TWO VS. FOUR-CHANNEL SOUND EVENT LOCALIZATION AND DETECTION

Julia Wilkins[1], *Magdalena Fuentes*[1], *Luca Bondi*[2], *Shabnam Ghaffarzadegan*[2], *Ali Abavisani*[2], *Juan Pablo Bello* [1]

[1] New York University, New York, NY, USA,
[2] Bosch Research, Pittsburgh, PA, USA

jw3596@nyu.edu

https://github.com/juliawilkins/SELD-2v4-DCASE23/

ABSTRACT

Sound event localization and detection (SELD) systems estimate both the direction-of-arrival (DOA) and class of sound sources over time. In the DCASE 2022 SELD Challenge (Task 3), models are designed to operate in a 4-channel setting. While beneficial to further the development of SELD systems using a multichannel recording setup such as first-order Ambisonics (FOA), most consumer electronics devices rarely are able to record using more than two channels. For this reason, in this work we investigate the performance of the DCASE 2022 SELD baseline model using three audio input representations: FOA, binaural, and stereo. We perform a novel comparative analysis illustrating the effect of these audio input representations on SELD performance. Crucially, we show that binaural and stereo (i.e. 2-channel) audio-based SELD models are still able to localize and detect sound sources *laterally* quite well, despite overall performance degrading as less audio information is provided. Further, we segment our analysis by scenes containing varying degrees of sound source polyphony to better understand the effect of audio input representation on localization and detection performance as scene conditions become increasingly complex.

Index Terms— sound event localization and detection, sound source localization, spatial audio, explainability

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is the process of estimating the direction-of-arrival (DOA) and class of sound events over time, given an input audio signal. SELD systems can translate well to a variety of real-world applications, including navigation for autonomous systems and assistive robotic devices. SELD methods are rooted in traditional signal processing techniques for multichannel audio processing, such as Steered Response Power [1] and acoustic intensity vectors [2]. For human-inspired audio recordings (e.g. binaural recordings), interaural time difference (ITD) and interaural level difference (ILD) are commonly used to characterize the direction of arrival of sounds [3]. However, these cues alone have shown limitations in terms of localization accuracy in real-world scenes that are particularly noisy, reverberant, or polyphonic [4–6]. Deep learning approaches were recently popularized to address these challenges in the context of SELD tasks [7–11]; most systems still utilize signal processing-based features like generalized cross correlation (GCC) and Mel spectrograms but benefit from automatic feature learning to improve robustness in difficult scene conditions [7, 11–13]. For example, in [14], authors use a CRNN architecture with magnitude and phase spectrograms from multichannel audio to show accurate DOA estimation and multiple sound source detection in reverberant conditions.

In the DCASE 2022 SELD challenge (Task 3), models were evaluated using real multichannel sound recordings. Participants had access to real recordings for development and could also use additional synthetic or real data for training. The challenge operates in a multichannel setting, utilizing two formats of 4-channel recordings: first-order Ambisonics (FOA) and a tetrahedral mic array. We are interested in exploring the capabilities of current SELD systems using more commonly found 2-channel microphone setups, namely binaural and stereo, as typical consumer electronics devices lack such complex 4-channel configurations.

There is little prior research quantifying the effect of using various audio input representations (i.e. 2 vs. 4-channel audio) for SELD tasks in deep learning-based systems. In the psychoacoustics community, this effect is well-studied; it is known that there is a general loss in spatial understanding between 4-channel audio configurations (e.g. Ambisonics) vs. 2-channel configurations (e.g. binaural or stereo). [15,16]. Humans can localize lateral sound sources well in binaural and stereo settings, but front-back confusion may increase without sufficient spatial information [3, 17, 18]. Further, perceiving the elevation of sound sources when listening to stereo audio in particular has been shown to be very difficult, largely due to the lack of interaural cues present in this recording configuration unlike that of a binaural setup [16]. However, these phenomena are underexplored in the context of deep learning-based systems for SELD. In [19], authors compared sound event detection performance using synthetic FOA, binaural, and monaural audio data in a CRNN-based system. Our approach differs significantly in that we provide a quantitative analysis of localization *and* detection performance, we use a FOA dataset of real recordings in addition to synthetic and decode these recordings to binaural, and lastly we include the stereo audio configuration as a point of comparison as this is common in consumer electronics devices today.

