Region-Specific Audio Tagging for Spatial Sound

Jinzheng Zhao[1], *Yong Xu*[2], *Haohe Liu*[1], *Davide Berghi*[1], *Xinyuan Qian*[3], *Qiuqiang Kong*[4], *Junqi Zhao*[1], *Mark D. Plumbley*[1], *Wenwu Wang*[1]

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2]Tencent AI Lab, Bellevue, WA, USA
[3]Department of Computer Science and Technology, University of Science and Technology Beijing, China
[4]The Chinese University of Hong Kong (CUHK)

***Abstract*—Audio tagging aims to label sound events appearing in an audio recording. In this paper, we propose region-specific audio tagging, a new task which labels sound events in a given region for spatial audio recorded by a microphone array. The region can be specified as an angular space or a distance from the microphone. We first study the performance of different combinations of spectral, spatial, and position features. Then we extend state-of-the-art audio tagging systems such as pre-trained audio neural networks (PANNs) and audio spectrogram transformer (AST) to the proposed region-specific audio tagging task. Experimental results on both the simulated and the real datasets show the feasibility of the proposed task and the effectiveness of the proposed method. Further experiments show that incorporating the directional features is beneficial for omnidirectional tagging.**

1. INTRODUCTION

The task of audio tagging aims to identify the sound events present in a sound clip. This topic has been a popular area in audio processing, playing an important role in applications, such as audio classification [1] and information retrieval [2]. In the current task setting, in general, all sound events are labeled regardless of the location of the sound events. In surveillance applications, however, sound events located in a specific spatial region may be more important than others. Therefore, it is of practical interest to study the problem of region-specific audio tagging, i.e., labelling audio events that are present in a specific spatial region. Tagging sound events according to the location can help the separation and make people focus on sound from a specific region. In addition, region-specific audio tagging allows people to attach importance to sounds, e.g. a warning from behind, thereby improving the safety.

In the traditional audio tagging task, deep learning based methods are popular choices. For a convolutional neural network (CNN) based system, pretrained audio neural networks (PANNs) [3] is a model pretrained on AudioSet [4], and shows promising performance on audio pattern recognition tasks such as audio tagging and acoustic scene classification. The system of pretraining, sampling, labeling, and aggregation (PSLA) [5] employs EfficientNet as the backbone and improves the model performance by ImageNet pretraining, balanced sampling and model aggregation. For a transformer-based [6] system, the audio spectrogram transformer (AST) [7] follows the vision transformer [8], and takes mel-spectrogram patches as input. The AST model can outperform PANNs and PSLA, when built with a large amount of data. The above methods only use spectral features. In [9], both spectral and spatial audio features are used with a gated CNN architecture. The incorporation of audio features further improves the model performance. In addition to the CNN and transformer-based methods, there are also methods based on graph neural network (GNN), such as the work proposed in [10], which leverages several GNNs to model the relationships between different feature patches and different labels for tagging. In our work, we extend these state-of-the-art models to the region-specific audio tagging problem.

This research was supported by Tencent AI Lab Rhino-Bird Gift Fund and University of Surrey. This work was also supported by the Engineering and Physical Sciences Research Council [grant numbers EP/T019751/1, EP/Y028805/1]. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. The codes and dataset are available at https://github.com/KawhiZhao/AudioTagging.

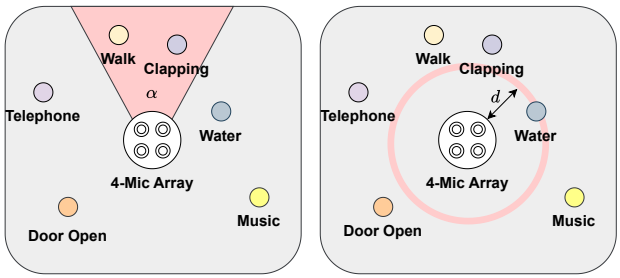

Fig. 1: The task scenario of region-specific audio tagging. Left: Query by horizontal angular regions. Right: Query by distance.

