Direction of Arrival Estimation of Noisy Speech using Convolutional Recurrent Neural Networks with Higher-Order Ambisonics Signals

Nils Poschadel (iD), Robert Hupke (iD), Stephan Preihs (iD), and Jürgen Peissig (iD)
Institute of Communications Technology
Leibniz University Hannover, Hannover, Germany
Email: {poschadel, hupke, preihs, peissig}@ikt.uni-hannover.de

Abstract—Training convolutional recurrent neural networks on first-order Ambisonics signals is a well-known approach when estimating the direction of arrival for speech/sound signals. In this work, we investigate whether increasing the order of Ambisonics up to the fourth order further improves the estimation performance of convolutional recurrent neural networks. While our results on data based on simulated spatial room impulse responses show that the use of higher Ambisonics orders does have the potential to provide better localization results, no further improvement was shown on data based on real spatial room impulse responses from order two onwards. Rather, it seems to be crucial to extract meaningful features from the raw data. First order features derived from the acoustic intensity vector were superior to pure higher-order magnitude and phase features in almost all scenarios.

Index Terms—Direction of arrival estimation, higher-order Ambisonics, convolutional recurrent neural network, spherical harmonics.

I. INTRODUCTION

Estimating the direction of arrival (DOA) of sound/speech is a key problem in acoustic signal processing. Neural networks have been shown to be superior to classical parametric approaches in this task, especially in reverberant, noisy, and low-SNR environments [1]–[4]. Recently, DOA estimation based on first-order Ambisonics (FOA) signals has been the subject of much attention [4]–[7]. Due to the flexibility and generalizability of the Ambisonics approach, it more or less enables microphone-array-independent DOA estimation models.

Perotin et al. [4], [8], [9] investigated the effect of different parameters when training convolutional recurrent neural networks (CRNNs) on FOA data for the DOA estimation of noisy speech. They proposed the usage of features derived from the sound intensity vector as input for the training, achieving greater accuracy in DOA estimation than with using pure magnitude and phase information [8]. Furthermore, they showed that a regression approach is at least as suitable as a classification interpretation for a single-source DOA estimation with diffuse interference and that a CRNN trained on spherical coordinates performs worse than a network trained on Cartesian coordinates when using the mean squared error (MSE) or angular distance as loss function [9].

However, despite the increasing availability of higher-order ambisonics (HOA) microphones, very little research is con-

ducted on the performance of DOA estimators based on HOA signals.

There are some results from other applications where the usage of HOA signals is advantageous over the use of FOA signals. Pointer experiments with subjects showed a positive influence of the order of the Ambisonics signal on perceptual localization accuracy in a loudspeaker reproduction of a sound field [10]. Similarly, the higher-order model of directional audio coding (HO-DirAC) achieved a higher reproduction accuracy than the first-order DirAC in a perceptual evaluation [11], [12]. Investigations on spherical harmonic (SH) beamforming with unsupervised peak clustering [13] also showed an improvement in localization accuracy with increasing Ambisonics order. However, to our knowledge, this topic has not yet been investigated or even quantified for state of the art deep learning approaches for DOA estimation.

This work therefore is the first to apply the idea of CRNN-based DOA estimation to HOA signals and to investigate whether or how much the additional spatial information contained in HOA can improve the estimation accuracy. We thereby compare our HOA models with FOA models based on both magnitude/phase spectrograms and spectrogram features derived from the acoustic intensity vector.

To the best of our knowledge, there is no sufficiently large dataset of HOA speech signals or impulse responses available. Therefore, we had to create a suitable dataset of noisy speech data with different orders of Ambisonics, taking inspiration from the procedure used in [9] for creating a FOA dataset. Due to the way we parameterize the impulse response simulation, this dataset can not only be used for training deep learning models for DOA estimation. The dataset also contains labels regarding room size/geometry as well as acoustic properties such as reverberation time and absorption/scattering coefficients and will serve as the basis for a number of studies in the context of acoustic analysis based on HOA signals.

