Comparison of Convolution Types in CNN-based Feature Extraction for Sound Source Localization

Daniel Krause
AGH University of Science and Technology
Department of Electronics
Kraków, Poland
danielkrause2h@gmail.com

Archontis Politis
Tampere University
Faculty of Information Technology and Communication Sciences
Tampere, Finland
archontis.politis@tuni.fi

Konrad Kowalczyk
AGH University of Science and Technology
Department of Electronics
Kraków, Poland
konrad.kowalczyk@agh.edu.pl

***Abstract*—This paper presents an overview of several approaches to convolutional feature extraction in the context of deep neural network (DNN) based sound source localization. Different ways of processing multichannel audio data in the time-frequency domain using convolutional neural networks (CNNs) are described and tested with the aim to provide a comparative study of their performance. In most considered approaches, models are trained with phase and magnitude components of the Short-Time Fourier Transform (STFT). In addition to state-of-the-art 2D convolutional layers, we investigate several solutions for the processing of 3D matrices containing multichannel complex representation of the microphone signals. The first two proposed approaches are the 3D convolutions and depthwise separable convolutions in which two types of filters are used to exploit information within and between the channels. Note that this paper presents the first application of depthwise separable convolutions in a task of sound source localization. The third approach is based on complex-valued neural networks which allows for performing convolutions directly on complex signal representations. Experiments are conducted using two synthetic datasets containing noise and speech signals recorded using a tetrahedral microphone array. The paper presents the results obtained using all investigated model types and discusses the resulting accuracy and computational complexity in DNN-based source localization.**

***Index Terms*—sound source localization, sound feature extraction, convolutional neural networks, complex convolutions, depthwise convolutions**

I. Introduction

Sound Source Localization (SSL) constitutes an important element of a wide variety of applications including speech recognition and separation [1, 2], audio-driven robots [3], surveillance systems [4], and teleconferencing [5]. Recent research shows that combining localization with Sound Event Detection (SED) can produce benefits in terms of both accuracy and system complexity [6–8]. In many of the aforementioned applications, estimation of a sound object's position typically concerns the task of the direction-of-arrival (DOA) estimation. Many classical DOA estimation methods have been established over the years including generalized cross-correlation with phase transform (GCC-PHAT) [9], Steered Response Power (SRP) [10], and subspace-based approaches such as MUSIC [11] and ESPRIT [12].

More recently, research on DOA estimation has switched to machine learning with the hope to overcome some of the limitations of classical model-based methods. To this end, supervised learning, which in contrast to deterministic algorithms enables adaptation of the models to various acoustic conditions and increases their robustness against noise and reverberation provided sufficient data is available for training. Early solutions are based on simple models such as Gaussian Mixture Models (GMM) [13], support vector machines (SVM) [14], and kernel estimators [15]. Current research focuses mainly on Deep Neural Network (DNN) based solutions [16–18], as they can utilize more general signal representations through encapsulating feature extraction into the training process. In DNN-based localization, feature extraction is most commonly performed using Convolutional Neural Networks (CNN) [6, 7, 17] with phase and magnitude of the STFT representation treated as independent channels.

In this paper, we aim to compare different types of deep convolutional layers in the context of extracting features from complex representation of the microphone signals in the Short-Time Fourier Transform (STFT) domain. In addition to commonly used two-dimensional (2D) CNNs applied independently to phase [17] or both magnitude and phase components [7], we explore several solutions that jointly exploit the time-frequency and inter-channel relations. The investigated techniques include (i) three-dimensional (3D) convolutions, (ii) depthwise separable convolutions as a novel way of processing multichannel audio data with reduced model complexity (i.e. significantly lower number of network parameters). (iii) We utilize complex CNNs which operate directly on the complex signal spectrum in the STFT domain. Comparative experiments are performed using two synthetic datasets in which a single sound source, either white Gaussian noise or speech, is localized from recordings of a tetrahedral microphone array.

