Audio Engineering Society

Convention Paper 10101

Presented at the 145th Convention
2018 October 17 – 20, New York, NY, USA

Speech Classification for Acoustic Source Localization and Tracking Applications using Convolutional Neural Networks

Jonathan D. Ziegler[1,2], Andreas Koch[1], and Andreas Schilling[2]

[1] *Stuttgart Media University, Institute for Electronic Media, Stuttgart, Germany*
[2] *Eberhard Karls University Tübingen, Visual Computing, Tübingen, Germany*

Correspondence should be addressed to Jonathan D. Ziegler (`zieglerj@hdm-stuttgart.de`)

ABSTRACT

Acoustic Source Localization and Speaker Tracking are continuously gaining importance in fields such as human computer interaction, hands-free operation of smart home devices and telecommunication. A set-up using a Steered Response Power approach in combination with high-end professional microphone capsules is described, and the initial processing stages for detection angle stabilization are outlined. The resulting localization and tracking can be improved in terms of reactivity and angular stability by introducing a Convolutional Neural Network for signal/noise discrimination tuned to speech detection. Training data augmentation and network architecture are discussed, classification accuracy and the resulting performance boost of the entire system are analyzed.

1 Introduction

For the scenario discussed in this paper, a Steered Response Power (SRP) algorithm, combined with a coincident microphone array is used to track speakers in a conference environment. As SRP is an energy-based detection algorithm, no distinction between a desired signal (i.e. human speech) and undesired interference (i.e. office or traffic noise) can be made. Some improvement in direction of arrival (DOA) estimation can be achieved by applying detection filters[1] prior to the SRP processing, thus only registering energy in a frequency range relevant to human speech. This approach is relatively limited, as many types of noise show a wide frequency range, often overlapping that of speech signals. A more sophisticated sound source discrimination is described using a Convolutional Neural Network (CNN) for sound source classification. Spectral and temporal information is processed by the CNN, using spectrograms of buffers spanning 128 ms and 75 frequency bands, in a frequency range of 200 Hz to 8000 Hz.

2 Methods

2.1 Microphone Array Configuration

The task of Acoustic Source Localization and tracking of a moving acoustical source can be approached in many different ways, the use of linear or circular spaced

[1] The results presented in this paper were obtained using a band-pass detection filter in the range of 200 Hz to 4000 Hz.

arrays being favored in many consumer-grade applications [1]. For audio capturing, the disadvantage of conventional spaced arrays, compared to coincident microphone configurations, is the inferior audio quality of the created beam. Spaced arrays are prone to distorted frequency responses, due to the fact that the created beam patterns are frequency-dependent [2]. Some recent advances have been made, although satisfactory results require an upper frequency limit of 8 kHz [3]. The audio quality of beams created by coincident microphone arrays solely depends on the quality of the microphone capsules used, thus resulting in a more linear frequency response, even with respect to moving beams required for source tracking. However, higher-order beams can not be achieved using first-order coincident arrays [4]. For Machine Listening applications, the requirements regarding audio quality are often relatively low and are defined by the algorithms used. Often a frequency range of 100 Hz to 8000 Hz is chosen. In other cases the bandwidth of telephone conversations (5 Hz to 3700 Hz) is sufficient [5]. Because the array described in this paper is used for audio capturing in conference environments, optimal sound quality is required. Therefore, a configuration consisting of three high-end microphone capsules is chosen. Due to hardware considerations, a Double-M/S configuration is used, consisting of two Schoeps CCM-4 cardioid capsules and a Schoeps CCM-8 figure-of-eight capsule. One cardioid c_f faces 0°, while the other cardioid c_r faces 180° and the figure-of-eight f_8 is positioned facing ±90°.

