Regression and Classification for Direction-of-Arrival Estimation with Convolutional Recurrent Neural Networks

Zhenyu Tang, John D. Kanu, Kevin Hogan, Dinesh Manocha

University of Maryland, United States

zhy@cs.umd.edu, jdkanu@cs.umd.edu, khogan@cs.umd.edu, dm@cs.umd.edu

Abstract

We present a novel learning-based approach to estimate the direction-of-arrival (DOA) of a sound source using a convolutional recurrent neural network (CRNN) trained via regression on synthetic data and Cartesian labels. We also describe an improved method to generate synthetic data to train the neural network using state-of-the-art sound propagation algorithms that model specular as well as diffuse reflections of sound. We compare our model against three other CRNNs trained using different formulations of the same problem: classification on categorical labels, and regression on spherical coordinate labels. In practice, our model achieves up to 43% decrease in angular error over prior methods. The use of diffuse reflection results in 34% and 41% reduction in angular prediction errors on LOCATA and SOFA datasets, respectively, over prior methods based on image-source methods. Our method results in an additional 3% error reduction over prior schemes that use classification networks, and we use 36% fewer network parameters.

Index Terms: speech recognition, sound propagation, direction of arrival estimation, data augmentation

1. Introduction

Estimating the direction-of-arrival (DOA) of sound sources has been an important problem in terms of analyzing multi-channel recordings [1, 2]. In these applications, the goal is to predict the azimuth and elevation angles of the sound source relative to the microphone, from a sound clip recorded in any multi-channel setting. One of the simpler problems is the estimation of the DOA on the horizontal plane [3]. More complex problems include DOA estimation in three-dimensional space or the identification of both direction and distance of an audio source. Even more challenging problems correspond to performing these goals in noisy and reverberant environments.

To analyze spatial information from sound recordings, at least two microphones with known relative positions must be used. In practice, various spatial recording formats including binaural, 5.1-channel, 7.1-channel, etc. have been applied to spatial audio related systems [4]. The Ambisonics format decomposes a soundfield using a spherical harmonic function basis [5]. Compared with its alternatives, Ambisonics has the advantage of being hardware independent–it does not necessarily encode microphone specifications into the recording.

Recent work [6] has applied the Ambisonics format to DOA estimation and trained a CRNN classifier that yields more accurate predictions than a baseline approach using independent component analysis. While a regression formulation seems more natural for the problem of DOA estimation, some recent work [3] suggests a regression formulation may yield worse performance than that of the classification formulation for multi-layer perceptrons. In this work, we present a novel learning-based approach for estimating DOA of a single sound source from ambisonic

audio, building on an existing deep learning framework [6]. We present a CRNN which predicts DOA as a 3-D Cartesian vector. We introduce a method to generate synthetic data using geometric sound propagation that models specular and diffuse reflections, which results in up to 43% error reduction compared with image-source methods. We conduct a four-way comparison between the Cartesian regression network, two classification networks trained with cross-entropy loss, and a regression network trained using angular loss. Finally, we investigate results on two test datasets: LOCATA [7] and SOFA [8], where our best model reduces prediction error by 43% compared to prior methods.

Section 2 gives an overview of prior work. We propose our method in Section 3. Section 4 presents our results on two benchmarks and we conclude in Section 5. Project website https://gamma.umd.edu/pro/speech/doa

2. Related Work

2.1. Overview

One classic approach to DOA estimation is to first determine the time delay of arrival (TDOA) between microphone array channels, which can be estimated by generalized cross correlation [9] or least squares [10]. The DOA can be computed from known TDOA and the array layout directly. Another approach is to use the signal subspace, as in the MUSIC algorithm [11]. With some restrictions on operating conditions, these techniques are very effective. However, they do not perform well in highly reverberant and noisy environments, or when the placement of signal sources is arbitrary [12]. More recently, researchers have applied modern machine learning techniques to speech DOA estimation with the goal of improving performance in noisy, realistic environments, which can be categorized into *classification* and *regression* networks. For both networks, angular error for a single example is proportional to the angular distance between the predicted and actual DOA.

