# Overlapped Sound Event Classification via Multi-Channel Sound Separation Network

Panagiotis Giannoulis[1], Gerasimos Potamianos[2], and Petros Maragos[1]
[1]*School of ECE, National Technical University of Athens, 15773 Athens, Greece*
[2]*Department of ECE, University of Thessaly, 38221 Volos, Greece*
pangian@cs.ntua.gr, gpotam@ieee.org, maragos@cs.ntua.gr

***Abstract*—Overlapped sound event classification (SEC) can be a challenging task, especially in scenarios where the number of possible event classes or the number of simultaneous events occurring (polyphony level) are large. In such cases, the effective training of a multi-label SEC neural network can be challenging, as enough and diverse data need to be available for each of the combinatorially many possible event sets. To alleviate this problem, we examine in this paper the combination and joint training of a multi-channel sound source separation network with a multi-label SEC network. With the separation module acting as a pre-processing step, the task can be approximately reduced to isolated SEC, therefore avoiding the training complexity of overlapped scenarios. In addition, we introduce a multi-channel polyphony detection module that is trained to selectively apply the separation network only in overlapping instances during testing. We evaluate our approaches on a multi-channel dataset of overlapping sound events originating from 50 different classes. Under moderate reverberation conditions, the proposed method achieves up to 7.7% absolute improvement in terms of Fscore in the overlapped scenarios, compared to the baseline approach with traditional multi-label training.**

***Index Terms*—Sound event classification, multi-channel, overlapping events, universal sound separation**

## I. Introduction

Sound event classification (SEC) is a rapidly growing research area with many applications, including smart-home environments [1], [2], monitoring for healthcare [3], multimedia indexing and retrieval [4], and surveillance [5], [6]. In recent years, SEC has been the subject of multiple evaluation campaigns in the literature, including the well-established DCASE Challenges [7]. In the context of these challenges, several tasks related to SEC have been studied, including isolated and overlapped scenarios in single or multi-channel setups, joint SEC and localization, audio tagging, etc.

In this paper we focus on the task of overlapped SEC in a multi-channel setup. This problem has attracted significant interest in the literature, with several deep-learning based methods successfully proposed, including deep neural networks [8], convolutional neural networks (CNNs) [9], convolutional recurrent neural networks [10], and transformers [11]. The standard training approach for overlapped SEC in deep learning-based systems is to feed a multi-label neural network with overlapped instances that either exist in training or are artificially generated from the available isolated instances. However, the number of possible event combinations that need to be modeled grows rapidly as the number of event classes or the polyphony level increase. In such cases, efficient training of the network can be problematic, as it depends on the existence or generation of sufficient overlapped data, thus rendering this approach not scalable.

An alternative approach that mitigates this issue is to employ a sound source separation network as a pre-processing step to SEC, aiming in this way to approximately transform the overlapped task into the isolated one. Significant progress has been made in the domain of sound source separation in recent years, including mostly works on speech separation [12]–[15], and lately also on universal sound separation [16], [17]. Based on the above, some works employ such systems, reporting improved results for the single-channel overlapped SEC task [18], [19]. Also in [20], in a multi-channel setup, the authors train their network using beamformed signals from various directions of arrival with respect to the microphone array.

In our work, we propose for the first time the combination of a multi-channel sound separation network with a multi-label SEC system for addressing the overlapped SEC task when the number of different event classes is large. In such a scenario, we examine how the proposed approach can reduce the performance gap of a SEC system between the isolated and the more challenging overlapped cases. In particular, we employ a state-of-the-art multi-channel sound separation network in order to exploit, additionally to spectral content, the spatial discrimination of the events present in a mixture clip, while for the SEC module we employ a CNN-based architecture suitable for SEC. For the resulting pipeline, we examine both sequential and end-to-end joint re-training of the two modules, with the latter achieving the best performance. In addition, we propose the incorporation of a polyphony detection network, which can selectively apply the proposed system only to the overlapped instances during testing. Although our system is scalable to an arbitrary polyphony level, in this study, we examine the case of overlap with up to 2 simultaneous events. For our experiments we employ the ESC50 data collection [21], as it provides balanced data from a large variety of different event classes (50), and in order to design a multi-channel dataset, we combine it with real impulse responses from the DIRHA smart-home dataset [22]. Our results show that in this challenging overlapped scenario, and under moderate reverberation conditions, the proposed system can provide significant improvements over a baseline

CNN-based SEC network trained with the standard multi-label training approach.

The remainder of the paper is organized as follows: Section II provides the description of the several modules employed in our approaches; Section III describes the database and experimental framework used and reports our results; and, finally, Section IV concludes the paper.

