Audio Engineering Society

Convention Express Paper 138

Presented at the 155th Convention
2023 October 25–27, New York, USA

Loudspeaker position identification using human speech directivity index

Adrian Celestinos[1], Carren Zhongran Wang[1], and Victor Manuel Chin Lopez[2]

[1]*DMS Audio, Samsung Research America, USA.*
[2]*SRT, Samsung, Mexico.*

Correspondence should be addressed to Adrian Celestinos (a.celestinos@samsung.com)

ABSTRACT

A regular user of a multichannel loudspeaker system in typical living rooms sets the loudspeakers in a non-uniform manner, with angles and distances that don't necessarily follow the recommended ITU-R BS.2159-4 standard. Assuming a multichannel audio system equipped with N number of loudspeakers and M very near-field (NF) microphones attached to each loudspeaker, the user location with respect to the loudspeakers can be estimated by utilizing a supervised machine learning (ML) model. Two neural networks (NN) were trained with the human speech directivity index (DI) computed by room simulations, where the sound source was the typical directivity radiation pattern of human speech, and the receivers were the NF microphones attached to the loudspeakers. The distances between loudspeakers and the DI data was combined as input for the two NN models. One network was dedicated to estimate distances from loudspeaker to user, and the other network was dedicated to the angle estimation. The results shown a 95% confidence interval (CI) of $\pm$1.7 cm and a CI of $\pm$7 degrees for the incidence angle.

1 Introduction

Often a typical user of a multichannel loudspeaker sound reproduction in regular living rooms locates the loudspeakers in a non-uniform manner, with angles that do not necessarily follow the recommendation ITU-R BS.2159-4 standard [1], and with inconsistent distances from each loudspeaker to the user. By identifying the physical loudspeaker location, a spatial correction can thus be applied to recreate the artistic intention of the producer. For example, the differences in arrival time from each loudspeaker to the user can be compensated adding digital delays so all loudspeakers sound waves arrive a the same time to the user.

More elaborated methods can be applied to correct for the spatial perception for non-standard loudspeaker positions [2, 3]. In [4], Moulin et al. presented a perceptual evaluation of the compensation suggested by the MPEG-H 3D audio encoder, detailed in standard ISO/IEC: 23008-3 [5], for loudspeaker misplacement from the standard positions. The purpose of this study is to find out if it is possible to exploit the human speech

directivity pattern in order to locate the position of loudspeakers and user head orientation for calibration of a standard multichannel sound reproduction system. It is well known that humans communicate better by talking facing to each other, this is because we do not radiate sound uniformly around the head, more energy is radiated forward than backwards due to the mouth location in the head. In [6], detailed measurements of sound fields measured in anechoic conditions around human talkers were reported. In a later paper the dynamic directivity of humans when speaking or singing was studied [7]. Several studies have been carried out analysing the directivity patterns of human speech.

In this study it was assumed a multichannel audio sound reproduction system equipped with N number of loudspeakers and M very near-field (NF) microphones attached to each loudspeaker placed in a room around a human speaker as source. Room simulations were carried out to recreate human voice commands in different rooms using not standard loudspeaker positions around the user. Then the DI was extracted from a voice command recorded in anechoic conditions and convolved with the simulated RIRs. Machine learning (ML) was employed to learn the relationship between the DI and the position of the loudspeakers with respect to the user orientation.

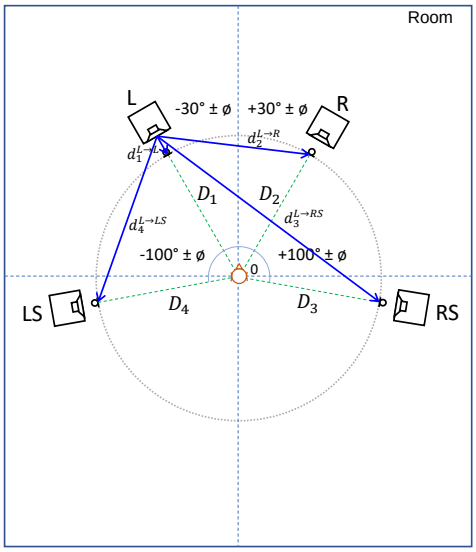


Fig. 1: Loudspeaker, room and NF microphones setup. Blue arrows represent the distance from loudspeaker L to NF microphones.

