JOINT DIRECTION AND PROXIMITY CLASSIFICATION OF OVERLAPPING SOUND EVENTS FROM BINAURAL AUDIO

Daniel Aleksander Krause[1], *Archontis Politis*[2], *Annamaria Mesaros*[3]

Faculty of Information Technology and Communication Sciences, Tampere University
Korkeakoulunkatu 7, 33720, Tampere, Finland,
[1] daniel.krause@tuni.fi, [2] archontis.politis@tuni.fi, [3] annamaria.mesaros@tuni.fi

ABSTRACT

Sound source proximity and distance estimation are of great interest in many practical applications, since they provide significant information for acoustic scene analysis. As both tasks share complementary qualities, ensuring efficient interaction between these two is crucial for a complete picture of an aural environment. In this paper, we aim to investigate several ways of performing joint proximity and direction estimation from binaural recordings, both defined as coarse classification problems based on Deep Neural Networks (DNNs). Considering the limitations of binaural audio, we propose two methods of splitting the sphere into angular areas in order to obtain a set of directional classes. For each method we study different model types to acquire information about the direction-of-arrival (DoA). Finally, we propose various ways of combining the proximity and direction estimation problems into a joint task providing temporal information about the onsets and offsets of the appearing sources. Experiments are performed for a synthetic reverberant binaural dataset consisting of up to two overlapping sound events.

Index Terms— binaural audio, binaural localization, distance estimation

1. INTRODUCTION

Acoustic environments consist of numerous types of sounds that are produced by sources distributed across space. Research in acoustic scene analysis has led to the formation of several inter-related audio tasks, including sound event detection and acoustic scene classification [1]. Localization and tracking of acoustic sources [2] are some of the oldest and most researched tasks, and constitute an important part of many practical applications like surveillance systems [3], audio-driven robotics [4], teleconferencing [5], or speech recognition and enhancement [6,7]. Recent research has led to an increased focus on merged tasks utilizing spatial recordings, such as sound event detection and localization (SELD) [8]. Although source distance and direction-of-arrival are both estimated from spatial information captured between the microphones, the former has been investigated to a much lesser extent [9], due to its estimation being crippled for distances that are a few times larger than the array size. In the case of binaural localization, this region extends to no more than about a meter from the listener, and understanding whether a source is proximal to the listener or not is still of interest in many applications.

The majority of localization methods rely on multi-microphone arrays with more than two microphones, employing linear, circular, or spherical arrangements, depending on the application. However, two channel binaural recordings are an important format for acoustic scene analysis, mainly because they are based on the same spatial cues as the auditory system, they match the audio perspective of the human or human-like recorder, and they are a popular option in anthropomorphic robotic processing [10, 11].

Early works on binaural localization and distance estimation have focused on statistical methods using binaural cues [12, 13]. More recent studies utilize deep neural networks to overcome the limitations of model-based approaches and improve robustness to interfering noise and reverberant conditions [14,15]. However, most studies investigating joint localization and distance estimation have focused on the azimuth plane only [16, 17], since inter-channel binaural cues contain inherent ambiguities with elevation due to the cone-of-confusion effect [18]. Some studies analyze performance on elevation classification separately from azimuth [19], whereas other studies include joint azimuth and elevation localization, but restricted to certain ranges, e.g., the frontal area [20]. Analogously, most work focused on distance estimation is restricted to sources at fixed positions [21] or varying in the azimuth plane only [22].

In this paper, we propose methods for obtaining joint proximity and direction information from binaural audio. We treat both problems as coarse classification tasks, in which proximity is described in a binary way, i.e., as near or far. For direction, two ways of splitting the sphere into angular areas are proposed, in order to obtain a set of rough classes - left, right, front, back, top and bottom. In the first part of our study, we investigate several ways of performing direction classification utilizing a single-label and a multi-label approach. In the second part, different techniques of combining the proximity and direction classifications tasks are investigated. Compared to existing studies on these topics, we extend our research to a scenario with temporally overlapping sound events spread across the whole sphere, and diverse reverberation conditions. Hence, models are trained and evaluated at a frame level to provide information about the onsets and offsets of the sound sources. To the authors' knowledge, this scenario has not been studied so far for binaural audio.