## II. System Description

### A. Baseline SEC network

The architecture of the baseline SEC network is depicted in Fig. 1. Given an input audio signal $\mathbf{s} \in \mathbb{R}^N$ (with $N$ denoting the number of signal samples), the feature extraction stage computes 64-band Log-Mel filter-bank energies (logFBE) and their Deltas using 0.4 sec Hanning windows with 0.2 sec shift, producing the feature matrix $\mathbf{X}^{(\mathbf{s})} \in \mathbb{R}^{128\times T}$, where $T$ is the number of resulting time frames. The feature matrix $\mathbf{X}^{(\mathbf{s})}$ is fed to the network which is comprised of a 5-layer CNN block, followed by 2 fully-connected linear layers. The output $\mathbf{y} \in \mathbb{R}^C$ has dimension equal to the number $C$ of event classes and is expected to have high values at the indexes of activated events. For the multi-label training we employ the binary cross-entropy loss function. During testing, active sources are decided by applying a threshold on the output.

### B. Multi-channel separation network

In our work we employ a multi-channel separation network originally proposed for speech separation in [23]. In this method the authors essentially improve their previous work on FaSNet [24], which is a multi-channel filter-and-sum neural beamforming network operating in the time domain. The improvements include (a) the incorporation of a transform-average-concatenate (TAC) module that makes the network invariant to the permutation and the number of microphones, and (b) the transition to a single-stage architecture where the filters for all channels are jointly estimated.

The network takes as input time-domain mixture signals from $M$ microphones and outputs $K$ time-domain separated signals. Regarding the loss function, similarly to [24], we use the mean squared error (MSE) between the FBE representations of the original sources, as captured by a reference microphone, and the reconstructed sources at the output of the network. In our case, as reference microphone we consider the central microphone of a 3-channel linear array (see Section III-A).

### C. Proposed system

The proposed system, as shown in Fig. 2, combines the separation and the SEC networks in a cascade. In particular, we employ the separation network as a pre-processing step which provides the SEC network with $K$ separated signals in place of the original mixture. The idea is that given a well-performing separation network, the overlapped task can be approximately reduced to classification of a set of isolated instances, therefore improving the performance of the system.

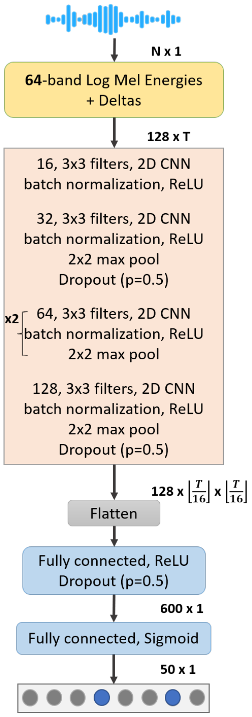

Fig. 1. Single-channel baseline architecture for sound event classification.

The SEC network used in the proposed system applies a SEC module with identical architecture with the baseline system for each of the $K$ separated inputs $\hat{\mathbf{s}}_k$, $k = 1, \dots, K$, and then averages their output vectors. For training the proposed pipeline, we examine two approaches:

- **Sequential training**
  In this case, we first train the separation network with mixtures that are artificially generated by the available isolated instances as described in Section III-B. Then we train the SEC network on the separated signals that result from the output of the separation network for the various mixtures.
- **Joint training**
  In this case the training consists of two stages. The first stage is the same with the sequential training, except that the SEC network is trained on ground-truth separated signals. In the second stage, the two networks are jointly re-trained, using as input the mixture signals from the microphone array and as loss function the binary cross-entropy on the final output. In this way, the parameters of both networks can be fine-tuned towards the final objective of event classification.

Finally, we also examine the ensemble of the baseline SEC network with the proposed system, by performing linear late fusion on their outputs, followed by thresholding.

### D. Polyphony network

The proposed method is designed to operate on audio segments with overlapped events. In order to evaluate it in a realistic scenario with both isolated and overlapped instances,

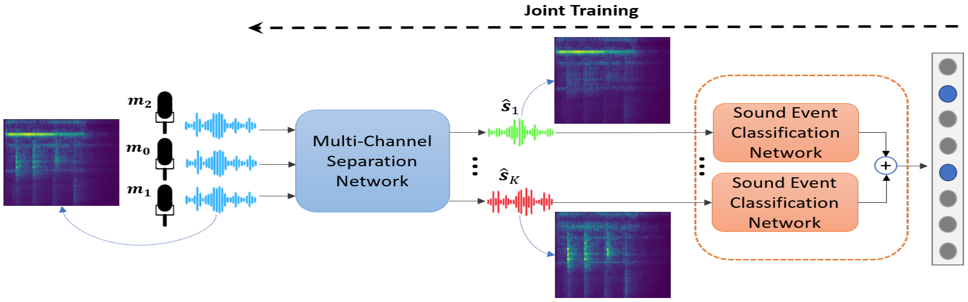
Joint Training
$m_2$
$m_0$
$m_1$
Multi-Channel Separation Network
$\hat{s}_1$
$\hat{s}_K$
Sound Event Classification Network
Sound Event Classification Network
+