Our proposed task is inspired by region-specific speech processing [11]–[13] including automatic speech recognition (ASR) and speech separation for a given direction or distance. Beamforming aims to locate the signal from a given direction by attenuating unwanted sound sources. Traditional methods like the minimum variance distortionless response (MVDR) [14] beamformer can minimize the total signal power and maintain a distortionless response for a given direction. Recently, deep learning based methods are also used for region-specific automatic speech recognition and separation. For region-specific ASR, in [11], long short term memory (LSTM)-based architecture is used for multi-channel overlapped ASR with the input of the concatenation of spectral, spatial and angle features. Directional features are proposed as the angle features, encoding the information for regions of interest. For region-specific speech separation, in [12], directional features are used as a condition to indicate the speaker for multi-modal target speech separation. In [13], a multi-channel band-split recurrent neural network (RNN) model is proposed for angular-query, spherical-query and conical-query based speech separation. In addition, the field of view (FOV) feature is proposed in [15] for audio zooming. The FOV feature can represent the property of an angular space while directional features can represent the property of an azimuth.

2. PROPOSED METHOD

Inspired by the previous work in region-specific speech processing, we study a new task, i.e. region-specific audio tagging, and explore the use of directional features and FOV features for this task. Following previous work [13], we explore this task in two paradigms, query by horizontal angular regions and query by distance, as shown in Fig. 1. We examine three types of features, i.e. spectral, spatial, and

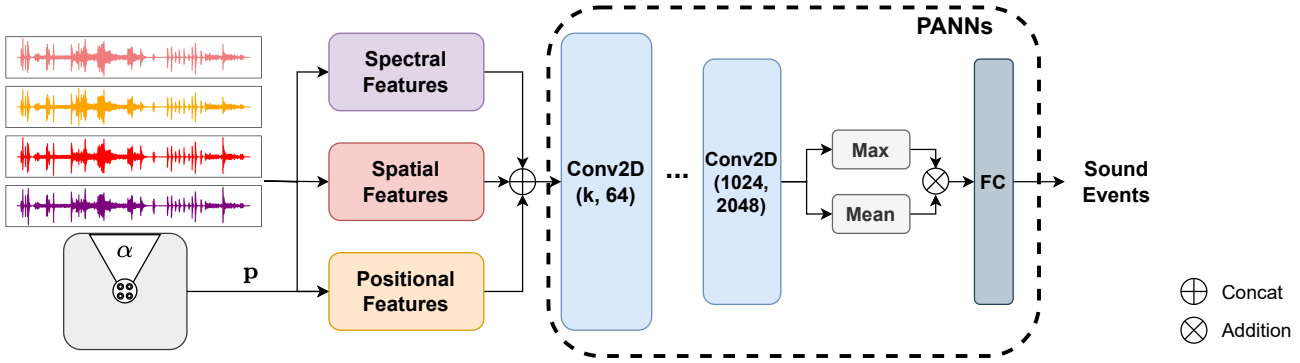

Fig. 2: The proposed multi-channel PANNs for region-specific audio tagging.

positional features. The spectral features are used to identify sound sources. Spatial features can give information for all potential sound events and positional features are used to describe a potential region-specific sound event.

Our contributions can be summarized as follows. Firstly, we propose, for the first time, the region-specific audio tagging task and build a benchmark. Secondly, we extend the state-of-the-art audio tagging models for this task and study the performance achieved by a different combination of spectral, spatial and positional features. Finally, we extend the proposed method to omnidirectional audio tagging, which aims to tag the events in the whole space instead of a region. Experimental results show that the proposed fixed-region and location-aware system are superior to the omnidirectional tagging systems.

2.1. Problem Definition

Given a four-channel audio in tetrahedral microphone array (MIC) format $\mathbf{x} \in \mathbb{R}^{4\times L}$ (where L denotes the audio length) and a given area $\mathbf{p}$, in the form of angular range $[\theta_{\text{begin}}, ..., \theta_{\text{end}}]$ or distance d from the microphone array to the sound events, the objective of region-specific audio tagging is to identify the sound sources located in $\mathbf{p}$. It is worth noting that the proposed new task can be achieved by a beamformer and an audio tagging model. However, this cascaded format is not end-to-end and the audio tagging model can only learns the semantic representation.