II. Common basic network architecture

In order to facilitate a proper comparison of different realizations of convolution, we use the same DNN architecture for all considered approaches. As a baseline model, the DNN architecture from [19] is adopted as it has been shown to provide good localization results for both real and synthetic

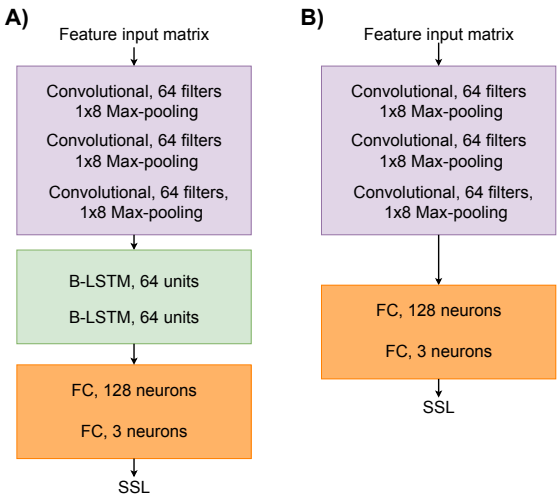

Fig. 1. Baseline model architecture A) using B-LSTM layers, B) without recurrent layers

data. This architecture, depicted in Fig. 1A, consists of 3 convolutional layers (of different type), followed by 2 recurrent layers that exploit temporal information, and 2 fully connected layers. In addition, we considered the second common architecture, shown in Fig. 1B, in which the recurrent layers are removed for simplification of the model.

The basic model is trained with a 8x24x512 time-frequency (TF) matrix, where 24 is the number of time frames in the sequence, 512 the number of frequency bins, and 8 depicts the doubled channel dimension, as for each of the four microphones phase and magnitude spectrograms are represented separately. The input is filtered by 3 convolutional layers, each consisting of 64 3x3 kernels. Each CNN layer is followed by a rectified linear unit (ReLU) activation function and batch normalization. MaxPooling in the frequency domain is performed with pooling ratios of 8, 8 and 4 for each respective layer. In the basic version including the recurrent layers, the extracted features are fed to 2 subsequent 64-unit bi-directional Long Short-Term Memory (B-LSTM) layers with tanh activation that improve the localization performance by exploiting longer temporal dependencies than the receptive field covered by the CNNs. In both variations of the network, the final output is produced by 2 fully-connected (FC) layers. The first FC layer consists of 128 linear neurons, whereas the last one produces 3 linear outputs corresponding to three Cartesian coordinates of the unit vector pointing towards the source (i.e., it represents the unit-norm DOA vector). Finally, we apply smoothing over 24 frames to the output of the network by replacing the angular values with their median.

III. Investigated Convolution Types

In this section, we provide an overview of the state-of-the-art 2D convolution techniques, and present three types of convolutions for complex spectra of the microphone signals.

A. State-of-the-art 2D convolutions

In DNN-based audio signal processing, a common approach is to represent the complex spectrum of a single audio signal as a phase-spectrogram (PS) and a magnitude-spectogram (MS), i.e. by representing the complex spectrum using phase and magnitude treated as independent channels. The advantage of such an approach is that both channels contain real numbers and thus standard 2D convolutions can be applied. Commonly used 2D CNN layers perform convolution by processing each channel separately and summing the results over all channels for each filter. This is a simple and efficient method to model time-frequency dependencies, however, it may lack important information about inter-channel relations. In the following, we describe three alternative types of convolution that can be applied to multichannel audio data for DOA estimation.

B. Complex convolutions

In deep learning, PS and MS are commonly used to represent complex-valued signals. Although spectro-temporal information is exploited in this approach, the disadvantage is that the number of feature channels is doubled and the magnitude and phase information is mixed already after the first CNN layer.

To this end, we directly use complex representation of the microphone signals with the aim to inherently exploit the entire information about the magnitude and phase contained in complex signal spectra. By allowing the network to self-exploit the relations between the complex spectra, rather than feeding the network with phase and magnitude extracted from the microphone signals, we expect to capture spatial information even more accurately. In this work, we follow the complex convolution processing proposed in [20], which is given by:

$$\begin{bmatrix}\Re(\mathbf{H}*\mathbf{X})\\ \Im(\mathbf{H}*\mathbf{X})\end{bmatrix} = \begin{bmatrix}\Re(\mathbf{H}) & -\Im(\mathbf{H})\\ \Im(\mathbf{H}) & \Re(\mathbf{H})\end{bmatrix} * \begin{bmatrix}\Re(\mathbf{X})\\ \Im(\mathbf{X})\end{bmatrix}, \quad (1)$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary components of a matrix, $i = \sqrt{-1}$, and $\mathbf{X} = \Re(\mathbf{X}) + i\,\Im(\mathbf{X})$ denotes the complex spectrogram which is convolved with a 3x3 kernel $\mathbf{H} = \Re(\mathbf{H}) + i\,\Im(\mathbf{H})$ containing complex filter weights. The features are normalized after each layer using Complex Batch Normalization, as described in [20]. Furthermore, we use the $\mathbb{C}$ReLU activation function given by

$$\mathbb{C}ReLU(z) = ReLU(\Re(z)) + ReLU(\Im(z)) \quad (2)$$

to handle complex numbers $z = \Re(z) + i\,\Im(z)$, which has been shown in [20] to give the best results amongst compared variants. MaxPooling is applied to the real and imaginary components separately.

