INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Binaural Sound Localization in Noisy Environments Using Frequency-Based Audio Vision Transformer (FAViT)

Waradon Phokhinanan[1,2], *Nicolas Obin*[2], *Sylvain Argentieri*[1]

[1]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
[2]Sorbonne Université, CNRS, STMS lab, IRCAM, Paris, France
waradonai@gmail.com, nicolas.obin@ircam.fr, sylvain.argentieri@sorbonne-universite.fr

Abstract

Binaural sound source localization (BSSL) aims to locate sound as the way human does, but it falls short due to acoustic interferences. While Convolutional Neural Networks (CNNs) have shown promise in localizing sounds corrupted by noise, their large parameter and training data requirements make them unsuitable for real-time processing on devices like hearing aids and robots. In this paper, we propose an adapted Vision Transformer (ViT) model for BSSL in noisy environments. Inspired by the Duplex Theory, our model uses selective attention mechanisms to the frequency range of binaural features to aid in sound localization. Our model outperformed recent CNNs and standard audio ViT models in localizing speech in unseen noises and speakers, even in challenging conditions with low training data and parameters. The attention heatmap results suggest differences in how humans and machines process binaural cues, opening up for further investigation.
Index Terms: sound source localization, binaural audition

1. Introduction

Sound source localization (SSL) involves accurately determining the direction of a target sound in 3D space using azimuth and elevation angles. While using additional sensors can enhance accurate performance in multiple sources, it may not always be feasible in devices like humanoid robots, hearing aids, or EarPods [1, 2]. In such cases, binaural SSL (BSSL) methods, which rely on two sensors, offer advantages over array-based methods and have demonstrated high performance in tasks such as [3, 4, 5, 6]. BSSL is also useful in real-world applications of speech technology, such as source separation [7, 8] and speech enhancement [9], by enabling the distinction of interferences.

However, BSSL accuracy decreases with multiple sound sources or strong reverberation [3, 4], and worsens with unseen noises with a low signal-to-noise ratio (SNR). To address these challenges, deep learning-based models (DLMs) like Multi-Layer Perceptrons (MLPs) [10, 4], Convolutional Neural Networks (CNNs) [5, 6], and Convolutional Recurrent Neural Networks (CRNNs) [11] significantly improve localization accuracy and robustness in interference-prone environments compared to non-DLMs with signal processing techniques. These models are trained on features inspired by the Duplex Theory [12], with many different implementation approaches developed, which suggests that BSSL relies on a combination of interaural time/phase differences (ITD/IPD) and interaural level differences (ILD). For Humans, ITD is known useful for low frequencies, while ILD is more adapted for high frequencies. However, most applications use both ITD and IPD on all frequency ranges, and only a few studies have investigated the effects of using different frequency ranges.

2. Related work

Comparing BSSL models using DLMs across experiments can be challenging due to the various objectives and setups, such as aiming for vertical or full-sphere localization [13]. DLMs are widely used due to their capability to learn from large datasets and perform well on learned data. However, they may struggle with new patterns that are not part of the training data. The initial DLMs utilized in BSSL were MLPs [10], which were often used as alternatives to statistical learning models like Gaussian Mixture Models. Ma et al. [4] may have been the first to train MLPs using extracted binaural features for each frequency band, a technique later employed by [14] with promising results. These findings suggest that selective frequency techniques could be effective, although it is currently unclear how each frequency band influences the final localization outcome.

Later, CNNs emerged as the preferred choice for BSSL due to their higher performance on 2D spectrogram-based features [15, 16]. However, their superior performance with noise and reverberation often requires complex architectures that may not be practical for real-time applications. Yang et al. [5] investigated the performance of IPD using the full frequency range versus only the low-frequency range for full-sphere BSSL. Surprisingly, they found that utilizing the full range of IPD led to better performance in both azimuth and elevation, contradicting the previous belief based on the Duplex Theory.

