BLIND ROOM VOLUME ESTIMATION FROM SINGLE-CHANNEL NOISY SPEECH

Andrea F. Genovese[1]*, *Hannes Gamper*[2], *Ville Pulkki*[3]†, *Nikunj Raghuvanshi*[2], *Ivan J. Tashev*[2]

[1]New York University, NY, USA
[2]Microsoft Research Redmond, WA, USA
[3]Aalto University, Espoo, Finland

ABSTRACT

Recent work on acoustic parameter estimation indicates that geometric room volume can be useful for modeling the character of an acoustic environment. However, estimating volume from audio signals remains a challenging problem. Here we propose using a convolutional neural network model to estimate the room volume blindly from reverberant single-channel speech signals in the presence of noise. The model is shown to produce estimates within approximately a factor of two to the true value, for rooms ranging in size from small offices to large concert halls.

Index Terms— Room acoustics, room size, non-intrusive parameter estimation, signal processing, convolutional neural network

1. INTRODUCTION

A current challenge in audio processing is the dynamic parameterization of the local acoustic space of a listener. The parameters that describe the acoustic character of a user environment can be used to model or design audio filters for various applications. Knowledge of the local room acoustics can be used to improve the plausibility of immersive audio in mixed reality applications that aim at blending real and virtual sound sources together into a cohesive auditory scene [1]. Speech-processing applications may use the room impulse response (RIR) parameters to enhance a speech signal or aid dereverberation algorithms, for the purpose of word recognition or communication clarity [2, 3]. Audio forensics strategies may benefit from this information for room identification tasks [4, 5].

Measured RIRs can be used to derive parameters such as reverberation time (T_{60}) and direct-to-reverberant ratio (DRR). The RIR is composed of early, low-order reflections that are dependent on the source and receiver positions, followed by position-independent diffuse reverberation [6]. Given a description of an acoustic space in terms of these acoustic parameters, it is possible to model the response of a room for which a measured RIR is not available [7]. Another position-independent parameter that is useful for modeling the character of a room is the geometric room volume. Room volume has been linked to the estimation of the *critical distance*, which is the distance at which the direct and reverberant portions of a sound source hold equal power, i.e., the point at which the DRR is 0 dB. An approximation of critical distance as a function of the room volume is given by Sabine's equation [6]:

$$d_c(m) = \frac{1}{4}\sqrt{\frac{QA}{\pi}} \cong 0.1\sqrt{\frac{QV}{\pi \cdot T_{60}}}, \tag{1}$$

where Q denotes the wave-source directivity, A is the room surface area, and V is the room volume in cubic meters. Given measured or estimated values of T_{60}, V, and Q, (1) can be used to determine whether a virtual source should be rendered with a DRR smaller or greater than 0 dB, which can serve as a distance cue for the listener [6]. Similarly, the mixing time, another perceptually relevant parameter, has been related to cubic volume via $M_t = \sqrt{V}$ [1].

Room volume has also been proposed as a key part of the "reverberation fingerprint" of a room [7, 1, 4]. This fingerprint is limited to the diffuse part of the reverberation as it characterizes a room in isolation from the orientation and directivity of sources and receivers. Room volume can be used to retrieve initial diffused power, which together with the T_{60} as a function of frequency, describes the energy decay relief (EDR). The power of the diffuse reverberation is inversely proportional to the cubic volume [7]. This relationship can also be used to adapt a known RIR to a new room:

$$P_{\text{local}}(f) = P_{\text{ref}}(f)\frac{V_{\text{ref}}}{V_{\text{local}}}, \tag{2}$$

where P is the initial power spectrum of the EDR.

In practice, typically little or no prior information is available about a user's local acoustic environment, and measuring RIRs in situ is often not an option. Furthermore, variations of the user's environment over time may require updating local acoustic parameters dynamically. Therefore, blind prediction methods are necessary to estimate acoustic parameters on the fly. This would allow for adaptive processing schemes which can enable users to freely roam the world while the audio rendering engine system adaptively updates according to the present space.

