LOCALIZATION, DETECTION AND TRACKING OF MULTIPLE MOVING SOUND SOURCES WITH A CONVOLUTIONAL RECURRENT NEURAL NETWORK

Sharath Adavanne, Archontis Politis, and Tuomas Virtanen

Audio Research Group, Tampere University, Finland, firstname.lastname@tuni.fi

ABSTRACT

This paper investigates the joint localization, detection, and tracking of sound events using a convolutional recurrent neural network (CRNN). We use a CRNN previously proposed for the localization and detection of stationary sources, and show that the recurrent layers enable the spatial tracking of moving sources when trained with dynamic scenes. The tracking performance of the CRNN is compared with a stand-alone tracking method that combines a multi-source direction of arrival estimator and a particle filter. Their respective performance is evaluated in various acoustic conditions such as anechoic and reverberant scenarios, stationary and moving sources at several angular velocities, and with a varying number of overlapping sources. The results show that the CRNN manages to track multiple sources more consistently than the parametric method across acoustic scenarios, but at the cost of higher localization error.

Index Terms— Multiple object tracking, recurrent neural network, sound event detection, acoustic localization

1. INTRODUCTION

Sound event localization, detection, and tracking (SELDT) is the combined task of identifying the temporal onset and offset of potentially temporally-overlapping sound events, recognizing their classes, and tracking their respective spatial trajectory when they are active. Performing SELDT successfully provides an automatic description of the acoustic scene that can be employed by machines to interact naturally with their surroundings. Applications such as teleconferencing systems and robots can use this information for tracking the sound event of interest [1–6]. Furthermore, smart cities and smart homes can use it for audio surveillance [7–9].

The joint localization and detection in static scenes with spatially stationary sources have been studied with different parametric [5, 8, 10, 11] and deep neural network (DNN) [12] based methods. However, these methods do not employ any temporal modeling required for the tracking of moving sources in dynamic scenes. Recently, we proposed a convolutional recurrent neural network (SELDnet) that was shown to perform significantly better localization and detection than the only other existing DNN-based method [12]. SELDnet's capabilities to localize events in full azimuth and elevation under matched and unmatched acoustic conditions, and without relying on features dependent on specific microphone arrays, were studied and presented in [13]. However, all the existing DNN-based methods including[12, 13] have only studied static scenes.

On the other hand, stand-alone tracking methods have been widely studied for both stationary and moving sources based on spatial information only [14–20], additional spectral information [21, 22], or in conjunction with visual information [23]. Such parametric methods often require manual tuning of multiple parameters corresponding to the scene composition and dynamics, and new sets of parameters have to be identified manually for different sound scenes. Furthermore, tracking usually focuses on distinguishing source trajectories, with no regard to source signal content. In the case of temporally overlapping trajectories, track identities are assigned to individual trajectories, but these identities are not source dependent and are generally re-used for trajectories from different sources across the audio recording. A balance between consistent association and localization determines the tracker's performance in most cases. Alternatively, a detect-before-track approach, as in the proposed SELDnet, circumvents the association problem by first detecting the active sound events, and then assigning a track to each detected event. As long as such a system is able to react to time-varying conditions, with temporally and spatially overlapping sound events from both stationary and moving sources, it is also able to detect and track the sound events of interest.

In this work, we study the multi-source tracking capabilities of a detection and localization system based on our recently proposed SELDnet [13]. To the best of the authors knowledge, this is the first DNN-based SELDT studies. We show that training the SELDnet with dynamic scene data results in tracking, in addition to localization and detection. This tracking ability is enabled by the recurrent layers of the SELDnet that can model the evolution of spatial parameters as a sequence prediction task given the sequential features and their spatial trajectory information. We show that the recurrent layers are crucial for tracking, and in comparison to stand-alone trackers they additionally perform detection. Unlike the parametric tracking methods discussed earlier, the recurrent layer is a generic tracking method that learns directly from the data without manual tracker-engineering. Finally, we show that the tracking performance of SELDnet is comparable with stand-alone parametric tracking methods through evaluation on five datasets, representing scenarios with stationary and moving sources at different angular velocities, anechoic and reverberant environments, and different numbers of overlapping sources. The method and all the studied datasets are publicly available[1].

2. METHOD

The block diagram of SELDnet [13] is illustrated in Figure 1. The input to SELDnet is a multichannel audio recording, from which a feature extraction block extracts the phase and magnitude components of the spectrogram from each channel. The SELDnet maps the input spectrogram of T-frames length to two outputs of the same length – sound event detection, and tracking; together they

This work has received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND.