2. METHOD

2.1. Model and features

To enable concurrent spatial information processing and temporal detection of sound sources, we utilize a convolutional recurrent neural network (CRNN) architecture shown in Fig. 1. This kind of model architecture has been proven to be effective in similar audio tasks [8, 23]. The first block of the model consists of three convolutional layers with 3x3 kernels. After each convolutional part,
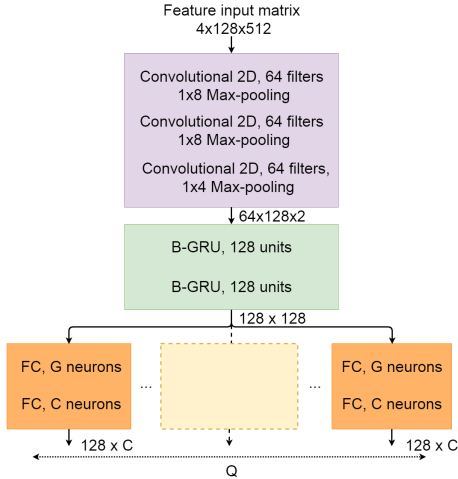
Figure 1: Baseline architecture used across all experiments.

Table 1: Randomization of parameters for data generation.

Parameter	Random range
Room width and length	[6.0 10.0] m
Room height	[2.5 6.0] m
RT	[0.3 0.9] s
Source distance (near)	[0.4 2.0] m
Source distance (far)	[3.0 8.0] m

Table 2: Definition of the equal and unequal sphere division types.

Division	Direction	Azimuth [$^\circ$]	Elevation [$^\circ$]
Unequal	Front	[-45, 45]	[-35, 35]
	Back	[135, -135]	[-35, 35]
	Left	[45, 135]	[-35, 35]
	Right	[-45, -135]	[-35, 35]
	Top	[-180, 180]	[35, 90]
	Bottom	[-180, 180]	[-90, -35]
Equal	Front	[-90, 90]	[-90, 90]
	Back	[90, -90]	[-90, 90]
	Left	[0, 180]	[-90, 90]
	Right	[-180, 0]	[-90, 90]
	Top	[-180, 180]	[0, 90]
	Bottom	[-180, 180]	[-90, 0]

ReLU activation outputs are followed by batch normalization and MaxPooling across frequency. The convolutional block is followed by two bidirectional gated recurrent units (B-GRU), each scaled by a hyperbolic tangent function. The bottom part of the model consists of Q fully connected (FC) branches. Each branch contains a G-dimensional linear layer, followed by an output layer with C neurons and a sigmoid activation function.

Models are trained using a set of three different feature representations. Firstly, we compute the complex spectrogram using a short-time Fourier transform with a Hamming window of $40ms$ length and 50% overlap. Next, to represent the phase shift between binaural channels we compute the sine and cosine values of interaural phase differences (**sin&cos**). These features have been shown to outperform direct use of phase differences in both speech separation [24] and sound event localization [23], as they avoid phase wrapping and provide a smoother representation of highly varying phase values. Furthermore, we utilize interaural level differences (**ILDs**) which, considering head shadowing effects, become significant at frequencies above 1.5kHz and constitute one of the two major cues in binaural localization [18]. The second major binaural localization cue, time-difference-of-arrival, which is mostly dominant at frequencies below 1kHz, is represented in the features by the phase differences to which it relates. Interaural level differences can flatten source-related magnitude patterns, which may contain important information for temporal detection of events. Therefore, a single magnitude spectrogram is added to offset this deficiency, making for 4 feature channels in total. The modelled sequence consists of 128 frames, which results in a 4x128x512 input matrix. DNNs are trained using the Adam optimizer [25] and the Keras library [26], with binary-cross entropy used as the loss function. A single output is considered to be positive when its value is greater than or equal to 0.5.

