp}$ in the training phase and $\mathbf{r}_{eval}$ in the testing phase.
The Green's Function is used also to model the virtual sources in (7).

4.3. Dataset Generation

To train the network we consider a dataset of $\mathcal{S}$ virtual sources, with a cardinality of $\#\mathcal{S} = 1500$. We then randomly sample from $\mathcal{S}$ to generate two datasets $\mathcal{S}_{train}$ and $\mathcal{S}_{val}$ for training and validation, respectively. These two datasets have a cardinality of $\#\mathcal{S}_{train} = 1200$ and $\#\mathcal{S}_{val} = 300$. The sources of $\mathcal{S}$ are placed in a rectangular area covering the range $[-3.75m, -1.75m] \times [-1.5m, 1.5m] \times 2m$, with a spacing of 0.04 m along the x axis and a spacing of 0.1 m along

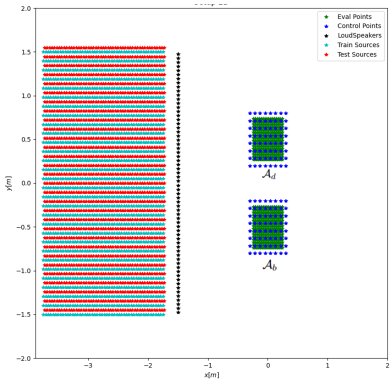

Figure 4: Virtual sources distribution for the generation of train set and test set.

the y axis. A last test dataset $\mathcal{S}_{test}$ of cardinality $\#\mathcal{S}_{test} = 1500$ is created by shifting the $\mathcal{S}$ by $0.02m$ on the x axis and by 0.05 on the y axis, as shown in Fig. 4. The signals emitted by the virtual sources are sinusoids, with $K = 64$ frequency values linearly spaced between 23.4375 Hz and 1500 Hz. We train our model for 5000 epochs and apply early stopping with a patience of 100 epochs, tracking the value of the loss of $\mathcal{S}_{val}$. Finally, we set the parameters for the loss (12) $\lambda_{abs} = 25$ and $\lambda_{dark} \approx 1$ and we adopt the Adaptive Moment (Adam) optimiser initialising the learning rate $lr = 0.001$.

4.4. Discussion

In the following discussion, we'll show the resulting pressure field for virtual source located outside the considered reproduction zone $\mathcal{Q}$ at position $\mathbf{r}_{\hat{s}} = [-3.75m, 1.5m, 2m]$ emitting a spherical wave at frequency $f_k = 961\ Hz$. From Fig. 5 we can see how the proposed method is able to accurately reproduce the desired pressure field in $\mathcal{A}_b$, while also achieving a high acoustic contrast. PM and AM tend to better focus on the area under study, but at cost of a lower acoustic contrast. ACC achieves a high acoustic contrast, but synthesised sound field is completely different from the desired one. In Figs 6, 7 and 8, we show a more quantitative comparison by averaging over all frequencies to show metric's behaviour as a function of the distance to the line that connects the centres of $\mathcal{A}_b$ and $\mathcal{A}_d$, and by averaging over all test positions to show metric's behaviour as a function of frequency f_k. The plots show how our approach has a reproduction error lower w.r.t. the other approaches in both cases, as a function of frequency and position.

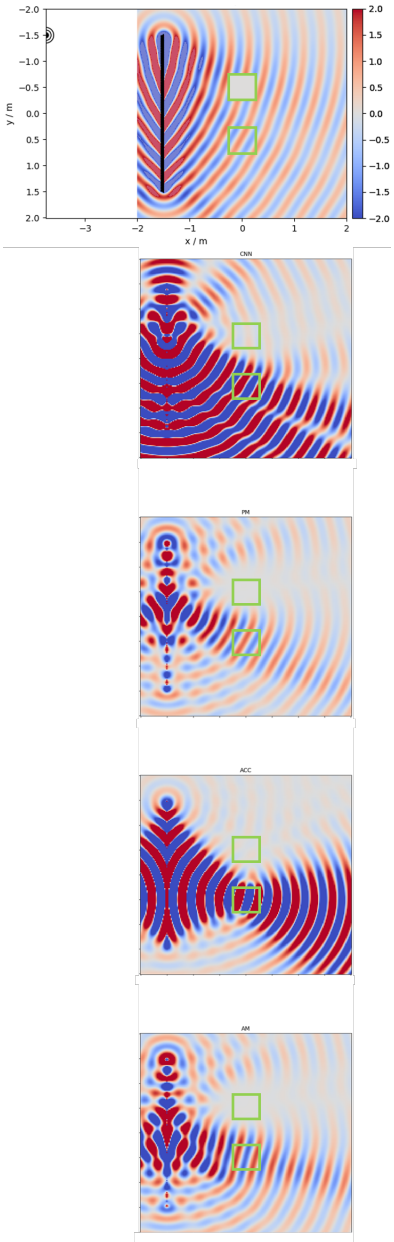

Figure 5: Pressure fields emitted by a virtual point source at position $\mathbf{r}_{\hat{\mathbf{s}}}$. In top-bottom order, the desired sound field and acoustic fields of MZ-DLPM, PM, ACC, AM. Each pressure field has been normalised w.r.t. their amplitude at position $\mathbf{r}_0$.