[1] https://github.com/sharathadavanne/seld-net

https://doi.org/10.33682/xb0q-a335

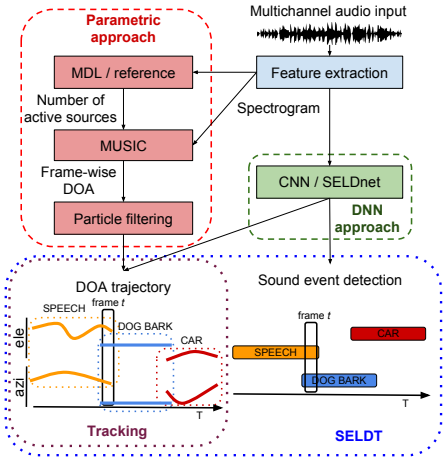

Figure 1: Workflows for the parametric tracking and DNN-based SELDT approaches. The sound class coloring and naming for the tracking task is only shown here to visualize the concept better. In practice tracking methods do not produce sound class labels as shown in Figure 3.

produce the SELDT output. The detection output is the class-wise probabilities for the C classes in the dataset of dimension $T \times C$, and is obtained as a multiclass multilabel classification task. The tracking output is a single direction of arrival (DOA) estimate per time frame for each sound class C as a multi-output regression task. Thus, when multiple instances of the same sound class are temporally overlapping, the SELDnet tracks only one instance or oscillate between the multiple instances. The estimated DOA is represented using 3D Cartesian coordinates of a point on a unit sphere around the microphone. The overall tracking output is of dimension $T \times 3C$, where $3C$ represents the three axes of the 3D Cartesian coordinates of a DOA for each class in C. Finally, to obtain the SELDT results, the class-wise probabilities of the detection output are binarized with a threshold of 0.5, anything greater represents the presence of the sound class and smaller represents the absence. The presence of a sound class in consecutive frames gives the onset and offset times, and the corresponding frame-wise DOA estimates from the tracking output when the sound class is active gives the DOA trajectory.

The SELDnet architecture used in this paper is identical to [13], with three convolutional layers of 64 filters each, followed by two-layers of 128-node gated recurrent units. The convolutional layers in the SELDnet are used as a feature extractor to produce robust features for detection and tracking. The recurrent layers are employed to model the temporal structure and the trajectory of the sound events. The output of the recurrent layers is shared between two branches of dense layers each with 128 units producing the detection and tracking estimates. The training and inference procedures of SELDnet are similar to [13] and is identical for both static and dynamic scenes, i.e., the same SELDnet designed for static scenes performs tracking when trained with moving scene data.

The recurrent layers utilize the current input frame along with the information learned from the previous input frames to produce the output for the current frame. This process is similar to a particle filter, which is a popular stand-alone parametric tracker and is also used as a baseline in this paper (see Section 3.3). The particle filter prediction at the current time frame is influenced by both the knowledge accumulated from the previous time frames and the input at the current time frame. For the tracking task of this paper, the particle filter requires the specific knowledge of the sound scene such as the spatial distribution of sound events, their respective velocity ranges

Table 1: Summary of Datasets

Sources	Sound scene	Impulse response	Acronym
Stationary [13]	Anechoic	Synthetic	ANSYN
	Reverberant		RESYN
		Real-life	REAL
Moving	Anechoic	Synthetic	MANSYN
	Reverberant	Real-life	MREAL

when active, and their probability of birth and death. Such concepts are not explicitly modeled in the recurrent layers used in SELDnet, rather they learn equivalent information directly from the input convolutional layer features and corresponding target outputs in the development dataset. In fact, recurrent layers have been shown to work as generic trackers [24] that can learn temporal associations of the target source from any sequential input features. Unlike the particle filters that only work with conceptual representations such as frame-wise multiple DOAs for tracking, the recurrent layers work seamlessly with both conceptual and latent representations such as convolutional layer features.

Finally, by training the recurrent layers in SELDnet using the loss calculated from both detection and tracking, the recurrent layers learn associations between DOAs from neighboring frames corresponding to the same sound class and hence produce the SELDT results. In general, unlike the parametric trackers, the recurrent layers perform similar tracking of the frame-wise DOAs in addition to also detecting their corresponding sound classes. Further, the recurrent layers do not need complicated problem-specific tracker- or feature-engineering that are required by the parametric trackers. A more theoretical relationship between recurrent layers and particle filter is presented in [25].

