Accepted by IEEE for publication at ICASSP 2021.

SYNTHETIC DATA FOR DNN-BASED DOA ESTIMATION OF INDOOR SPEECH

Femke B. Gelderblom[⋆]*, Yi Liu*[†]*, Johannes Kvam*[†]*, Tor Andre Myrvoll*[⋆]

[⋆]NTNU & SINTEF, Norway, [†]SINTEF, Norway

ABSTRACT

This paper investigates the use of different room impulse response (RIR) simulation methods for synthesizing training data for deep neural network-based direction of arrival (DOA) estimation of speech in reverberant rooms.

Different sets of *synthetic* RIRs are obtained using the image source method (ISM) and more advanced methods including diffuse reflections and/or source directivity. Multi-layer perceptron (MLP) deep neural network (DNN) models are trained on generalized cross correlation (GCC) features extracted for each set. Finally, models are tested on features obtained from *measured* RIRs.

This study shows the importance of training with RIRs from directive sources, as resultant DOA models achieved up to 51% error reduction compared to the steered response power with phase transform (SRP-PHAT) baseline (significant with $p << .01$), while models trained with RIRs from omnidirectional sources did worse than the baseline. The performance difference was specifically present when estimating the azimuth of speakers not facing the array directly.

***Index Terms*—** synthetic data, speech source localization, direction of arrival estimation, room impulse response, deep neural network, generalized cross correlation features

1. INTRODUCTION

DNN-based methods are nowadays successfully applied to many different tasks in the field of speech processing. For training such methods, there are large datasets available, containing annotated single microphone recordings of clean speech. These datasets can be converted into multichannel datasets for microphone array processing by convolving the clean speech with recorded room impulse responses (RIRs) specific for each array element and acoustic setting.

However, learning-based methods can only be expected to be widely applicable in realistic settings if they are trained for exactly that. This issue is two-fold: first of all, to ensure results apply to a wide range of rooms of varying acoustical characteristics, the training set needs to contain a similar variety [1], and secondly, the training data must approach reality as much as possible.

While recorded RIRs are a direct reflection of reality, it quickly becomes too difficult or expensive to record a sufficient number of RIRs from many different environments. Instead models can be trained on single channel recordings augmented with synthetic RIR data.

Here it is common to rely on the relatively simple image source method (ISM) room impulse response (RIR) simulation technique [2], where scattering effects that cause the late reflections of the diffuse field are ignored for simplicity. Additionally, all sources are assumed to behave in an omnidirectional manner, while a speaking person is a directive source.

This paper therefore investigates how more advanced RIR simulation methods can affect final model performance on real data. We have chosen to do this through the DOA estimation task, because of its central role in multi-channel speech processing. The ability to discriminate on where speech originates from is crucial for applications like multi-channel speech enhancement, speaker identification and automatic speech recognition.

Classic approaches to DOA estimation include multiple signal classification (MUSIC) [3], the least squares (LS) method [4], multi-channel cross correlation (MCCC) [5], and the steered response power with phase transform (SRP-PHAT) [6]. A main challenge is the multipath propagation effect where microphone sensors not only receive the direct-path signal, but also attenuated signals due to both the specular and diffuse reflections.

Inspired by the success of DNNs in many fields, several such approaches have been proposed for sound/speech source localisation (SSL) [7, 8, 9, 10, 11, 12, 13, 14].

Research based on training data generated from measured RIRs is automatically constricted to a severely limited number of rooms [7, 8]. Others rely on the simulation of just one or two acoustical environments [9, 10]. Xiao *et al.* and Perotin *et al.* simulated more varied data for DOA estimation of speech [11, 12, 13], but they, as is common practise, relied on ISM with omnidirectional sources for RIR simulation.

Only recently have researchers attempted to improve deep learning model performance in speech processing tasks, by improving the quality of the RIRs used for synthesizing data. Tang *et al.* found significant performance increases on an automatic speech recognition and keyword spotting task in [15] by using an acoustic simulation method that includes diffuse reflections. Using the same method, Tang *et al.* also observed improved performance at a DOA estimation task [14].