2.2. Classification Formulations

In the classification formulation, DOA is encoded using a *categorical* representation: an approximately uniform mesh-grid defines the score for each of a finite set of possible categories, where each category corresponds to a unique region of the continuous DOA space. The mesh grid is defined by subdividing the DOA space at a given resolution. The DOA is decoded as the direction associated with the bin with highest score. The categorical formulation uses a discrete encoding, lending itself to a class-based formulation of DOA estimation. Generalized cross correlation (GCC) feature vectors of a microphone array input have been fed to a multilayer perceptron classifier, which predicts a DOA in one angular dimension [3] and show superior performance over the classic least square method [10] in both simulated and real rooms of various sizes. Perotin et al. [6] cal-

culate acoustic intensity vectors using a first-order Ambisonics representation of audio. This representation serves as input to a CRNN, which predicts a DOA in two angular dimensions. Their CRNN yields more accurate predictions than a baseline approach using independent component analysis. CRNNs have also been used in [13] to identify the DOA for overlapping sound sources, in two angular dimensions.

2.3. Regression Formulations

In the regression formulation, two representations of DOA are commonly used, which we refer to as *Cartesian* and *spherical*. With the Cartesian representation, DOA is encoded as a three-dimensional vector in Cartesian (x, y, z) coordinates, pointing toward the source. With the spherical representation, DOA is encoded as a two-dimensional vector of azimuth (θ) and elevation (ϕ) angles. Both formulations encode DOA in continuous space, leading to a regression formulation of DOA estimation. In prior work, regression formulations have not shown superior empirical results for DOA estimation. Higher angular errors for regression than for classification is claimed in [3]. CNN regression has been used in [14] to estimate the Cartesian coordinates of a sound source in 3-D space. CRNN regression has been used in [13] and higher angular error is observed for DOA estimates for regression than for classification. Similar to this result, our experiments show a higher angular error for regression on spherical DOA than for classification. However, we discover a *lower* angular error for regression on Cartesian DOA, than for both classification on categorical and regression on spherical DOA.

3. Proposed Method

3.1. Data Preparation

Our DOA estimation network also relies on a large amount of labeled training data. However, collecting Ambisonic recordings and manually labeling them for training is tedious and time-consuming. Therefore, in speech/audio related training, the common practice is to use image-source methods to generate synthetic impulse responses for augmenting the training data [15]. However, the distribution of synthetic data may not match that of the real data well enough, which can cause large generalization error when applying a synthetically trained DNN to real test data. To overcome this issue of domain mismatch, a more accurate approach for generating training data is needed.

Sound propagation methods compute the reflection and diffraction paths from the sound sources to a listener in the virtual environment. Image-source methods do not model sound scattering or diffuse reflections, which are important phenomenons in acoustic environments. We utilize the state-of-the-art geometric sound propagation method [16, 17, 18] to generate synthetic data that is more accurate than image-source methods. Its benefit has also been observed in other speech tasks [19].

Following the suggested procedure in [6], we generated 42,000 rectangular room configurations with dimensions uniformly and independently sampled between $2.5m \times 2.5m \times 2m$ and $10m \times 10m \times 3m$. Under each room configuration, we randomly populate three paired source-listener locations, both at least $0.5m$ away from walls. The geometric sound propagation method based on path tracing is used on each source-listener pair to generate its spatial room impulse response (SRIR). Then we convolve each SRIR with a randomly selected one-second clean speech sample from the Libri ASR corpus [20] to generate realistic reverberant speech recordings in Ambisonic format. Babble and speech shaped noise [21] are added to the convolved sound
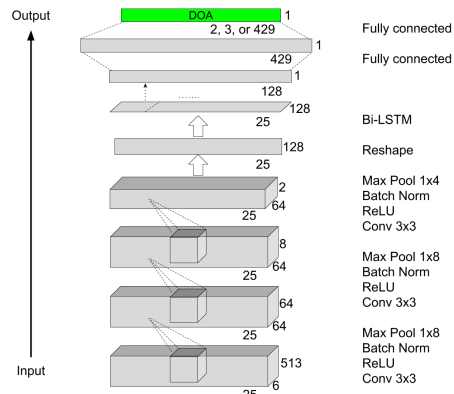