3. EVALUATION PROCEDURE

3.1. Datasets

The performance of SELDnet is evaluated on five datasets that are summarized in Table 1. We continue to use the stationary source datasets: ANSYN, RESYN and REAL from our previous work [13] to evaluate the tracking performance of the parametric tracker that was missing in [13], and compare with SELDnet. The recordings in ANSYN and RESYN are synthesized in anechoic and reverberant environments respectively. The recordings in REAL are synthesized by convolving isolated real-life sound events with real-life impulse responses collected at different spatial locations within a room. Further, we create moving-source versions of the ANSYN and REAL datasets, hereafter referred as MANYSYN and MREAL, to evaluate the performance on moving sources. The recordings of all datasets are 30 seconds long and captured in the four-channel first-order Ambisonics format [26]. Each dataset has three subsets with no temporally overlapping sources $O1$, maximum two $O2$, and maximum three temporally overlapping sources $O3$. Each of these subsets has three cross-validation splits consisting of 240 recordings for development and 60 for evaluation. All the synthetic impulse response datasets (ANSYN, RESYN and MANYSN) have sound events from 11 classes and DOAs with full azimuth range and elevation range $\in [-60, 60)$. The real-life impulse response datasets (REAL and MREAL) have 8 sound event classes and DOAs in full azimuth range and elevation range $\in [-40, 40)$. During the synthesis of stationary source datasets, all the sound events are placed in a spatial grid of 10° resolution for both azimuth and elevation angles. We refer the readers to [13] for more details on these datasets.

The anechoic moving source dataset MANSYN has the same sound event classes as ANSYN and is synthesized as follows. Every event is assigned a spatial trajectory on an arc with a constant distance from the microphone (in the range 1-10 m) and moving

with a constant angular velocity for its duration. Due to the choice of the ambisonic spatial recording format, the steering vectors for a plane wave source or point source in the far field are frequency-independent. Hence, there is no need for a time-variant convolution or impulse response interpolation scheme as the source is moving; the spatial encoding of the monophonic signal was done sample-by-sample using instantaneous ambisonic encoding vectors for the respective DOA of the moving source. The synthesized trajectories in MANSYN vary in both azimuth and elevation, and are simulated to have a constant angular velocity in the range $\in [-90^\circ, 90^\circ]/s$ with $10^\circ/s$ steps. Similarly, the MREAL dataset was synthesized with real-life impulse responses from [13] that were sampled at 1° resolution along azimuth only. Hence, unlike MANSYN, the sound events in MREAL (that are identical to REAL) have motion only along the azimuth with a constant angular velocity in the range $\in [-90^\circ, 90^\circ]/s$ and $10^\circ/s$ steps.

3.2. Metrics

The evaluation of the SELDT performance is done using individual metrics for detection and tracking identical to [13]. As the detection metric, we use the F-score and error rate calculated in segments of one-second with no overlap [27]. An ideal detection method will have an F-score of one and an error rate of zero. As the tracking metric, we use two frame-wise metrics: the frame recall and DOA error. The frame recall gives the percentage of frames in which the number of predicted DOAs is equal to the reference. The DOA error is calculated as the angle in degrees between the predicted and reference DOA. In order to associate multiple estimated DOAs with the reference, we use the Hungarian algorithm [28] to identify the smallest mean angular distance and use it as DOA error. An ideal tracking method has a frame recall of one and DOA error of zero (see [13] for more details).

3.3. Baseline Method

In the absence of publicly available implementations of multiple moving sound sources trackers, we use a combination of MUSIC [29] and an RBMCDA particle filter [30] to obtain tracking results in a similar fashion as in [15] and further made it publicly available [2]. The workflow of the baseline method is shown in Figure 1. MUSIC is a widely used [13, 31] subspace-based high-resolution DOA estimation method that can detect multiple narrowband sources. It relies on an eigendecomposition of the narrowband spatial covariance matrix computed from the multichannel spectrogram, and it additionally requires a source number estimate in order to distinguish between a signal and noise subspace. Herein, the number of active sources is taken from the reference of the dataset. To obtain broadband DOA estimates, the narrowband covariance matrices are averaged across three consecutive frames and frequency bins from 50 Hz to 8 kHz. We perform 2D spherical peak-finding on the resulting MUSIC pseudospectrum generated on a 2D angular grid with a 10° resolution for stationary and 1° for moving sources, in both azimuth and elevation. The final output of MUSIC MUS_{GT} is a list of frame-wise DOAs corresponding to the highest peaks equal to the number of active sources in each frame.