2.2. Overall System Architecture

To this end, we present a system, as shown in Fig. 2. The system is composed of the modules including feature extraction and audio tagging. The features indicate the region of interests while the tagging model predicts the sound events. Compared to previous tagging systems, the proposed system predicts the sound events in the given region specified by a user, instead of all events present in microphone recordings. Compared to the cascaded model combining a beamformer and an audio tagging model, the proposed model can learn both semantic and spatial information simultaneously.

2.3. Extraction of Features

Following previous work for region-specific speech separation [12], [13], we also use the combination of spectral, spatial and positional features.

2.3.1. Spectral Features: Spectral features can help classify the audio events. Here we use logarithm power spectrum (LPS) $\mathbf{L}$ of the first-channel audio $\mathbf{x}$. $\mathbf{L}$ is calculated based on short-time Fourier transform (STFT) $\mathbf{X}$ of $\mathbf{x}$, as follows,

$$\mathbf{L} = \log(|\mathbf{X}|^2) \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{T\times F}$ and $\mathbf{L} \in \mathbb{R}^{T\times F}$, and T and F are the number of time frames and frequency bins, respectively.

2.3.2. Spatial Features: Spatial features are useful for localizing sound events [16]. Here, we explore two features, generalized cross-correlation phase transform (GCCPHAT) $\mathbf{G}$ and inter-channel phase difference (IPD) $\mathbf{I}$.

GCCPHAT is widely used in sound source localization [17] and speaker tracking [18], which estimates the time difference of arrival between a pair of microphones, as follows,

$$\mathbf{G}^n(\tau) = \int_{-\infty}^{+\infty} \frac{\mathbf{X}^{n_1}(f)\mathbf{X}^{n_2*}(f)}{|\mathbf{X}^{n_1}(f)\mathbf{X}^{n_2*}(f)|} e^{i2\pi f\tau} df \tag{2}$$

where $n = (n_1, n_2)$ indexes the microphone pair, $*$ denotes complex conjugate, f denotes the frequency, and τ denotes the time delay. It is normalized to mitigate the impact of changing amplitudes. We stack GCCPHAT from selected pairs as the spatial features.

IPD is calculated as the phase difference between two audio channels.

$$\mathbf{I}^n = \text{angle}(\mathbf{X}^{n_1}) - \text{angle}(\mathbf{X}^{n_2}) \tag{3}$$

where $\text{angle}(\cdot)$ denotes the phase of a signal. IPD shows effectiveness in multi-channel speech separation [12]. IPDs from selected microphone pairs are stacked.

2.3.3. Positional Features: In this section, we discuss the extraction of positional features from the angular region or a distance, respectively.

Angular Region For the horizontal angular areas from -180° to 180°, we detect the sound events within the region $\Theta = [\theta_{\text{begin}}, ..., \theta_{\text{end}}]$ by attenuating events from the unselected areas. We choose the FOV feature proposed in [15] for this task.

To calculate the FOV feature, we first calculate directional features (DF) $\mathbf{D} \in \mathbb{R}^{T\times F}$ for each angle in Θ following [19], as follows,

$$\mathbf{D}(\theta) = \sum_n \cos(\mathbf{I}^n - \mathbf{P}^n(\theta)) \tag{4}$$

$$\mathbf{P}^n(\theta) = 2\pi f \phi^n \cos(\theta) f_s / c \tag{5}$$

where $\mathbf{P}^n$ is the target-dependent phase difference calculated at the n-th microphone pair with frequency f, ϕ^n is the distance between the n-th microphone pair, f_s is the sampling rate, and c is the sound velocity. Then, the FOV feature in the view of Θ is calculated as:

$$\mathbf{F}_{\text{in}} = \max(\mathbf{D}(\theta)), \theta \in \Theta \tag{6}$$

The feature out of the view Θ is calculated as:

$$\mathbf{F}_{\text{out}} = \max(\mathbf{D}(\theta)), \theta \notin \Theta \tag{7}$$

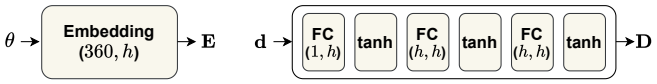

Fig. 3: The learning-based method for obtaining angle features (left) and distance features (right), where h is the hidden dimension and FC denotes the fully connected layer.