C. Depthwise separable convolutions

Another interesting approach is to use the so-called depthwise separable (DS) convolutional neural networks which split the convolution process into two parts [21]. As depicted in Fig. 2, the first step consists in applying a single 3x3 depthwise convolution matrix to each channel separately, keeping the channel dimension unchanged. Note that this is analogous to

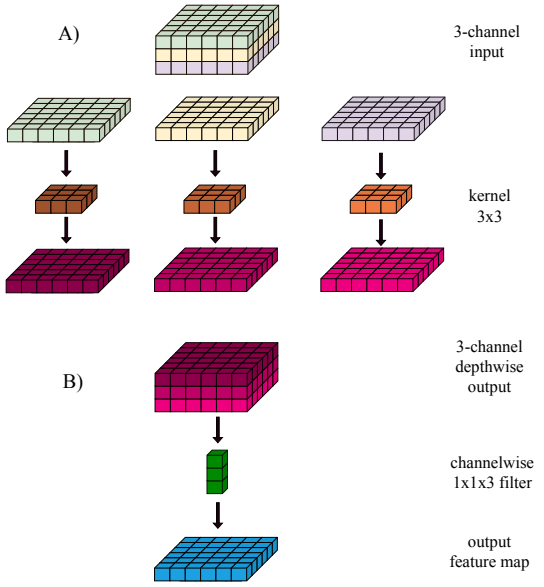

Fig. 2. Depthwise separable convolution steps for a 3-channel input map: A) depthwise convolution along the time-frequency axes; B) pointwise convolution along the channel axis.

the classical convolutional layers, in which 2D filters are used to find the time-frequency dependencies. In depthwise separable convolutions usually only a single kernel is used [21]. The second step consists in applying 1D filters that extract inter-channel dependencies and lower the channel dimension size to one. Since typically several inter-channel filters are used, the resulting feature maps are mainly exploring the channel dimension, and less of time-frequency dependencies. This approach enables efficient multichannel information extraction in conjunction with a significant complexity reduction. In this paper, we investigate depthwise separable CNNs with one (**DS-1**), two (**DS-2**), four (**DS-4**) and eight (**DS-8**) time-frequency kernels with the aim to achieve a potential improvement in localization accuracy. The number of inter-channel filters is fixed and set to 64. To the best of the authors' knowledge, this kind of convolutions has not yet been applied in the context of sound source localization.

D. 3D convolutions

3D CNNs are a three-dimensional expansion of standard 2D convolutional layers. Adding channels as a third dimension of the input matrix allows the network to learn features that capture inter- and intra-channel dependencies simultaneously. In this work, we propose three ways to do that:

- **3D-1** is a method in which all PS and MS are stacked to create a single 3D matrix, which is then fed as a single channel to the first layer. This results in a 1x8x24x512 input matrix. To compute features that capture dependencies in all dimensions, a 3x3x3 kernel is used.
- **3D-2** is a method in which feature matrices are stacked, however the resulting PS and MS 3D-arrays are processed separately, i.e. a 2x4x24x512 input multidimensional array is created allowing the network to exploit phase and magnitude information independently. Similarly to the first method, for each of them a 3x3x3 kernel is used.
- **3D-4** is a method in which we pair PS and MS matrices for each signal into separate 3D arrays, obtaining a 4x2x24x512 input for 4 microphone signals. The idea is to extract features characteristic of each audio channel. Since the feature channel dimension is reduced to 2, a kernel of size 2x3x3 is used.

IV. Performed experiments

A. Datasets

Experiments are performed using two synthetic datasets with the source signals being random white Gaussian noise in the first, and speech in the second. Both consist of 4 splits which are used for fold-wise cross-validation. Each split consists of 400 recordings. The noise dataset is created using 3 different 10 seconds long random noise signals, whilst the speech data consists of 5 male and 5 female speech signals, each of 3 second duration. Speech signals are from the TIMIT Acoustic-Phonetic Continuous Speech Corpus [22]. The room impulse responses are simulated using the image source method [23]. The input audio data is synthesized by convolving the recordings of individual sources with multichannel spatial room impulse responses of 3 shoebox rooms with reverberation times of 0.35, 0.6 and 0.85 s for a tetrahedral microphone array of four cardioid microphones, so that both inter-channel time and magnitude differences encode directional information. Both datasets are created by using randomly picked source signals, randomly selected rooms and randomly selected source positions at a fixed distance of 2 m from the array.

