ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/348733585

Feature Overview for Joint Modeling of Sound Event Detection and Localization Using a Microphone Array

Conference Paper · January 2021

DOI: 10.23919/Eusipco47968.2020.9287374

CITATIONS
17

READS
138

3 authors, including:

Daniel Aleksander Krause
Tampere University
19 PUBLICATIONS **107** CITATIONS

SEE PROFILE

Archontis Politis
Tampere University
135 PUBLICATIONS **2,761** CITATIONS

SEE PROFILE

All content following this page was uploaded by Daniel Aleksander Krause on 02 April 2021.

The user has requested enhancement of the downloaded file.

Feature Overview for Joint Modeling of Sound Event Detection and Localization Using a Microphone Array

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*— In this paper, we present a comparative study of a number of features and time-frequency signal representations for the task of joint sound event detection and localization using a state-of-the-art model based on a convolutional recurrent neural network. Experiments are performed for a dataset consisting of the recordings made using a tetrahedral microphone array. Several feature inputs specific to the task of sound event detection and sound source localization are combined and subsequently tested, with the aim to achieve joint performance of both tasks for multiple overlapping sound events using a single model based on a deep neural network architecture. Apart from providing a comprehensive comparison of various state-of-the-art acoustic features such as generalized cross-correlation, and inter-channel level and phase differences, we propose new features that have not been used for this task before such as eigenvectors of the microphone covariance matrix or sines and cosines of phase differences between the channels. Results for all combinations of input features are analyzed and discussed, followed by conclusions.**

Index terms — sound event detection, sound source localization, convolutional recurrent neural networks, feature extraction

I. INTRODUCTION

Sound event detection (SED) is one of the hot topics in current audio research. Automatization of the process of detecting and classifying signals present in an acoustic environment is of interest to numerous applications such as audio-driven robots [1], surveillance systems detecting hazardous audio events [2], and multimedia annotation [3]. Therefore combining the tasks of sound event localization and detection (SELD) is a natural step forward in the research on machine learning for audio applications.

Although much effort has been put into both tasks, sound event detection and sound source localization (SSL) are typically treated as separate problems. Currently, most of the SED systems are built upon deep neural networks (DNNs) [4-11]. Recurrent neural networks (RNNs) are used to utilize temporal relations between frames [5-6]. Convolutional neural networks (CNNs) enable efficient data filtering and statistical information retrieval from more general signal representations [7-8]. Recently, both model types have been combined to form convolutional recurrent neural networks (CRNNs) [9-11]. Such systems are typically trained based on amplitude-based feature representations such as Mel-band energies [12-14], constant-Q transform (CQT) features [15], or spectrograms obtained using short-time Fourier transform (STFT). Further improvement in detection performance can also be achieved by an additional exploitation of the phase information [16-17]. For the SSL task, the generalized cross correlation (GCC) with phase transform (PHAT) is a popular method [18-20]. Localization can also be performed using DNN-based approaches formulated as a classification or a regression problem. Classification-oriented SSL methods assign the signals to a discrete set of angles [21], while the regression methods enable continuous direction of arrival (DOA) estimation [22]. Joint SELD using DNNs has recently become an emerging field of audio research [21-24].

In this study, we compare several feature extraction algorithms used in sound event detection or sound source localization. In order to search for optimal combinations for the joint task, we perform cross-merging between input types from both tasks and investigate the SELD task performance. Investigated features include magnitude and phase spectrograms, phase differences, logmel spectrograms and GCC-PHAT. In addition, we propose solutions that have not been yet applied in the context of joint detection and localization, namely the inter-channel level differences, and sines and cosines of phase differences. We also investigate three methods of exploiting information from the inter-channel covariance matrix by employing it directly or through the eigenvalue decomposition. To avoid excessive variables in the presented analysis, we use a single model for all feature types. For that purpose, the so-called SELDnet architecture is used, which is a current state-of-the-art neural network in this field [22]. This network was proven to outperform classical SED and SSL solutions used separately, therefor in our study we focus on the joint task only. Experiments are performed for the TAU Spatial Sound Events 2019 - Microphone Array dataset, which consists of recordings from a tetrahedral microphone array.