2.2. Data

Experiments are performed using a synthetic dataset. In order to ensure a diverse representation of sound sources, data is created using isolated sound events derived from several datasets: NIGENS [27], DESED [28] and TUT Rare Sound Events 2017 [29], containing 18 total sound classes, namely: alarm, baby, blender, cat, crash, dishes, dog, engine, fire, footsteps, glassbreak, gunshot, knock, phone, piano, scream, speech, water. The data is split into two subsets, one of which contains up to two overlapping sound events, whereas the other one consists of single sources only. Each subset contains 400 audio files, divided into 4 equal splits for fold-wise cross-validation.

In order to provide a possibly large and general representation of localization scenarios, we simulate a random shoe-box room for each file separately. We randomize both the reverberation time (RT) and physical dimensions of the room, as well as the receiver position. For each sound source, a separate room impulse response (RIR) is synthesized by selecting a random location inside the room. Randomization parameters are summarized in Table 1. Binaural RIRs are simulated using the image source method as implemented in [30], where individual image sources are convolved with head-related impulse responses (HRIRs) from the respective source direction. We employ an available HRIR set that includes near-field HRIRs [31], measured at every 10cm radiuses away from the head, from 20cm up to 1.5 m. In that way it is possible to include all the relevant near-field distance cues of proximal sources. If the source is closer than 1.5 m the respective distance-dependent HRIR is convolved with the direct sound part, while the rest of the image sources are convolved with the far-field 1.5 m HRIR set. Each source and receiver is forced to be at least 1 m apart from the walls, ceiling, and floor. The recordings are finally synthesized by convolving randomly picked sound events with the corresponding RIRs, with their temporal onset placed randomly in the recording. Each audio file is 15 seconds long with frequency sampling of 24 kHz and 16 bit resolution. The dataset is available at Zenodo [1].

3. EXPERIMENTAL SETUP

We investigate several methods to obtain joint information about the proximity and direction of sound sources from binaural audio. We treat both problems as coarse classification tasks with predefined

---

[1] 10.5281/zenodo.5118587

classes. Proximity is expressed by two classes: near and far. The **near** class is defined for sources appearing in the proximal region around the head in the range of [0.4, 2.0] m. After some initial experiments, a buffer zone of 1m is added to avoid issues with sources appearing at the border of two classes. Therefore, the **far** class covers the distance from 3m upwards.

The direction task is defined for a coarse set of classes, corresponding to six basic ranges of DoAs - **left**, **right**, **top**, **bottom**, **front** and **back**. We propose two ways of defining those, as summarized in Table 2:

- **Unequal:** this division type separates the sphere into six disjunctive cones, which are supposed to roughly reflect the intuitive human perception of each direction.
- **Equal:** this method divides the sphere into three sets of hemispheres defined around each of the Cartesian planes, instead of creating exclusive angular areas. Every source can be essentially described in each plane by one of two disjunctive classes - left or right, front or back and top or bottom, allowing for multi-label classification.

In the experiments, we investigate joint modelling of proximity and direction classification. Among these two tasks, classifying the direction appears to be more difficult due to its more complex nature and the spatial limitations of binaural recordings. Therefore, we split our experiments into two consecutive stages:

1. **Direction classification:** In this stage, we separately investigate direction classification. We compare the performance of several ways to utilize both sphere divisions and perform the task using a single-label and multi-label approach.
2. **Joint proximity and direction classification:** Here we combine the solutions analyzed in the previous steps with proximity classification. The joint task is performed as a single task for the unequal sphere split and using a multi-task approach for both division variations.