In this work we present a novel comparative analysis of the DCASE 2022 SELD baseline model across FOA, binaural, and stereo audio input representations. To the best of our knowledge, this is the first work quantifying the effect of these audio configurations on both localization and detection performance in a deep-learning based SELD system. We show that lateral sound source localization remains fairly accurate in the 2-channel settings despite an overall degradation in SELD performance, and provide an analysis of performance in scenes of varying levels of polyphonic sound source complexity.

2. PROBLEM FORMULATION

In this manuscript, we examine the problem of Sound Event Localization and Detection (SELD) under different audio input representations: first-order Ambisonics (FOA), binaural, and stereo record-

ings. In this context, *detection* refers to determining the number of active sound sources per class over time, while *localization* aims at identifying the azimuth and elevation angle for each of the active sources over time. While Ambisonics recordings provide state-of-the-art performance in SELD [20], in practical applications we hypothesize that binaural and stereo recordings are more accessible.

We rely on the most popular framework used by participants in the DCASE 2022 Challenge Task 3. A multichannel audio recording is fed as input to a Convolutional Neural Network (CNN), whose output is a 4-dimensional matrix arranged according to the Multi-Activity Coupled Cartesian DOA (ACCDOA) format [21]. For a given class, time instant, and sound source index, the model arrives at a three-dimensional vector (x, y, z) whose orientation represents the direction of arrival of the sound, and whose intensity is directly proportional to the likelihood of a sound of that class being present at a given time.

First-order Ambisonics (FOA): FOA is a 4-channel, 3D audio recording format. In FOA, each channel corresponds to a spherical harmonic component representing a change in sound pressure in a specific direction [22]. The channels *W, Y, Z, X* map to the omnidirectional, left-right, vertical, and front-back directions of sound pressure change, respectively.

Binaural: The binaural recording technique aims to capture 3D audio in just two channels, ideally simulating the experience of a human experimncing auditory cues. Binaural audio is typically recorded using two microphones placed in the ears of a dummy head (e.g. Neumann KU100), or synthesized using the head-related transfer functions (HRTFs) of such a dummy head [23]. Binaural recordings deliver immersive spatial sounds containing amplitude, time and timbral differences of two channels vs. traditional stereo recordings where only amplitude and time differences are available.

Stereo: In stereo recordings, two microphones are used to capture the left and right audio channels independently. This differs from binaural recordings; in the binaural configuration the goal is to simulate a human's listening experience. Critically, in a stereo setup, elevation differentiation cannot be perceived; binaural recordings contain the filtering effect of the head, ear pinna, and torso and this is not present in a stereo recording configuration [16].

3. EXPERIMENTAL SETUP

3.1. Datasets

Following the setup of the DCASE 2022 Task 3 challenge, we rely on the STARSS22 dataset [24], together with a synthetic mixture (SYNMIX) for baseline training[1] provided by the organizers of the challenge. The STARSS22 dataset is comprised of 121 recordings of various lengths of real sound scenes across 13 sound event classes, with around 5 hours of audio recordings in 4-channel FOA format and an interpolated tetrahedral microphone array. At the time of this work, the evaluation set was not yet released, so we use the "development" partition of train and test, consisting of 67 and 54 recordings, respectively. The dataset contains instances with up to 5 simultaneous sound sources, and up to 4 simultaneous sources of the same class, though 2-source polyphony is much more frequent.

Due to the small size of the STARSS22 dataset, a base set of synthetic data was also provided to participants (SYNMIX). This data is synthesized using audio samples from FSD50k [25] convolved with Spatial Room Impulse Responses from the TAU-Nigens Spatial Sound Events 2020 [26] and 2021 [27]. The dataset contains 1200, 1-minute synthesized FOA recordings across classes mapped to the classes present in STARSS22, and maximum polyphony of 2 sources.

Both datasets are annotated at 100ms resolution with labels of sound source class, azimuth, and elevation as well as additional flags for overlapping sound events. The azimuth angles $\phi \in [-180°, 180°]$, and elevation $\theta \in [-90°, 90°]$, with $0°$ at front. Note that azimuth angles increase counterclockwise.

3.2. Input representations

To fairly compare the three multichannel audio representations, we look at the problem of sound localization on the horizontal plane only by removing the elevation component, thus fixing elevation to $0°$ in the ground truth. We train and test separately for each input representation using the same acoustic scenes, simply replacing the original FOA audio representation with binaural or stereo audio, as per following procedures.

