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

Nils Poschadel
Leibniz University Hannover
Hannover, Germany
poschadel@ikt.uni-hannover.de

Stephan Preihs
Leibniz University Hannover
Hannover, Germany
preihs@ikt.uni-hannover.de

Jürgen Peissig
Leibniz University Hannover
Hannover, Germany
peissig@ikt.uni-hannover.de

***Abstract*—Convolutional recurrent neural networks provide state of the art results in direction of arrival estimation based on first-order Ambisonics signals, especially in the presence of noise and/or interfering sound sources. In this work, we investigate whether increasing the order of Ambisonics up to the fourth order further improves the estimation results in a challenging multi-speaker setting with two or three simultaneously active speakers. Our results show that each additional order of the Ambisonics representation further improves the localization performance for both speech signals based on simulated and real measured spatial room impulse responses. The greatest gains in accuracy can be observed in the particularly demanding scenarios with three speakers and poor signal-to-interference-ratio.**

***Index Terms*—Multi-source direction of arrival estimation, higher-order Ambisonics, convolutional recurrent neural network, spherical harmonics.**

I. Introduction

Direction of arrival (DOA) estimation is a well-known task in audio signal processing and an important component of many applications such as speech separation [1] or speech enhancement [2]. Neural networks have been shown to be superior to classical parametric approaches in DOA estimation, especially in very demanding reverberant, noisy, and low-SNR environments [3]–[6]. In addition, Ambisonics-based audio signal processing is becoming increasingly popular due to the flexibility and generalizability it enables. Therefore, DOA estimation based on first-order Ambisonics (FOA) signals has been the subject of much attention [6]–[9].

Perotin et al. [6], [10], for example, 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. Furthermore, they stated that a classification approach seems to be more robust to localized interference than a regression interpretation [10].

While much attention has been focused on FOA signals for deep learning based DOA estimation, only little research is conducted on the performance of DOA estimators based on higher-order Ambisonics (HOA) signals. Pointer experiments with subjects demonstrated a positive influence of the order of the Ambisonics signal on the perceived localization accuracy in a loudspeaker reproduction of a sound field [11]. In addition, we investigated in a previous study [12] whether increasing the order of the Ambisonics signal has a positive effect on the localization results on single-speaker signals, especially compared to the first-order intensity features proposed in [6]. Our results showed that the CRNN trained on the first-order intensity features outperformed every model trained on higher-order magnitude and phase information. Nevertheless, the potential of the HOA signals has been seen as the estimation accuracy increased with each additional order for the simulated spatial room impulse responses (SRIRs) and at least from order 1 to 2 for the real SRIRs. Investigations on spherical harmonic (SH) beamforming with unsupervised peak clustering [13] showed a similar improvement in estimation accuracy with increasing Ambisonics order, also observing the greatest improvement when increasing from order 1 to 2.

In real world applications, however, scenarios with several simultaneously active speakers are particularly relevant. This work therefore is the first to apply the idea of CRNN-based DOA estimation to multi-speaker HOA signals of two or three speakers and to investigate whether or how much the additional spatial information contained in HOA signals can improve the estimation accuracy. We especially compare our results to the well performing single-speaker approach based on features derived from the sound intensity vector. Since the HOA models in [12] performed comparatively well in the acoustically most challenging scenarios of low SNR and because of the physical motivation and interpretation of the sound intensity features, it can be suspected that the higher-order models are superior to the Intensity-CRNN in the multi-speaker scenarios investigated here.

For training our neural networks, we use the same dataset of simulated SRIRs as in [12] and adapt it for the multi-speaker setting. 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 [6], [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 $f(\theta, \phi) = f(\Omega) \in L^2\left(S^2\right)$ on the unit 2-sphere $S^2 := \left\{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 = 1\right\}$, 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

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

where $\int_{\Omega \in S^2} \mathrm{d}\Omega = \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi$. 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 $\mathrm{ACN} = n^2 + n + m$ and normalised by the SN3D normalisation $N_n^{|m|} = \sqrt{\frac{2-\delta_m}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}}$.