3.1. Stage 1: Direction classification

For both types of sphere divisions we propose several ways to perform direction classification:

- **UneqOne:** using the unequal split, we train a single model to perform classification for all 6 directions classes using a single output layer. In this case $Q = 1, G = 128, C = 6$.
- **EqSepMod:** for the equal sphere division, we train three distinct models to perform direction detection in each hemisphere separately, resulting in a multi-label task as a whole. Hence, an exclusive binary classifier determines whether a sound source appears in the left or right area, while the other ones analyse the front/back and top/bottom planes. Since each output represents only two classes, we set lower values for the parameters: $Q = 1, G = 64, C = 2$.
- **EqSepBran:** in this method, we use the same multi-label approach as in EqSepMod, however instead of training three models, we utilize a single DNN with separate output branches, each corresponding to a different hemisphere division. Hence, $Q = 3, G = 64, C = 2$.
- **EqOne:** we use the equal division to separate the sphere into disjunctive eighths, each being a combination of the hemispheric classes, e.g. top-front-left. This concept is closer to UneqOne, since it is defined as single-label classification of exclusive directions that serve as a reference for results obtained with the multi-label approach. We perform classifica-

tion for each of the 8 disjunctive classes using a single-label approach, hence $Q = 1, G = 128, C = 8$.

We note that EqSepMod and EqSepBran give ambiguous information about the location of the sound source. To enable direct comparison of all methods, we combine the three independent outputs into 8 disjunctive classes as defined for EqOne. To achieve this, we use a frame-level probability product:

$$P_c[n] = \sqrt[3]{P_{lr}[n] * P_{fb}[n] * P_{tb}[n]} \geq 0.5. \tag{1}$$

$P_c[n]$ denotes the output values for the n-th frame and c-th single-label class defined on the equal sphere division. Each value of c corresponds to a unique combination of lr, fb and tb, which are associated with probabilities obtained in the left/right, front/back and top/bottom planes respectively.

3.2. Stage 2: Joint task

In the second phase of our experiments, we focus on joint modelling of proximity and direction classification. Here, we introduce the following methods of performing this task:

- **UneqSingle:** for the unequal sphere division, we combine all proximity and direction categories into a single-task set of 12 classes (e.g., near-left). Hence, $Q = 1, G = 128, C = 12$.
- **UneqMulti:** alternatively, we propose a multi-task approach by utilizing a single fully connected layer with outputs corresponding to each task separately. Since there are 2 proximity classes and 6 direction classes, this results in the following parameters: $Q = 1, G = 128, C = 8$.
- **EqSepMod-J:** this method combines the separate models trained for EqSepMod in the previous stage and utilizes an analogous fourth DNN to perform proximity classification.
- **EqSepBran-J:** similarly to the previous one, we use a model derived from EqSepBran with another output branch associated with proximity classes. Thus, $Q = 4, G = 64, C = 2$.
- **EqOne-J:** analogously to UneqMulti, we combine proximity and direction classes into a multi-task output branch, utilizing the equal sphere division. $Q = 1, G = 128, C = 10$.

4. RESULTS AND DISCUSSION

4.1. Direction classification

Table 3 and 4 show results obtained for data with one source and two concurrent sources. Model performance is measured using the one second segment-based F_1 score averaged over all cross-validations folds.

As can be observed for the results obtained on the unequal sphere split (UneqOne), there is a significant imbalance between the model's performance on different classes. With an overall F_1 measure of 55.30% (Table 4), the model scores at 71.77% for the left/right classes, whereas for front/back the performance decreases to 48.35%. Overall performance is strongly affected by the top/bottom directions, for which the accuracy scored at 26.40%. This contrast is inherently related to the binaural format qualities. Largest differences between channels appear in the left/right plane, which allows easier localization. For the front/back plane the distinction becomes less pronounced due to diminishing interaural differences. The top and bottom directions are affected most by binaural limitations, with elevation localization cues known to be weaker than lateralization cues [18].

Table 3: F_1 score [%] obtained for a single source scenario.

	UneqOne	EqOne	EqSepBran	EqSepMod
Left/Right	71.39	-	88.25	**89.82**
Front/Back	39.91	-	60.54	**65.75**
Top/Bottom	28.66	-	60.53	**63.75**
Single-label	**51.98**	30.44	29.66	34.37
Multi-label	-	-	68.76	**73.11**

Table 4: F_1 score [%] obtained for two concurrent sources.