We present the details on the generation of our training, validation, and testing data in Sec. III after a brief introduction to the fundamentals of Ambisonics and SH in Sec. II. The configuration of the trained model and the metrics are described in Sec. IV. Finally, the results based on simulated and measured data are compared and discussed in Sec. V and summarized in Sec. VI.

II. Ambisonics

Ambisonics is a 3D audio surround representation and rendering approach based on the spatial decomposition of the sound field in the orthonormal basis of SH [4], [14]. This section gives an overview of the mathematical principles of Ambisonics. This condensed description of the SH decomposition is based on the more detailed presentation in [15], [16].

In the following, the Cartesian $(x, y, z) \in \mathbb{R}^3$ and the spherical $(r, \theta, \phi) = (r, \Omega) \in [0, \infty) \times (-\frac{\pi}{2}, \frac{\pi}{2}] \times [-\pi, \pi]$ coordinate systems are used. The x-, y- and z-axes point to the front, left and top, respectively. The angle ϕ is the azimuth, which is zero at the frontal direction and increasing counterclockwise; θ is the elevation, which is zero at the horizontal plane and positive above, and r is the radius.

Consider a function $f(\theta, \phi) = f(\Omega) \in L^2(S^2)$ on the unit 2-sphere $S^2 := \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 = 1\}$, then the SH decomposition of f is given by

$$f(\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} Y_n^m(\Omega), \tag{1}$$

where Y_n^m is the *spherical harmonic* of order n and degree m. The coefficients f_{nm} are calculated by

$$f_{nm} = \int_{\Omega \in S^2} f(\Omega) Y_n^{m*}(\Omega) \, d\Omega, \tag{2}$$

where $\int_{\Omega \in S^2} d\Omega = \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} \sin\theta \, d\theta \, d\phi$. Equations (1) and (2) show that any square-integrable function on the unit 2-sphere can be approximated by a linear combination of the SH. This approximation even becomes exact for an infinite number of SH. In this paper, the ambiX format [14] is used for the (real) SH Y_n^m:

$$Y_n^m(\theta, \phi) = N_n^{|m|} P_n^{|m|}(\sin(\theta)) \begin{cases} \sin(|m|\phi), & \text{for } m < 0 \\ \cos(|m|\phi), & \text{for } m \geq 0 \end{cases}$$

with the Legendre-functions P_n^m. To build the set of Ambisonics signals according to ambiX, the channels corresponding to the SH are ordered by the Ambisonics channel number $\text{ACN} = n^2 + n + m$ and normalised by the SN3D normalisation

$$N_n^{|m|} = \sqrt{\frac{2 - \delta_m}{4\pi} \frac{(n - |m|)!}{(n + |m|)!}}.$$

In the special case of FOA, the channels 1-4 according to ACN are often referred to as W, Y, Z, X.

III. Data

A. Simulated SRIRs

The training, validation and testing data was generated from a set of spatial room impulse responses (SRIRs) simulated with the MCRoomSim toolbox [17] as Ambisonics signals up to fourth order corresponding to the ambiX format. The approach was inspired by the procedure described in [4]. Alltogether we generated 8000, 500, and 500 rooms with random dimensions in $[3, 20] \times [3, 20] \times [3, 5]$ m for the training, validation, and testing set, respectively. The acoustic properties