2.2 Acoustic Source Localization

From the Double-M/S configuration, a horizontal Ambisonics B-format can be decoded [6]:

$$W = c_f + c_r \tag{1}$$

$$X = c_f - c_r \tag{2}$$

$$Y = f_8 \tag{3}$$

Using the WXY-decoded signals, any arbitrary first-order microphone pattern $M(\theta, p)$ can be synthesized on the horizontal plane [7, 8]:

$$M(\theta, p) = pW + (1-p)\left(X\cos\theta + Y\sin\theta\right), \tag{4}$$

with p representing the polar pattern shape between $p = 0$ (figure-of-eight) and $p = 1$ (omnidirectional), and θ describing the orientation on the horizontal plane.

Using (4), n_M virtual cardioid microphone signals[2] can be synthesized. The virtual microphone with the highest relative RMS level indicates the Direction of Arrival of the sound source θ_{DOA}:

$$\theta_{DOA} = \underset{\theta_i}{argmax}\left(\overline{M}(\theta_i, p = 0.5)\right), i = 1, ..., n_M. \tag{5}$$

Under certain conditions reflected sound can surpass the original source in sonic energy. Currently no scenarios have been recorded in which a significant performance decrease could be attributed to false DOA detection due to reflections.

2.3 Confidence Weighting

Building on the SRP maximization described in section 2.2, additional angular stabilization is applied. This is achieved using exponential smoothing [9]:

$$s_t = \alpha x_t + (1-\alpha)\, s_{t-1}, \tag{6}$$

with x_t and s_t representing the input and smoothed output angle for time frame t. 2π-wrapping of the angle is addressed in a separate function.

Using the smoothing factor $\alpha \in [0, 1]$ creates a static smoothing effect which does not reflect any characteristics of the processed signal buffer. To achieve variable smoothing, the coefficient α is dynamically assigned, depending on a set of signal quality metrics. In the following paragraphs, this will be called confidence weighting, which consists of four types of confidence indeces C:

- Directivity weighting C_d – the level of anisotropy of the detected sound indicates whether an actual sound event is detected.
- Level weighting C_l – if a buffer contains a low relative sound level, no relevant sound events are expected.
- Long-term weighting C_{lt} – if sound events have frequently been detected from a direction, a quasi-static sound source such as a speaker at a table can be assumed.

[2] $p = 0.5$

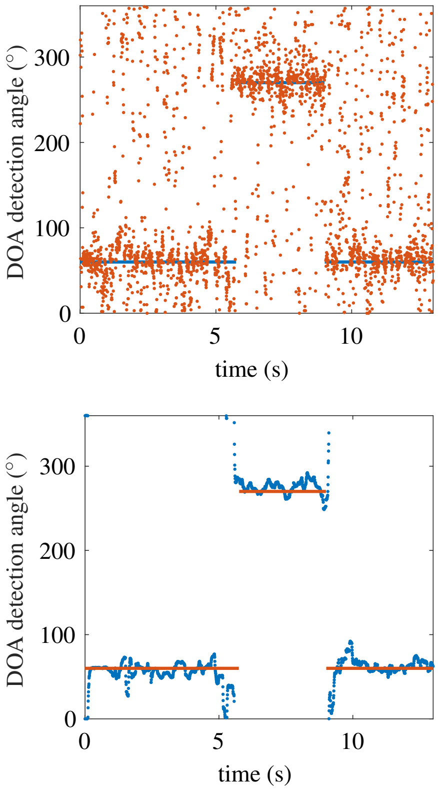

Fig. 1: Comparison of tracker output with and without angular stabilization. The average error can be reduced from 25.68 % to 6.46 % when using a dynamic smoothing coefficient α. Solid lines represent reference position.

- Speech detection C_s– if a buffer is not classified as speech, no relevant sound events are expected.

The last contribution to the confidence index is determined using a Convolutional Neural Network (CNN), trained to discriminate between speech and non-speech.

The confidence indeces are combined to create the dynamic weighting factor α:

$$\alpha = \left(\kappa C_d + (1-\kappa) C_d C_{lt} C_s^2\right) C_l \,, \qquad (7)$$

using the empirically determined mixing factor κ.