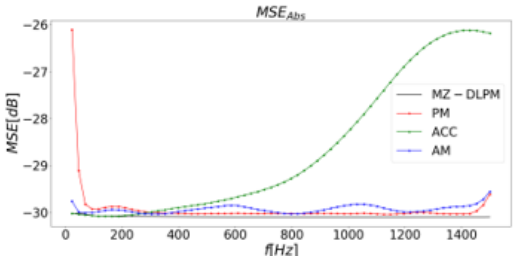

Figure 6: MSE of the absolute values as a function of frequency in $\mathcal{A}$.

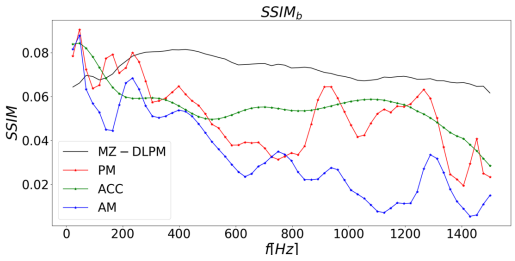

Figure 7: SSIM as a function of the frequency in $\mathcal{A}_b$.

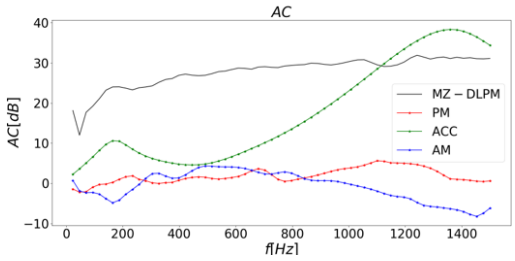

Figure 8: AC as a function of the frequency in $\mathcal{A}$.

When considering MSE as a function of position, we can see how there is a tendency to degradate as we approach the evaluation zones. This could be due to fact that pressure waves from farther virtual sources are more attenuated when they arrive to the bright zone, hence a lower acustic contrast is necessary, which leads to a better accuracy. Also SSIM shows how our technique is able to reproduce wavefronts that better resemble those of the ground truth. We can see how as the wavelength decreases, for all methods also decreases the similarity to the statistical distribution of the pressure values. Finally, for what concerns AC, our approach is able to surpass in some cases ACC, which in previous studies - as confirmed by our analysis - was demonstrated to be by far the one that achieves a higher contrast. ACC starts outperforming the proposed method as we approach higher frequencies or represent virtual sources near the ULA. By comparing this plot with the MSE we can deduct that our approach's performance tends to degrade both in terms of accuracy and acoustic contrast as virtual sources approach approach the ULA.

5. Conclusions

In this thesis's summary we have presented a technique for multi-zone sound field synthesis by means of deep neural network. Setting two zones with high and low acoustic potential energy - namely, *bright* and *dark* - we retrieve the desired driving signals by feeding the ground truth sound field at a series of control points in the bright zone into a convolutional neural network and computing the loss separately for the the amplitude and phase of the bright zone, and amplitude of the dark zone. Results demonstrate the validity of the proposed method and its ability to overcome the MSE-AC trade-off. Future works could include reverberation and noise in the environment, and evaluation of the proposed system as we increase the number of target zones or reduce the number of loudspeakers.

6. Acknowledgements

I would like to deeply thank my advisor Prof. Fabio Antonacci for giving me the possibility to develop this project for my Master thesis and my co-advisor Luca Comanducci for the advices throughout all the phases of the development.

References

[1] Takumi Abe, Shoichi Koyama, Natsuki Ueno, and Hiroshi Saruwatari. Amplitude matching for multizone sound field control. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:656–669, 2023.

[2] Jung-Woo Choi and Yang-Hann Kim. Generation of an acoustically bright zone with an illuminated region using multiple sources. *The Journal of the Acoustical Society of America*, 111:1695–700, 05 2002.

[3] Luca Comanducci, Fabio Antonacci, and Augusto Sarti. A deep learning-based pressure matching approach to soundfield synthesis. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5, 2022.

[4] Shoichi Koyama, Keisuke Kimura, and Natsuki Ueno. Sound field reproduction with weighted mode matching and infinite-dimensional harmonic analysis: An experimental evaluation, 2021.

[5] mark poletti. an investigation of 2-d multizone surround sound systems. *journal of the audio engineering society*, october 2008.