2 Methods

In this section the methods utilized in this study are presented, first the directivity extraction is explained, after the machine learning networks and training, then the room simulations, loudspeaker setup and flow processing are detailed.

2.1 Directivity Index

The human voice presents a unique directivity pattern which is frequency, angle/direction and distance dependent; the computed DI carries that information out. The DI represents an acoustical energy ratio of one specific direction to all directions. In Eq. 1, (adapted from [8]) the directivity index is described,

$$DI(w) = 10 \cdot log_{10} \frac{|H_0(w)|^2}{\frac{1}{N}\sum_{n=0}^{N-1} |H_n(w)|^2}, \qquad (1)$$

where the nominator is the acoustical energy computed from H sound pressure, radiated towards 0°, and the denominator is the average energy radiated from all n directions. N is the total number of directions and w is the angular frequency $w = 2\pi f$, were f are discrete frequency bands.

2.2 Neural Networks

Typically, the term artificial intelligence (AI) is used when a machine emulates cognitive functions that humans associate with other human minds, such as learning and problem solving. In this paper we explore the use of machine learning to solve and learn the relationship between the human speech directivity index (DI) recorded by microphones attached to loudspeakers in a typical multichannel surround loudspeaker setup. The positions of the loudspeakers were expressed in polar coordinates using incidence angle and distance to the user which is the origin of the polar coordinate system. More specifically we used two feed forward neural networks in MATLAB [9], to automatically estimate the position of the loudspeakers around the user in the room.

2.2.1 Feed-forward Neural Network

The Feed-forward neural network (FFNN) is one of the first successful artificial neural networks and is known for its simplicity. The information is only processed

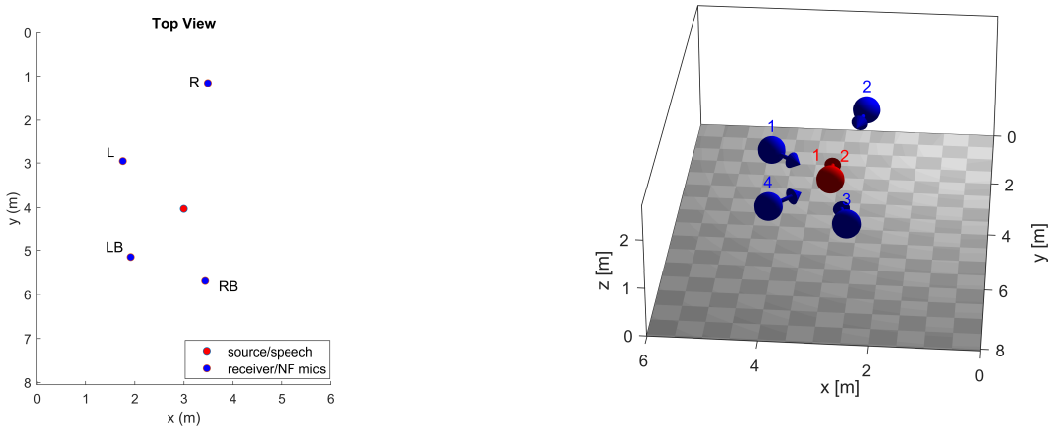


Fig. 2: Room simulation setup. Left plot, coordinate system view. Right plot, 3D room source and receivers locations. The red dot indicates user position, blue dots receiver positions.

forward in the network. As the universal approximation theorem describes, using one single hidden layer with enough hidden neurons can approximate any continuous function [10]. The neural network (NN) applied here consists of two network models, one for the distance to the user, and one for the incidence angle. The supervised training of the NN was performed with DI data computed from human speech in-room simulations and the actual distance and angle coordinates from each loudspeaker to the user.

2.3 Room Simulation Model

A simulation model was utilized to replicate a typical human speech recorded by receivers placed in typical loudspeaker positions around the source (user).The simulation software `MCRoomSim` described in [11] models both specular and diffuse reflections in a shoe-box type. The model runs in Matlab and is based on the image source method algorithm which provides accurate direction and timing of the primary reflections. With this model it is possible to simulate room impulse responses (RIR) and microphone arrays with arbitrary directional sensitivity and large numbers of receivers. The software package includes typical female and male directivity which was included for the source in the simulations. Since the NF microphones are attached to the loudspeakers very close to the driver, their response would be affected by the loudspeaker baffle. In the simulation model the NF microphone directivity was included. This was determined by using a finite element model of the loudspeaker geometry.