Figure 1: *General network architecture for each regression network and classification network. The dimensionality of the output vector (shown in green) is 2 for spherical formulation, 3 for Cartesian, and 429 for classification. Note that our implementation of the classifier is equivalent to the implementation in [6], but our regression networks differ in the size of the output layer, and use 36% fewer trainable parameters.*

at signal-to-noise ratios (SNRs) following a normal distribution centered at $15dB$ with a standard deviation of $1dB$ as recommended by [22]. A short time fourier transform (STFT) is used to convert speech waveforms to spectrogram, and the features are extracted according to Section 3.2.

3.2. Ambisonic Input Features

In theory, Ambisonics of an infinite number of bases can reproduce the recorded soundfield with no error. As a practical approximation, we use the first four bases/channels necessary for first-order ambisonics (FOA). The FOA channels are denoted by W, X, Y, Z, where W channel contains the zeroth-order coefficients that represents the omnidirectional signal intensity, and X, Y, Z channels contain the first-order coefficients that encode direction modulated information. For a plane wave with azimuth θ and elevation ϕ creating a sound pressure p, the complex FOA components are:

$$\begin{bmatrix} W(t,f) \\ X(t,f) \\ Y(t,f) \\ Z(t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{3}\cos\theta\cos\phi \\ \sqrt{3}\sin\theta\cos\phi \\ \sqrt{3}\sin\phi \end{bmatrix} p(t,f), \quad (1)$$

where t and f are time and frequency bins. We follow the approach in [6] to construct input features from raw FOA audio. The active and reactive intensity vectors are encoded as:

$$\mathbf{I}(t,f) = \begin{bmatrix} W(t,f) * X(t,f) \\ W(t,f) * Y(t,f) \\ W(t,f) * Z(t,f) \end{bmatrix}, \quad (2)$$

$$\mathbf{I}_a(t,f) = \mathcal{R}\{\mathbf{I}(t,f)\}, \mathbf{I}_r(t,f) = \mathcal{I}\{\mathbf{I}(t,f)\}, \quad (3)$$

where $\mathcal{R}\{\cdot\}$ and $\mathcal{I}\{\cdot\}$ extract the real and imaginary components of a complex signal respectively. Both feature vectors are divided by $|W(t,f)|^2 + \frac{1}{3}(|X(t,f)|^2 + |Y(t,f)|^2 + |Z(t,f)|^2)$ to have a uniform range for deep neural network training.

3.3. Output and Loss Formulation

We compare Cartesian (x, y, z) and Spherical (θ, ϕ) output representations in addition to the common Categorical represen-

tation in this work. A stacked CRNN formulation using Categorical outputs has been proposed in [6]. We use this network structure and derive a set of Categorical, Cartesian, and Spherical forms from it, differing only in the size of the output layer, our independent variable. Maintaining a high similarity among the network architecture enables us to conduct well controlled comparisons between output representations for DOA estimation. We visualize our network architecture for the following three output formulations in Fig. 1. Key differences between the formulations are summarized in Tab. 1.

Table 1: *Three types of output representations studied in this paper. d is the number of classes into which a spherical surface is discretized. In our experiment, $d = 429$.*

Output Type	Activation	Dimension	Loss Function
Categorical	Sigmoid	$\mathbb{R}^d$	Cross-entropy
Cartesian	Linear	$\mathbb{R}^3$	MSE
Spherical	Linear	$\mathbb{R}^2$	Haversine

Categorical Outputs   We discretize the continuous DOA space into some number of possible categorical outputs. The angular resolution is chosen to be $10°$, which results in 429 direction classes. Each training DOA label is assigned to the direction class that has the smallest angular difference from itself. We use this labeling to train a classifier that outputs a sigmoid vector of the direction class using cross-entropy loss.