The Vision Transformer (ViT) [17] was introduced in the field of image processing with the notion that the entirety of an image may not necessarily contain relevant information, but the actual spatial position within an image could still be significant, and long-range dependencies may exist between spatial regions. To achieve this, the full image is divided into sub-blocked patches with position embeddings, forming a sequence. These sequential patches can be processed with attention mechanisms (AM), similar to the language model proposed by [18]. In other words, the AM enables the ViT to identify which part of the image (patch) contains relevant information for decision-making, leading to better performance than CNNs trained on large datasets and parameters. This idea has also been adapted to audio processing with the development of Audio ViT (AViT), which utilizes spectro-temporal regions for decision-making [19]. Although AViT has shown improved performance in audio classification, it has not yet been applied to BSSL[1].

Although the AM has been incorporated prior AViT into CRNN-based SSL, it is typically used for sound event detection [20] or to attend to a target speaker in multisource scenarios [21]. However, it has not been employed to attend to specific frequency bands for sound localization.

[1]This study does not review several preprint AViT models, including the AViT model for BSSL for high reverberance conditions.

10.21437/Interspeech.2023-2015

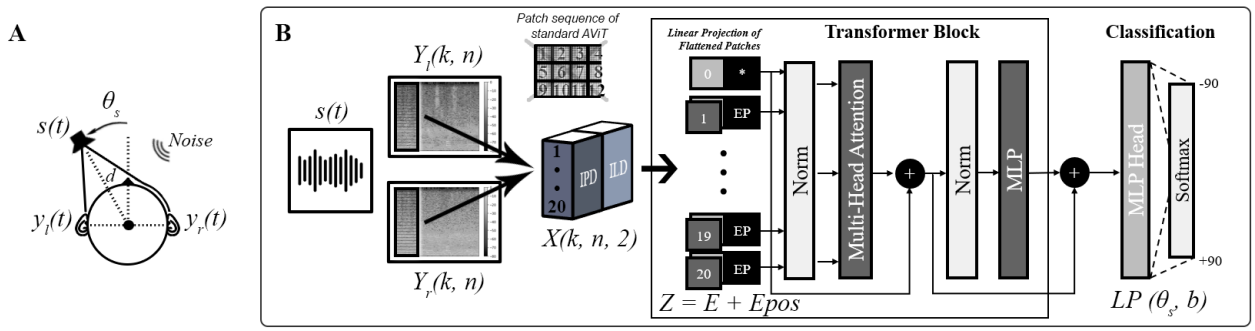

Figure 1: *(A) An illustration of the binaural system from the sound source position represented by $s(t)$. (B) The architecture of the FAViT model. The diagram shows the FAViT patch embedding sequence compared to the standard AViT, the transformer block, and the final classification layer.*

3. Proposed Architecture

This study proposes a novel approach for applying AViT to BSSL, focusing on localizing single speech sources in the frontal horizontal plane of noisy environments. Rather than using all patches from the spectrogram-based binaural feature map, we only utilize vertical patches, each corresponding to a specific frequency band ranging from high to low, as shown in Figure 1B. This approach, which we call FAViT, significantly reduces computational loads compared to CNNs and the original ViT while allowing us to investigate how the attention mechanism exploits specific frequency bands of Interaural Phase Difference (IPD) and Interaural Level Difference (ILD), an area that has not been explored previously.

In this section, we introduce our FAViT model, which consists of three components: binaural feature extraction, ViT patch encoding, and ViT classifier decoding. We then provide details on our experiments and discuss the outcomes of our approach, comparing them to existing BSSL techniques to demonstrate the effectiveness of our proposed method.

3.1. Front-end: Binaural features extraction

This study aims to localize a single sound source emitting a signal $s(t)$ from a location (d, θ, ψ) relative to the center of the binaural sensor, as shown in Figure 1A.

The focus is on horizontal angular localization, with the azimuth θ_s being the only positional parameter of interest, measured in radians. As the sound source signal propagates and interacts with the head, it generates two binaural signals, $y_l(t)$ and $y_r(t)$, in the left and right microphones which can be used to determine the source location. The relationship between the time-frequency domain of binaural signals Y and the source position θ_s in this study is given by

$$\begin{cases} Y_l(k,n) = H_l(\theta_s,k)S(k,n) + N_l(k,n) \\ Y_r(k,n) = H_r(\theta_s,k)S(k,n) + N_r(k,n) \end{cases}, \tag{1}$$