In this study we further investigate the effect of RIR sim-

ulation methods on final DOA model performance. Our study is unique in that we are, as far as we know, the first to investigate the effect of simulating speakers as directive sources. Like Tang *et al.* we also study the effect of diffuse reflections, but we rely on the GCC speech features and the MLP architecture proposed in [11], instead of ambisonic features and CRNN architecture. We focus only on reverberance (no noise added), and use our own dataset, which includes two test sets that allow us to differentiate between results for speakers looking directly at the array, and the more challenging situation where speakers face the array at a 90° angle.

2. DATA ACQUISITION

2.1. Synthetic RIRs for training

We simulated RIRs with four different simulation methods using the MATLAB package MCRoomSim [16]:

- **ISM-omni**: the basic RIR generated by ISM where sources are modelled as omnidirectional. No scattering and no diffuse field.
- **ISM-dir**: Like ISM-omni, but now sources are modelled as directive speakers, with either an average male or female directivity. No scattering and no diffuse field.
- **WithDiffuse-omni**: An advanced RIR with not just specular reflections, but also a diffuse field due to scattering, where sources are modelled as omnidirectional.
- **WithDiffuse-dir**: Like WithDiffuse-omni, but sources are again modelled as directive speakers.

For each method, 18 000 training and 6000 validation RIRs were simulated from three random source positions in 6000 and 2000 virtual rooms. Each room was randomly configured with parameters drawn from the uniform distributions specified in Table 1, ensuring evenly distributed target DOAs in all directions. The average absorption of a room was determined from the drawn reverberation time with Eyring's [17] algorithm with air absorption taken into account.

Table 1. Details of random virtual room configuration

Item	Parameter	Min.	Max.
Room size	width	3 m	8 m
	length	3 m	10 m
	height	2.5 m	6 m
	RT60	0.2 s	1 s
	scattering coefficient	0	1
Array position	from walls	1 m	-
	from floor	0.6 m	0.9 m
Speaker position	from walls	0.5 m	-
	from floor	1 m	1.8 m
	from array	0.5 m	-
	yaw (directive speakers only)	-180°	180°

2.2. Measured RIRs for testing

To create realistic test data, RIRs were measured manually with a 9-channel circular array (planar) with 4 cm radius, positioned on a table approximately in the middle of a typical rectangular meeting room with dimensions 4.5 x 3.8 x 2.6 m, and $RT60_{1kHz}$ of 0.3. An NTi TalkBox was used to produce the sinusoidal sweeps required for RIR measurements. This loudspeaker has human head-size like dimensions and is specifically designed for human speech measurements.

Of the measured RIRs, 47 where obtained with the speaker facing towards the array (the 'Easy' set), and 107 with the speaker rotated at 90° (the 'Challenging' set). The true DOAs were measured with an uncertainty of $\pm$ 1° at random angles uniformly distributed around the array, at a distance varying between 1 and 2 m (above critical distance).

2.3. Obtaining Speech Features

Our preprocessing steps are inspired by [11], but the specifics differ. We used 'NB Tale', a Norwegian speech database. This database contains circa 19 hours of training data and circa 5 hours of validation data from a total of 380 speakers.

First the speech files were passed through the open source voice activity detector from WebRTC with a hop length of 30 ms, zero minimum silence length and strength 3. They were then convolved with (simulated or measured) RIRs to create a reverberant multichannel speech sample, which was resampled from 48 kHz to 16 kHz. We then selected a random 1 s long segment.

Lastly, GCC vectors with PHAT weighting were obtained for each pair of microphone channels. For our array, the maximum distance between a pair of microphones is 8 cm, which represents a maximum delay of 4 (0.08 m / 340 m/s $\times$ 16 000 Hz) time samples of each GCC vector. Hence, the GCC vector was truncated to the 9 centre time samples for each microphone pair. From the 9 channels, we have 36 possible microphone pairs, giving us 36 GCC vectors. Each of the vectors was scaled so that its max value became 1, and then stacked to obtain a single model input sample.

Due to the random selection of the speech segment, diffuse reflections of earlier speech affect the model input sample, even if vector truncation removes later reflections. This can be seen in Figure 1, which shows examples of the synthetic input training samples for each simulation method, given the same room size, source and array location. Less aggressive truncation did not improve final model performance.