of the walls (frequency dependant scattering and absorption coefficients) were set to plausible, randomly chosen surfaces of the GRAP database [18]. For every room, one receiver was randomly positioned with a minimum distance of 1.5 m to the walls. Furthermore, one source was randomly positioned at 8 different locations such that the DOAs in the dataset are uniformly distributed. The distance from the source to the receiver was chosen randomly, ensuring that the source and the receiver are at least 1 m apart from each other and that the source is at least 49 cm from a wall. With this setup, we simulated 64 000, 4000 and 4000 fourth-order Ambisonics SRIRs. Although the experiments in this paper were conducted using speech signals with a sampling rate of 16 kHz, the SRIRs were simulated with a sampling frequency of 48 kHz to be able to expand the methods of this paper to general audio/music signals using the same database. After resampling, the SRIRs were convolved with a randomly chosen sentence from the TIMIT database [19]. This database contains a total of 6300 sentences, 10 sentences spoken by each of the 630 speakers (192 female, 438 male) from eight major dialect regions of the United States. The TIMIT database was split into training, testing and validation sets resulting in 462 (136 female, 326 male), 88 (30 female, 58 male), and 80 (26 female, 54 male) speakers, respectively. The training set corresponds to the recommended one by the authors of the TIMIT database. The test set includes the recommended core test set and it is ensured that there is at least one female/male speaker per dialect in the validation and test set, respectively.

Furthermore, we added ambient noise to the speech signals similar to the procedure in [4]. Therefore, we generated single-channel babble noise by overlaying 50 sentences of the respective sets. This babble noise was then convolved with a diffuse SRIR, which was generated by averaging three simulated diffuse parts of SRIRs with a receiver placed in the middle of a random room and a randomly positioned source. This ambient noise was added to the speech signal at a signal-to-noise ratio (SNR) between 0 and 20 dB. Finally, these sentences were cut to one-second-sequences which led to 164 303, 10 285 and 10 394 sequences for the training, validation and testing set, respectively.

B. Real SRIRs

For the analysis of DOA estimation performance in a more realistic scenario, we measured real SRIRs in the Immersive Media Lab (IML) [20] at the Institute of Communications Technology. We measured the SRIRs from each of our 36 KH120 loudspeakers to an em32 Eigenmike® [21] microphone at nine different positions, each with two different heights and eight different orientations of our microphone. In total, the described procedure led to 5184 measured SRIRs in the IML, which were afterwards encoded to a fourth-order Ambisonics signal using the EigenUnit-em32-encoder[1]. These measured SRIRs were used according to the same procedure as for the simulated SRIRs to generate HOA multispeaker signals which

---

[1] https://mhacoustics.com/eigenunits

resulted in 13 414 sequences for the testing set based on real SRIRs.

IV. DOA ESTIMATION FRAMEWORK

A. Networks and metrics

Our trained networks follow a similar basic CRNN structure compared to the ones in [3], [4]. A detailed overview of the network's architecture is given in Table I, where the final normalization layer scales the prediction to lie on the unit 2-sphere. We formulated this task as a regression problem with the MSE loss function and the Nadam optimizer [22]. For training the network, we used the TensorFlow platform [23]. Since we use a time-distributed output layer and assume

Layer	Details	Output Shape
Input	Spectrograms	$(50, 512, dim_{in})$
Conv2D	3×3	$(50, 512, n_{filter})$
BatchNorm		$(50, 512, n_{filter})$
Activation	elu	$(50, 512, n_{filter})$
MaxPooling	1×8	$(50, 64, n_{filter})$
Dropout	0.2	$(50, 64, n_{filter})$
Conv2D	3×3	$(50, 64, n_{filter})$
BatchNorm		$(50, 64, n_{filter})$
Activation	elu	$(50, 64, n_{filter})$
MaxPooling	1×8	$(50, 8, n_{filter})$
Dropout	0.2	$(50, 8, n_{filter})$
Conv2D	3×3	$(50, 8, n_{filter})$
BatchNorm		$(50, 8, n_{filter})$
Activation	elu	$(50, 8, n_{filter})$
MaxPooling	1×4	$(50, 2, n_{filter})$
Dropout	0.2	$(50, 2, n_{filter})$
Reshape		$(50, 2 \cdot n_{filter})$
BiLSTM		$(50, 2 \cdot n_{filter})$
BiLSTM		$(50, 2 \cdot n_{filter})$
Time-Dist. Dense	elu	$(50, 2 \cdot n_{filter})$
Dropout	0.2	$(50, 2 \cdot n_{filter})$
Time-Dist. Dense	linear	$(50, 3)$
Normalization		$(50, 3)$