	UneqOne	EqOne	EqSepBran	EqSepMod
Left/Right	71.77	-	87.45	**88.29**
Front/Back	48.35	-	73.88	**74.76**
Top/Bottom	26.40	-	73.10	**74.24**
Single-label	**55.30**	34.11	46.79	47.90
Multi-label	-	-	78.05	**79.11**

Notable improvements can be obtained by using a multi-label approach and the equal sphere division. Utilizing separate models (EqSepMod) for each plane results in a high score for the left/right directions ($88.29\% F_1$). The front/back and top/bottom models achieve 74.76% and 74.24%, outperforming the UneqOne model by 26.41 p.p. and 47.84 p.p. respectively. We note that in this method each class includes a whole hemisphere of cues, contrary to the unequal sphere division in which all directions are defined by narrower cones. Interestingly, a single model with separate branches (EqSepBran) shows just a very slight decrease in performance of about 1 p.p. in each plane, indicating that very similar results can be achieved with a joint DNN. EqSepBran and EqSepMod obtain overall multi-label scores of 78.05% and 79.11%. However, combining the results for each plane into a single-label grid using (1) shows a respective drop to 46.79% and 47.90%. Despite efficient direction classification in each dimension, different onsets and offsets might create a challenge with temporal matching of independently obtained output sequences. This approach might therefore benefit from a more intelligent approach than a frame-level probabilistic product, which achieves slightly worse results than UneqOne. Training a model directly with a single-label approach (EqOne) results in $34.11\% F_1$ measure, which is the worst overall score. Compared with UneqOne, the performance might be affected by a different sector orientation and a finer division of the sphere. A difference of over 13 p.p. and high accuracy for the front/back and top/bottom classes show the potential of multi-label training for future research.

Results obtained for both data subsets show coherent relations between methods for the single- and two-source case. No significant differences can be observed for the left and right directions, however we note a significant drop in the front/back and top/bottom planes for a single source scenario. We link it to a higher number of false positives caused by switching between opposite classes due to the cone of confusion. This effect might be partly masked in a double source scenario by the appearance of two sound events located on opposing sides of the head.

4.2. Joint proximity and direction classification

Results obtained for the joint task of proximity and direction classification are summarized in Table 5. Combinations marked with a "-J' ending refer to models evaluated in the previous stage with

Table 5: F_1 score [%] obtained for the joint task. Direction performance is described for a double source single-label scenario. The last column shows separate direction classification for reference.

Model	Proximity	Direction	Sep. direction
UneqSingle	50.63		55.03
UneqMulti	82.54	51.03	
EqOne-J	83.45	30.31	34.11
EqSepBran-J	**83.92**	**45.39**	46.79
EqSepMod-J	**83.97**	**47.90**	47.90

an additional branch performing the proximity task. UneqMulti and UneqSingle are models described in section 3.2.

As a reference for the tested network variations, we use results derived from separate DNNs. A model trained for proximity classification alone achieves a score of **83.97**%, which shows that coarse proximity classification in conjunction with temporal detection of sound sources can be learnt very efficiently using binaural audio. Considering the connection of proximity information with other audio tasks, our results set a space for wider research in the future.

A decline in performance of both tasks when combined into a joint model is well expected, however the differences in our experiments are not major. Specifically, the proximity task shows a maximal drop of only 1.34 p.p. when combined using the multi-task method (UneqMulti). The lowest performance difference for both tasks can be observed using separate multi-label branches (EqSepBran-J). Compared with the baseline models, F_1 score decreases by 0.05 p.p. and 1.4 p.p. for proximity and direction, respectively. These results are comparable with the separate model approach (EqSepMod-J). We notice a more significant decrease of direction classification for a single-label approach, for both equal (EqOne-J) and unequal sphere division (UneqMulti). This might suggest that multi-label learning is less prone to be affected by joint tasks. Finally, utilizing a single-task output for the equal sphere split (UneqSingle) results in $50.63\% F_1$ score, which is mostly affected by poor direction classification. Since this approach disables direct separation of information about proximity and direction, its low efficiency makes it the least practical solution in this setting.