III. Data

A. Simulated SRIRs

The training, validation and testing data was generated from a set of SRIRs simulated with the MCRoomSim toolbox [17] as Ambisonics signals up to fourth order as described in [12]. 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 SRIRs were convolved with a randomly chosen sentence from the TIMIT database [18]. For the 3-source case, another SRIR was selected belonging to the same room but to a different source and having an angular distance of at least 15° from the first source. This SRIR was then convolved with a different speech sample from the TIMIT database. The second HOA speech signal was added to the first HOA signal at a random signal-to-interference-ratio (SIR) between 0 and 10 dB.

For the 3-source case, this procedure was repeated, maintaining an angular distance of at least 15° from each of the two sources. The third HOA speech signal was than added to the other two, again at a random SIR between 0 and 10 dB relative to the first source.

The signals were cut to the minimum length of the respective individual speech signals, such that the respective target number of speakers is active the entire duration of the signal.

Furthermore, we added ambient noise to the speech signals similar to the procedure in [6]. 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 of 20 dB. Finally, these sentences were cut to one-second-sequences which led to 134 391, 8288 and 8461 sequences for the training, validation and testing set in the 2-source-case, respectively and 121 397, 7465 and 7715 sequences in the 3-source-case.

B. Real SRIRs

For the analysis of DOA estimation performance in a more realistic scenario and to assess the generalization quality of our models trained on simulated data, we measured real SRIRs in the Immersive Media Lab (IML) [19] at the Institute of Communications Technology. We measured the SRIRs from each of our 36 KH120 loudspeakers to an em32 Eigenmike® [20] microphone at nine different positions, each with 2 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 multi-speaker signals which resulted in 11 004 and 9943 sequences for the testing set based on real SRIRs for two and three sources, respectively.

IV. Networks and Metrics

Our CRNNs follow the same basic structure as the ones in [12]. A detailed overview of the network's architecture is given in Table I. We formulated the task as a classification problem, i.e ., as the task of estimating whether or not each point on a predefined grid corresponds to the direction of an active source or not, assuming that the number of active sound

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

sources is known. Here, we use the following quasi-uniform grid on the unit 2-sphere [6]:

$$\begin{aligned} \theta_i &= -90 + \frac{i}{I} \cdot 180 \qquad\qquad \text{, with } i \in \{0, \ldots, I\}, \\ \phi_j^i &= -180 + \frac{j}{J^i + 1} \cdot 360 \quad \text{, with } j \in \{0, \ldots, J^i\}, \end{aligned} \tag{3}$$

with $I = \lfloor \frac{180}{\alpha} \rfloor$, $J^i = \lfloor \frac{360}{\alpha} \cos(\theta_i) \rfloor$, and a grid resolution parameter α which results in a grid of $n_{\text{grid}} = \sum_{i=0}^{I} \left(J^i + 1\right)$ points. In this paper we set $\alpha = 10$ which leads to a grid of $n_{\text{grid}} = 425$ points. According to this classification setting, the target of our CRNNs is a multi-hot-encoded vector of size n_{grid}, where each index corresponds to a DOA according (3). For each active speaker in the scene, the entry in the target vector corresponding to the direction closest to the respective DOA is set to one.

Since we use a Time-Distributed Dense layer as output layer, we get a n_{grid}-dimensional output vector for each time frame. However, we assume that the sources are static during each one-second-signal. Therefore, we compute the average over all time frames for each grid point to get a single output vector for the whole sequence. From this vector, we choose the n_{sources} largest values that correspond to the respective DOAs. In contrast to the results in [6], smoothing the output vector over neighboring directions did not improve the estimation accuracy in our case. For comparing the predicted DOA $(\hat{\theta}, \hat{\phi})$ with the reference (θ, ϕ) used to synthesize the dataset, we compute the *angular distance* $\delta\big[(\hat{\theta}, \hat{\phi}), (\theta, \phi)\big]$ defined by

$$\begin{aligned} \delta\big[(\hat{\theta}, \hat{\phi}), (\theta, \phi)\big] = \arccos\big[&\sin(\hat{\theta}) \sin(\theta) \\ &+ \cos(\hat{\theta}) \cos(\theta) \cos(\hat{\phi} - \phi)\big]. \end{aligned}$$

Since there is no direct mapping of target and estimation sources due to the classification approach, the estimation needs to be associated with the corresponding target. We chose the permutation that minimized the mean angular distance.

TABLE I. Architecture of the CRNNs for DOA estimation.