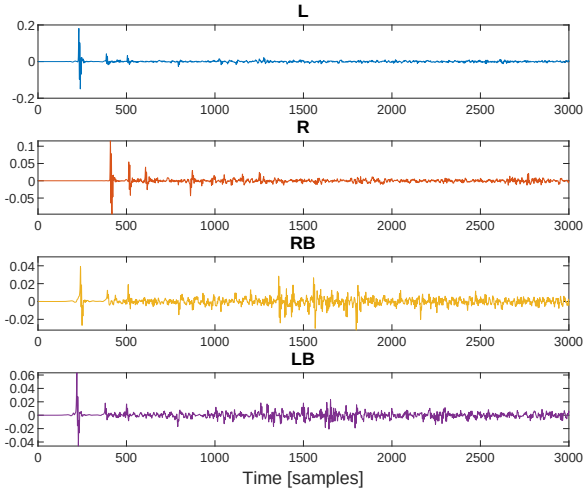


Fig. 3: Simulated RIR corresponding to the Figure 2 room-receiver setup.

A customized room generator was used to create shoebox room setups of various sizes. Each setup has material absorption coefficients chosen from a selection pool, with randomized receiver and source locations within some limits.

A total of 39 rooms were simulated. Among these simulations there were three room sizes from 80 to 300 cubic meters. In Figure 3, an example of simulated RIR are presented, the simulation IR length was set to 1 second, in the graph the RIRs were cropped to the first

Table 1. Room dimensions, volume and materials used for the simulations.

Room dimensions	Room Volume	Materials
Height = [2.7 - 3.0 m]	13 small size [80 - 150m^3]	Plasterboard on frame 100 mm cavity;
Width = [4.0 - 8.0 m]	13 medium size [150 - 220m^3]	Mineral wool in cavity, surface painted;
Length = [6.0 - 12.0 m]	13 large size [120 - 300m^3]	Double glazing, 2-3mm glass, 10mm air gap
		Plywood, hardwood panels over 25mm airspace;
		Wooden floor on joists;
		Rubber floor tiles;
		Carpet, thin, over thin felt on wood floor;

3000 samples to show the corresponding arrival time.

The RIRs at the NF microphone were convolved with anechoic male and female monoaural recordings. Thus, the convolved audio is the result from the simulation representing the voice command that the loudspeaker multichannel NF microphones are supposed to "record" in each simulation case.

2.4 Loudspeaker Setup

For this study we have utilized four loudspeakers in a multichannel configuration, two front Left and Right loudspeakers to reproduce left, right and phantom center signals. And one Rear Left and Rear Right loudspeakers to reproduce left surround and right surround signals from a 5.1 program material respectively. A compact loudspeaker prototype was designed for the purpose of this study. Four 11 $\times$ 11 $\times$ 11 cm sealed boxes with a 51 mm full-range driver each, and a 0.6 L volume were built. A miniature MEMS microphone was attached with a mechanical fixture in front of the driver at approx. 2 cm from the diaphragm to record the human voice as seen in Figure 4.

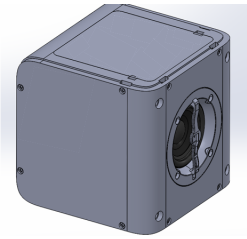


Fig. 4: Loudspeaker prototype with NF microphone.

2.5 Flow Processing

In this section the audio processing and data generation are detailed.

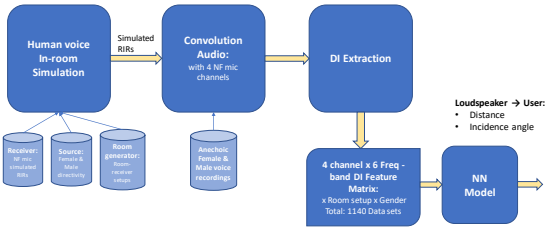


Fig. 5: Flow processing block diagram.

2.5.1 Data Generation

The data generation starts with the human voice in-room simulation. The simulation consisted on creating source and receiver setups where the location of each loudspeaker/receiver is randomly placed within a limited sector of the room. The user/source was positioned at the center of the room and from there each loudspeaker/receiver location was generated. The orientation of the user was always facing the front wall at 0° as seen in Figure 1.