The second stage of the parametric method involves a particle filter that produces tracking results by processing the frame-wise DOA information of MUSIC MUS_{GT}. The particle filter assumes that the number of sources at each time frame is unknown and tracks them with respect to time using a fixed number of particles. At each time frame, the particle filter receives multiple DOAs and, based on knowledge accumulated from the previous time frames, it assigns each new DOA to one of the existing trajectories, clutter (noise), or a newborn source. Additionally, it also decides if any of the existing trajectories have died. The final output of the particle filter MUS^{PF}_{GT} produces the temporal onset-offset and the DOA trajectory for each of the active sound events. We refer the reader to [30] for the details of this approach.

[2] https://github.com/sharathadavanne/multiple-target-tracking

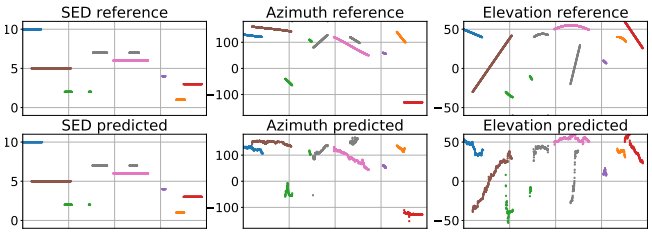

Figure 2: Visualization of the SELDnet predictions and its respective reference for a MANSYN $O2$ dataset recording. The horizontal-axis of all sub-plots represents the same time frames. The vertical-axis represents sound event class indices for the detection subplots, and DOA azimuth and elevation angles in degrees for remaining subplots.

3.4. Experiments

In all our experiments, the baseline particle filter parameters and the sequence length of input spectrogram for SELDnet was tuned using the development set of the respective subset. The performance of the tuned method was tested on the evaluation set of the subset, and the respective metrics averaged across the three cross-validation splits of the subset are reported.

Unlike the DNN-based method, the parametric method requires additional information on the number of active sources per frame to estimate the corresponding DOAs. However, SELDnet obtains this information from the data itself. In order to have a fair comparison, we used the minimum description length (MDL) [32] principle to estimate the number of sources from the input spectrogram and use it with MUSIC, resulting in the MUSIC output of MUS_{MDL} and the corresponding particle filter output of MUS^{PF}_{MDL}.

Finally, we studied the importance of recurrent layers for the SELDT task by removing them from SELDnet and evaluating the model containing only convolutional and dense layers, referred to as CNN hereafter. The best CNN architecture across datasets had five convolutional layers with 64 filters each.

4. RESULTS AND DISCUSSION

On tuning the input sequence length for SELDnet, it was observed that a sequence of 256 frames gave the best scores for the reverberant datasets, and 512 frames gave the best scores for the anechoic datasets. The SELDnet predictions and the corresponding references are visualized in Figure 2 for a respective 1000 frame test sequence from MANSYN $O2$ dataset. Each sound class is represented with a unique color across subplots. We see that the detected sound events are accurate in comparison to reference. The DOA predictions are seen to vary around the reference trajectory with a small deviation. This shows that SELDnet can successfully track and recognize multiple overlapping and moving sources.

Figure 3 visualizes the tracking predictions and their respective references for SELDnet and the baseline method MUS^{PF}_{GT}. In gen-

Table 2: Evaluation results on different datasets. Since the number of active sources information is used in $\mathrm{MUS_{GT}}$, the frame recall is always 100% and hence not reported. DE: DOA error, FR: Frame recall, F: F-score, SCOF: Same class overlapping frames