Angular stabilization is essential for this type of acoustic source localization. Figure 1 shows a comparison of the tracker output with and without stabilization.

2.4 Mel-Scale Spectral Analysis

Convolutional Neural Networks provide excellent processing capabilities on two-dimensional arrays, such as images. For training and classification, the audio stream is processed via Fourier Transform to Log-Mel-scale spectrograms, using the feature extraction toolbox provided by the University of Oldenburg [10]. The Mel scale is used, since it closely resembles human perception of sound and has proven effective in combination with Neural Networks for audio classification and speech detection [11, 12]. Buffers of 2048 samples are analyzed at $Fs = 16$ kHz sampling rate[3], resulting in 128 ms of audio per buffer. The spectral transform is performed using a window size of 28 ms, which is successively shifted by 10 ms. The processed frequency range is between 200 Hz and 8 kHz, divided into 75 Mel-bands. Examples of extracted spectrograms can be seen in Figure 3.

2.5 Neural Network Architecture

As the entire signal processing chain was created in MATLAB, the use of MATLAB's Neural Network Toolbox for the speech detector ensures a seamless integration and easy fine-tuning of the processing.

The processing steps presented in section 2.3 and 2.4 output Log-Mel-scale spectrograms of the dimension 75x11x1. These define the dimensions of the input layer of the CNN. Two-dimensional convolution is applied, using 8 5x5 filter matrices and zero-padding to maintain the input layer dimension ("same" padding) [13, 14]. The convolution output is then shrunk by choosing the maximum value of every 2x2 subset. This operation is called Max-pooling with a pool size of 2 and a stride of 2, and is used to transform the matrix to a dimension of 38x6x8[4]. The next convolution operation uses 16 3x3 filters and *same* padding. Combined with a max-pooling operation with a pooling size and stride of 2, the dimensions are transformed to 19x3x16. The last convolution uses 32 3x3 filters and *same* padding, resulting in 1824 inputs for the first fully connected layer, which outputs

[3] The entire tracking algorithm runs at 48 kHz. The decision to down-sample by a factor of 3 is the result of the data set used for augmentation, as described in section 2.6, and the chosen frequency range with an upper limit of 8 kHz.

[4] To achieve the desired output dimensions, max-pooling is padded with $p_{bottom} = 1$ and $p_{right} = 1$. Details are discussed in section 2.7.

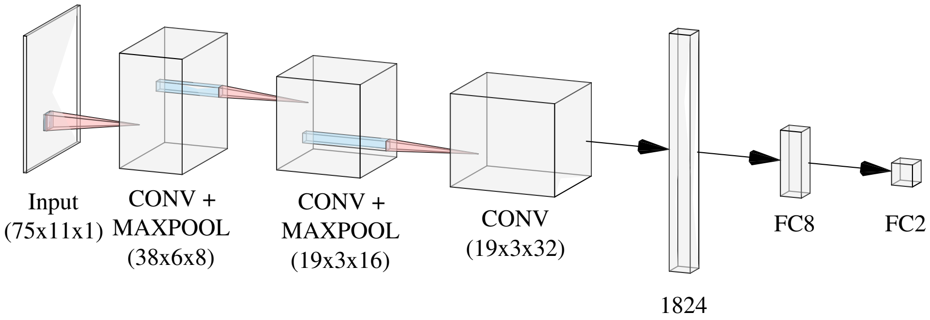

Fig. 2: Architecture of the neural network used as speech detector. The input Mel spectrogram measures 75x11x1 pixels. Using three convolution layers, two max-pool layers and two fully connected layers, validation accuracy is 91.23 %.