While complex multimedia systems may be able to leverage depth sensors or cameras for estimating the local room volume, here we target scenarios where the available input data are limited to one or multiple microphone signals. In a real-world situation the captured audio signals may be affected by different types of environmental noise which may degrade the performance of an estimation algorithm. Past work regarding blind room parameter estimation in the presence of noise has focused on T_{60} and DRR. The 2015 ACE challenge [8, 9] set the bar for estimating T_{60} and DRR from speech signals in the presence of ambient, babble, or fan noise, with signal-processing based techniques achieving the best results in terms of the mean-squared estimation error and correlation between true and estimated parameters.

Recent work on T_{60} estimation showed promising results using deep neural networks [10, 11]. Here we propose using a deep convolutional neural network (CNN) to blindly estimate room volume from noisy speech signals recorded in a user's acoustic environment. The network is evaluated on a set of publicly available RIRs with known room volume information. Results indicate that room vol-

*The work was done as a research intern at Microsoft Research Labs in Redmond, WA, USA.

†This work was done as a consulting researcher at Microsoft Research Labs, Redmond, WA, USA.

To appear in *Proc. ICASSP 2019, Brighton, United Kingdom*

ume can be estimated blindly within an error of about a factor of two of the true volume.

2. RELATED WORK

Room volume estimation has often been approached as a classification problem. In the field of audio-forensics, early attempts used MFCC-based features to identify the room of a previous recording from a closed set [4]. Overall accuracy reached 84%. However, the method was not evaluated on unseen rooms. Peters et al. studied room identification using machine learning and a corpus of measured RIRs [12]. The study collected a total of 168 RIRs for 7 unique rooms, convolved with anechoic speech to create reverberant examples. A room classifier was implemented with Gaussian Mixture Models using MFCC acoustic features. An identification accuracy of 85% was achieved although the method was not evaluated with respect to additive noise or rooms not contained in the training set. Shabtai et al. [13] also propose a room volume classification paradigm but working directly on RIRs rather than reverberant signals, effectively making the method non-blind.

Similarly, Murgai et al. propose treating abrupt speech stops in reverberant speech signals directly as the RIR to semi-blindly estimate the T_{60}, the clarity index (C_{50}), early reflections density, and others, including room volume [14]. While the results for volume estimation through regression were promising, the data set used consisted of RIRs obtained via the image source method (ISM), which may not reflect the complexity of real, measured RIRs. Furthermore, the robustness of the method to environmental noise was not reported.

3. PROPOSED APPROACH

The present study focuses on the blind estimation of the local room volume from single-channel speech signals in the presence of different forms of noise, with no prior knowledge of the RIR, dry speech signal or room geometry. We propose formulating the estimation as a regression problem by using a convolutional neural network (CNN) model trained with examples including different types of noise at various SNRs levels, using measured RIR data spanning a wide range of room sizes. To avoid overfitting, the test set used to evaluate the model performance does not contain any RIRs measured in rooms that are part of the training set.

3.1. Data Collection

A corpus of measured RIRs was collected with the aim of training and testing the proposed CNN model over a wide range of room volumes and types. Table 1 summarizes the data corpus. RIRs for which room volume information was available were drawn from 11 public data sets as well as a proprietary database, resulting in a total of 83 unique rooms. All RIRs were re-sampled at 16 kHz. To reduce possible imbalance and overfitting, the representation of each unique room was capped to a maximum of 100 examples [15]. Figure 1 illustrates the distribution of the unique room volumes on a logarithmic scale. The bimodal pattern and center gap in the histogram of the real room measurements reflects the nature of the collected datasets; in the field of room acoustics, small rectangular rooms or large concert halls are often the object of study.

Dataset	# rooms	sampling rate [kHz]
ACE [8]	7	48
AIR [16]	4	48
CHRG [17]	13	12.78
ECHOTHIEF [18]	1	44.1
MARDY [19]	1	48
OPENAIRLIB [20]	22	various
PORI [21]	2	48
QMUL [22]	3	96
REVERB2014 [3]	3	16
SMARD [23]	1	48
SOFA [24]	1	44.1
(proprietary)	25	48
Total	83	

Table 1. Summary of selected RIR databases.