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, $2 \cdot n_{\text{filter}}$)
BiLSTM		(50, $2 \cdot n_{\text{filter}}$)
BiLSTM		(50, $2 \cdot n_{\text{filter}}$)
Time-Dist. Dense	elu	(50, $2 \cdot n_{\text{filter}}$)
Dropout	0.2	(50, $2 \cdot n_{\text{filter}}$)
Time-Dist. Dense	sigmoid	(50, 425)

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. Due to the grid resolution of the classification setup, an arbitrary point on a 2-sphere can have an angular distance of up to around 7° to the nearest point on the grid. Therefore, we will also evaluate the so-called *classification accuracy*, which is the proportion of samples assigned to the correct or closest discrete grid point.

Our networks are trained on pure magnitude and phase spectrograms as done before by [5], [12]. The network trained on Ambisonics signals of order $n = 1, \ldots, 4$ is referred to as HOA-n-CRNN. We compare the HOA-n-CRNNs to an approach proposed by Perotin et al. [6]. They used spectrograms of features derived from the FOA sound intensity vector as input (Intensity-CRNN). For more information on this particular approach, please refer to [6]. The input shape of all the different networks is $(50, 512, dim_{\text{in}})$, where 50 is the number of frames, 512 the number of frequency bins, and dim_{in} the number of input channels with $dim_{\text{in}} = 2(n+1)^2$ for the HOA-n-CRNNs and $dim_{\text{in}} = 6$ for the Intensity-CRNN. The Short-time Fourier transform 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}), 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 128 for all networks in the 2-source case and the Intensity-, HOA-1-, and HOA-2-network in the 3-source case as well as 256 for the HOA-3- and HOA-4-CRNN in the 3-source case. For training the neural network, we used the binary-crossentropy loss function together with the Nadam optimizer [21] within the TensorFlow platform [22].

V. Results

As can be seen in Table II, the accuracy of the estimation improves with each additional Ambisonics order, for both two and three sources as well as simulated and real data. The representation using the FOA intensity features is thereby ranked between HOA-1 and HOA-2 in all cases and for all metrics. This underlines the results in [6] and [12] that the features derived from the intensity vector are a very suitable input feature for FOA-CRNNs in this task. However, in contrast to the easier single-speaker scenario in [12], this does not seem to compensate for the additional higher-order spatial information. For example, when comparing the Intensity-CRNN with the HOA-4-CRNN, the mean angular distance decreases by about 21 % for both simulated and real data and two sources. Similarly, for three sources, there is an improvement in mean angular distance by 39 % for simulated data and 31 % for real data. However, this higher relative improvement in localization accuracy for three speakers occurs, as expected, at an overall worse localization level than for two sources. As shown in Fig. 1, the sources in the more demanding scenario with three speakers are on average localized worse by each model than

TABLE II. Accuracies (%) given error tolerances as well as mean/median angular distances (°) and classification accuracies (%) for the different networks with two/three sources and simulated/real SRIRs. The best result for each case is shown in bold.

(a) Two sources with simulated SRIRs.

Network	Accuracy				Ang. distance		Classif.
	$< 3°$	$< 5°$	$< 10°$	$< 15°$	mean	median	
Intensity	19.87	46.77	81.18	87.21	13.55	5.28	57.08
HOA-1	16.92	40.57	75.40	83.45	16.25	5.91	50.51
HOA-2	20.19	48.60	83.18	88.80	12.65	5.11	58.95
HOA-3	22.97	55.24	87.01	91.02	11.30	4.67	66.25
HOA-4	**23.92**	**57.56**	**88.32**	**91.46**	**10.67**	**4.55**	**68.79**

(b) Two sources with real SRIRs.

Network	Accuracy				Ang. distance		Classif.
	$< 3°$	$< 5°$	$< 10°$	$< 15°$	mean	median	
Intensity	11.62	30.57	72.60	85.65	14.98	6.92	39.11
HOA-1	10.11	27.05	66.79	82.53	17.07	7.53	35.09
HOA-2	11.42	29.75	72.62	87.60	13.36	6.99	37.90
HOA-3	12.54	32.54	76.83	90.01	12.28	6.67	40.81
HOA-4	**13.00**	**33.15**	**78.34**	**90.70**	**11.78**	**6.52**	**41.71**

(c) Three sources with simulated SRIRs.