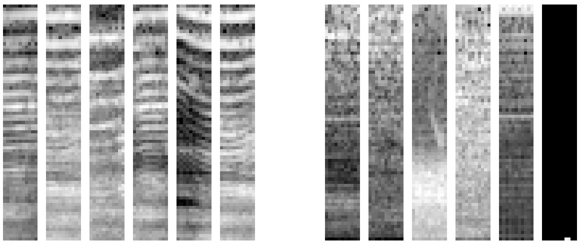

Fig. 3: Log-Mel-scale spectra created for training and classification using a CNN. 128 ms of audio are processed using 75 Mel-bands, spanning 200 Hz to 8000 Hz. The resulting spectrograms are displayed as gray-scale images of 75x11 pixels. *Left:* spectrograms of audio buffers containing speech. *Right:* spectrograms of buffers without speech.

8 activations for the final layer. This layer, using a softmax activation function, discriminates between *speech* and *non-speech*. An Adam optimization algorithm was used to train the network [15].

2.6 Training Data and Augmentation

The network was trained using 30866 speech spectra and 29360 noise spectra. The validation set consisted of 2×2822 labeled samples. Lacking sufficient training data, the dataset was augmented using the Musan dataset [16]. Within the dataset, the Librivox speech files and the Free-Sound noise samples were used. To create training data more similar to the test data, the relatively direct recordings of the dataset needed to be placed in virtual rooms. Room impulse responses (RIR) were created using the image method described by Allen and Berkley [17], implemented in the RIR-generator, provided by the International Audio Laboratories Erlangen [18]. To prevent the Neural Network from overfitting to a specific room dimension, random room dimensions were chosen to create impulse responses of virtual rooms similar in size to a potential application environment. For every audio file of the dataset, room dimensions were varied from 2 m to 7 m. Within these randomly chosen room dimensions, the sound-source and sound-detector were randomly positioned. Once the RIR was created, a convolution with the audio file from the dataset created a reverberant version of the file. This reverberant audio was then divided into frames of 2048 samples and transformed into the Log-Mel-spectrograms described in section 2.4. The validation set consisted of 2822 spectrograms for *speech* and *non-speech*, respectively. The spectrograms labeled *speech* in the validation set were obtained from recordings of the virtual conference described in section 3, using speech-only scenarios. To obtain the maximum possible number of spectrograms from the recordings, all individual microphone streams, as well as the combined omnidirectional and virtual cardioid signals, were analyzed individually and used for training and cross validation.

2.7 Real-Time Classification

The spectral analysis described in section 2.4 requires 128 ms of audio per spectrogram. With the main tracking algorithm running at 48 kHz, this is equivalent to 6144 samples. To maintain the low-latency operation of the tracking algorithm, which runs at 256 samples, classification is performed on the current audio buffer, in combination with the 23 previous buffers. To give the

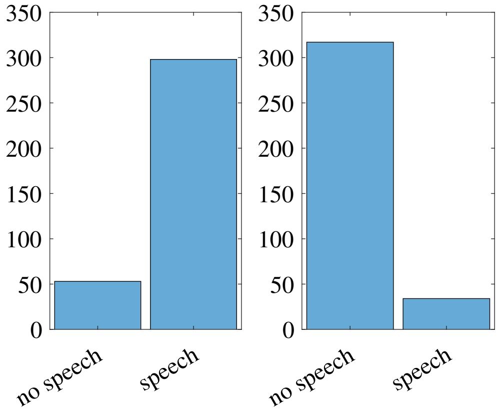

Fig. 4: Results of the decomposed test files. *Left:* Only analyzing the clean speech file results in 298 *speech* classifications and 53 *non-speech* classifications. *Right*: Analyzing the noise components returns 317 *non-speech* classifications and 34 *speech* classifications. The overall accuracy in this test case is 87.61 %.