II. MODEL USED FOR THE SELD TASK

In this section, we present the architecture of the existing SELDnet network, which has been proposed in [22]. Due to the reduced number of variables in the model and providing good reproducible results, this network architecture is selected in this study for comparing different types of input features. The architecture is depicted in Figure 1.

The model takes an HxTxF matrix as input, with H, T and F denoting the number of feature channels, temporal length of the modeled sequence and the feature dimension, respectively. The first part of the model is composed of three convolutional layers, each consisting of P 3x3 filters. After each convolutional part, ReLU activation outputs are normalized using batch normalization. Dimensionality in the frequency domain is reduced to 4 using Max-pooling with values MPi which depend on the input feature dimension size. For magnitude spectrograms they are equal to MPi=[8, 8, 4], whereas for logmel spectrograms MP_i=[4, 3, 2]. The convolutional layers are followed by two bidirectional GRUs, each consisting of Q units and a tanh activation function. The network is then split into two branches, of which one is performing SED classification, whilst the other is performing SSL. Both branches consist of two fully connected layers, the first of which consists of G linear neurons. The SED branch ends with C neurons and a sigmoid activation function to obtain the probability values, with C denoting the number of sound events classes. Localization is performed as a regression task, using the elevation and azimuth angles as the DOA output. For this reason, the last SSL layer consists of 2C linear neurons to be able to output the pair of DOAs per class. In this work, we use the following parameter values: T=128, P=64, Q=128, R=128, C=11. H is dependent on the feature combination used for the model input.

III. Features under investigation

We split the investigated features into three groups, depending on the kind of information they relay. In the following paragraphs we describe the features based on the magnitude, phase and the covariance matrix, which cover both types of information.

A. Amplitude based features

In this study we use three types of purely magnitude-based signal representations.

1) Magnitude spectogram: As the most general, we use the magnitude spectrogram in the STFT domain. The matrix with time-frequency representation is extracted using a 20ms long Hamming window with the DFT length of 2048, resulting in feature dimension of F=1024.

2) Logmel spectrogram: As the second feature, a logmel spectrogram is used due to its popularity of compressing the energetic information from the entire spectrum. It is obtained by filtering the spectrum of each frame, using a triangular Mel filter bank, followed by a logarithm operation [25]

$$X_{mel}[n,m] = \log\left(\sum_{k=0}^{K-1} |X[n,k]| H_{mel,m}[k]\right), \quad (1)$$

where $X[n,k]$ denotes the matrix with complex coefficients in the STFT domain, $H_{mel,m}[k]$ is the m-th Mel filter, n and k denote the time and frequency indices, respectively. In this study we use Mel filters for F=96 subbands.

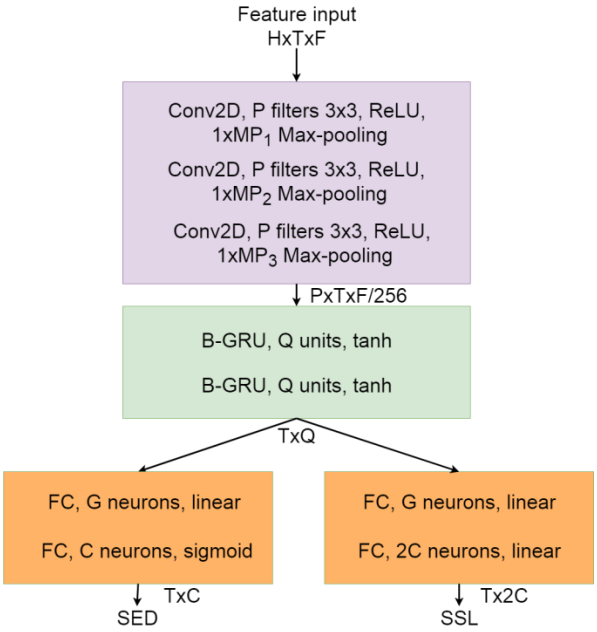

Fig. 1. SELDnet architecture.