Using the above procedure, we created *synthetic* training and validation sets for each of the RIR simulation methods, with 18 000 training and 6000 validation samples per set. The same procedure was also applied using the two types of *recorded* RIRs to create two *measured* test sets called 'Easy' (speaker facing directly towards the array) and 'Challenging' (speaker at a 90° angle away from the array). The final test sets had 517 'Easy' and 1177 'Challenging' input samples.

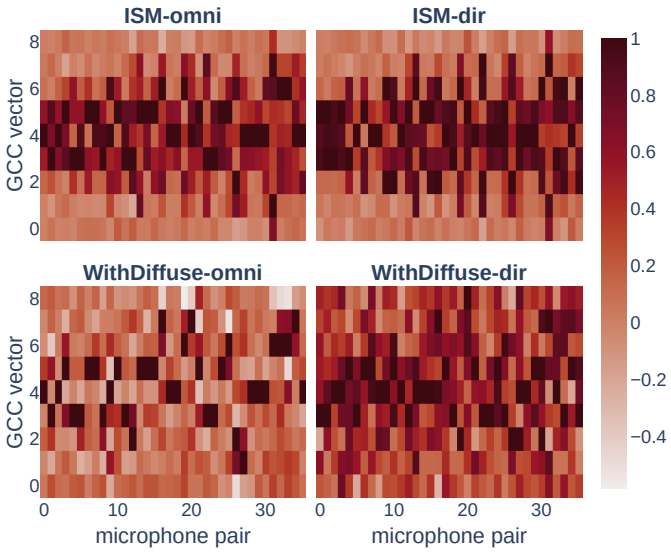

Fig. 1. Examples of the GCC input feature for each method

3. DOA ESTIMATION MODEL

The DOA estimation task is most intuitively formulated as a regression task where the continuous azimuth variable is directly predicted from the input features. However, others have noted advantages from converting the task into classification, where possible azimuths are separated into discrete bins [11, 14]. In this paper, we include both.

For the regression formulation, we investigated two loss functions, which we call the angular mean square error:

$$\mathrm{MSE}_{\measuredangle} = \frac{1}{N}\sum_{n=1}^{N}\Big(\operatorname{atan2}\left(\sin\left(\hat{y}-y\right),\cos\left(\hat{y}-y\right)\right)\Big)^2, \tag{1}$$

and the angular mean absolute error:

$$\mathrm{MAE}_{\measuredangle} = \frac{1}{N}\sum_{n=1}^{N}\Big|\operatorname{atan2}\left(\sin\left(\hat{y}-y\right),\cos\left(\hat{y}-y\right)\right)\Big|, \tag{2}$$

where $\hat{y}$ and y are the true and estimated DOA respectively, and the $\operatorname{atan2}$ operator computes the arctangent of the element-wise division of its first and second argument, respecting signs of the arguments.

These are based on the general mean squared error (MSE) and mean absolute error (MAE) loss functions, but ensure the calculation is always based on the minimal error between two angles, be it clockwise or anticlockwise. Output layers of both regression formulations were given linear activation.

For the classification formulation, we used the standard categorical crossentropy loss with either 72 (5° per bin) or 360 (1° per bin) classes. Both classification models were given an output layer with softmax activation.

As model we chose the MLP neural network. First a wide hyperparameter search was conducted for all datasets using the tree-structured parzen estimator (TPE) approach [18], to determine a single model topology that worked well for all datasets. This search included varying the number of hidden layers, number of nodes per layer, type of activation, rate of dropout, $\ell 1$ or $\ell 2$ regularization, batch normalization and learning rate for the Adam optimizer.

From this, a general model with 3 hidden layers, each with 3072 hidden nodes and relu activation, was chosen for all datasets and problem formulations. No batch normalisation was applied. A new optimisation process was then started for each combination of the 4 datasets and 4 loss functions. Now only the learning rate and level of dropout was varied to find the best model for each set, to ensure that results would be directly comparable. Classification models converged best with high levels of dropout (circa 0.8), while regression models did best without dropout.