		ANSYN			RESYN			REAL			MANSYN			MREAL		
Tracking results		*O*1	*O*2	*O*3	*O*1	*O*2	*O*3	*O*1	*O*2	*O*3	*O*1	*O*2	*O*3	*O*1	*O*2	*O*3
$\mathrm{MUS_{GT}}$	DE	1.3	5.0	12.2	21.7	28.9	32.5	15.1	33.9	44.1	0.6	14.8	28.0	16.4	34.1	43.9
$\mathrm{MUS^{PF}_{GT}}$	DE	0.1	1.1	2.3	4.0	5.2	6.1	3.3	8.8	12.0	0.2	4.2	8.0	3.6	8.1	11.9
	FR	97.0	88.5	74.3	83.8	55.6	37.3	93.0	71.0	44.7	98.7	92.3	75.1	91.0	69.9	48.3
Methods estimating the number of active sources directly from input data																
$\mathrm{MUS_{MDL}}$	DE	0.5	14.2	24.0	22.3	31.9	38.5	25.3	36.2	44.1	4.2	17.8	28.5	26.5	35.9	44.9
	FR	93.9	**89.4**	**86.7**	61.7	45.6	52.5	53.6	35.7	**57.5**	63.8	48.1	51.85	53.4	35.2	**58.9**
$\mathrm{MUS^{PF}_{MDL}}$	DE	**0.1**	**4.4**	**7.2**	**6.4**	**10.6**	**12.7**	**9.3**	**10.9**	**13.7**	**3.5**	**6.8**	**8.0**	**13.6**	**11.2**	**13.6**
	FR	96.3	83.5	67.7	52.0	34.1	24.2	52.7	40.1	29.6	64	49.9	39.8	58.7	34.4	27.5
CNN	DE	25.7	25.2	26.9	39.1	35.1	31.4	32.0	34.9	37.1	26.1	25.8	28.2	36.6	39.3	40.2
	FR	80.2	45.6	32.2	69.5	45.8	29.7	45.1	28.4	16.9	83.7	58.1	38.3	44.5	26.2	16.3
SELDnet	DE	3.4	13.8	17.3	9.2	20.2	26.0	26.6	33.7	36.1	6.0	12.3	18.6	36.5	39.6	38.5
	FR	**99.4**	85.6	70.2	**95.8**	**74.9**	**56.4**	**64.9**	**41.5**	24.6	**98.5**	**94.6**	**80.7**	**69.6**	**42.8**	28.9
Detection results																
CNN	ER	0.52	0.46	0.51	0.44	0.45	0.54	0.52	0.51	0.51	0.59	0.47	0.48	0.46	0.49	0.52
	F	70.1	66.5	68	57	54.9	42.7	50.1	49.5	48.9	65.6	62.7	60.1	55.4	50.9	48.8
SELDnet	ER	**0.04**	**0.16**	**0.19**	**0.1**	**0.29**	**0.32**	**0.4**	**0.49**	**0.53**	**0.07**	**0.1**	**0.2**	**0.37**	**0.45**	**0.49**
	F	**97.7**	**89**	**85.6**	**92.5**	**79.6**	**76.5**	**60.3**	**53.1**	**51.1**	**95.3**	**93.2**	**87.4**	**64.4**	**56.4**	**52.3**
SCOF (in %)		0.0	4.2	12.1	0.0	4.2	12.1	0.0	7.6	23.0	0.0	3.0	9.1	0.0	7.1	20.9

eral, the performance of the two methods is visually comparable. Both methods are often confused in similar situations, for example in the intervals of 4-5 s, 10-13 s, and 23-25 s.

The SELDnet, by design, is restricted to recognize just one DOA for a given sound class. But in real life, there can be multiple instances of the same sound class occurring simultaneously. This is also seen in the datasets studied, the last row (SCOF) in the Table 2 presents the percentage of frames in which the same class is overlapping with itself. In comparison, the parametric method has no such restriction by design and can potentially perform better than SELDnet in these frames (even though, highly correlated sound events coming from different DOAs can easily degrade the performance of parametric methods such as MUSIC). The performance of the two methods in such a scenario can be observed in the 10-13 s interval of Figure 3. The SELDnet tracks only one of the two sources, while the parametric method tracks both overlapping sources and introduces an additional false track between the two trajectories.

Table 2 presents the quantitative results of the studied methods. The general trend is as follows. The higher the number of overlapping sources, the lower the tracking performance by both SELDnet and the parametric method. Across datasets, the DOA error improves considerably with the use of the temporal particle filter tracker, but at the cost of lower frame recall. By using MDL instead of reference information for the source number, the overall performance of the parametric approach reduces ($MUS^{PF}_{GT} > MUS^{PF}_{MDL}$). This reduction is especially observed in the frame recall metric, that drops significantly for reverberant and moving source scenario datasets, indicating the need for more robust source detection and counting schemes.