3) Inter-channel level difference (ILD): Finally, we use inter-channel level differences as a set of features that might provide useful information for both detection and localization tasks. This feature can be computed for each channel pair as the difference between the magnitude values in dB or directly from the STFT representation as

$$ILD_{i,j}[n,k] = \frac{|X_i[n,k]|}{|X_j[n,k]|}, \quad (2)$$

with i and j denoting the i-th and j-th channel. Since the level differences can flatten source-related magnitude patterns, we add a single magnitude spectrogram of the first channel to the input. This results altogether in 7 channels of features, contrary to the other two magnitude-based features that consist of just 4 channels.

B. Phase based features

In this section, we describe three phase-based features.

1) Phase spectrogram: As the first and most basic phase-based feature, we extract the phase information directly from the STFT representation, which results in 4 feature channels.

2) Inter-channel phase difference (IPD): In order to help the model to recognize the inter-channel time dependencies, we test two types of input features derived from the phase information present in the STFT time-frequency bins. The first one is using the inter-channel phase difference (IPD) between each (i, j) channel pair

$$IPD_{i,j}[n,k] = \arg(X_i[n,k]) - \arg(X_j[n,k]), \quad (3)$$

which results in 6 feature channels.

3) Sine & cosine of phase differences: Alternatively, we use a feature presented in [26], which takes the sine and cosine values of the phase differences

$$sinIPD_{i,j}[n,k] = \sin\big(IPD_{i,j}[n,k]\big), \quad (4)$$

$$cosIPD_{i,j}[n,k] = \cos\big(IPD_{i,j}[n,k]\big). \quad (5)$$

The purpose of this operation is to provide a smoother representation of the highly varying phase values and to emphasize those frequency bands which show the most notable inter-channel phase differences. This feature has been successfully applied in speaker separation [26], therefore utilizing it for DOA estimation may well be expected to yield promising results.

4) Generalized cross correlation (GCC): Apart from the phase-based features in the STFT domain, we study another method that is based on the complex spectrum, namely the GCC-PHAT method [19], which is given by

$$GCC_{i,j}[n,l] = F^{-1}\left\{\frac{X_i[n,k]\cdot X_j^*[n,k]}{\big|X_i[n,k]\big|\big|X_j[n,k]\big|}\right\}, \quad (6)$$

where $F^{-1}\{.\}$ denotes the inverse Fourier transform and index l denotes the time-lag between the channels. This technique is based on cross-power spectra between the respective microphone channels with whitening applied in order to level off the magnitude information [19].

All four considered phase-based features are tested in combination with all three aforementioned magnitude features. To ensure the same feature dimension size, the majority of phase-based features are filtered with the Mel filterbank when combined with the logmel spectrogram. In addition, the matrix with GCC-PHAT results is cut to the appropriate number of delay time-lags. For the particular array used in experiments, selecting 96 delay time-lags is enough to cover the maximum inter-channel delay, which results in the same feature dimensions as the number of 96 Mel frequency bands.

C. Covariance matrix based features

Finally, we propose to use the covariance matrix of the microphone signals or its principle eigenvectors as input features for both detection and localization, as the covariance incorporates magnitude and phase information of the microphone signals as well as inter-channel information.

1) Covariance matrix: For each complex time-frequency bin, we compute the covariance matrix, averaged over 5 adjacent time frames. Since the covariance matrix is Hermitian, we use diagonal values and real and imaginary parts of half of its coefficients as features. The outcome is a total of 16 channels.

2) Principle eigenvector: Alternatively, we present two methods of decomposing the information derived from the covariance matrix to make it easier to interpret for the neural network. The first approach is to use the real and imaginary part of the principal eigenvector, compressing the information to 8 feature channels.

3) *Two eigenvectors*: Secondly, we use the first and second eigenvectors and multiply them by the corresponding eigenvalues, which is expected to help the network distinguish between the time frames with 1 or 2 active acoustic events. Similarly to the basic method, the number of feature channels for the two eigenvectors remains the same. i.e. it amounts to 16. This feature is expected to provide better results for two sources than a full covariance matrix.