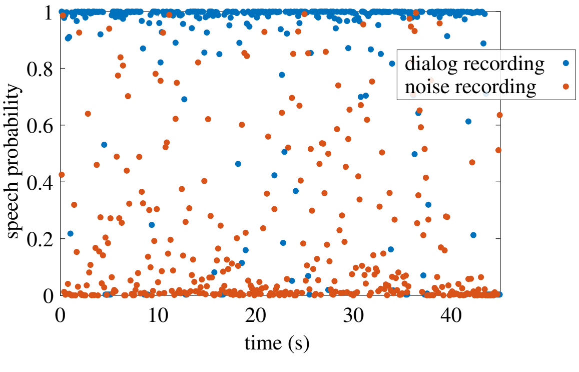

Fig. 5: Probability of a buffer containing speech, analyzed over the entire test file. As in Figure 4, the file was decomposed in *speech* and *non-speech* components which were analyzed individually.

current buffer a higher weight, the first max-pooling layer is padded only on the right, thus reducing the importance of the left-most (oldest) part of the spectrogram. The choice of audio stream for the classification operation is still under investigation. The most promising choices are a virtual omnidirectional microphone

$$S_o = W \tag{8}$$

and a virtual supercardioid microphone signal S_p ($p = 0.34$), aimed at the detected direction of the previous buffer θ_{DOA}:

$$S_p = 0.34 \cdot W + 0.66 \cdot (X \cos\theta_{DOA} + Y \sin\theta_{DOA})\,. \tag{9}$$

The results presented in section 3 were gathered using S_o. Current measurements show no performance gain[5] when using the computationally more expensive S_p.

3 Results

The tracker performance was evaluated using a multichannel playback system, reproducing a virtual conference scenario. The set-up was placed within a large, acoustically untreated room[6] with 8 loudspeakers arranged in two concentric rings of 1.5 m and 2.5 m around the microphone array. One additional loudspeaker was placed at 0.5 m distance from the array, another at 4 m. Microphone and loudspeaker hight were chosen to realistically match a real-world scenario. While the microphone and close-range loudspeaker were placed at the hight of a table-top (800 mm and 700 mm, respectively), the loudspeakers placed at 1.5 m and 2.5 m distance were set to the height of the mouth of a seated person (1240 mm and 1390 mm, respectively). The distant loudspeaker was placed at the approximate height of the mouth of a standing speaker (1700 mm). All heights were measured from the center of the tweeter. A prepared scenario was played back[7], consisting of male and female speech in German and English. Additionally, a variety of non-speech signals were played back, such as cell phone ring-tones, moving chairs, et cetera, combined with recordings of construction sites and office noise. The recording and playback format of the multichannel noise recordings were chosen to be identical. Two scenarios were played back, once exclusively using speech signals, once containing additional noise. The introduction of speech

[5] Measured performance gain was < 0.1 %.

[6] 8.3 m × 8.2 m × 3.8 m, $RT60 \approx 2.3$ s.

[7] Recording and playback format: 48 kHz, 24 Bit.

	$S1$	$S1_{noise}$	$S2$	$S2_{noise}$
accuracy gain	$-1.2\,\%$	$-1.8\,\%$	$-0.4\,\%$	$-6.6\,\%$
stability gain	$1.2\,\%$	$5.4\,\%$	$2.4\,\%$	$4.2\,\%$

Table 1: Measured performance gain when using speech classification as part of the angular stabilization process.

detection decreased the average accuracy by 2.5 % and increased the angular stability by 3.3 %. Table 1 shows increased smoothing especially in noisy environments. To further evaluate the classifier performance, the multichannel scenario was split into speech and non-speech components and rendered to mono-files. The classifications throughout the speech and non-speech files can be seen in Figures 4 and 5. The sum test accuracy in this case is 84.61 %. Because the split was performed on the near-anechoic scenario without being played back in the virtual conference environment, the comparison is skewed, with both real-time application and training being performed on reverberant signals. Additional testing is described in section 4. Within the tracking scenario, the addition of the CNN classifier results in a performance boost. Increased angular stability during speech, combined with less erratic movement in periods without speech, improve the audio quality of beamforming algorithms being driven by the tracked position data. A beamformer tuned to the signals of the array in use is described by Runow et al. [19]. Figure 6 shows the classifier-induced performance boost for a virtual conference recording. Close sources with a high signal to noise ratio can be tracked well without the need of speech classification. During the second half of the recording, the sources are played back on the mid-range[8] and far-range[9] loudspeakers, with a larger amount of ambient noise. Here, the discrimination between *speech* and *non-speech* (desired signal and noise) increases tracking stability. Since this test was designed for general tracking performance evaluation, and not for speech classification evaluation, additional testing will be required to better assess the added confidence factor. In real-time tests, a clear improvement of speaker tracking can be observed, with office and traffic noise, as well as structural vibration being rejected well beyond the level achieved when using only the detection filter described in section 1.