where the left $Y_l(k,n)$ and right $Y_r(k,n)$ spectrograms are obtained through STFT with frequency index k and time index n in frames. They are generated from the source signal $S(k,n)$ combined with the Head-Related Transfer Functions (HRTFs) for a source at azimuth angle θ_s, represented by $H_l(\theta_s,k)$ and $H_r(\theta_s,k)$, respectively, which highlight the effect of the head on the signals. Additionally, noise is present in the binaural signals, represented by $N_l(k,n)$ and $N_r(k,n)$ in the frequency domain. Then, the ILD and IPD that will be used as inputs to the model in the next subsection are defined by

$$\begin{aligned} \text{ILD}(k,n) &= 20\log\frac{|Y_l(k,n)|}{|Y_r(k,n)|}, \\ \text{IPD}(k,n) &= \angle\frac{Y_l(k,n)}{Y_r(k,n)}. \end{aligned} \tag{2}$$

3.2. Vision Transformer Patch Encoder

Our proposed model is based on the Vision Transformer (ViT) by [17], but with a modification in the way sequences of patches are embedded. Instead of embedding entire squared ILD and IPD maps, we embed patches along the frequency bins (vertically, top-down) as shown in Figure1B. This approach enables us to observe how the attention mechanism behaves for different frequency bins and compare it to the Duplex theory. Moreover, this modification significantly reduces the number of parameters, which may facilitate real-time tracking of the source's moving position. The model's input is a two-dimensional tensor X that stacks the binaural cues $\text{ILD}(k,n)$ and $\text{IPD}(k,n)$ defined in Equation (2), with

$$X(k,n,2) = (\text{IPD}(k,n), \text{ILD}(k,n)), \tag{3}$$

where frequency k and time-frame n is resolution of binaural feature size with 2 channels input. Then, this tensor X is divided into multiple patches X_p defined as

$$X_p(M,(p,p,2)) = X(k,n,2) \tag{4}$$

where $(p \times p)$ is the resolution of each patch, and M is the total number of patches calculated from $M = k/n$.

Therefore, to capture positional information and differentiate between high and low-frequency bins, position embeddings (also known as patch encoders) are utilized. The tensor Z represents the position embeddings and is composed of two parts: E and E_{pos}. E is created by flattening each patch of X and concatenating its channels into a single vector, followed by a Feed-Forward dense layer with a linear activation function. This results in a dimensional vector of shape (D,1), which is a learnable linear projection. On the other hand, E_{pos} is a learnable position encoding represented as a vector of shape (M + 1,1). Together, these parts are used to feed the transformer decoder and enable sound localization.

3.3. Transformer Decoder Classifier

The Transformer decoder architecture is composed of several identical blocks stacked together. Each block comprises a self-

attention mechanism, implemented using a Multi-Head Self-Attention (MSA) layer, followed by a feed-forward Multilayer Perceptron (MLP). Layer normalization is applied before and residual connections after each block. Based on [18], the primary component of the MSA layer is composed of multiple Scaled Dot-Product Attention mechanisms. Within each of these mechanisms, the attention matrix $A(Q, E, V)$ is computed by

$$\begin{aligned} Q, E, V &= ZW_q, ZW_e, ZW_v \\ G &= \text{Softmax}\left(\frac{QE^T}{\sqrt{dim_E}}\right)V \end{aligned} \tag{5}$$

where the input vector Z undergoes a transformation process where it is passed through three separate weight matrices W_q, W_e, and W_v to create queries (Q), keys (E), and values (V). The resulting Q and E are then multiplied and divided by the square root of the key dimension (dim_E), and the resulting matrix is processed by a Softmax function. The V are then multiplied to produce the attention head (G).

This process is repeated for all attention heads, and the resulting matrices are concatenated and passed through an MLP layer. The output of the MLP layer is the output vector C (6), which is used as the new Z input for the next recursive cycle of the transformer decoder blocks. Finally, the classifier MLP Head applies a Softmax function with Cross-Entropy Loss function to the final vector C to produce the class probabilities $LP(\theta)$ for each localization azimuth.

$$\begin{aligned} C(M, (1, D)) &= \text{concat}(G_1, ..., Gi) \\ LP(\theta) &= \text{Softmax}(C), \end{aligned} \tag{6}$$

3.4. ViT Implementation details

IPD and ILD were extracted from binaural signals using a STFT with a Hamming window and an overlap of $50ms$, with $n = 640$ samples and $f = 320$ frequency bins, and a sampling frequency of $f_s = 16kHz$. For FAViT, we selected only 16 time-frames instead of the entire spectrogram, resulting in input features of size $(16 \times 320 \times 2)$, which were then converted into 20 patches. In contrast, the standard full AViT (referred to as 3-AViT in the results section) contained full frames.