IV. EXPERIMENTS AND EVALUATION

A. Dataset and system setup

Experiments are performed using the TAU Spatial Sound Events 2019 - Microphone Array provided by the DCASE2019 challenge organizers. The dataset consists of 5 splits, four of which are used for cross-validation and one for the evaluation. Each split contains 100 one-minute recordings sampled at 48 kHz. The data is synthesized using recordings of individual audio events and multichannel spatial room impulse responses of 5 different environments. Overall it consists of 11 classes, namely: clearing throat, coughing, door knock, door slam, drawer, human laughter, keyboard, keys, page turning, phone ringing and speech. Half of the sound events are temporally overlapping with the maximum number of simultaneous events set to two. Recordings include added ambient noise so that the signal-to-noise ratio is equal to 30 dB [27].

Experiments are performed using the Keras [28] and Theano [29] libraries; the code is based upon the baseline system provided by the DCASE2019 organizers [22].

B. Evaluation measures

Systems are evaluated using separate metrics for SED and DOA estimation. We measure the SED performance using F1-score and Error Rate (ER) computed in one-second segments [30], while DOA error and frame recall evaluate the localization performance [31]. For a recording of length N frames, the DOA error is defined as

$$DOA\ error = \frac{1}{\sum_{n=0}^{N-1} D_E^n}\sum_{n=0}^{N-1} H(DOA_R^n, DOA_E^n), \quad (7)$$

where D_E^n is the number of all estimated DOAs in the n-th frame, $H(.)$ denotes the Hungarian algorithm, which solves the problem of matching the reference DOAs with the estimated ones. This is done using a spherical distance σ between the respective angles:

$$\sigma = cos^{-1}(sin\,\theta_E\,sin\,\theta_R + cos\,\theta_E\,cos\,\theta_R\,cos(|\phi_R - \phi_E|) \quad (8)$$

where θ and ϕ denote the azimuth and elevation angles, and subscripts E and R denote the estimated and reference DOAs. Frame recall is used as a measure of how often the system produces the correct number of DOAs, counting the number of frames where the number of estimated and the number of reference DOAs are equal. It is calculated as

$$frame\ recall = \sum_{n=0}^{N-1}\frac{\mathbb{1}(D_R^n = D_E^n)}{N}, \quad (9)$$

where $\mathbb{1}(.)$ denotes an indicator function returning one if the condition is met and zero otherwise. Finally, the overall SELD score is obtained using the following formula

$$SELD = \frac{[(1-F_1) + ER + \frac{DOA\ error}{180^\circ} + (1 - Recall)]}{4}. \quad (10)$$

Metrics are calculated over the statistics from all folds, as it yields the most reliable outcome [32]. Models are trained with the Adam optimizer for 300 epochs. Training is stopped after 50 epochs of no improvement in both SELD score and training loss. The SED output uses the binary cross-entropy (BCE) loss function, whilst SSL uses mean squared error (MSE) as DOA estimation is treated as a regression problem.

C. Results and discussion

Since the results for both cross-validation set and evaluation set turn out to be consistent, for clarity we show the results for the evaluation part only. The study outcome is depicted in Table 1.

As can be observed, there is no clear best performing method amongst the magnitude-based features for the detection part, and their performance depends strongly on the phase features with which they are combined. However, both magnitude (*mag*) and logmel (*mel*) spectrograms can be used effectively for SELD, while the ILD features seem to consistently provide slightly worse performance.

Comparing the localization performance expressed in terms of both DOA error and frame recall, IPD feature outperforms pure phase spectrogram in combination with the majority of magnitude-based features. A lower frame recall when using mel bands is a minor exception. However, taking the sine and cosine of IPD (*sin&cos*) improves the results even more for all tested configurations, resulting in 4°, 7.5° and 5.2° DOA error decrease for the magnitude spectrograms, logmel spectrograms, and ILDs, respectively. The lowest DOA error amongst all tested feature combinations is obtained for the magnitude and sines and cosines feature set. We associate that with the smoothing qualities of the sine and cosine functions, which puts additional emphasis on inter-channel phase differences, making them easier for the network to recognize. It is worth noting that this feature set has not been so far used in the literature for the SELD task.

Another feature which significantly improves the localization results in comparison with the phase spectrogram is the GCC-PHAT (*GCC*). Although the results for GCC are noticeably worse than for sin&cos when combined with magnitude or ILD, we observe that GCC provides a very good localization accuracy amongst the Mel-based configurations. That might be due to cutting the GCC spectrogram to 96 delayed samples, feeding the network only with its most pivotal part. Interestingly, when using GCC features a consistent improvement in detection performance is also observed, leading to the best detection results among all magnitude-based representations. When using GCC in conjunction with the logmel spectrogram, the best overall SELD score equal to 0.143 is obtained.