Table 2 shows the MAE results for all model types and all simulation methods, obtained for a validation test set specific for each simulation method. These errors do not reflect real-life performance, but performance on synthetic validation set that was created in the same way as the training set used to train each model. Therefore, the consistently lower MAE for methods with omnidirectional sources merely shows that these tasks are easier to learn, but it is not an indication of how the resulting MLPs will deal with real data.

Table 2. MAE for each method's *synthetic* validation set

	Regression		**Classification**	
	$\mathrm{MSE}_{\measuredangle}$	$\mathrm{MAE}_{\measuredangle}$	1° bins	5° bins
ISM-omni	2.4°	1.8°	1.6°	2.3°
ISM-dir	5.5°	5.0°	4.6°	4.7°
WithDiffuse-omni	2.0°	1.4°	1.1°	2.0°
WithDiffuse-dir	6.3°	4.3°	4.0°	4.4°

4. RESULTS

All final models were tested with the exact same two *measured* test sets ('Easy' and 'Challenging'), and performance was evaluated with MAE for all models (independent of the training loss function used!), to allow for direct comparison. For Table 3, test samples are based on RIRs where the speaker was facing directly towards the array. Table 4 shows the results for RIRs where the speaker faced past the array at a 90° angle. Testing with MSE or accuracy within 5° or 10° instead of MAE resulted in the same trends, and are therefore not included in this paper.

In our application, the variance of the error from the true direction indicates system performance (assuming zero mean error). We therefore apply the Brown-Forsythe statistical test [19], which tests the variance of the distributions without a strong assumption of normality. We report the test's probability results p, for relevant pairs of systems, in Section 5.

Table 3. MAE for the 'Easy' test set, where speakers face directly towards the array

	Regression		Classification	
	$MSE_{\measuredangle}$	$MAE_{\measuredangle}$	1° bins	5° bins
SRP-Phat	1.5°			
ISM-omni	2.2°	2.1°	1.4°	1.3°
ISM-dir	3.0°	2.1°	1.5°	1.5°
WithDiffuse-omni	2.8°	1.1°	1.3°	1.4°
WithDiffuse-dir	3.8°	1.4°	1.1°	**0.9**°

Table 4. MAE for the 'Challenging' test set, where speakers face 90° away from the array

	Regression		Classification	
	$MSE_{\measuredangle}$	$MAE_{\measuredangle}$	1° bins	5° bins
SRP-Phat	16.5°			
ISM-omni	18.2°	18.2°	19.1°	18.8°
ISM-dir	12.7°	11.5°	8.9°	**8.1**°
WithDiffuse-omni	19.7°	19.6°	18.6°	17.9°
WithDiffuse-dir	13.0°	10.5°	9.9°	10.1°

5. DISCUSSION

From Table 3 we observe that for the relatively easy task of finding the correct azimuth of a speaker facing the array, all models are able to estimate the DOA with high accuracy.

The training data simulation method starts to matter when testing with samples where speakers looked past the array, giving increased confounding reflections. In this case (see Table 4) all directional data based MLPs outperformed their omnidirectional equivalents and the SRP-Phat baseline method significantly ($p << .01$). Simulating with directional sources also increased the difficulty of the task given to the SSL method as evident from the increase in validation error (see Table 2). As such, results show that the MLPs were able to learn relevant information from the directional simulations that turned out to be applicable on measured data.

This is crucial given that we found no studies that simulated directive sources to train learning-based SSL models. Also, given the importance of localisation for many other speech processing tasks like speech recognition and speech enhancement, the conclusion may be valid for many other multichannel speech applications.

We observe that for each DNN topology, either the simulation methods ISM-dir or WithDiff-dir leads to the highest performance, and overall the performance difference between the two was insignificant ($p > .01$). Adding a diffuse field when simulating sources as omnidirectional also did not have a significant effect ($p > .01$).

As such, in contrast with [14], we do not find benefit (nor deterioration) from adding the diffuse field. However, this may simply be because the chosen preprocessing steps to generate speech features may have stopped the models from learning relevant information from the diffuse field. We also have to be careful to draw conclusions based on measurements taken in a single meeting room, as its diffuse field is not representative for all meeting rooms.