Finally, the FOV feature $\mathbf{F} \in \mathbb{R}^{T*F}$ is defined as:

$$\mathbf{F} = \begin{cases} \mathbf{F}_{\text{in}}, & \text{if } \mathbf{F}_{\text{in}} > \mathbf{F}_{\text{out}}, \\ -1, & \text{if } \mathbf{F}_{\text{in}} \leq \mathbf{F}_{\text{out}}. \end{cases} \tag{8}$$

Distance For obtaining a distance feature given a distance d from the microphone array to the sound events, we use learning-based methods following [13], as shown in the right part of Fig. 3. The distance feature is repeated to match the time dimension of spectral or spatial features.

2.4. Model

We extended PANNs for region-specific multi-channel audio tagging, as shown in Fig. 2. The spectral, spatial and positional features are concatenated on the channel dimension, with k being the number of concatenated channels. PANNs takes the concatenated features and employs a sequence of convolutional layers to downsize the time-frequency dimension while increasing the channel dimension from k to 2048. Then the time-frequency dimension is pooled and the channel dimension is retained. Finally, the system outputs the probabilities of the sound event classes. The binary cross-entropy loss is used to optimize the model.

We also extended the other two models, PSLA [5] and AST [7], to the proposed task. For PSLA, the EfficientNet backbone is trained from scratch with 2560 hidden dimension in the multi-head attention module. For AST, we use the *small224*[1] version.

3. EXPERIMENT SETUP

3.1. Dataset

We use the STARSS23 dataset [20] for the proposed task and this dataset is also used in DCASE 2024 task 3. However, the scenario presented in this dataset is relatively simple for region-specific audio tagging, since the average number of overlapping sound events is 1.32. Thus, we use SpatialScaper [21] to simulate a new dataset, named Spatial Region-Specific Audio Tagging (SRSAT). The room is randomly chosen from the given room set[2]. The mean and standard deviation of the number of sound events are set to 25 and 3. Each audio clip is in the format of a tetrahedral microphone array, 60 seconds long, sampled at 24 kHz. We generate 1,000 audio clips for training, 60 audio clips for validation, and 60 audio clips for testing, respectively. There are 13 sound classes (female speech, male speech, clapping, telephone, laughter, domestic sounds, walk, door, music, musical instrument, water tap, bell and knock) from FSD50K [22] in both the STARSS23 and SRSAT datasets. In each time step, the annotation contains the classes and positions (azimuth, elevation, and distance) of the sound events.

3.2. Implementation Details and Evaluation Metrics

A 2-second audio clip is randomly extracted from the 60-second audio signals. We ensure that the extracted audio clip has at least one sound event. During training, we perform audio channel swapping data augmentation using the method proposed in [23] to enlarge the dataset by a factor of eight. STFT is calculated with a window of 512 samples and a hop size of 256 samples. Both IPD and directional features are calculated with microphone pairs of $(0, 0), (0, 1), (0, 2)$ and $(0, 3)$ [11]. We use the CNN-14 variant of PANNs as the model backbone. Each model is trained using the Adam optimizer with a learning rate 10^{-5}. The model is trained for 50 epochs with an early stop mechanism at a patience of 10 epochs. Apart from the features mentioned above, we explore the learning-based angular features shown in the left part of Fig. 3. A learnable embedding layer is used to extract features E for different input angles which is jointly trained with PANNs. In addition, for spectral and spatial features, we explore spatial cue-augmented log-spectrogram features (SALSA) proposed in [24]. We use the angular region of 60° as the region of interest. For DF and learning-based features, we use the middle angle of the region to extract the features. For the FOV feature, we calculate each DF in a resolution of 5°. Following the previous work [7], [9], we use both mean average precision (mAP) and equal error rate (EER) as the metric to evaluate the model performance.

[1]https://github.com/YuanGongND/ast/blob/master/src/models/ast_models.py
[2]https://github.com/iranroman/SpatialScaper

Table 1: Experimental results for different features. Bold numbers indicate the best performance.

Spectral	Spatial	Angle	mAP (↑)	EER (↓)
LPS	IPD	DF	0.473	0.253
LPS	IPD	FOV	**0.485**	0.260
LPS	IPD	learned	0.416	0.260
LPS	GCCPHAT	DF	0.479	**0.246**
SALSA	SALSA	DF	0.455	0.260

Table 2: Experimental results for different models.