A total of 1140 setups consisting of 570 $\times$ two gender sets of RIRs data were computed, (see Table 1). For each loudspeaker a base angle was used being the optimal positioning angle according to the ITU-R BS.775-1 standard [1], then a random value was added on a range of +/- ϕ°. For the distance D_n from loudspeaker to the user a random number between the range of $[D_{min}, D_{max}]$ was set as detailed in Figure 1. The distance from each loudspeaker to each NF microphone is calculated from the driver's coordinate to the NF microphone coordinate and save for each setup. On a real loudspeaker setup these distances would be calculated from the RIR measured with the NF microphones.

For each loudspeaker location two coordinates were generated, one for the driver and one for the NF microphone. Each setup includes four simulated RIRs by gender. Then the RIRs are convolved with the anechoic recordings. Next the audio is passed through a high-pass filter at 100 Hz to remove unwanted low frequency noise.

2.5.2 DI extraction

The DI is computed from the four anechoic recordings convolved with the RIRs, using a $1/3^{\text{rd}}$ octave-band filter. The values at $f_1 = 250$ Hz, $f_2 = 500$, $f_3 = 1$ kHz, $f_4 = 2$ kHz, $f_5 = 4$ kHz and $f_5 = 8$ kHz frequency-bands were obtained thus generating 4×6 DI matrix per gender and per loudspeaker setup. In Figure 6 a polar plot example of extracted DI from the simulated setup shown in Figure 2 is shown.

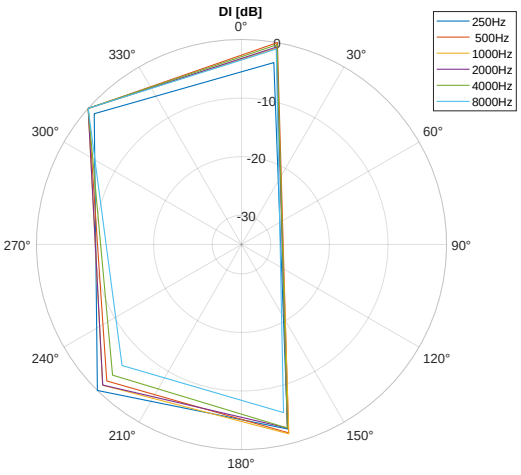


Fig. 6: Extracted DI from a simulated setup.

In Figure 7 the average $1/3^{\text{rd}}$ octave band energy example extracted from the convolved audio with RIRs simulated in setup of Figure 2 is shown. The DI is normalized per frequency-bands over all channels. Each frequency-band has a maximum of 0 dB DI.

2.5.3 NN Training

A supervised training was was carried out for the distance and angle models. The 4×6 DI matrix

$$\begin{pmatrix} DI_1^{L-f_1} & DI_2^{L-f_2} & \dots & DI_6^{L-f_6} \\ DI_7^{R-f_1} & DI_8^{R-f_2} & \dots & DI_{12}^{R-f_6} \\ DI_{13}^{RS-f_1} & DI_{14}^{RS-f_2} & \dots & DI_{18}^{L-f_6} \\ DI_{19}^{LS-f_1} & DI_{20}^{LS-f_2} & \dots & DI_{24}^{L-f_6} \end{pmatrix}, \quad (2)$$

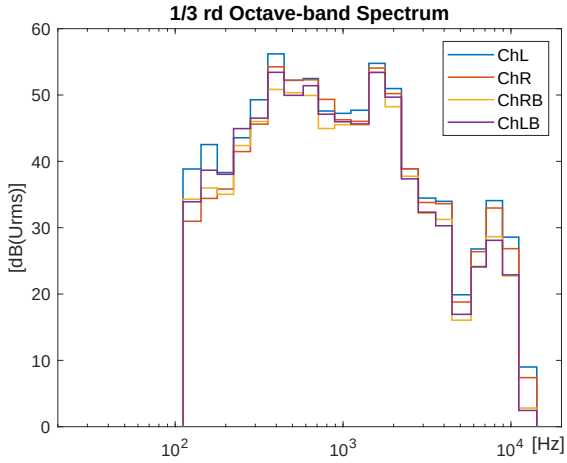


Fig. 7: Average $1/3^{\text{rd}}$ octave-band energy example extracted from the convolved audio with RIRs.