TABLE I. Architecture of the CRNNs for DOA estimation.

the sources to be static over the whole duration of the signal, we first average the network outputs for each axis over time. We then compare the predicted DOA $(\hat{\theta}, \hat{\phi})$ with the reference (θ, ϕ) used to synthesize the dataset, using the *angular distance* $\delta\left[(\hat{\theta}, \hat{\phi}), (\theta, \phi)\right]$ defined by

$$\delta\left[(\hat{\theta}, \hat{\phi}), (\theta, \phi)\right] = \arccos\left[\sin(\hat{\theta})\sin(\theta) \right.$$
$$\left. + \cos(\hat{\theta})\cos(\theta)\cos(\hat{\phi} - \phi)\right].$$

For additional evaluation, we further define the so-called *accuracy* as the proportion of samples for which the prediction has an angular distance below a given error tolerance.

B. Input features

The input features of the networks based on HOA signals are pure magnitude and phase spectrograms. In the following, we will call these networks HOA-n-CRNN with n being the order of the HOA signal. We compare our HOA-n-CRNNs to two other published approaches for FOA DOA estimation with CRNNs. On the one hand, Adavanne et al. [3] used pure FOA magnitude and phase spectrograms (FOA-CRNN). Of course, HOA-1-CRNN and FOA-CRNN are identical and will

be referred to as HOA-1-CRNN in the following. On the other hand, Perotin et al. [4] proposed using spectrograms of 6-channel features derived from the FOA sound intensity vector according to (3) as input to the CRNN (Intensity-CRNN). By using these features, they were able to significantly improve the localization performance compared to using magnitude and phase spectrograms.

$$\frac{-1}{C(t, f)} \begin{bmatrix} \mathbf{I}_a(t, f) \\ \mathbf{I}_r(t, f) \end{bmatrix} \tag{3}$$

$\mathbf{I}_a(t, f)$ and $\mathbf{I}_r(t, f)$ describe the active and reactive intensity vector as a Short-time Fourier transform (STFT) expression of the FOA channels and $C(t, f)$ is a normalization term. They can be computed according to (4), (5), (6). For further details on acoustic intensity see [4], [24], [25].

$$\mathbf{I}_a(t, f) = -\begin{bmatrix} \text{Re}\{W(t, f)X^*(t, f)\} \\ \text{Re}\{W(t, f)Y^*(t, f)\} \\ \text{Re}\{W(t, f)Z^*(t, f)\} \end{bmatrix} \tag{4}$$

$$\mathbf{I}_r(t, f) = -\begin{bmatrix} \text{Im}\{W(t, f)X^*(t, f)\} \\ \text{Im}\{W(t, f)Y^*(t, f)\} \\ \text{Im}\{W(t, f)Z^*(t, f)\} \end{bmatrix} \tag{5}$$

$$C(t, f) = |W(t, f)|^2 + \frac{1}{3}(|X(t, f)|^2 + |Y(t, f)|^2 + |Z(t, f)|^2) \tag{6}$$

The input shape of all the different networks is $(50, 512, dim_{in})$, where 50 is the number of frames, 512 the number of frequency bins, and dim_{in} the number of input channels with $dim_{in} = 2(n + 1)^2$ for the HOA-n-CRNNs and $dim_{in} = 6$ for the Intensity-CRNN. The STFT for the creation of the spectrograms was performed on 640 samples, zero-padded to 1024 samples with a hop-size of 320 samples. For identifying the optimal number of filters n_{filter} (and thus the number of parameters), different values ranging from 32 to 1024 were tested for each network and the value which resulted in the lowest error on the validation set was chosen. The best values were 256 for the HOA-1-CRNN and HOA-2-CRNN and 512 for all the other networks.