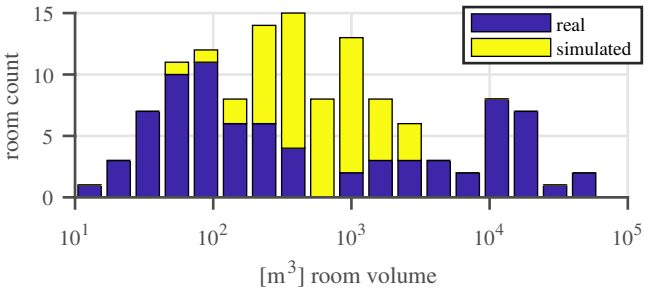

Fig. 1. Room volume distributions for real and synthetic rooms.

3.2. Data Augmentation

To mitigate the lack of available data around approximately 1000 m^3, simulated RIRs were added to the training corpus. The simulations were obtained using adaptive rectangular decomposition (ARD), a time-domain spectral wave solver [25]. Synthetic RIRs for a single source position and 50 random receiver positions were computed for 50 polygonal rooms of irregular, convex geometry. The rooms spanned a volume range of 59 m^3 to 2715 m^3 and simulations were band-limited to 2 kHz. These choices were found to sufficiently augment the data while managing computational requirements, which scale linearly with volume and the fourth power of the upper frequency limit.

A wave-propagation model based on direct solution of the wave equation is more realistic than geometric approximations, such as ISM, due to the ability to model the response of arbitrary geometries and account for diffraction and scattering effects. Early experiments indicated that augmenting the data with synthetic RIRs reduced the volume prediction error. Synthetic RIRs lack some expected irregularity features of measured RIRs, causing risk of overfitting if the dataset is largely synthetic [15]. Therefore, the simulations were only used for data augmentation and not included in the test set.

3.3. Stimuli Generation

To produce simulated examples of speech in noisy environments, the RIRs were convolved with anechoic speech signals (male and female speakers), following a similar protocol to the ACE challenge [8].

Each example y was created as:

$$y[n] = \mathbf{h}_1^T \mathbf{x}[n] + \mathbf{h}_2^T \left(\frac{v[n]\sqrt{\xi_s}}{\sqrt{\xi}} \right), \quad (3)$$

where $\mathbf{h}_1$ and $\mathbf{h}_2$ are the RIR vectors used for the speech signal and noise signal respectively (belonging to the same room), $\mathbf{x}$ is the speech vector, ξ is the desired SNR level and ξ_s is the power ratio of the clean speech and noise signal v. The SNR levels used were +Inf (no noise), +20 dB, +10 dB and 0 dB.

Similarly to earlier work [10], the noise was simulated by shaping Gaussian noise with the spectra of ambient noise recorded using a spherical microphone array in various environments. Speech samples were drawn randomly from a data set of over 900 semi-anechoic recordings and convolved with RIRs, to yield a total of 100 samples for each unique room in the data corpus. The resulting noisy speech signals were split into frames with a fixed duration of 4 seconds with an overlap of 1 second. The level of each chunk was normalized using A-weighting. The total number of fixed-duration examples was 23072.

3.4. Training and test sets

All noisy reverberant speech samples were assigned to either the training set or the test set based on the RIRs used to generate them, such that no room would be part of both sets. The 7 rooms from the ACE dataset [8], 2 large rooms from the OpenAir dataset [20], and one concert hall from the CHRG dataset [17] were selected for the test set. To monitor performance during training and prevent overfitting, the training set was further split into training and validation sets using a 9:1 ratio. The data split is summarised in Table 2.

Set	# examples	real rooms	simulated rooms
Training	19 608	66	47
Validation	1713	7	3
Test	1751	10	-

Table 2. Summary of data split.

4. FEATURE REPRESENTATION

Exploratory listening tests looked at the effect of room volume on sound in order to obtain insights about which representations were able to capture its influence on the signal. RIRs with similar T_{60}, but different volume, were convolved with a test sound. This process highlighted the possible impact of low frequency effects. We computed a set of exploratory features to provide the network with a variety of spectro-temporal representations of the input signals. This included Gammatone features, which were used in prior work on T_{60} estimation[10].