and the 4×4 loudspeaker distances matrix

$$\begin{pmatrix} d_1^{L\to L} & d_2^{L\to R} & d_3^{L\to RS} & d_4^{L\to LS} \\ d_5^{R\to L} & d_6^{R\to R} & d_7^{R\to RS} & d_8^{R\to LS} \\ d_9^{RS\to L} & d_{10}^{RS\to R} & d_{11}^{RS\to RS} & d_{12}^{RS\to LS} \\ d_9^{LS\to L} & d_{10}^{LS\to R} & d_{11}^{LS\to RS} & d_{12}^{LS\to LS} \end{pmatrix}, \quad (3)$$

were combined and reshaped into a one dimensional array of size N = 40 which is used as the raw input data for the NN model training. The prediction target is the actual distance and incidence angle from loudspeaker to user. The DI data was passed into a principal component analysis (PCA) process to exclude redundant data information [12]. The DI dB units data was converted to linear amplitude values before entered into the PCA block in order to facilitate the method. Then the 1140 data cases were split into 80% training, 10% test and 10% validation respectively. Two NN models were trained, one for distance and one for angle prediction. The distance prediction model contained an input layer with M features, a hidden layer with 13 neurons and a Tanh activation layer.

From there a second hidden layer with 95 neurons and Tanh activation was connected before the output layer that contained 4 neurons, see Figure 8. The following parameters were used for the distance prediction network:

- Max Epochs: 20,000 (Finished at 16054)
- Max fail: 5,000

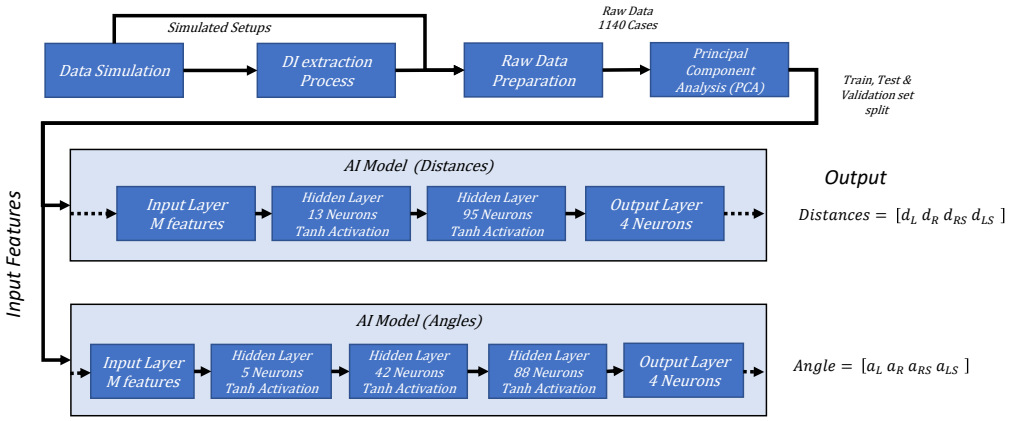


Fig. 8: NN training process block diagram.

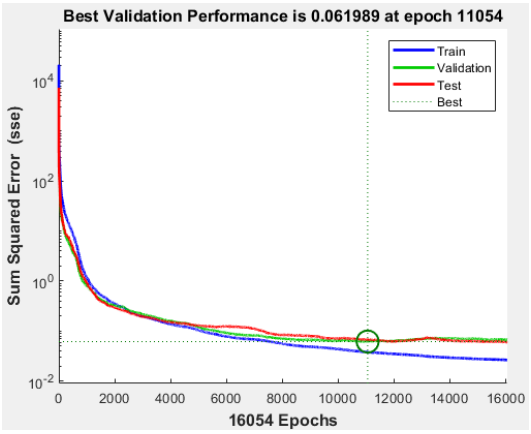


Fig. 9: Training performance graph. Distance prediction model.