V. RESULTS

As expected, the results belonging to the simulated SRIRs are overall slightly better than those belonging to the real SRIRs. Nevertheless, all models show a good and reliable generalization ability. Alltogether, the results presented in Fig. 1 and 2 show that the Intensity-CRNN provides the best localization accuracy on both simulated and real SRIRs. This underlines the statement of Perotin et al. [4] that their intensity features are very well suited for deep learning based DOA estimation

Nevertheless, it can be seen in Fig. 1a and 2a, that the HOA-n-CRNNs perform better with increasing order n on the simulated data. Both the median and the IQR of the angular distance become smaller with each additional order. In particular, the additional orders of the SH seem to allow a better fine localization. Thus, only about 70 % of the predictions of the HOA-1-CRNN lie within the error tolerance of $4°$,

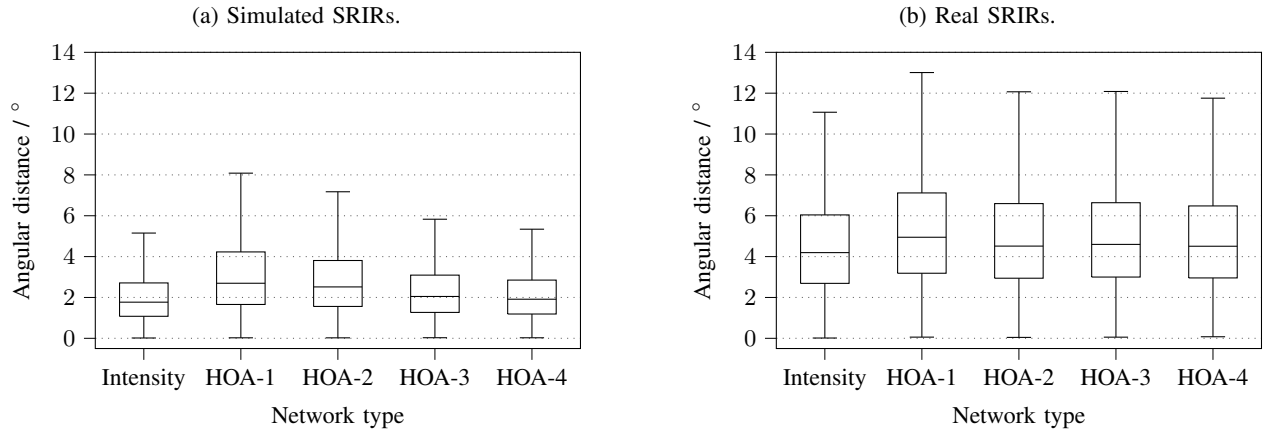Fig. 1. Box plot of angular distances (°) for the five different networks using simulated (a) and real (b) SRIRs. The boxes are drawn from the first to the third quartile. The horizontal line shows the median. The whiskers go from the lowest data still within 1.5 interquartile range (IQR) of the lower quartile to the highest data within 1.5 IQR of the upper quartile.
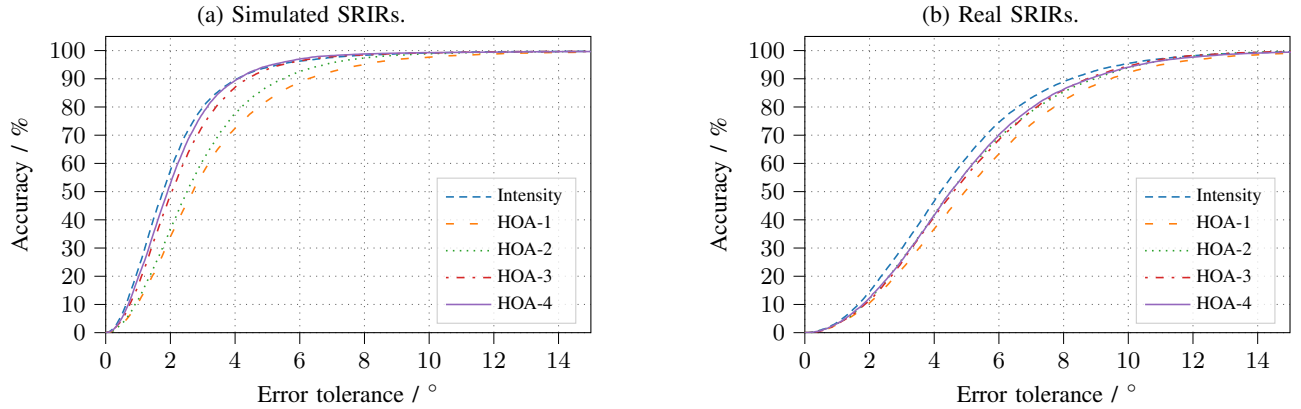