FOA → Binaural: To decode the original FOA audio from the STARRS22 and synthetic datasets to binaural, we used the BinauralDecoder plug-in from the IEM Plug-In Suite[2]. This decoder uses pre-processed Neumann KU100 dummy head HRTFs via the magnitude least-squares (MagLS) method proposed in [28]. We apply this binaural decoding to all FOA audio used in training and testing, yielding 2-channel binaural audio for our experiments [3].

FOA → Stereo: To convert our FOA audio to stereo, we used a very simple translation: *left* $= W + Y$ and *right* $= W - Y$, following [29]. Note that W is the omnidirectional signal and Y is the first-order horizontal (left-right) component. An increase in air pressure from left causes an increase in values of Y and an increase in pressure from the right causes a decrease in values of Y. Because of this, the simple translation above allows us to move easily from FOA to left and right channels yielding 2-channel stereo audio.

3.3. Baseline model

The model used for our analysis is the DCASE 2022 Task 3 Baseline model[4]. The architecture is similar to the CRNN-based model initially proposed in [7], with extensions to accommodate simultaneous sources of the same class in the Multi-ACCDOA format [21]. The input to the model is the multichannel audio, segmented into 5-second chunks, yielding a sequence of 50 x 0.1 second frames. In the FOA configuration, Mel spectrogram features are used to capture frequency information and intensity vectors provide spatial information. In the binaural and stereo settings, we modify the model slightly to use Mel spectrograms and GCC features. GCC features are commonly used in 2-channel localization settings to capture Time Difference of Arrival (TDOA) information between two microphones. Audio is resampled to 24kHz, and 64 Mel coefficients are computed from an STFT on windows of 1024 samples with a hop size of 480 samples. The model has 604.5K trainable parameters. Models are trained for a multi-output regression task, with a mean-squared-error loss, for 200 epochs using 1 RTX 8000 GPU, in batches of 64 samples with a learning rate of 10^{-3}. The model checkpoint with the lowest validation loss is selected.

3.4. Data augmentation via Audio Channel Swapping (ACS)

An initial exploration of the STARSS and SYNMIX datasets revealed that the distribution of azimuth angles across sound sources

[1] https://zenodo.org/record/6406873#.Y_-SBuzMK2o.

[2] https://plugins.iem.at/docs/plugindescriptions/#binauraldecoder.

[3] https://github.com/juliawilkins/ambisonics2binaural_simple.

[4] https://github.com/sharathadavanne/seld-dcase2022.

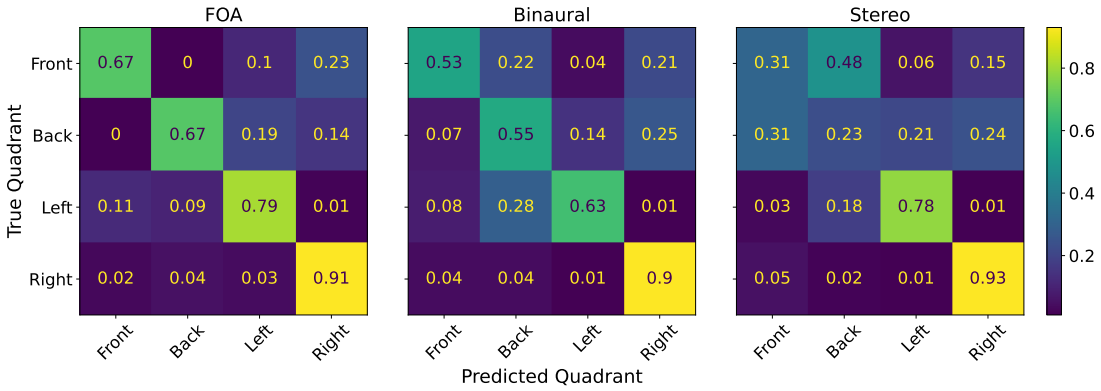

Figure 1: Normalized confusion matrices showing true vs. predicted quadrant of sources across audio configurations. The FOA model performs near-perfect at distinguishing front and back sources, while front and back sources are commonly confused in binaural and stereo settings. Quadrants of size 90° are defined based on the azimuth angle of a sound source: Front $\in [-45^\circ, 45^\circ]$, Left $\in [45^\circ, 135^\circ]$, Back $\in [135^\circ, \pm 180^\circ] \cup [\pm 180^\circ, -135^\circ]$, Right $\in [-135^\circ, -45^\circ]$