4 Discussion

Because the available test data were not recorded for the specific purpose of evaluating speech detection, additional testing is needed to assess the full benefit of the added confidence weighting. Initial real-time tests indicate a considerable performance boost; quantitative measurements are the next step. With the small amount of training data requiring additional synthetic data, the training and validation sets do not come from the same data distribution. This is not ideal, but could not be prevented without recording and labeling large amounts of additional data. To ensure satisfactory generalization of the trained net within the intended application, most of the recorded data was used for cross validation. Initial training of the CNN indicated overfitting, which has been countered with the use of stronger L2-regularization [20]. This suggests that test accuracy will profit from additional training data recorded in environments more similar to the final application. The test environment used for evaluation was considerably larger than the virtual rooms used for data augmentation, which were chosen to be closer to the final application environment. A performance gain is expected when using more realistic surroundings for further testing. If the desired increase in performance is not observed, additional training rounds will contain a larger variety of virtual spaces.

5 Summary

A system for Acoustic Source Localization and Tracking is described, which is capable of locating and tracking speech sources in real-time. The main system is set up using an algorithmic approach, with Steered Response Power maximization as the direction-of-arrival estimator and a series of weighting factors for variable exponential smoothing of the detected angle. Additionally, a Convolutional Neural Network is used for speech detection. Discrimination between *speech* and *non-speech* events enables the system to effectively reject sound sources which are not of relevance for the the application of speaker tracking, increasing the performance beyond that of the purely algorithmic approach. Initial tests show high classification accuracy within the final application, and additional data promise still higher accuracy.

[8] $r = 2.5\,\mathrm{m}$

[9] $r = 4\,\mathrm{m}$

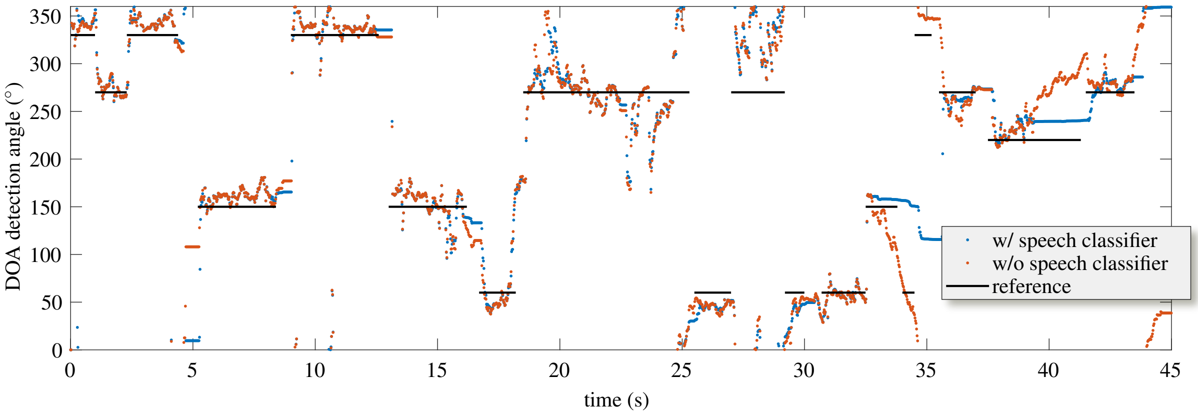

Fig. 6: Output of tracking algorithm with and without CNN speech detection. The first half of the test file is played back at a distance of $r = 1.5\,\mathrm{m}$ around the microphone array. At this distance, confidence weighting works well and the speech detection has a negligible impact on performance. Towards the end of the sample, playback distance is increased to a distance of 2.5 m to 4 m around the microphone array and the SNR is reduced. The increased levels of non-directional reverberation and noise components considerably reduce the tracker's performance. Using the speech detector, a higher level of stability can be maintained.