The resolution specifications for the spectro-temporal features were designed with the intent of keeping a low amount of trainable network parameters. A low complexity model can yield more generalizable results as well as shorter training times [15]. Thus, an ERB Gammatone filterbank with 20 frequency bands from 50 Hz to 2 kHz was used. Features were computed as the log-energy of frames of 64 samples, with a hop size of 32. This resulted in a 20×1991 feature matrix for each 4-seconds clip. Other features able to capture the low-frequency behaviour were investigated. Perhaps because of higher room modes activity and interactions, large room volumes pointed to higher low-frequency energy and higher cepstrum energy associated with frequency notch patterns. Pilot experiments were conducted with a number of feature candidates. The finalized set of features was concatenated into a single stack, as described in Table 3. The final number of trainable parameters was 12556.

Feature	Dimensions
Gammatone filterbank	20×1991
DFT (up to 500 Hz)	1×1991
Magnitude-sorted DFT	1×1991
Cepstrum	1×1991
Envelope Follower	1×1991
Time-domain signal (low-passed)	1×1991

Table 3. Network Model Feature Stack.

	conv1	conv2	conv3	conv4	conv5	conv6
size	(1,10)	(1,10)	(1,10)	(1,10)	(3,9)	(3,9)
stride	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)
$\text{avgpool}_{\text{size}}$	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(2,2)
$\text{avgpool}_{\text{stride}}$	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(2,2)
# filters	30	20	10	10	5	5

Table 4. Parameters of the convolutional layers used. The input feature size was $[25 \times 1991]$. The network had a total of 12556 trainable parameters.

5. EXPERIMENTAL EVALUATION

5.1. Model Architecture

The proposed model architecture was based on CNNs due to their suitability for capturing 2-dimensional time-frequency signal patterns. The final model comprises six convolutional layers followed by an average pooling layer, one dropout layer (50% rate), and a final, fully connected layer with a single output node. Table 4 and Figure 2 illustrate the full architecture employed. The filters were designed to first convolve across time bins in one dimension and then combine the spectral features in the last two layers.

5.2. Evaluation metrics

The problem was formulated as a regression problem on the log-10 of the room volume, ensuring that the estimation error would be related to its order of magnitude. Given the large range of room sizes, a logarithmic estimate is deemed as more appropriate than a linear one. The performance was evaluated in terms of the mean squared error (MSE), the mean error (bias), and the Pearson's correlation coefficient (ρ). Another metric used for this specific study is based on the mean absolute logarithm of the ratio between the estimated volume $\hat{L}$ in m^3 and the ground truth L:

$$MeanMult = e^{\wedge} \left(\frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} \left| \ln \left(\frac{\hat{L}_n}{L_n} \right) \right| \right), \quad (4)$$

where N denotes the number of test samples and $\ln$ is the natural logarithm. This metric summarizes the error in terms of the average multiple of the estimated volume in m^3 compared to the true volume.

5.3. Results

The model was implemented using the Microsoft Cognitive Toolkit (CNTK [26]) and employed stochastic optimization with a squared

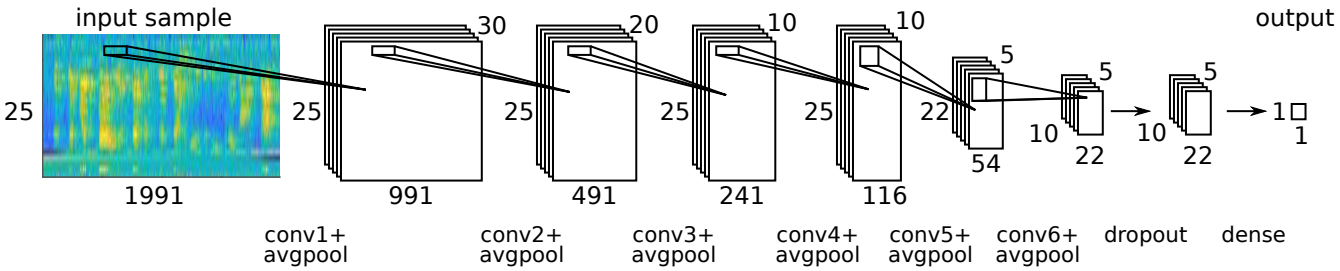

Fig. 2. CNN architecture used. The dropout layer rate was 50%, followed by a fully connected layer with one output node.

error loss function [27] over 1000 epochs. The model was trained and evaluated on the log10 of the room volume. Experiments were performed with different feature combinations and different variations of the network specifications. The best output on the test set created was given by the model hereby presented. To further validate the results against non-simulated noise recordings, the model was also tested directly on the development and evaluation sets used for the ACE challenge [8]. These sets were generated by combining real, recorded ambient noise and anechoic speech convolved with measured RIRs. The rooms measured as part of the ACE corpus ranged from 47.3 m^3 to 364.6 m^3.