was largely imbalanced, with far more sound sources in the front and right regions than in the left and back. Following [30], we hypothesize that localization performance on the real test dataset could be improved by balancing this distribution. To do so, we use a data augmentation technique known as Audio Channel Swapping (ACS) [31]. We perform 3 transformations involving azimuth to simulate the rotation of sound sources by 90°, 180°, and 270°. We performed different permutations of swapping and negating the X and Y of FOA channels directly. This simple augmentation strategy not only quadruples our overall dataset size but more importantly gives us a uniform distribution of azimuth angles. We show that this augmentation has a significant impact on localization performance in Table 1. Please refer to [31] for more details on ACS.

3.5. Evaluation metrics

We use the joint localization and detection metrics as defined by the DCASE 2022 Task 3 SELD Challenge in our proceeding analysis. The F-Score and error rate (ER) capture location-dependent detection. True Positives (TP) and False Positives (FP) are considered with a tolerance 20° in the direction of arrival. Class-dependent localization error (LE) and localization recall (LR) measure localization performance without considering the spatial threshold. See [32] for more details on SELD metrics.

4. RESULTS

4.1. A baseline model for FOA input

Prior to evaluating the impact of different input representations, we first assess the performance of a baseline model trained and evaluated on FOA input using varied training data configurations. The STARSS22 and SYNMIX dataset are both quite imbalanced in terms of distribution of sound source across azimuth angles. As described in Section 3.4, we use Audio Channel Swapping (ACS) to mitigate this problem and balance the distribution at train time.

Table 1 reports results for 5 training data configurations: **A**: training and evaluating only in azimuth using STARSS22 dataset; **B**: adding SYNMIX dataset to A's training; **C**: adding ACS augmentation to B's training, $\mathbf{B}^{+E}$: training and evaluating B in both azimuth and elevation; $\mathbf{C}^{+E}$: training and evaluating C in both azimuth and elevation. Note that B^{+E} and C^{+E} help us to understand the impact of removing elevation in the overall metrics. By comparing C^{+E} and C, we see how removing elevation improves all metrics, as one could imagine given less degree of freedom in the predictions. Moreover, we see an improvement in the joint localization and detection metrics across the board with the addition of the augmented data. Hence, we use **C** as our reference configuration to assess the impact of the input representation in proceeding sections.

Conf.	SELD ↓	ER↓	F ↑	LE ↓	LR ↑
A	0.65	0.73	15.3%	53.7°	27%
B	0.47	0.62	34.5%	22.5°	51%
C	0.42	0.56	43.3%	16.9°	54.1%
B^{+E}	0.53	0.70	27.3%	26.1°	47.5%
C^{+E}	0.48	0.62	33%	22.7°	51%

Table 1: Results with **FOA** input across different configurations; A: STARSS22; B: A + SYNMIX; C: B with ACS; B^{+E} and C^{+E}: B and C are trained and evaluated using both azimuth and elevation. Results are reported on the STARSS22 DCASE dev-test set. ↓ indicates metrics that are better when value is lower, ↑ viceversa.

4.2. Comparing audio input representations

Table 2 reports results when changing input representation, moving from the highly-privileged FOA representation, to binaural, and stereo audio. Our experiments show that as one moves from FOA to binaural and stereo, overall SELD model performance degrades. While this is to be expected because binaural and stereo audio are not designed to capture full spatial audio, this is the first quantification of deep learning-based SELD performance across these audio input representations on real multichannel recordings lays the groundwork for our deeper proceeding analysis.

4.3. Localization error by sound source quadrant

We are also interested in dissecting localization performance to understand where key success and failure points occur in terms of sound source position and polyphonic scene conditions.