Fig. 2. Pipeline of the proposed system for overlapped sound event classification.

we need a module able to detect the polyphony level and selectively apply it only in the overlapped cases.

Polyphony classification modules based on deep learning have been recently employed with success in the literature [25], [26]. In our work, we implement a polyphony classification network that exploits both the spectral and the spatial information by using logFBE features in conjunction with Generalized Cross Correlation (GCC-PHAT) based features, computed for different pairs of microphones of the array. Similarly with [27], we consider the GCC-PHAT features as GCC spectrograms which are concatenated with the logFBEs to form the final feature matrix. In our case that we use a 3-microphone array, the network takes as input the feature matrix $\left[\mathbf{X}^{(\mathbf{s}_0)};\mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_1)};\mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_2)}\right] \in \mathbb{R}^{384\times T}$, where $\mathbf{s}_0$ is the signal captured by the central microphone $m_0$, and outputs a $P$-dimensional vector $\mathbf{y} \in \mathbb{R}^P$, where $P$ denotes the maximum possible degree of polyphony (in this study, $P = 2$). For the polyphony network, we use the same architecture with the baseline SEC network (just changed the output dimension of the last linear layer), and the cross-entropy as loss function.

## III. Experiments

### A. Database

For our experiments we employ the environmental sound classification (ESC50) dataset [21]. ESC50 contains 2000 5 sec-long audio clips from 50 different event classes, belonging to various sound categories such as animal sounds, natural soundscapes, human (non-speech) sounds, domestic sounds, and urban noises.

In order to create a multi-channel dataset, we convolve the audio clips with real room impulse responses (RIRs) from the DIRHA smart-home dataset [22]. In particular, we use a linear microphone array with 3 omni-directional microphones (spaced 15 cm apart) placed inside the living room of the DIRHA smart home, and 12 different locations with 2 possible orientations each for the event sound sources. With respect to the central microphone, the $T_{60}$ reverberation times for the different source locations range from 0.58 to 0.83 sec, while their distances from 0.72 to 3.2 m.

### B. Experimental setup

At first, all audio clips from ESC50 and RIRs from DIRHA are downsampled to 16 kHz. Before the convolution with the DIRHA RIRs, we pre-process the weakly-labeled audio clips of ESC50 as follows: similarly to [28], we first remove silent areas using an energy thresholding criterion, and then we split them to 1-sec segments with 80% overlap, thus producing about 34k clips in total. In this way, we obtain more samples to train our network, and also our system can operate at a finer temporal resolution. These audio clips are then split into training, validation, and test sets at a 8:1:1 ratio. In the split we ensure that different sets do not contain clips from the same recording.

In order to simulate a realistic scenario, we assume that for each set, 50% of their clips are observed as isolated instances and 50% as parts of overlapped instances. The audio clips are then convolved with RIRs to produce 1.5-sec long segments (by truncating longer parts). In the case of overlapped instances, we randomly choose a location and orientation for each event and mix them at SNRs between -2 and 2 dB. Overall, we end up with approximately 13.5k isolated and 6.5k overlapped instances in the training set, and 1.8k isolated and 0.8k overlapped for each of the validation and test sets. Also, by following the standard data augmentation paradigm, we further generate artificial mixtures from the observed isolated instances of each set by superposition. In this way, we also generate 30k overlapped instances for the training set (resulting in 36.5k total), and 2.2k for each of the validation and test sets (3k total each).

Regarding the evaluation metrics, for the multi-label SEC task we employ the Fscore metric, while for the performance of the polyphony network we use the classification accuracy.

### C. Network training details

For training the networks, the Adam optimizer is used [29], with initial learning rate set to 0.001 and decreased to half every 30 epochs. All the networks are trained for 100 epochs, except the joint network that is re-trained for 30 epochs. In the end, the epoch with best performance on the validation set is kept. The batch size for the separation network is set equal to 20, while for the SEC networks is set to 150. Finally, the separation network is trained on the set of 30k generated mixtures where separated ground-truth signals can be considered as known, and for the overlapped task the SEC baseline network is trained on both 30k and 6.5k overlapped instances of the training set.