Network	Accuracy				Ang. distance		Classif.
	$< 3°$	$< 5°$	$< 10°$	$< 15°$	mean	median	
Intensity	15.02	35.51	65.56	73.96	23.63	6.76	46.89
HOA-1	13.44	31.58	60.63	70.94	25.57	7.59	42.47
HOA-2	16.72	39.69	69.86	77.56	20.86	6.07	51.29
HOA-3	20.24	48.89	77.98	82.85	16.73	5.07	61.61
HOA-4	**22.48**	**54.24**	**82.70**	**86.23**	**14.46**	**4.73**	**66.98**

(d) Three sources with real SRIRs.

Network	Accuracy				Ang. distance		Classif.
	$< 3°$	$< 5°$	$< 10°$	$< 15°$	mean	median	
Intensity	8.52	21.42	55.12	70.50	25.63	8.99	29.91
HOA-1	7.49	19.14	50.47	67.48	26.93	9.90	27.04
HOA-2	9.32	23.48	59.01	73.90	23.41	8.40	32.49
HOA-3	11.43	27.68	66.67	79.79	19.52	7.46	37.09
HOA-4	**12.08**	**30.39**	**70.31**	**82.68**	**17.68**	**7.05**	**39.87**

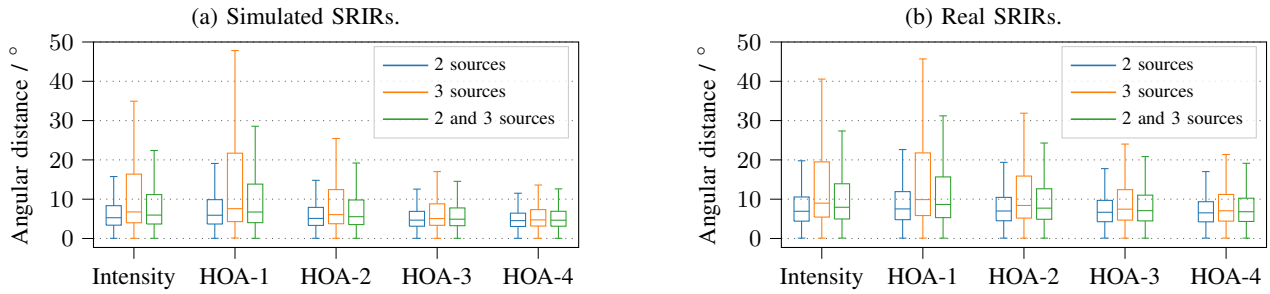

Fig. 1. Box plot of angular distances (°) for the five different networks using simulated (a) and real (b) SRIRs. The boxes are drawn from the first to the third quartile. The horizontal line shows the median. The whiskers go from the lowest data still within 1.5 IQR of the lower quartile to the highest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

those in the 2-source case for both simulated and real SRIRs. This can be partly explained by the fact that the SIR level for three speakers is lower overall due to the way it is synthesized. In the 2-speaker scenario, the SIR ranges from -10 to 10 dB, while in the 3-speaker scenario it ranges from -13 to 7 dB.

The results in Fig. 2 allow a finer analysis of the localization results as a function of the SIR. It can be seen clearly that the median accuracy of the estimate considerably decreases and that the variance increases with decreasing SIR in all scenarios (simulated/real SRIR and two/three sources). In addition, there is significant improvement using higher orders for all SIR conditions, with particularly large improvements for poor SIR. For example, the median angular distance decreases from about 11.1° for the Intensity-CRNN to 8.0° for the HOA-4-CRNN (an improvement of about 28 %) for a low SIR between -13 and -9 dB using real data with three speakers. At the same time, the interquartile range (IQR) decreases by about 70 % from 37.6° (Intensity-CRNN) to 11.3° (HOA-4) in the same SIR range. In the high SIR scenarios (SIR between 3 and 7 dB), the Intensity-CRNN already achieves rather a small median angular distance and IQR. Nevertheless, the usage of HOA further improves estimation accuracy from 7.4° to 6.4° (improvement of about 14 %) and the IQR from 7.4° to 5.3° (28 %) for the real data and three sources.

Overall, our evaluations show a very reliable generalization of our network trained on simulated data to speech signals synthesized with real SRIRs (some of the results in the low-SIR-scenarios are even more accurate for the real SRIRs). Of course, this conclusion is based on only a small sample and needs to be strengthened with additional real rooms and/or speech recordings.