Experiments are performed using the Keras [24] and Theano [25] libraries; the code is based upon the baseline system provided by the DCASE2019 task 3 organizers [7].

B. Description of experiments and evaluation measures

The following experiments are performed:

- **Experiment 1:** In this experiment, we compare the localization performance using the DNN with and without the recursive layers. We compare standard 2D CNNs, DSCNNs with a single TF output, complex CNNs, and 3D-4. The better performing structure (with RNNs as will be shown in Sec. 5) is then selected for further analysis.
- **Experiment 2:** In this experiment, we investigate all types of considered convolutions in a common network structure shown in Fig. 1 using the noise-source dataset. In addition to convolutions studied in the first experiment, we add 3D-1 and 3D-2. Furthermore, we verify the performance of DSCNNs for 2 and 4 time-frequency outputs instead of just one.
- **Experiment 3** - In this experiment, we repeat experiment 2 but this time speech is used instead of the noise signals.

Models are evaluated using the DOA error measure [18]. For a signal of length N frames, the DOA error is defined as

$$E_{\text{DOA}} = \frac{1}{\sum_{n=0}^{N-1} D_E^n} \sum_{n=0}^{N-1} \mathcal{H}(DOA_R^n, DOA_E^n), \quad (3)$$

where D_E^n denotes the number of all estimated DOAs in the n-th frame with $n = 0, 1, \ldots, N-1$, $\mathcal{H}(\cdot)$ denotes the Hungarian algorithm which solves the problem of matching the reference DOAs with the estimated DOAs. This is achieved by using a Cartesian distance between the respective direction vectors according to

$$\sigma = \frac{360^\circ}{\pi} \sin^{-1}\left(\frac{\sqrt{(x_E - x_R)^2 + (y_E - y_R)^2 + (z_E - z_R)^2}}{2}\right), \quad (4)$$

where x, y, z denote the coordinates of the DOA vector, and subscripts E and R denote the estimated and reference DOA values. The final results presented in tables to follow are calculated over all folds [26].

Models are trained using Adam optimizer and mean square error (MSE) loss. We train the DNNs with noise signals for 300 epochs, with stopping criterion after 50 epochs of no improvement in both training loss and DOA error. For speech signals, we used 600 epochs with 100 epochs patience to offset the more challenging character of the dataset.

V. Results and evaluation

A. Experiment 1

Table I presents the results of Experiment 1 for 4 example CNN models (one from each convolution type), with and without the use of 2 recurrent layers in the network architecture. As can be clearly observed, an inclusion of bidirectional LSTMs into the network causes a significant increase in localization accuracy. The largest improvement is observed for depthwise separable CNNs with a drop of 9.96° in DOA error; for all methods the relative gain exceeds the value of 40%. Including RNNs results in a significant improvement of the localization performance for all considered methods, and hence the networks studied in Experiments 2 and 3 always contain the recursive layers.

B. Experiments 2 and 3

In this section, we compare the localization accuracy and computational complexity of the investigated models, where complexity is defined as the number of trainable network parameters. Table II presents the results for the noise source for the standard 2D convolutions, complex convolutions, depthwise separable CNNs with 1, 2, 4 time-frequency filters, and three types of 3D convolutions as described in Sec. IIID. The results for the speech dataset for standard, complex, depthwise separable with 2, 4, 8 time-frequency filters and three types of 3D convolutions are depicted in Table III.

The first and most natural approach is to use complex networks that operate directly on complex spectra of the microphone signals. In comparison with standard 2D CNNs, their computational complexity is increased by 55% to process both real and imaginary parts. Unfortunately, complex convolutional CNNs achieve the worst results among all considered models for both speech and noise sources, which may be attributed to a large number of parameters as well as the fact that complex networks are still in an initial development phase, and hence e.g. the non-linear nature of MaxPooling is not tailored to the complex processing.