Model	# Params	mAP (↑)	EER (↓)
PSLA [5]	64.1M	0.384	0.283
AST [7]	22.8M	0.360	0.308
PANNs [3]	79.7M	**0.473**	**0.253**

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1. Impact of Features

We show the experimental results on SRSAT using different feature combinations for 60-degree angular region-specific audio tagging in Table 1. Experimental results show that using LPS and IPD with FOV features achieves the best performance of mAP. One possible reason is that FOV features can focus more on the desired area than other angle features, as DF and learning-based can only attend to a single azimuth. However, the FOV feature is more computationally expensive than DF. With a resolution of 5°, the FOV features require 72 times computational costs than DF. Thus, we used DF for the remaining experiments, given its competitive performance over the FOV feature. For other spatial features, GCCPHAT has competitive performance with IPD.

4.2. Impact of Models

The experimental results for different models are demonstrated in Table 2 and it shows that PANNs achieve the best performance. AST and PSLA do not perform well giving lower mAP and higher EER than PANNs, which shows that PANNs can be adapted from traditional audio tagging to region-specific audio tagging effectively. The potential reason of AST not performing good is that the transformer-based method generally requires a large amount of data for good results [8].

4.3. Compared with Omnidirectional Tagging Systems

We compare the region-specific tagging systems with the systems for omnidirectional audio tagging, i.e., tagging sound events in the whole

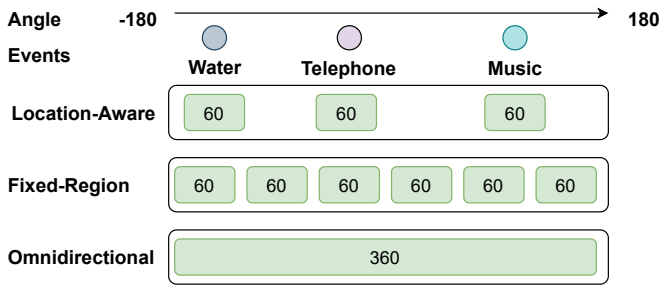

Fig. 4: Illustration of different tagging system. The numbers are in degrees.

Table 3: Experimental results for omnidirectional (OD), fixed-region (FR) and location-aware (LA) audio tagging systems.

Spectral	Spatial	Angle	System	mAP (↑)	EER (↓)
LPS	IPD	-	OD	0.371	0.267
LPS	IPD	FOV	OD	0.370	0.274
LPS	GCCPHAT	-	OD	0.376	0.263
SALSA	SALSA	-	OD	0.373	0.265
LPS	IPD	DF	FR	0.653	0.252
LPS	IPD	DF	LA	**0.667**	**0.244**

space instead of a specific region. We explore three systems as shown in Fig. 4. The omnidirectional audio tagging system does not use the angle information. Both the location-aware system and the fixed-region system are based on a trained region-specific system. The location-aware system has the prior knowledge of the locations of all sound sources. The system is operated in the region of 60° centered at the azimuth of the sound events. We filter out overlapped regions, which are defined as two regions overlapping by 30°, to reduce unnecessary computational cost. The fixed-region tagging system is operated on six fixed regions from -180° to 180° at an interval of 60°. For the location-aware and fixed-region systems, the system output and the ground truth from different regions are aggregated by the operation of maximum before calculating the metrics.

The experimental results are shown in Table 3. Similar to the previous experiments, we compare the model performance achieved with different feature combinations for omnidirectional audio tagging. All omnidirectional systems have mAP between 0.37 and 0.38. It can be observed that adding the angle feature does not improve the model performance, as indicated by the third system. It is demonstrated that the fixed-region system outperforms the omnidirectional tagging one, which shows that the fixed-region system benefits from the region division. If the positions of the sound events are known, the model performance can be further improved with mAP from 0.653 to 0.667 and EER from 0.252 to 0.244.

4.4. Impact of Angular Range

We explore the influence of different angular range and show the results in Table 4. We can see that the proposed task becomes harder as the angular range becoming larger as more sound events would fall into the selected area, making the task more challenging. When the angular range is 300°, mAP is 0.370, which is close to the model performance in the omnidirectional tagging scenario shown in Table 3.