As can be observed, higher level features such as the covariance matrix yield significantly worse results than the low level features. The basic covariance matrix method (*cov. matrix*) exhibits the worst overall performance, which might be caused by the high number of feature channels, which in turn causes difficulty for the model to extract important information. Extracting the principal eigenvector (*principal eigvec*) as a feature vector leads to a very large improvement in both detection and localization compared with the covariance matrix method, which means that efficient data compression from the covariance matrix can lead to obtaining acceptable results. However, we observe worse results when using two eigenvectors (*two eigvecs*), which suggests that the applied model is not capable of extracting well the information from 16 channels. The overall SELD score of the principal eigenvector method is equal to 0.2, which however is still worse than most of the magnitude and phase feature combinations. On the other hand, it provides one of the lowest DOA errors among all tested methods.

TABLE I.
EVALUATION RESULTS FOR ALL FEATURE INTPUT COMBINATIONS.

Input features	ER	F_1 score [%]	DOA error [°]	Frame recall [%]	SELD
mag+phase	0.278	84.771	35.161	84.849	0.194
mag+IPD	0.297	84.492	23.482	86.514	0.179
mag+sin&cos	0.286	84.868	**19.561**	**87.838**	0.167
mag+GCC	**0.247**	**86.422**	26.029	86.973	**0.164**
mel+phase	0.277	84.578	29.931	84.370	0.189
mel+IPD	0.373	79.794	29.878	80.532	0.234
mel+sin&cos	0.286	83.219	22.293	86.506	0.178
mel+GCC	**0.226**	**87.738**	**20.420**	**89.000**	**0.143**
ILD+phase	0.280	84.090	31.279	84.393	0.192
ILD+IPD	0.278	84.746	27.897	85.600	0.182
ILD+sin&cos	0.303	83.308	**22.668**	**86.101**	0.184
ILD+GCC	**0.262**	**85.065**	25.537	85.173	**0.175**
cov. matrix	0.756	60.797	36.263	65.804	0.423
principal eigvec	**0.335**	**81.558**	**23.146**	**84.950**	**0.200**
two eigvecs	0.469	74.354	36.309	81.033	0.279

V. CONCLUSIONS

In this paper, we present a wide overview of input features that can be used to train a neural network with the aim to perform multi-label overlapping sound event detection and localization. In performed experiments, we focus on features that are applicable to processing the multichannel microphone array signals using the state-of-the-art CRNN model. We present several combinations of the existing and novel features in the context of the SELD task, and compare their performance in terms of localization, detection, and a joint task. Furthermore, we propose to exploit the spatio-spectral information inherently present in the microphone covariance matrix and its principle eigenvectors. The presented results indicate that the combination of the selected low-level features enables to obtain more accurate results in the SELD task than using a single high-level feature, e.g. the principle eigenvector, when the same low-complexity DNN architecture is applied.

ACKNOWLEDGMENT

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

REFERENCES

[1] S. Chu, S. Narayanan, C. C. J. Kuo, and M. J. Mataric, „Where am I? Scene Recognition for Mobile Robots Using Audio Features", *IEEE International Conference Multimedia and Expo (ICME),* 2006, pp. 885-888.

[2] V. Carletti, P. Foggia, G. Percanella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio Surveillance Using a Bag of Aural Words Classifier", *IEEE International Conference on Advanced Video and Signal Based Surveillance,* 2013, pp. 81-86.

[3] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio Key-words Generation for Sports Video Analysis", *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM),* vol. 4, no. 2, 2008, no. 11.

[4] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic Sound Event Detection Using Multi Label Deep Neural Networks," *2015 International Joint Conference on Neural Networks (IJCNN).* IEEE, 2015, pp. 1–7.

[5] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola and T. Virtanen, "Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 6-10.

[6] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux and K. Takeda, "Bidirectional LSTM-HMM Hybrid System for Polyphonic Sound Event Detection", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016*, 2016, pp. 35-39.