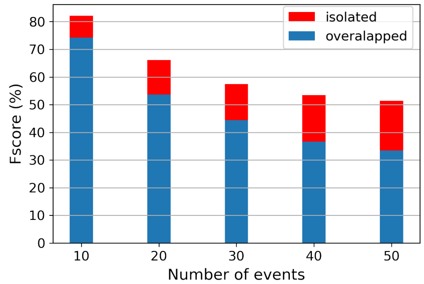

Fig. 3. Performance of the SEC baseline network for isolated and overlapped tasks for event sets of various numbers of classes.

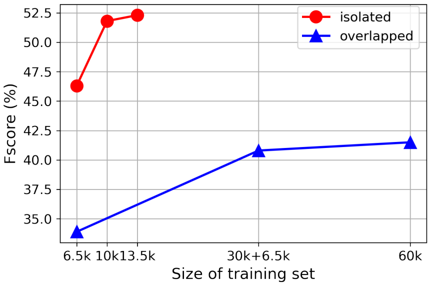

Fig. 4. Performance of the SEC baseline network for isolated and overlapped tasks for various sizes of the training set.

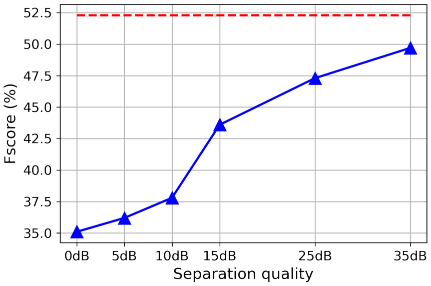

Fig. 5. Performance of the SEC network of Fig. 2 in the overlapped task (blue) for various levels of separation quality (measured in dB). The red dashed line corresponds to the performance of the baseline SEC network on the isolated task, which can be considered as an upper limit.

### D. Results

In Fig. 3, we compare the performance of the baseline SEC network for the overlapped and isolated tasks as the number of event classes considered increases. While the performance clearly degrades in both tasks, their gap progressively increases as the number of events adds complexity to the overlapped task. Given a well-performing separation network, our proposed pipeline aims to reduce this gap.

One way to improve performance in overlapped scenarios is to increase the training size. In Fig. 4, the performance of the SEC baseline network for both isolated and overlapped tasks is depicted for different sizes of their training sets. As we can see, the performance in the overlapped scenario improves as we add more data to the training set, but at a decreased rate compared to the isolated scenario. Although we can artificially generate infinite overlapped examples, the contribution of the augmented data saturates at some point, as the diversity of produced mixtures from a given set is limited. On the other hand, in the isolated task higher Fscore values are achieved for quite smaller training set sizes.

TABLE I
PERFORMANCE OF THE VARIOUS SYSTEMS FOR THE OVERLAPPED-EVENT SCENARIO, IN TERMS OF FSCORE.

| System | Fscore (%) | |
|---|---|---|
| | $\overline{T}_{60}$=0.61s | $\overline{T}_{60}$=0.80s |
| (A) Baseline (1 channel) | 41.26 | 39.05 |
| (B) Baseline (3 channels) | 41.45 | **39.33** |
| (C) Proposed - Sequential | 44.72 | 38.41 |
| (D) Proposed - Joint | **47.46** | 38.75 |
| Late Fusion (B+C) | 46.20 | 41.52 |
| Late Fusion (B+D) | **48.95** | **41.95** |

In Fig. 5, in an oracle experiment, we examine the performance of the SEC network of Fig. 2 in the overlapped task (using 10k training samples and sequential training), in relation to several hypothetical levels of separation quality provided by the separation network. To simulate the outputs of the separation network, we artificially mix the isolated sources at different SNRs. While this experiment ignores the possible distortion artifacts that can be inserted by the separation network, it provides evidence that even when residuals of the undesired source are present in each separated input signal, the separation module can significantly boost the performance of the SEC network, provided that its separation quality exceeds a certain level ($\sim$10 dB). Indeed, it can be seen that as the separation quality increases, the overlapped task performance (in blue) approximates the isolated task performance (in red) of the baseline network.