Cartesian Outputs   We define the output of the Cartesian network as a 3-D vector, representing DOA in Cartesian coordinates. Training labels are unit vectors in $\mathbb{R}^3$ pointing toward the source. We use mean-squared error (MSE) as the loss to train our networks. Note that the output of the network is not constrained to lie on the unit sphere. As a result, a hypothetical output that is a large scalar multiple of the DOA label will result in a large loss, despite the perfect alignment of the output vector with the label. In practice, this property does not prevent our formulation from generating an accurate predictor, as shown in Section 4.

Spherical Outputs   In contrast to the Cartesian formulation, the spherical formulation encodes DOA as a 2-D vector representation azimuth (θ) and elevation (ϕ) angles. This representation has only 2 degrees of freedom in 3-D, which means the Cartesian representation has added a redundant dimension to this learning problem. One issue with using this form is that the periodicity of spherical angles makes distance computation between two angles more complicated than in Cartesian coordinates. Conventional mean squared loss is discontinuous, and therefore non-differentiable, over predicted azimuth and elevation, which eliminates the guarantee of convergence of Gradient Descent. Instead, we compute the great-circle distance on a sphere's surface using the haversine formula, which is differentiable:

$$h = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\theta_2 - \theta_1}{2}\right),$$
$$D = 2r\arcsin(\sqrt{h}), \tag{4}$$

where r is the radius of the sphere, and D is the great-circle distance between azimuth-elevation angles (θ_1, ϕ_1) and (θ_2, ϕ_2). By setting $r = 1$, we define D as our *Haversine loss*.

3.4. Model Architecture and Training Procedure

As described in Section 3.2, the inputs to these networks are active and reactive intensity vectors. The size of the input is

6x25x513, containing 513 frequency bins (16kHz sample rate), for each 6-D intensity vector, computed at each of 25 frames.

The CRNN architecture is shown in Fig. 1. There are three convolutional layers, each consisting of 2-D convolution, a rectified linear unit (ReLU), batch normalization, and max pooling. The outputs of each of the three layers are 64x25x64, 64x25x8, and 64x25x2, respectively. The output of the last layer is flattened to a 128-D vector for each of the 25 frames. Each frame's vector is fed into a two-layer bi-directional LSTM, and the output of the LSTM for each frame is fed through two time-distributed, fully-connected linear layers, generating a DOA estimate for each frame. As described in Section 3.3, we generate three forms of outputs depending on implementation: 2-D azimuth-elevation angles, 3-D Cartesian coordinates, or 429 DOA classes. During training, losses are computed at each frame and error is backpropagated through the network. During evaluation, a single DOA estimate is taken as the uniformly weighted average of all estimates across frames.

4. Experiment and Results

4.1. Benchmarks

We evaluate each model on 1,189 samples from three static-source microphone signals in the third-party sound localization and tracking (LOCATA) challenge dataset [7], and a dataset of ambisonic RIRs accompanying the Spatially Oriented Format for Acoustics (SOFA) convention [8]. No real RIRs are involved during the training phase. The noise and clean speech used for training and test are from different datasets. We use Eq. (4) as our angular error metric for visualization in Fig. 2.

Each signal in the LOCATA dataset is a real-world ambisonic speech recording with optically tracked azimuth-elevation labels. In theory, ambisonic coefficients up to the 4th order can be captured by an Eigenmike microphone, but we only use its first order components.

From the SOFA dataset, we extract 225 SRIRs recorded in the Alte Pinakothek museum using Eigenmike®, Sennheiser AMBEO®, and SoundField®microphones. Positions and rotations of all loudspeaker and microphones are provided by measurements using laser meter and pointers. A reverberant test set is generated by convolving each SOFA SRIR with a random 1-second clip from the CMU Arctic speech databases [23]. Recorded background noise from the LOCATA dataset is added at a mean SNR level of $10dB$ with a standard deviation of $5dB$.