Apart from complex CNNs, 3D convolutions offer the most flexible and general processing of 3D data since filtering in the channel, time and frequency axis is performed concomitantly. Three-dimensional kernels result in very high model complexity, with 3D-1 model reaching the top value of 857,763 parameters. On the other hand, it also brings about the most accurate localization results among all considered approaches, with DOA error equal to 9.39° and 10.29° for noise and speech signals respectively. Splitting 3D matrices into subchannels effectively lowers the complexity, e.g. 3D-4 has 2 times less parameters than 3D-1, however this comes at a cost of lowering the localization accuracy. In fact, 3D-4 offers comparable accuracy with standard 2D convolutions,

TABLE I
DOA error results for the exemplary 4 CNN model types with and without the RNN layers.

CNN	DOA error [°] without RNNs	DOA error [°] with RNNs
Standard	19.35	**10.10**
Complex	22.44	**15.65**
DS-1	23.10	**13.14**
3D-4	18.75	**10.08**

TABLE II
DOA error results for the noise source.

CNN	Complexity	DOA error [°]
Standard	254,435	**10.10**
Complex	394,475	15.65
DS-1	186,027	13.14
DS-2	195,955	11.27
DS-4	215,811	**10.25**
3D-1	857,763	**9.39**
3D-2	597,347	10.46
3D-4	393,699	**10.08**

TABLE III
DOA error results for the speech signals.

CNN	Complexity	DOA error [°]
Standard	254,435	12.20
Complex	394,475	19.52
DS-2	195,955	**12.08**
DS-4	215,811	**11.92**
DS-8	255,523	12.38
3D-1	857,763	**10.29**
3D-2	597,347	**10.64**
3D-4	393,699	12.49

while it has a higher computational complexity.

Among considered approaches, the lowest complexity can be achieved using depthwise separable convolutions which put more emphasis on the inter-channel point-wise convolutions and reduce the number of parameters by splitting the process into two independent parts. The DS-1 with a single time-frequency filter is characterized by the least parameters, however, the DOA error is higher than for the standard 2D and 3D approaches. The depthwise separable network with 4 time-frequency filters (DS-4) seems to be the optimum choice since it offers complexity reduction in comparison with the standard 2D convolutions, whilst its performance can be considered comparable with the standard model, outperforming slightly the standard 2D CNNs for speech sources. Thus the depthwise separable convolutions seem to be the optimum choice to exploit inter-channel dependencies, which are critical in source localization task, while preserving low network complexity.

VI. Conclusions

In this paper, we present a comparative study of different types of convolutional layers in the context of sound source localization using convolutional recurrent neural networks. We evaluate the performance of the proposed 3D convolutions and depthwise separable convolutions with complex convolutional networks and standard 2D convolutions, in terms of the localization accuracy and model complexity. 3D CNNs are shown to achieve the highest accuracy at a high computational cost, whilst the depthwise separable convolutions offer a good balance between the accuracy and complexity of the model. As this is the first time when DSCNNs are used for the sound source localization task, they are a worthwhile proposition for further research in this field of study.

Acknowledgment

This research was supported by the National Science Centre under grant number DEC-2017/25/B/ST7/01792.

References

[1] M. Wölfel and J. W. McDonough, *Distant speech recognition*. Wiley, 2009.

[2] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*. Wiley, 2012.

[3] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots-building audio-motor maps based on the HRTF," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 1170–1176.

[4] K. Łopatka, J. Kotus, and A. Czyżewski, "Application of vector sensors to acoustic surveillance of a public interior space," *Archives of Acoustics*, vol. 36, no. 4, pp. 851–860, 2011.

[5] S. Aoki and M. Okamoto, "Audio teleconferencing system with sound localization effect," in *Joint Meeting ASA/EAA*, 1999.

[6] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, 2015.

[7] 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, 2018.

[8] C. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, "Sound based localization and identification in industrial environments," in *43rd Annual Conference of the IEEE Industrial Electronics Society (IECON)*, 2017, pp. 6119–6124.

[9] 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.

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

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

[12] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[13] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[14] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 99–103.

[15] N. Roman, D. Wang, and G. J. Brown, "A classification-based cocktail-party processor," in *Advances in neural information processing systems (NIPS)*, 2004, pp. 1425–1432.

[16] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 603–609.

[17] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 136–140.

[18] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466.

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

[20] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," *arXiv:1705.09792*, 2017.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[22] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[23] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Aalto University, 2016, https://github.com/polarch/shoebox-roomsim.

[24] F. Chollet *et al.*, "Keras," GitHub, Web Download, 2015.

[25] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky *et al.*, "Theano: A python framework for fast computation of mathematical expressions," *arXiv:1605.02688*, 2016.

[26] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 49–57, 2010.