In Figure 1, we show a set of confusion matrices illustrating the distribution of true quadrants of sound sources vs. predicted quadrants across audio input representations. We segment the 90°

Input	SELD ↓	ER ↓	F ↑	LE ↓	LR↑
FOA	0.42	0.56	43.3%	16.9°	54.1%
Binaural	0.50	0.67	33.9%	30.1°	49.2%
Stereo	0.60	0.76	21.7%	42.9°	38.8%

Table 2: Results for models trained using STARSS22 + SYNMIX using ACS, with different audio input representations. Results are reported on the STARSS22 DCASE development-test set. ↓ indicates metrics that are better when value is lower, ↑ viceversa.

quadrants as follows, based on azimuth angle: Front $\in [-45°, 45°]$, Left $\in [45°, 135°]$, Back $\in [135°, \pm 180°] \cup [\pm 180°, -135°]$, Right $\in [-135°, -45°]$. Notably, using the FOA representation, the model has near-perfect performance in terms of distinguishing front vs. back sources. In the binaural setting, we see an increase in front-back confusion, and in the stereo setting this error is glaring as 48% of sources in the front are predicted in the back quadrant. In fact, this is a well-studied topic in psychoacoustics related to the cone of confusion phenomenon, which occurs when a sound source is equidistant to both the left and right ears [33–35]. Thus, it is difficult for the listener to distinguish whether a sound source is in front or behind them. It is likely that our binaural model is affected by this as well. Across audio input representations, the accuracy of source detection in the left and right quadrants is fairly consistent, showing reliability in terms of lateral sound source detection given 2- or 4-channel audio input.

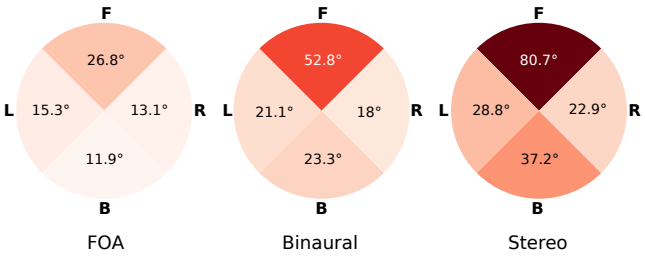

Figure 2: Average localization error across audio representations, based on ground truth sound source quadrant position. Results are normalized by number of instances of sound sources per quadrant.

In Figure 2, we analyze average localization error (LE) based on the quadrant of the ground truth sound sources. In the FOA setting, the difference of LE between the left, right, and back quadrants is quite small, however the error for sources in the front is nearly double that of the other quadrants. In the binaural setting, LE increases in the front and back quadrants, approximately doubling that of the FOA setting, though this increase is much less notable in the lateral (left-right) regions. Further, in the stereo context, we find similar trends but with overall poorer performance. The front and back LE are over three times that of the FOA model, with less significant degradation in the performance of the left and right quadrants. Here, we crucially observe that despite the binaural and stereo models struggling to localize sources in the front quadrant in particular compared to the FOA system, these 2-channel models are still able to localize sources laterally quite well.

4.4. SELD performance in polyphonic conditions

The DCASE SELD challenge is unique in that the test dataset contains real audio recordings with multiple overlapping sound sources. Hence, investigating SELD model performance in complex polyphonic conditions can help us better understand how these systems handle more complex scene conditions that are closer to reality. In Figure 3, we analyze localization recall (LR) of the FOA, binaural, and stereo models in the presence of 1, 2, 3, and 4 simultaneous sources (this encapsulates both simultaneous sources of the same or different classes). Note that approximately 56% of frames contain 1 source, 31% contain 2, 10% contain 3, and 3% contain 4 or more simultaneous sources, so we normalize by source count accordingly in Figure 3. We show that LR steadily decreases in all audio configurations as the number of polyphonic sound sources increases in Figure 3. The model struggles to detect the correct number of sources as the scene conditions become increasingly complex, though proportionally the decrease in recall is relatively similar across audio contexts as polyphony increases.

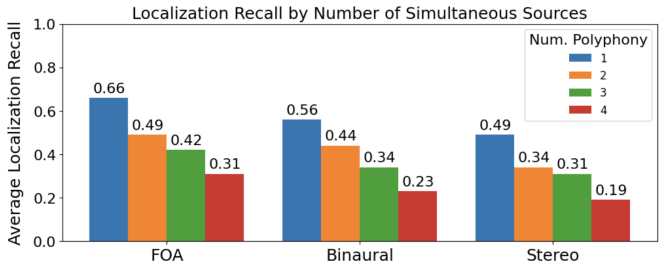

Figure 3: Localization recall in multiple audio representations, segmented by number of simultaneous sources in the test data and normalized by number of sources satisfying each condition.