4.2. Results and Analysis

4.2.1. LOCATA Dataset

We compute average error along the temporal axis, for each static-source signal in LOCATA Task 1. Estimates of DOA are generated for each frame in the microphone signal using a sliding window. The resulting estimates are interpolated to the timestamps provided in the LOCATA dataset. Prediction error is computed as the angular distance between the prediction and the ground-truth DOA. Angular tracking error is visualized in Fig. 2, for each predictor. Average angular error is computed over 234, 439, and 512 timestamps for Recordings 1, 2, and 3, respectively. Timestamps are selected to compute angular error if each algorithm makes a prediction for that timestamp, and the timestamp is located inside a voice-active region.

To generate a prediction at frame i using a neural network, we feed in the sequence of 25 frames centered at i, and generate a sequence of 25 outputs of the network. If the output is Cartesian
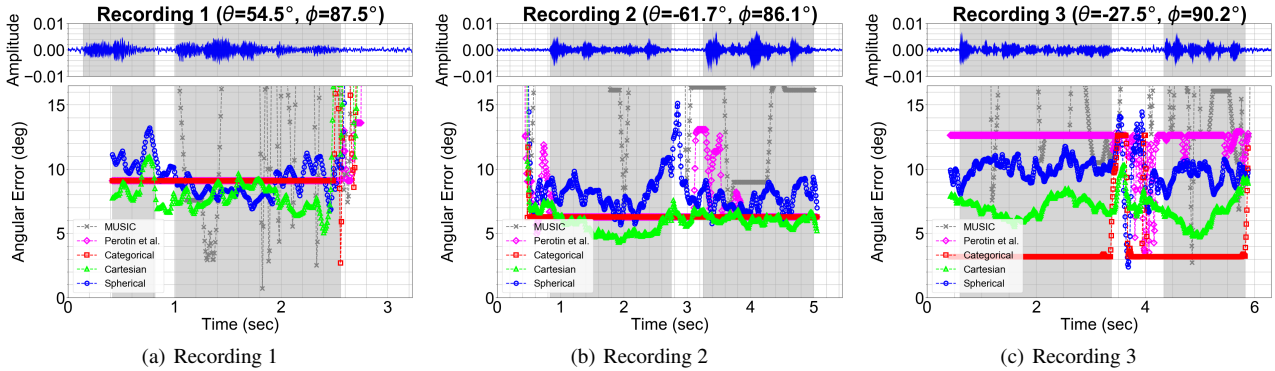
Figure 2: *Waveforms (top row) and angular tracking error (bottom row) for Recordings 1-3 in LOCATA Task 1. Shaded regions in the waveform indicate voice-active regions, while shaded regions in angular tracking error indicate the intersection of voice activity and regions containing predictions for all models. Each model must wait for a complete input window to make a prediction, hence the regions are not always identical between waveform and tracking error. Cartesian and Categorical models trained on our synthetic dataset achieve consistently lower tracking error compared with the classifier trained by Perotin et al. [6] and the MUSIC algorithm [11].*

or spherical, we estimate DOA at i as the average of the outputs at each frame. If the output is a classification grid, we average the output grid over the frames, to produce a cumulative score for each DOA, and choose the DOA with highest score as the prediction. When running the MUSIC algorithm, we restrict it to 4-channel recordings, as well, for fair comparisons.

Table 2: *Average angular tracking error within voice-active regions of LOCATA Task 1 recordings. Best performance in each column is highlighted in **bold**. All models are trained on data using specular and diffuse reflections in the geometric propagation algorithm, except for the MUSIC algorithm, which does not rely on training data. Perotin et al. [6] refers to the Categorical model trained on data generated by the image-source method.*

Model	Recording 1	Recording 2	Recording 3
MUSIC	18.6°	16.9°	17.5°
Perotin et al.	9.1°	6.7°	12.5°
Categorical	9.3°	6.3°	**3.2°**
Cartesian	**8.5°**	**5.8°**	6.8°
Spherical	9.2°	7.9°	9.9°

We observe that in Fig. 2 and Tab. 2, our Cartesian model achieves consistently lowest error on Recording 1 and 2, while our categorical model shows best performance in Recording 3. However, the Spherical model yields higher error than the other two models. We also observe that re-training the original model from Perotin et al. [6] using data generated by geometric method to augment data results in lower tracking error compared with the use of the image-source method.