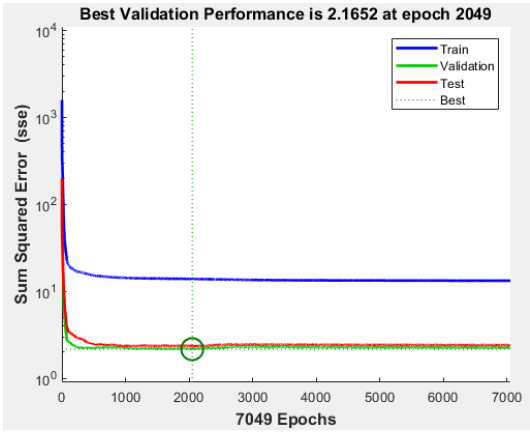


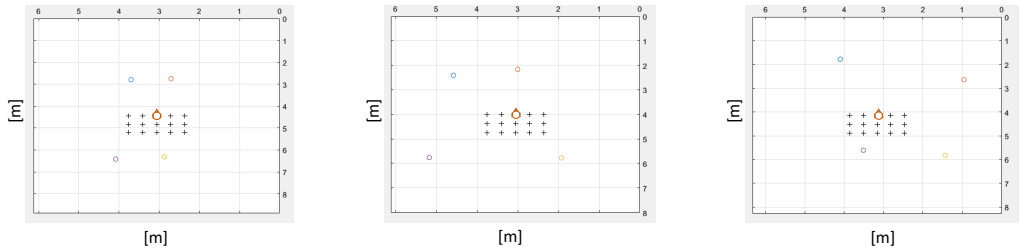
Fig. 10: Training performance graph. Angle prediction model.

- Optimizer: Scaled Conjugated Gradient (SCG)
  - Goal: 0
  - Minimum gradient: 1×10^{-6}
  - μ: 5×10^{-3}
  - σ: 5×10^{-5}
  - λ: 5×10^{-7}
- Performance function: Sum of the square error

As observed in Figure 8, the angle prediction network consisted of an input layer with M features, then 3 hidden layers were connected in cascade with 5, 42 and 88 neurons respectively. The three hidden layers contained Tanh activation. Then the model was finalized with an output layer with 4 neurons.

In Figure 9 the distance performance model graph is shown. The same parameters used for the distance prediction model were utilized for the angle prediction network.

In Figure 10 the angle performance model is shown. The training of the distance model had its best validation performance on epoch 11054, while the angle prediction model had its best validation performance on the epoch 2049.



Loudspeaker-User		L	R	RS	LS
Distance	*Expected*	1.7714	1.7335	1.8758	2.219
(m)	*Predicted*	1.7703	1.7417	1.8807	2.2186
	Mean error				0.0029
Angle	*Expected*	-21.047	11.8857	174.152	-152.64
(deg)	*Predicted*	-19.325	13.0779	175.632	-150.17
	Mean error				1.7174°

Loudspeaker-User		L	R	RS	LS
Distance	*Expected*	2.2111	1.8428	2.0993	2.7479
(m)	*Predicted*	2.2147	1.8472	2.0974	2.7419
	Mean error				4E-05
Angle	*Expected*	-43.529	1.7956	147.422	-129.67
(deg)	*Predicted*	-44.428	0.8276	145.978	-130.92
	Mean error				1.1406°

Loudspeaker-User		L	R	RS	LS
Distance	*Expected*	2.5435	2.6693	2.4102	1.5086
(m)	*Predicted*	2.5461	2.6747	2.4132	1.5063
	Mean error				0.0022
Angle	*Expected*	-21.6	55.8047	134.204	-166.53
(deg)	*Predicted*	-16.728	60.8042	137.061	-162.6
	Mean error				4.1666°

Fig. 11: Results example of three loudspeaker layouts. Predicted, expected and mean error shown.

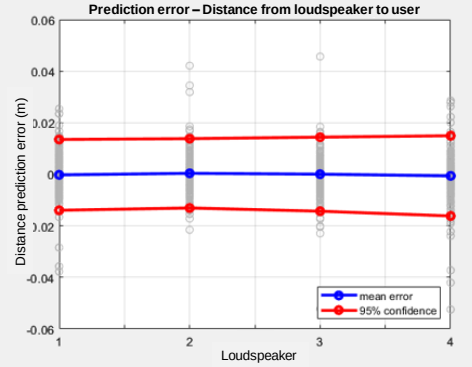


Fig. 12: Distance prediction error.

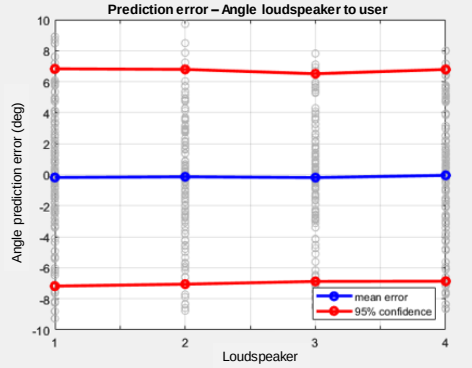


Fig. 13: Angle prediction error.