The model architecture consists of eight Transformer decoder blocks, each containing four MSA layers and two MLP layers, with a dropout rate of 0.1. The MLP layers within each Transformer block have 4 heads and an internal projected dimension of 20 hidden units (D), followed by layer normalization. The Gaussian Error Linear Unit (GeLU) activation function is applied to all MLP layers in the Transformer block and the classification block. The MLP layers have a learning rate of 0.001 with weight decays of 0.0001 and a batch size of 32, utilizing the Adam optimizer. Each experiment runs for a maximum of 500 epochs or until there is no further improvement in performance. The final classification block has two hidden dense layers with MLP layer sizes of 1024 and 512 respectively, each with a dropout rate of 0.5. The final layer performs classification of 37 azimuths. The repository can be found at `https://www.github.com/Senzt/FAViT`.

4. Experiments

4.1. Dataset and experimental setup

4.1.1. Benchmark

This study involved an experiment to compare the performance of two previously proposed approaches, general CNNs (1-CNNs) [6] and double CNNs (2-CNNs) [5], with a proposed model (4-FAViT) and full audio ViT (3-AViT) in localizing a simulated single sound source in the horizontal plane under both seen and unseen noisy environments. All models were trained and tested on the same binaural signals dataset generated based on Equation (1) with a sampling rate of $16kHz$.

4.1.2. Material

A pair of Head-Related Transfer Functions (HRTFs) $H_l(\theta, k)$ and $H_r(\theta, k)$ were selected from the MIT KEMAR database [22]. These HRTFs were chosen for azimuths $\theta = \theta_s$ covering a range from 0 to 180° with a 5° step, which provided a total of $N_\theta = 37$ angular positions. To create the training data, 30 audio signals were randomly selected from the TIMIT database [23], with an equal number of male and female speakers. To simulate the source position θ_s, each audio signal was spatialized to all 37 azimuths using the corresponding left and right Head-Related Transfer Functions (HRTFs) by performing frequency-domain multiplication. This process generated two binaural signals: $y_l(t)$ and $y_r(t)$. Next, noise was added independently to the left and right channels for each angular position.

To evaluate the performance of the model, 10 distinct unseen speaker TIMIT signals were each selected and divided into 10 groups, with an equal number of male and female speakers. These signals were preprocessed in the same manner as the training data. However, different Signal-to-Noise Ratio (SNR) levels in dB were used for the testing set: [-5,5,15,25], while SNRs of [0,10,20] were used for the training set. The varied SNRs were chosen to evaluate the model's robustness to noise and its ability to generalize to new, unseen SNRs.

Two categories of additive noises, $N_l(k, n)$ and $N_r(k, n)$, were used in the experiments: (i) spatially uncorrelated and almost stationary binaural noises from the Noisex92 database [24] and (ii) non-stationary noises taken from [25] and simulated as diffuse noises. The Noisex92 database includes white, pink, F16, and babbling binaural noises, while the second database provides non-stationary monaural ambient noises from a cafe, a car, a kitchen, or a street. These non-stationary noises are transformed into diffuse binaural noises by simulating 72 identical sources located around the head, and their contributions are summed to obtain the left and right simulated noise signals.

4.1.3. Methodology

Two main experiments were conducted. In the first experiment, the performance of all models was evaluated in both seen and unseen noise conditions with respect to signal-to-noise ratio (SNR) generalization. The training data only had type (i) noises, while the testing data contained both noise (i) and (ii) to assess the difference between seen and unseen noise performance. Special SNRs of [-7,-3,-1] were used to evaluate the model's performance when learning from corrupted binaural signals. The second experiment aimed to assess the model's performance when the training data was reduced from 100% to 75%, 50%, 25%, and 10%, respectively. For both experiments, the number of candidate azimuths for localization was 37 directions, ranging from 0 to 180° with a 5° step. The speakers used for training and testing in these experiments are distinct. The models' performance was evaluated based on localization accuracy (in %) and root mean squared error (RMSE in degrees).