Table I shows the performance of the various approaches for the overlapped task in terms of Fscore for two different reverberation scenarios. In particular, the locations and orientations of the event sources are selected such as the mean reverberation time is 0.61s and 0.80s respectively. As a multichannel extension of the baseline SEC network, we perform decision level fusion on the outputs of the three single-channel networks. In both scenarios, we observe that this multi-channel version of the baseline is only slightly better than the single-channel one, as the logFBE features are expected to be similar in adjacent microphones. For the lower reverberation case, we observe that both of the proposed methods outperform the baseline, with the jointly trained variant achieving the best performance (47.46%). Further improvements are observed with the fusion schemes (46.20% and 48.95%), which indicates that the SEC networks trained on the mixture signal and on the separated signals learn complementary information. This corresponds to 7.7% absolute improvement compared to the baseline (A). On the contrary, in the higher reverberation case, the proposed system (D) fails to improve the baseline, due to inadequate performance of the separation network. This is in agreement with the results of recent works on the performance of separation networks under high reverberation conditions [30] and also with our results in Fig. 5 which indicate that separation needs to exceed a certain quality

TABLE II
PERFORMANCE OF THE POLYPHONY CLASSIFICATION NETWORK FOR DIFFERENT FEATURE SETS, IN TERMS OF ACCURACY.

| Features | Notation | Accuracy (%) |
|---|---|---|
| logFBE | $[\mathbf{X}^{(\mathrm{s}_0)}]$ | 95.59 |
| GCC | $[\mathbf{GCC}^{(\mathrm{s}_0,\mathrm{s}_1)}; \mathbf{GCC}^{(\mathrm{s}_0,\mathrm{s}_2)}]$ | 98.68 |
| logFBE + GCC | $[\mathbf{X}^{(\mathrm{s}_0)}; \mathbf{GCC}^{(\mathrm{s}_0,\mathrm{s}_1)}; \mathbf{GCC}^{(\mathrm{s}_0,\mathrm{s}_2)}]$ | **99.27** |

to boost the overall performance. Nevertheless, the fusion schemes still achieve improvements over the baseline.

Table II provides the polyphony level classification accuracy of the proposed polyphony network for various choices of feature sets. We observe that while all feature sets achieve good performance, the best option is to combine the logFBE features with the GCC-based ones, leading to 99.27% classification accuracy. With such performance, it is guaranteed that our pipeline will be applied almost only on overlapped instances during testing.

## IV. CONCLUSIONS

In this paper, we examined the combination of sound source separation with overlapped sound event classification in a multi-channel setup with a large variety of event classes. Our results showcase the potential of incorporating separation methods in SEC systems, albeit high reverberation scenarios can be a limiting factor for the performance of the proposed pipeline. In future work, we plan to explore scenarios with polyphony of higher degree ($\geq 3$ simultaneous events). Also we will investigate the perspective of sound separation via a distributed microphone network, which could potentially further improve the separation quality.

## REFERENCES

[1] S. Krstulovic, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds., pp. 335–371. Springer, 2018.

[2] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2375–2379.

[3] D. Hollosi, J. Schröder, S. Goetze, and J.-E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," in *Proc. 3rd International Symposium in Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, 2010, pp. 1–5.

[4] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[5] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[6] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6460–6464.

[7] *DCASE: Detection and classification of acoustic scenes and events*, http://dcase.community/.

[8] I. Choi, K. Kwon, S. H. Bae, and N. S. Kim, "DNN-based sound event detection with exemplar-based approach for noise reduction," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 16–19.

[9] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *Signal Processing Letters (SPL)*, vol. 24, no. 3, pp. 279–283, 2017.

[10] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 5, no. 6, pp. 1291–1303, 2017.

[11] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution augmented transformer for semi-supervised sound event detection," in *Proc. Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE), Tech. Rep.*, 2020.

[12] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2016, pp. 545–549.

[13] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," *arXiv preprint arXiv:1910.06379*, 2019.

[14] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[15] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," *arXiv preprint arXiv:2003.01531*, 2020.

[16] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 96–100.

[17] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 175–179.

[18] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8677–8681.

[19] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," *arXiv preprint arXiv:2007.03932*, 2020.

[20] W. Xue, T. Ying, Z. Chao, and D. Guohong, "Multi-beam and multi-task learning for joint sound event detection and localization," in *Proc. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE), Tech. Rep.*, 2019.

[21] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Int. Conference Multimedia*, 2015, pp. 1015–1018.

[22] L. Christoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. Int. Conference Language Resources and Evaluation (LREC)*, 2014, pp. 2629–2634.

[23] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," *arXiv preprint arXiv:1910.14104*, 2020.

[24] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, pp. 260–267.

[25] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Proc. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE), Tech. Rep.*, 2019.

[26] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "CountNet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 268–282, 2019.

[27] Y. Cao, T. Iqbal, Q. Kong, M. B. Galindo, W. Wang, and M. D. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Tech. Rep.*, 2019.

[28] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2017, pp. 3107–3111.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation.," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 696–700.