3 Results

In this section the results of the study are presented. The neural networks were evaluated with the 10% of the data which was not used in the training. This is to test how well the model is able to generalize. In Figures 12 and 13, the 95% confidence interval and mean prediction error for distance and angle are shown respectively. An error of $\pm$ 1.7 cm, and $\pm$ 7° was achieved for the distance and incidence angle from the loudspeakers to the user. In Figure 11, an example of results of three loudspeaker layouts is shown.

4 Discussion

Results shown a very good distance prediction, which is crucial for spatial correction on an irregular loudspeaker layout. As explained in Section 1, the first possible spatial correction for a loudspeaker setup that was not correctly located with respect to the user, is to compensate with delays corresponding with the propagation distance from the closest loudspeaker to the further loudspeaker. If this parameter is well predicted, a reasonable improvement in the spatial audio quality can be obtained.

It would be desirable to obtain a better prediction angle, but it seems that the typical human speech directivty pattern due to the shape of the human head and the position of the mouth does not allow to refine the learning process. The lack of precision on the prediction angle can be also attributed to the limited amount of microphones around the user.

Future research will include the implementation of the recording voice command on the actual NF prototype and verification on a real case scenario. On these cases human head physical differences, background noise and other issues might play a significant role on the prediction performance. Another subject of discussion is that the receiver-room simulation is based on RIR, assuming that the human speech is produced by a sound source with linear time invariant properties. In reality the human voice has different dynamic features that are difficult to simulate. A topic not addressed in this study is the influence of the room on the DI extraction, since typically the DI is calculated in free field conditions, this can be also another subject for research.

5 Summary

Several room simulations of human speech on typical multichannel loudspeakers equipped with very near field microphones have been carried out in order to automatically estimate the loudspeaker position with respect to the user in the room. The simulations included the directivity sensitivity of the receivers due to the position of the microphone in front of the driver of the loudspeaker. The resulting 570×2 RIRs were convolved with male an female anechoic voice recording commands. The DI was extracted from the four audio processed channels. Machine learning in the form of two NN was utilized to predict the distance and incidence angle with respect to the user. Two FFNN were trained with the processed data. Results shown on the evaluation of the models a 95% confidence interval (CI) of ±1.7 cm for loudspeaker to user distance and a CI of ±7 degrees for the loudspeaker incidence angle.

Acknowledgements

Samsung Electronics and Samsung Research America supported this work. The authors would like to thank the entire staff of Samsung's US Audio Lab who contributed to this work with help and offered insightful suggestions.

References

[1] International Telecommunication Union, *ITU-R BS.775-1, Multichannel stereophonic sound system with and without accompanying picture*, 1992.

[2] Poletti, M., "Robust Two-Dimensional Surround Sound Reproduction for Nonuniform Loudspeaker Layouts," *J. Audio Eng. Soc.*, 55, pp. 598–610, 2007.

[3] Pulkki, V., "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc*, 45(6), pp. 456–466, 1997.

[4] Moulin, S., Pallone, G., Faure, N., and Bech, S., "Perceptual Evaluation of Loudspeaker Misplacement Compensation in a Multichannel Setup Using MPEG-H 3D Audio Renderer. Application to Channel-Based, Scene-Based, and Object-Based Audio Materials," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, 2019.

[5] International Organization for Standardization, *ISO/IEC: 23008-3, Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio*, 2022.

[6] Chu, W. and Warnock, A., "Detailed Directivity of Sound Fields Around Human Talkers," 2002, doi:10.4224/20378930.

[7] Pörschmann, C. and Arend, J., "Analyzing the Directivity Patterns of Human Speakers," 2020.

[8] Kocon, P. and Monson, B. B., "Horizontal directivity patterns differ between vowels extracted from running speech," *The Journal of the Acoustical Society of America*, 144(1), pp. EL7–EL12, 2018, ISSN 0001-4966, doi:10.1121/1.5044508.

[9] Deep Learning Toolbox version 14.1, "MATLAB," (R2020b), the MathWorks, Natick, MA, USA.

[10] Hornik, K., Stinchcombe, M., and White, H., "Multilayer feedforward networks are universal approximators," *Neural Networks*, 2(5), pp. 359–366, 1989, ISSN 0893-6080.

[11] Wabnitz, A., Epain, N., Jin, C., and van Schaik, A., "Room acoustics simulation for multichannel microphone arrays," 2010.

[12] Jolliffe, I., *Principal component analysis*, Springer Verlag, New York, 2002.