Fig. 2. Accuracies of the different networks as a function of the error tolerance for the simulated (a) and real (b) SRIRs.

whereas this is the case for about 90 % of the predictions of the HOA-4-CRNN. The rough direction, however, seems already to be well predictable with the HOA-1-CRNN. All considered networks have an accuracy of about 99 % with an error tolerance of $15°$.

However, the results belonging to the real SRIRs in Fig. 1b and 2b show that an improvement of the DOA estimates is only obtained when the order is increased from 1 to 2. The HOA-CRNNs of orders 2 to 4 achieve almost identical results.

In Fig. 3 the localization accuracy is evaluated as a function of the SNR of the respective speech signal. As expected, the localization becomes more accurate for each model with increasing SNR. For both simulated and real SRIRs, a slight trend can be seen that the advantage of the Intensity-CRNN over the HOA-n-CRNN of orders 3 and especially 4 mainly exists at relatively high SNR. In the case of poor SNR between 0 and 4 dB, the HOA-4-CRNN performs even sligthly better than the Intensity-CRNN. Otherwise, the respective order of localization accuracy among the models remains the same.

VI. CONCLUSION AND OUTLOOK

In this paper we investigated the influence of the order of HOA signals on the accuracy of single-speaker DOA estimation of noisy speech with CRNNs. We have shown that there is potential in using the additional spatial information of HOA signals for a CRNN-based DOA estimation. However, when evaluated on real data, it has been shown that the advantage of this additional information may possibly be reduced in practice due to effects such as a non-perfect simulation, a limited generalization capability of the models, or additional measurement noise. Rather, it became very clear that it is highly useful and advisable to extract the information present in the signals in a preprocessing step to make it more accessible for the network. Only in low-SNR conditions a slight improvement of the DOA estimation could be achieved by using fourth order Ambisonics signals comparing to the Intensity-CRNN.

Since the HOA models seem to perform comparatively well in acoustically challenging scenarios, we will also investigate the effect of the Ambisonics order on localization accuracy in multi-speaker DOA estimation scenarios in the future. Also
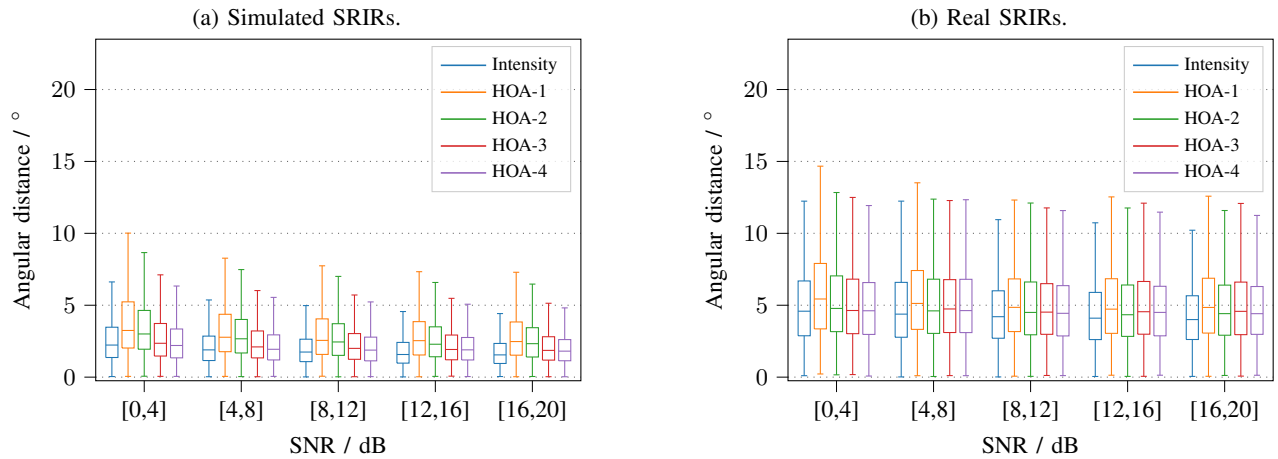
(a) Simulated SRIRs.   (b) Real SRIRs.