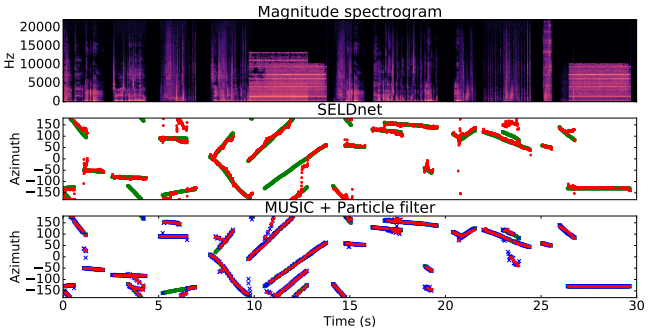

Figure 3: The tracking results of the two proposed methods are visualized for a MANSYN *O*2 dataset recording. The top figure shows the input spectrogram. The center and bottom figures show the output of SELDnet and $\mathrm{MUS^{PF}_{GT}}$ tracker in red, and the groundtruth in green. The blue crosses in the bottom figure represents the frame-wise DOA output of MUSIC

The frame recall of SELDnet is observed to be consistently better than MUS^{PF}_{MDL}, but the DOA estimation is poorer across datasets. A similar relationship is observed between SELDnet and MUS^{PF}_{GT} for all the datasets generated with simulated impulse responses, while for the real-life impulse response datasets the frame recall of SELDnet is poorer than MUS^{PF}_{GT}. That could indicate the need for more extensive learning for real-life impulse response datasets, with larger datasets and stronger models.

Using recurrent layers definitely helps the SELDT task. It was observed from visualizations that the tracking performance by the CNN was poor, with spurious and high variance DOA tracks, thus resulting in poor DOA error across datasets as seen in Table 2. This suggests that the recurrent layers are crucial for SELDT task and perform a similar task as an RBMCDA particle filter of identifying the relevant frame-wise DOAs and associating these DOAs corresponding to the same sound class across time frames.

5. CONCLUSION

In this paper, we presented the first deep neural network based method, SELDnet, for the combined tasks of detecting the temporal onset and offset time for each sound event in a dynamic acoustic scene, localizing them in space and tracking their position when active, and finally recognizing the sound event class. The SELDnet performance was evaluated on five different datasets containing stationary and moving sources, anechoic and reverberant scenarios, and a different number of overlapping sources. It was shown that the recurrent layers employed by the SELDnet were crucial for the tracking performance. Further, the tracking performance of SELDnet was compared against a stand-alone parametric method based on multiple signal classification and particle filter. In general, the SELDnet tracking performance was comparable to the parametric method and achieved a higher frame recall for tracking but at a higher angular error.

6. REFERENCES

[1] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[2] ——, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[3] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," in *Journal of Robotics and Mechatronics*, vol. 29, no. 1, 2017.

[4] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Int. Conf. on Robotics and Automation (ICRA)*, 2018.

[5] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *European Signal Processing Conference (EUSIPCO)*, 2011.

[6] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," in *IEEE Signal Processing Letters*, vol. 21, 2014.

[7] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," in *ACM Computing Surveys (CSUR)*, 2016.

[8] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound based localization and identification in industrial environments," in *IEEE Industrial Electronics Society (IECON)*, 2017.

[9] P. W. Wessels, J. V. Sande, and F. V. der Eerden, "Detection and localization of impulsive sound events for environmental noise assessment," in *The Journal of the Acoustical Society of America 141*, vol. 141, no. 5, 2017.

[10] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[11] K. Lopatka, J. Kotus, and A. Czyzewsk, "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.

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

[13] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, 2018.

[14] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of Multiple Moving Speakers With Multiple Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[15] J. M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[16] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[17] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2009.

[18] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1409–1415, 2012.

[19] J. Traa and P. Smaragdis, "Multiple speaker tracking with the Factorial von Mises-Fisher Filter," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

[20] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 392–402, 2014.

[21] J. Nix and V. Hohmann, "Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 995–1008, 2007.

[22] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 806–815, 2013.

[23] N. Strobel, S. Spors, and R. Rabenstein, "Joint Audio-Video Signal Processing for Object Localization and Tracking," in *Microphone Arrays*. Springer, 2001, pp. 203–225.

[24] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] Y. J. Choe, J. Shin, and N. Spencer, "Probabilistic interpretations of recurrent neural networks," *Probabilistic Graphical Models*, 2017.

[26] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*. John Wiley & Sons, 2017, pp. 89–140.

[27] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6, no. 6, 2016.

[28] H. W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, no. 2, 1955, p. 8397.

[29] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[30] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, no. 1, pp. 2–15, 2007.

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

[32] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.