Figure 3 illustrates the confusion matrices for the training set, the test set, and the ACE corpus. The room volume estimates are distributed around the ground truth across the range of tested room sizes. In the range where real room data had to be augmented with synthetic data, i.e., around 1000 m^3, the training error shows a lower variance, perhaps indicating a mismatch between the simulations and the complexity of real measurements. Performance on the test set was comparable to the training performance, indicating that the model did not appear to be overfitting. Figure 3 (right) illustrates the performance on the ACE corpus, i.e., the speech and noise samples provided with the ACE challenge [8]. Note that these are the same RIRs as in the test set below 1000 m^3. As can be seen, the variance of the error increases.

Table 5 shows the results of the prediction model for the metrics introduced in Section 5.2. For the test set, containing only unseen rooms, the model achieves an MSE of 0.19 and *MeanMult* factor of 2.27. For comparison, results are provided for the test set considering only the RIRs obtained from the ACE corpus. As can be seen, the estimation bias increases while correlation decreases, indicating that distinguishing small rooms (below 400 m^3) is challenging, perhaps due to insufficient training data. When evaluating the model on the ACE corpus, which consists entirely of unseen data, performance deteriorates. This may be due to a mismatch between the simulated noise conditions in our data set and the real, recorded noise in the ACE corpus. After excluding samples with SNR < 18 dB from the ACE corpus, the performance metrics are quite similar to the test results for ACE rooms only (see Table 5). This indicates that a mismatch between recorded and simulated noise may indeed explain some of the differences in performance between the ACE corpus, consisting entirely of unseen data, and the test set.

6. CONCLUSIONS AND FUTURE WORK

This paper proposes estimating geometric room volume blindly from noisy single-channel speech signals using a convolutional neural network. The proposed model was trained on a data corpus comprising measured as well as simulated RIRs. Unlike previous methods, the

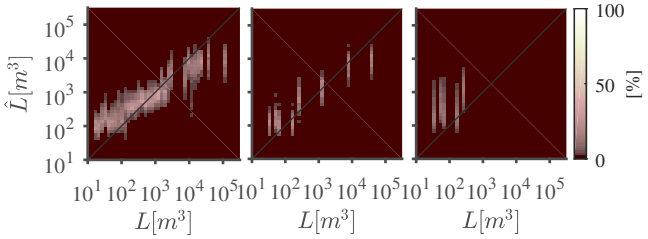

Fig. 3. Confusion matrices of the training set (left), test set (center), and the ACE corpus (right).

	MSE	Pearson's ρ	MeanMult	bias
training	0.20	0.89	2.25	0.02
test	0.19	0.90	2.27	-0.01
test (ACE rooms only)	0.16	0.41	2.16	-0.18
ACE corpus	0.43	0.28	3.31	-0.44
ACE corpus (high SNR)	0.19	0.39	2.28	-0.23

Table 5. Results with respect to log-10 of the volume for training and test set, as well as the public ACE corpus of noisy, reverberant speech recordings. For comparison, the performance of the test set considering only RIRs contained in the ACE corpus is provided.

estimation was formulated as a regression problem in the logarithmic domain, and results were obtained for a test set containing only unseen, measured RIRs. Results show that room volume can be estimated within approximately a factor of two of the true value, for a wide range of room sizes. Performance deteriorated when evaluating the model on a completely separate, measured data corpus of unseen rooms, speech, and ambient noise. However, when excluding examples with low SNR, results improved and were approximately in line with the performance on the test corpus simulated using the same rooms. This indicates that the data set generated here for model training and testing is useful and, to some extent, realistic, but should be extended to include a wider range of noise scenarios. Overall, these initial results are encouraging, and increasing the size and quality of the training data should improve the model's performance and ability to generalize.

Future work will address the uneven distribution of room sizes and the limited scope of noise scenarios in the current data set. Furthermore, alternative feature representations, e.g., the echo density profile or the phase spectrum [28] could be studied. Neural network explanatory techniques [29] may allow to identify the most suitable spectro-temporal signal representation for blind volume estimation.