Fig. 3. Box plots with angular distances of the different networks for different SNR regions and simulated (a) and real (b) SRIRs.

based on the physical motivation and interpretation of the sound Intenisty features, it can be suspected that the higher-order models are superior to the Intensity-CRNN there.

Furthermore, we want to strengthen our results by additional evaluations of our models on more data generated from real SRIRs and also on real recordings. In addition, we want to use our presented dataset to estimate additional parameters such as room volume, reverberation time and frequency-dependent absorption and scattering coefficients using HOA signals.

REFERENCES

[1] S. Chakrabarty and E. A. P. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[2] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 2814–2818, IEEE.

[3] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[4] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[5] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, 2019, pp. 119–123.

[6] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," Accessed on: Feb. 18, 2021. [Online], Available: http://arxiv.org/pdf/2009.02792v1, 2020.

[7] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, 2019, pp. 30–34.

[8] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 241–245.

[9] L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, "Regression versus classification for neural network based audio source localization," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 343–347.

[10] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[11] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015.

[12] V. Pulkki, S. Delikaris-Manias, and Politis. A., "Higher-order directional audio coding," in *Parametric Time-Frequency Domain Spatial Audio*, pp. 141–159. 2018.

[13] M. Green and D. Murphy, "Sound source localisation in ambisonic audio using peak clustering," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. 2019, pp. 79–83, New York University.

[14] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix - suggested ambisonics format," *Ambisonics Symposium 2011*, 2011.

[15] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[16] F. Zotter and M. Frank, *Ambisonics*, vol. 19, Springer International Publishing, Cham, 2019.

[17] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik, "Room acoustics simulation for multichannel microphone arrays," *International Symposium on Room Acoustics (ISRA) 2010*, 2010.

[18] D. Ackermann et al., "A ground truth on room acoustical analysis and perception (grap)," Technische Universität Berlin, 2018.

[19] J. S. Garofolo, *TIMIT: Acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, Philadelphia, Pa., 1993.

[20] R. Hupke, M. Nophut, S. Li, R. Schlieper, S. Preihs, and J. Peissig, "The immersive media laboratory: Installation of a novel multichannel audio laboratory for immersive media applications," *Journal of the Audio Engineering Society*, 2018.

[21] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, pp. II–1781–II–1784.

[22] T. Dozat, "Incorporating nesterov momentum into adam," Technical Report, Stanford University, 2015.

[23] A. Martín et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," Accessed on: Feb. 18, 2021. [Online], Available: http://download.tensorflow.org/paper/whitepaper2015.pdf, 2015.

[24] V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds., *Parametric time-frequency domain spatial audio*, Wiley, Hoboken NJ USA, 2018.

[25] F. Jacobsen, "A note on instantaneous and time-averaged active and reactive sound intensity," *Journal of Sound and Vibration*, vol. 1991, no. 147(3), pp. 489–496, 1991.