Table 4: Impact of angular range.

Spectral	Spatial	Angle	Range	mAP (↑)	EER (↓)
LPS	IPD	DF	60°	**0.473**	**0.253**
LPS	IPD	DF	180°	0.406	0.266
LPS	IPD	DF	300°	0.370	0.275

Table 5: Dataset statistics and performance comparisons between STARSS23 and SRSAT dataset.

	STARSS23	SRSAT
Avg. No. Events	1.320	2.233
Max. No. Events	6	10
Prop. One Event	0.623	0.287
Prop. Two Events	0.278	0.294
mAP of PANNs (↑)	0.938	0.473

Table 6: Model performance for a given distance.

Model	Param	mAP (↑)	EER (↓)
PANNs [3]	79.8M	**0.417**	**0.266**
AST [7]	23.0M	0.324	0.320
PSLA [5]	64.2M	0.399	0.274

4.5. Model Performance on STARSS23 Dataset

We further evaluate PANNs on the STARSS23 dataset and make comparisons with the model performance on SRSAT. The results are shown in Table 5. STARSS23 was originally used for sound source localization and detection. We can see that STARSS23 is much simpler than SRSAT for the proposed task, as indicated by the lower average number of events (Avg. No. Events) and maximum number of events (Max. No. Events) per frame. We show the proportions of frames containing one event (Prop. One Event) or two events (Prop. Two Events). It is demonstrated that STARSS23 dataset contains one event and two events at 90% of the time frames, which necessitates the creation of SRSAT dataset.

4.6. Model Performance for a Given Distance

In this part we explore the model performance for distance-guided audio tagging. We randomly select a sound event and use the ground truth distance as the condition. The performance of the model is reported in Table 6. For each model, we use LPS as the spectral feature and IPD as the spatial feature following the previous settings. The distance feature extraction network is jointly trained with the model. The PANNs model achieves the best performance while AST and PSLA do not perform well. Compared with the model performance on azimuth-queried tagging reported in Table 2, distance-queried tagging scenarios are more challenging. One possible reason is that the statistic-based DF reflects the confidence of the sound source coming from a given azimuth, which can capture the characteristics of the selected regions better than the learning-based distance feature.

5. CONCLUSION

We have presented a novel task named region-specific audio tagging for spatial sound. We extended the current advanced tagging models for this task using angular region-queried and distance-queried mode. We further explored the impact of different combinations of spectral, spatial and positional features. We find that using LPS and IPD combined with FOV features achieves the best mAP. When extending the region-specific tagging system to omnidirectional tagging, we find that the proposed fixed-regional and location-aware tagging system outperforms the omnidirectional tagging system. In the current setting, we only include 13 sound classes following the setting of DCASE Task 3. In future, we plan to include more sound classes from AudioSet to enhance the diversity of dataset.

REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[2] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[5] Y. Gong, Y.-A. Chung, and J. Glass, "PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3461–3466.

[10] S. Singh, C. J. Steinmetz, E. Benetos, H. Phan, and D. Stowell, "ATGNN: Audio tagging graph neural network," *IEEE Signal Processing Letters*, 2024.

[11] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[12] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[13] R. Gu and Y. Luo, "Rezero: Region-customizable sound extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2576–2589, 2024.

[14] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[15] M. Yu and D. Yu, "Deep audio zooming: Beamwidth-controllable neural beamformer," *arXiv preprint arXiv:2311.13075*, 2023.

[16] D. Berghi and P. J. Jackson, "Audio inputs for active speaker detection and localization via microphone array," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023, pp. 1–5.

[17] J. Zhao, Y. Xu, X. Qian, D. Berghi, P. Wu, M. Cui, J. Sun, P. J. Jackson, and W. Wang, "Audio-visual speaker tracking: Progress, challenges, and future directions," *arXiv preprint arXiv:2310.14778*, 2023.

[18] J. Zhao, Y. Xu, X. Qian, H. Liu, M. D. Plumbley, and W. Wang, "Attention-based end-to-end differentiable particle filter for audio speaker tracking," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 449–458, 2024.

[19] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust ASR," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5709–5713.

[20] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi *et al.*, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[21] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial Scaper: a library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1221–1225.

[22] 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.

[23] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[24] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.