Acknowledgments

This research was in part funded by the *Zentrales Innovationsprogramm Mittelstand*, a grant of the *Bundesministerium für Wirtschaft und Energie*, Germany.

This work was supported by *Kooperatives Promotionskolleg Digital Media* at Stuttgart Media University and the University of Tübingen.

References

[1] Reinette, A., Cornejo, M., Rouchon, C., and Fester, M., "Benchmarking Microphone Arrays: ReSpeaker, Conexant, MicroSemi AcuEdge, Matrix Creator, MiniDSP, PlayStation Eye," *Snips Labs*, 2017.

[2] Benesty, J., Jingdong, C., and Huang, Y., *Microphone Array Signal Processing*, Springer Berlin Heidelberg, 2008, ISBN 978-3-540-78612-2.

[3] Delikaris-Manias, S., Valagiannopoulos, C. A., and Pulkki, V., "Optimal directional pattern design utilizing arbitrary microphone arrays: A continuous-wave approach," in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.

[4] Benesty, J. and Jingdong, C., *Study and Design of Differential Microphone Arrays (Springer Topics in Signal Processing)*, Springer, 2012, ISBN 364233752X.

[5] Gruhn, R. E., Minker, W., and Nakamura, S., *Automatic Speech Recognition*, pp. 5–17, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, ISBN 978-3-642-19586-0, doi:10.1007/978-3-642-19586-0_2.

[6] Wittek, H., Haut, C., and Keinath, D., "Double M/S – a Surround recording technique put to test," in *Tonmeistertagung*, Verband Deutscher Tonmeister eV, 2006.

[7] Gerzon, M. A., "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," in *Audio Engineering Society Convention 50*, 1975.

[8] Freiberger, K., *Development and Evaluation of Source Localization Algorithms for Coincident Microphone Arrays*, diploma thesis, 2010.

[9] Brown, R. G., *Smoothing, forecasting and prediction of discrete time series*, Prentice-Hall Englewood Cliffs, N.J, 1963.

[10] Schädler, M. R., "Reference Matlab/Octave implementations of feature extraction algorithms," 2015, Carl von Ossietzky Universität Oldenburg, Department für Medizinische Physik und Akustik.

[11] Piczak, K. J., "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2015, ISSN 1551-2541, doi:10.1109/MLSP.2015.7324337.

[12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, 29(6), pp. 82–97, 2012, ISSN 1053-5888, doi:10.1109/MSP.2012.2205597.

[13] Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.

[14] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 86(11), pp. 2278–2324, 1998.

[15] Kingma, D. P. and Ba, J., "Adam: A Method for Stochastic Optimization," *CoRR*, abs/1412.6980, 2014.

[16] Snyder, D., Chen, G., and Povey, D., "MUSAN: A Music, Speech, and Noise Corpus," *CoRR*, abs/1510.08484, 2015.

[17] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979, doi:10.1121/1.382599.

[18] Habets, E., "RIR Generator," 2018, International Audio Laboratories Erlangen.

[19] Runow, B., Schilling, A., and Curdt, O., "Störgeräuschreduktion mit einer Mel-Filterbank in Verbindung mit koinzidenten Mikrofonarrays," *29. Tonmeistertagung des Verbandes Deutscher Tonmeister*, 2016.

[20] Schmidhuber, J., "Deep Learning in Neural Networks: An Overview," *CoRR*, abs/1404.7828, 2014.