5. CONCLUSIONS

In this paper, we investigate a scenario in which joint proximity and direction classification is performed in conjunction with temporal onset and offset detection of overlapping sound events. We propose two ways of splitting the whole sphere into coarse sets of direction classes, for which we propose a single-label and multi-label approach to perform classification. Finally, we present several ways to combine the proximity and direction classification tasks to perform them in a joint manner.

The main challenge when tackling the direction classification problem appears in the top/bottom and front/back planes, that are most affected by the cone of confusion. Our experiments show that with the proposed multi-label approach and a hemisphere division of classes the F_1 score for these directions can be increased significantly. In this case, combining independent output sequences into disjunctive areas can become another problem, which is a subject for future research. Finally, we show that joint proximity and direction classification can be achieved without significant losses in performance for both problems, with the multi-task approach standing out as the most efficient one.

6. REFERENCES

[1] T. Virtanen, A. Mesaros, T. Heittola, M. D. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology, 2016.

[2] M. Brandstein, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[3] J. Kotus, K. Łopatka, and A. Czyżewski, "Application of vector sensors to acoustic surveillance of a public interior space," *Archives of Acoustics*, pp. 851–860, 2011.

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

[5] S. Aoki and M. Okamoto, "Audio teleconferencing systems with sound localization effects," in *The Journal of the Acoustical Society of America*, 1999.

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

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

[8] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, p. 34–48, 2019.

[9] A. Brendel and W. Kellermann, "Distance estimation of acoustic sources using the coherent-to-diffuse power ratio based on distributed training," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 1–5.

[10] J. C. Murray, H. Erwin, and S. Wermter, "Robotics sound-source localization and tracking using interaural time difference and cross-correlation," in *AI Workshop on NeuroBotics*, 2004.

[11] I. Trowitzsch, J. Mohr, Y. Kashef, and K. Obermayer, "Robust detection of environmental sounds in binaural auditory scenes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1344–1356, 2017.

[12] F. Keyrouz and K. Diepold, "An enhanced binaural 3D sound localization algorithm," in *IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 662–665.

[13] T. Rodemann, "A study on distance estimation in binaural sound localization," *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 425–430, 2010.

[14] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, 2012.

[15] I. Trowitzsch, C. Schymura, D. Kolossa, and K. Obermayer, "Joining sound event detection and localization through spatial segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 487–502, 2020.

[16] M. Yiwere and E. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," in *International Journal of Applied Engineering Research*, 2017.

[17] J. Ding, Y. Ke, L. Cheng, C. Zheng, and X. Li, "Joint estimation of binaural distance and azimuth by exploiting deep neural networks." *The Journal of the Acoustical Society of America*, vol. 147, p. 2625, 2020.

[18] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[19] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial audio feature discovery with convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6797–6801.

[20] C. Pang, H. Liu, and X. Li, "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access*, vol. 7, pp. 40 725–40 737, 2019.

[21] M. Zohourian and R. Martin, "Binaural direct-to-reverberant energy ratio and speaker distance estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 92–104, 2020.

[22] ——, "Direct-to-reverberant energy ratio estimation based on interaural coherence and a joint ITD/ILD model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 611–615.

[23] D. Krause, A. Politis, and K. Kowalczyk, "Feature overview for joint modeling of sound event detection and localization using a microphone array," in *28th European Signal Processing Conference (EUSIPCO)*, 2020, pp. 31–35.

[24] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 2018, pp. 1–5.

[25] S. Ruder, "An overview of gradient descent optimization algorithms," in *arXiv:1609.04747*, 2017.

[26] F. Chollet *et al.*, "Keras," GitHub, Web Download: https://github.com/keras-team/keras, 2015.

[27] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, *NIGENS general sound events database*. Zenodo, 2019.

[28] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[29] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 85–92.

[30] J. B. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1976.

[31] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1124–1132, 2009.