We also analyze localization error (LE) across polyphonic conditions. Here we find that while on average LE increases as we use less-informative audio representations (i.e. stereo), it is not a fully monotonically increasing trend across polyphonic conditions. In the FOA setting, the LE is similar regardless of level of polyphony. In the binaural and stereo settings, there is a much larger spread of LE across conditions, however not in a monotonically increasing manner, e.g. in the stereo setting the average LE is $31.3°$ in the occurrence of 3 overlapping sources vs. $46.1°$ for 2 sources. We hypothesize that there are many interacting effects contributing to this, including but not limited to class imbalance in different polyphonic conditions, simultaneous sources of the same class, and the nature of the LE metric as it does not take false negatives into account.

5. CONCLUSION

This work presents a novel comparative analysis of the DCASE 2022 SELD baseline model across first-order Ambisonics, binaural, and stereo audio input representations. We show quantitatively that while localization and detection performance decreases given less informative audio representations, binaural and stereo-based SELD models are still able to localize lateral sound sources relatively well. These findings could be highly informative in the development of applications such as an audio-visual navigation system equipped with a stereo microphone configuration and a camera; if we are confident in lateral source localization based on auditory cues, we can lean more on visual cues for sources directly in front of the camera. Future work in this space could entail an investigation into the effect of sound source class or of overlapping sources of the same class on localization performance across polyphonic conditions and audio input representations.

6. REFERENCES

[1] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University, 2000.

[2] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 241–245.

[3] R. Stern, D. Wang, and G. Brown, "Binaural sound localization-chapter 5 from computational auditory scene analysis," 2006.

[4] C. Giguère and S. M. Abel, "Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 769–776, 1993.

[5] S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1956–1968, 2017.

[6] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 258–262.

[7] 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, pp. 34–48, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[8] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[9] F. Zhao, R. Li, and D. Pan, "Deep learning for binaural sound source localization with low signal-to-noise ratio," *Journal of Physics: Conference Series*, vol. 1828, p. 012017, 02 2021.

[10] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[11] P.-A. Grumiaux, S. Kiti'c, L. Girin, and A. Gu'erin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152 1, p. 107, 2021.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[13] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[14] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[15] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. ii: Psychophysical validation," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 868–878, 1989.

[16] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[17] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney, "Auditory localization in low-bitrate compressed ambisonic scenes," *Applied Sciences*, vol. 9, no. 13, p. 2618, 2019.

[18] L. Thresh, C. Armstrong, and G. Kearney, "A direct comparison of localization performance when using first, third, and fifth ambisonics order for real loudspeaker and virtual loudspeaker rendering," in *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.

[19] S. Adavanne, A. Politis, and T. Virtanen, "Multichannel sound event detection using 3d convolutional neural networks for learning inter-channel features," in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[20] N. Poschadel, S. Preihs, and J. Peissig, "Multi-source direction of arrival estimation of noisy speech using convolutional recurrent neural networks with higher-order ambisonics signals," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1015–1019.

[21] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022.

[22] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the audio engineering society*, vol. 21, no. 1, pp. 2–10, 1973.

[23] I. Engel, D. F. Goodman, and L. Picinali, "Assessing hrtf preprocessing methods for ambisonics rendering through perceptual models," *Acta Acustica*, vol. 6, p. 4, 2022.

[24] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," 2022. [Online]. Available: https://arxiv.org/abs/2206.01948

[25] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[26] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[27] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," 2021. [Online]. Available: https://arxiv.org/abs/2106.06999

[28] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Proceedings of the DAGA*, vol. 44, 2018, pp. 339–342.

[29] F. Zotter and M. Frank, *XY, MS, and First-Order Ambisonics*. Cham: Springer International Publishing, 2019, pp. 1–22. [Online]. Available: https://doi.org/10.1007/978-3-030-17207-7_1

[30] Q. Wang, L. Chai, H. Wu, Z. Nian, S. Niu, S. Zheng, Y. Wang, L. Sun, Y. Fang, J. Pan, J. Du, and C.-H. Lee, "The nerc-slip system for sound event localization and detection of dcase2022 challenge," DCASE2022 Challenge, Tech. Rep., June 2022.

[31] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," 2021. [Online]. Available: https://arxiv.org/abs/2101.02919

[32] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9306885

[33] L. Rayleigh, "Xii. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[34] O. Balan, A. Moldoveanu, and F. Moldoveanu, "A systematic review of the methods and experiments aimed to reduce front-back confusions in the free-field and virtual auditory environments." *RoCHI*, pp. 24–29, 2018.

[35] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, 2000.