[7] I. Jeong, S. Lee, Y. Han and K. Lee, "Audio Event Detection Using Multiple-input Convolutional Neural Network", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017),* 2017, pp. 51-54.

[8] D. Lee, S. Lee, Y. Han and K. Lee, "Ensemble of Convolutional Neural Networks for Weakly-supervised Sound Event Detection Using Multiple Scale Input", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017),* 2017, pp. 74-79.

[9] S. Adavanne and T. Virtanen, "Sound Event Detection Using Weakly Labeled Dataset with Stacked Convolutional and Recurrent Neural Network", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017),* 2017, pp. 12-16.

[10] E. Cakır and T. Virtanen, "Convolutional Recurrent Neural Networks for Rare Sound Event Detection", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017),* 2017, pp. 27-31.

[11] H. Lim, J. Park and Y. Han, "Rare Sound Event Detection Using 1D Convolutional Recurrent Neural Networks", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017),* 2017, pp. 80-84.

[12] Q. Kong, I. Sobieraj, W. Wang and M. Plumbley, "Deep Neural Network Baseline for DCASE Challenge 2016", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016),* 2016, pp. 50-54.

[13] J. Kürby, R. Grzeszick, A. Plinge and G. Fink, "Bag-of-Features Acoustic Event Detection for Sensor Networks", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016),* 2016, pp. 55-59.

[14] A. Pillos, K. Alghamidi, N. Alzamel, V. Pavlov and S. Machanavajhala, "A Real-Time Enviromental Sound Recognition System for the Android OS", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016),* 2016, pp. 75-79.

[15] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, P. Shaohu, "Acoustic Scene Classification Using Deep Convolutional Neural Network and Multiple Spectrograms Fusion", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017),* 2017, pp. 133-137.

[16] S. Adavanne, P. Pertilä and T. Virtanen, "Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, 2017.

[17] S. Adavanne, A. Politis and T. Virtanen, "Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features", *International Joint Conference on Neural Networks (IJCNN 2018)*, 2018.

[18] C. H. Knapp and G.C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976, pp. 320-327.

[19] M. S. Brandstein and H.F. Silverman, "A Robust Method for Speech Signal Time-delay Estimation in Reverberant Rooms", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997.

[20] J. Benesty, J. Chen, and Y. Huang, "Time-delay Estimation via Linear Interpolation and Cross Correlation", IEEE Trans. Speech Audio Process., 2004, vol. 12, no.5, pp. 509-519.

[21] T. Hirvonen, "Classification of Spatial Audio Location and Content Using Convolutional Neural Networks", *Audio Engineering Society Convention 138*, 2015.

[22] 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, vol. 13, pp. 34-38.

[23] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound Based Localization and Identification in Industrial Environments", *IEEE Industrial Electronics Society (IECON),* 2017.

[24] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, Classification and Localization of Acoustic Events in the Presence of Background Noise for Acoustic Surveillance of Hazardous Situations", Multimedia Tools and Applications Journal, vol. 75, no. 17, 2016.

[25] S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357-366, 1980.

[26] Z. Wang, J. Le Roux, and J. R. Hershey, "Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation", *Proceedings of The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, pp. 1-5, 2018.

[27] S. Adavanne, A. Politis, and T. Virtanen, "A Multi-room Reverberant Dataset for Sound Event Localization and Detection", *arXiv:1905.08546v2*, 2019.

[28] F. Chollet, "Keras", *GitHub*, 2015, https://github.com/fchollet/keras%7D%7

[29] Theano Development Team, "Theano: A Python Framework for Fast Computation of Mathematical Expressions", *arXiv:1605.02688*, 2016.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for Polyphonic Sound Event Detection", *Applied Sciences*, vol. 6, No. 6, 162, 2016.

[31] S. Adavanne, A. Politis, and T. Virtanen, "Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network", *2018 26th European Signal Processing Conference (EUSIPCO),* IEEE, 2018, pp. 1462-1466.

[32] G. Forman and M. Scholz, "Apples-to-apples in Cross-validation Studies: Pitfalls in Classifier Performance Measurement", *SIGKDD Explor. Newsl,*. 2010, pp. 49–57.

View publication stats