Observed trends are independent of the choice of loss function and whether the problem is formulated as a regression or classification task. This provides evidence that the obtained differences are indeed due to the different datasets used for training, and not due to effects of biased hyperparameter tuning.

Like others [11, 14], we note that defining the DOA estimation task as a classification task is advantageous as this formulation resulted in our best performing models. Especially the directive training sets contain samples that are too challenging for the network to learn. The regression network with $MSE_{\measuredangle}$ loss penalises large errors harshest, and as such the learning process focuses most on these outliers. The classification networks are on the other end of the spectrum - penalising all predictions outside the target bin equally, and as such, their training focuses on the more informative samples. Additionally, all classification networks required high levels of dropout, indicating that smaller networks may work equally well for this task formulation.

The focus of this study was on the effect of using more advanced RIR simulation techniques for generating better training data, and not on finding the best DOA estimator.

6. CONCLUSION

We synthesized different training sets to train MLP models for a DOA estimation task from 4 different RIR simulation techniques. The model trained on data from RIR simulation techniques with directive sources, achieved up to a 51% lower mean absolute error on a measurement-based test set than the industry standard SRP-PHAT method, while equivalent models trained on the standard image source method with omnidirectional sources performed worse than this baseline.

Results show that, for improved real-life performance, sources should be modelled as directive speakers, rather than omnidirectional sources, especially for the situation where the speaker is not directly looking at the array. This is an important conclusion given the widespread use of simple ISM RIRs, indicating that the complexity of the RIR simulation technique has been undervalued as a source of performance gain for learning-based SSL. We further speculate that the conclusion may hold true for other applications within multichannel speech processing.

7. ACKNOWLEDGMENTS

We thank the Research Council of Norway and Huddly for their support through project '256753 - Meet Easy'.

8. REFERENCES

[1] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 379–383.

[2] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] Yiteng Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[5] Jacob Benesty, Jingdong Chen, and Yiteng Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 509–519, 2004.

[6] Joseph Hector DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, Providence, Rhode Island, USA, 2000.

[7] Ryu Takeda and Kazunori Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016, pp. 405–409.

[8] David Diaz-Guerra and Jose R. Beltran, "Direction of Arrival Estimation with Microphone Arrays Using SRP-PHAT and Neural Networks," in *IEEE 10th Sensor Array and Multichannel Signal Processing Workshop*, Sheffield, UK, 2018, pp. 617–621.

[9] Zhaoqiong Huang, Ji Xu, and Jielin Pan, "A regression approach to speech source localization exploiting deep neural network," in *IEEE Fourth International Conference on Multimedia Big Data*, Xi'an, China, 2018, pp. 1–6.

[10] Soumitro Chakrabarty and Emanuel A. P. Habets, "Broadband DOA estimation using Convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York City, USA, 2017, pp. 136–140.

[11] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, 2015, pp. 2814–2818.

[12] Laureline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guerin, "CRNN-based Joint Azimuth and Elevation Localization with the Ambisonics Intensity Vector," in *International Workshop on Acoustic Signal Enhancement*, Tokyo, Japan, 2018, pp. 241–245.

[13] Laureline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guerin, "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.

[14] Zhenyu Tang, John D. Kanu, Kevin Hogan, and Dinesh Manocha, "Regression and Classification for Direction-of-Arrival Estimation with Convolutional Recurrent Neural Networks," in *INTERSPEECH*, Graz, Austria, 2019, pp. 654–658.

[15] Zhenyu Tang, Lianwu Chen, Bo Wu, Dong Yu, and Dinesh Manocha, "Improving Reverberant Speech Training Using Diffuse Acoustic Simulation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, 2020, pp. 6969–6973.

[16] Andrew Wabnitz, Nicolas Epain, Craig Jin, and André van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *International Symposium on Room Acoustics*, Melbourne, Australia, 2010.

[17] Carl F. Eyring, "Reverberation time in "Dead" rooms," *The Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 168–168, 1930.

[18] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl, "Algorithms for hyper-parameter optimization," in *24th International Conference on Neural Information Processing Systems*, Granada, Spain, 2011, NIPS 2011, pp. 2546–2554.

[19] Morton B. Brown and Alan B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.