7. REFERENCES

[1] Jean-Marc Jot and Keun Sup Lee, "Augmented reality headphone environment rendering," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2016.

[2] Patrick A. Naylor and Nikolay D. Gaubitch, *Speech dereverberation*, Springer Science & Business Media, 2010.

[3] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[4] Alastair H. Moore, Mike Brookes, and Patrick A. Naylor, "Room identification using roomprints," in *Audio Engineering Society Conference: 54th International Conference: Audio Forensics*. Audio Engineering Society, 2014.

[5] Matteo Mascia, Antonio Canclini, Fabio Antonacci, Marco Tagliasacchi, Augusto Sarti, and Stefano Tubaro, "Forensic and anti-forensic analysis of indoor/outdoor classifiers based on acoustic clues," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 2072–2076.

[6] Heinrich Kuttruff, *Room acoustics*, CRC Press, 2016.

[7] Jean-Marc Jot, Laurent Cerveau, and Olivier Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Audio Engineering Society Convention 103*. Audio Engineering Society, 1997.

[8] James Eaton, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor, "The ACE challenge: Corpus description and performance evaluation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*. IEEE, 2015, pp. 1–5.

[9] James Eaton, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 10, pp. 1681–1693, 2016.

[10] Hannes Gamper and Ivan J Tashev, "Blind reverberation time estimation using a convolutional neural network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 136–140.

[11] Myungin Lee and Joon-Hyuk Chang, "Deep neural network based blind estimation of reverberation time based on multichannel microphones," *Acta Acustica united with Acustica*, vol. 104, no. 3, pp. 486–495, 2018.

[12] Nils Peters, Howard Lei, and Gerald Friedland, "Name that room: Room identification using acoustic features in a recording," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 841–844.

[13] N. Shabtai, Boaz Rafaely, and Yaniv Zigel, "Room volume classification from reverberant speech," in *Proc. of intl Workshop on Acoustics Signal Enhancement, Tel Aviv, Israel*, 2010.

[14] Prateek Murgai, Mark Rau, and Jean-Marc Jot, "Blind estimation of the reverberation fingerprint of unknown acoustic environments," in *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[16] Marco Jeub, Magnus Schafer, and Peter Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009, pp. 1–5.

[17] J.S. Bradley, "Data from 13 North American concert halls," *Internal Report No. 668, Institute for Research in Construction, National Research Council Canada*, 1994.

[18] Chris Warren, "Echothief impulse response library," `http://www.echothief.com/`, Accessed: 2018-08-01.

[19] Jimi Y.C. Wen, Nikolay D. Gaubitch, Emanuel A.P. Habets, Tony Myatt, and Patrick A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *in Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC*. Citeseer, 2006.

[20] Damian T. Murphy and Simon Shelley, "OpenAir: An interactive auralization web resource and database," in *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[21] J. Merimaa, T. Peltonen, and T. Lokki, "Concert hall impulse responses-pori," *Finland, reference documentation. Published online at http://www. acoustics. hut. fi/projects/poririrs*, 2005.

[22] Rebecca Stewart and Mark Sandler, "Database of omnidirectional and b-format room impulse responses," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 165–168.

[23] Jesper Kjær Nielsen, Jesper Rindom Jensen, Søren Holdt Jensen, and Mads Græsbøll Christensen, "The single-and multichannel audio recordings database (SMARD).," in *IWAENC*, 2014, pp. 40–44.

[24] "Sofa general purpose database," `https://www.sofaconventions.org/mediawiki/index.php/Files`, Online; accessed May 2018.

[25] Nikunj Raghuvanshi, Rahul Narain, and Ming C. Lin, "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 789–801, 2009.

[26] Frank Seide and Amit Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2135–2135.

[27] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Mikko-Ville Laitinen, Sascha Disch, and Ville Pulkki, "Sensitivity of human hearing to changes in phase spectrum," *Journal of the Audio Engineering Society*, vol. 61, no. 11, pp. 860–877, 2013.

[29] Etienne Thuillier, Hannes Gamper, and Ivan J Tashev, "Spatial audio feature discovery with convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6797–6801.