4.2.2. SOFA Dataset

A larger scale test is performed using the SOFA dataset. MUSIC algorithm is not evaluated because this dataset does not provide microphone hardware configurations required by MUSIC. We compute the percentage of correctly predicted directions under error tolerances of 5°, 10°, and 15°, as well as each model's average angular error on the whole dataset. It can be seen from Tab. 3 that our Cartesian model consistently achieves the best performance under each column, outperforming the baseline model by 43% in terms of average prediction error.

During our training procedure, we notice that each model is able to converge within dozens of epochs. However, the number of trainable parameters in regression models is only 64% of that

Table 3: *Results on the SOFA dataset. First three columns show the percentage of DOA labels correctly predicted within error tolerances, followed by average angular errors, and %-improvement on baseline. Best performance in each column is highlighted in **bold**.*

Model	$< 5°$	$< 10°$	$< 15°$	Error	Improv.
Perotin et al.	11.9%	35.9%	73.2%	16.9°	-
Categorical	**24.4%**	58.2%	**88.7%**	9.96°	41%
Cartesian	**24.4%**	**66.3%**	88.2%	**9.68°**	**43%**
Spherical	18.2%	55.8%	82.5%	11.2°	34%

in the classification model. This suggests that regression models tend to have a hypothesis set with lower complexity, which results in lower generalization error when tested on real data. We also tested the option of letting all models have approximately the same amount of trainable parameters, which has degraded the performance of the regression models. In conclusion, we are able to train a regression model that has superior performance over its corresponding classification model, although we do not observe obvious benefits in using the Spherical formulation.

5. Discussion and Future Work

In this paper, we demonstrate the benefits of using a geometric sound propagation simulator, as compared with image source methods for training DOA estimation networks, by reporting a higher accuracy on evaluation data. We evaluate the performance of a CRNN model in three output formulations: categorical, Cartesian, and spherical. We test them on two 3rd-party datasets and show that our Cartesian regression model achieves superior performance over classification and spherical models.

Evaluating classification models involves tuning the classification grid resolution, which we kept fixed. Further, our work is limited to single-source localization problems, whereas in multi-source localization problems, classification models have intrinsic advantages over regression models. Lastly, we restricted our simulation to extremely simple room settings to guarantee a fair comparison with the image-source method. Future work may involve experimentation on more complex room configurations.

6. Acknowledgements

7. References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1997, pp. 375–378.

[3] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2814–2818.

[4] W. Zhang, P. Samarasinghe, H. Chen, and T. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Applied Sciences*, vol. 7, no. 5, p. 532, 2017.

[5] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[6] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector," in *IWAENC*, 2018.

[7] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, July 2018.

[8] A. Pérez-López and J. De Muynke, "Ambisonics directional room impulse response as a new convention of the spatially oriented format for acoustics," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[9] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume 24 Issue 4*, 1976.

[10] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," in *IEEE Transactions on Speech and Audio Processing, Volume 9 Issue 8*, 2001.

[11] R. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, 1986.

[12] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays, pp. 157-180*, 2001.

[13] 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*, 2018.

[14] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: from audio signal to source position coordinates," in *Sensors*, 2018.

[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE*, 2017, pp. 5220–5224.

[16] C. Schissler and D. Manocha, "Gsound: Interactive sound propagation for games," in *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society, 2011.

[17] C. Schissler, R. Mehra, and D. Manocha, "High-order diffraction and diffuse reflections for interactive sound propagation in large environments," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 39, 2014.

[18] C. Schissler and D. Manocha, "Interactive sound propagation and rendering for large multi-source scenes," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 1, p. 2, 2017.

[19] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, "Improving reverberant speech training using diffuse acoustic simulation," *arXiv preprint arXiv:1907.03988*, 2019.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[21] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[22] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng, and Y. Li, "Noisy training for deep neural networks in speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 2, 2015.

[23] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.