VI. Conclusion and Outlook

In this paper we investigated the influence of the order of HOA signals on the accuracy of multi-speaker DOA estimation of noisy speech with CRNNs. It has been shown that the average angular distance of the localization can be significantly improved by using HOA signals compared to FOA ones, for both simulated and real SRIRs. Furthermore, the largest improvements were achieved in the most demanding scenarios with three speakers and poor SIR. The Intensity-CRNN, which was very impressive in the single-speaker case, is always ranked between the HOA-1-CRNN and the HOA-2-CRNN.

In the future we want to further evaluate our model in more detail on additional data generated from real SRIRs and also on real recordings. Furthermore, we want to use our dataset to estimate other parameters such as room volume, reverberation time and frequency-dependent absorption and scattering coefficients using HOA signals.

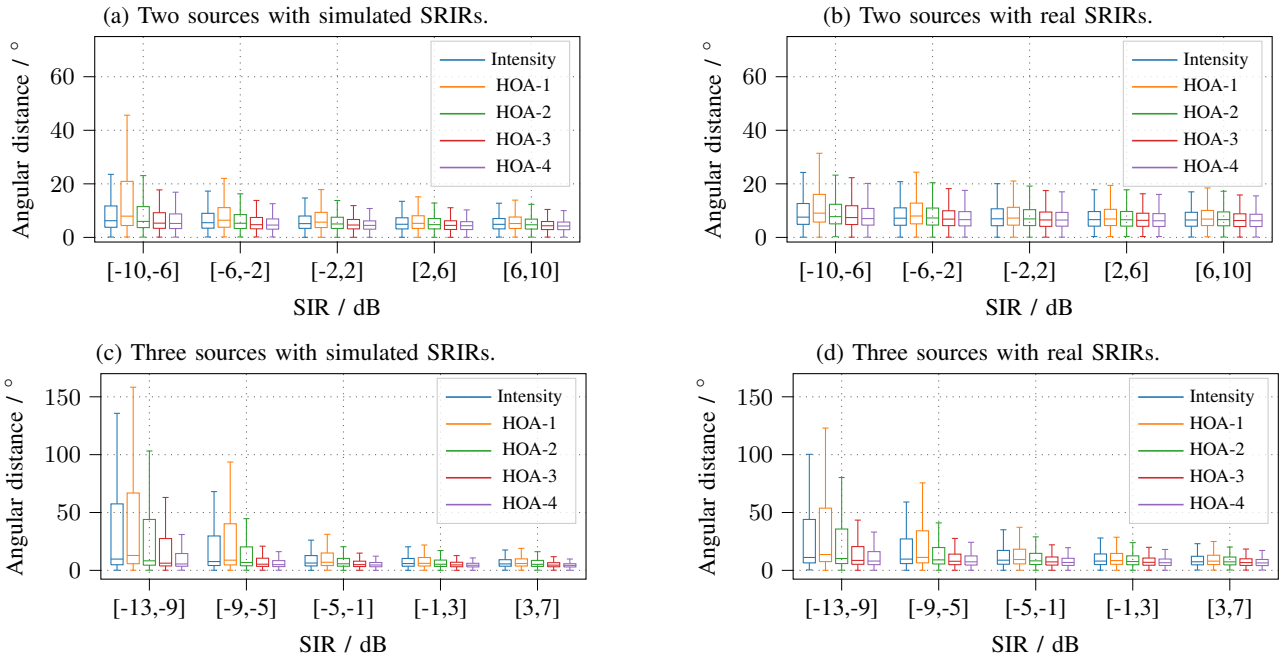

Fig. 2. Box plot with angular distances of the different networks for different SIR regions, two/three sources and simulated/real SRIRs. Note the different scaling of the axes for two and three sources.

References

[1] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[2] T. Lotter, H. W. Loellmann, and P. Vary, "Robust direction of arrival estimation for speech enhancement in noisy reverberant rooms," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2002.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] N. Poschadel, R. Hupke, S. Preihs, and J. Peissig, "Direction of arrival estimation of noisy speech using convolutional recurrent neural networks with higher-order ambisonics signals," Accessed on: Feb. 22, 2021. [Online], Available: http://arxiv.org/pdf/2102.09853, 2021.

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

[19] 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.

[20] 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.

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

[22] Martín A. 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.