Table 1: *The table presents the localization accuracy, tested at full SNR [-5, 5, 15, 25], as a percentage and root mean squared error in degrees for all models. Despite having significantly fewer parameters than the other three models, FAViT achieves the best performance.*

Model	Parameters	Training with low SNR				Training with full SNR									
		Seen Noises		Unseen Noises		Unseen Noises									
		100% Training		100% Training		100% Training		75% Training		50% Training		25% Training		10% Training	
		Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE
1-CNNs	2.7 M	74.8	5.3	47.9	4.7	87.8	3.4	83.7	3.1	86.2	2.6	72.8	4.6	48.9	6.5
2-CNNs	210.8 M	80.6	4.1	77.3	4.5	87.1	2.5	84.0	3.5	81.5	3.9	61.2	5.6	46.9	7.0
3-AViT	28.5 M	90.2	2.6	87.5	2.7	88.5	2.9	86.4	3.3	83.5	3.8	79.9	4.2	77.4	4.5
4-FAViT	**0.6 M**	**91.8**	**2.3**	**89.2**	**2.4**	88.3	3.0	**87.9**	**3.2**	**86.4**	**3.4**	**84.2**	**3.9**	**82.9**	**4.0**

5. Results and Discussion

5.1. Signal-to-noise ratio (SNR) generalization

To enable BSSL CNN models to perform in unseen SNRs, it is typically necessary to train them with diverse SNR data so that they can generalize well. However, Table 1 shows that training FAViT with a low SNRs group can achieve high performance of 89.2%, compared to 88.3% when training with the full range of SNRs. This suggests that the ViT model has high performance in SNR generalization and can localize clean speech by learning from noised, corrupted patterns of binaural cues.

5.2. Noise and speaker generalization

As shown in Table 1, both ViT models outperform CNN models in the unseen, non-stationary noise situations proposed by this study. Although the average accuracy is still under 90%, it shows potential for future improvement. Further investigation of this model could be conducted with high reverberation. In addition, given that all testing data are generated from unseen speakers, FAViT proves proficient in speaker generalization tasks amidst both seen and unseen noises.

5.3. The number of training data

Table 1 shows that 4-FAViT's average localization performance only slightly decreases when the amount of training data is reduced from 100% to 10%. Compared to 3-AViT, it has relatively similar performance, but the number of parameters is reduced by about 40 times. In contrast, the 1-CNNs and 2-CNNs models have significantly lower performance under the same conditions. This could be because specific neural network settings require a significant amount of training data. Typically, models with more parameters require larger amounts of training data. Interestingly, this suggests that the optimal amount of training data for ViT models may depend on the specific task they perform. The proposed ViT model's low data requirement makes it compatible with robot audition, where training data availability is often limited. Furthermore, its suitability for embedding in real-time processing devices, coupled with its low computational complexity, results in lower response delays. This feature could enable it to track moving sound sources and detect sound events more effectively. However, further improvements to the model with low training data requirements could be explored.

5.4. Interpretation of the attention map

We would expect the attention maps to follow the duplex theory, which suggests that IPD is more important for low frequencies, while ILD is more important for high frequencies. However, as shown in Figure 2, there are different patterns that can be interpreted as follows. The way the human auditory system processes ILD and ITD is still being determined. In other words, we do not actually know how to compute brain-like features of both binaural cues corresponding to each frequency. So, the IPD and ILD representations of this Transformer and the brain are different. So, this Transformer, in fact, might learn from what computed features they received. Another explanation could be that speech is more complex than pure tone. The duplex theory experiments did not investigate speech. So, people have yet to learn what fully attended speech frequencies should look like. Some experiments could support this argument, such as researchers finding the IPD/ITD sensitivity of speech across all frequencies [26] on both hearing and impaired listeners. Our current interpretation is that this ViT is more like they learn from the visual representation of the IPD map. As can be seen from Figure 2, at 0 azimuth, they should have no or very low phase difference. So, as a result, the Transformer does not attend to any specific frequencies compared to higher IPD in different azimuths.

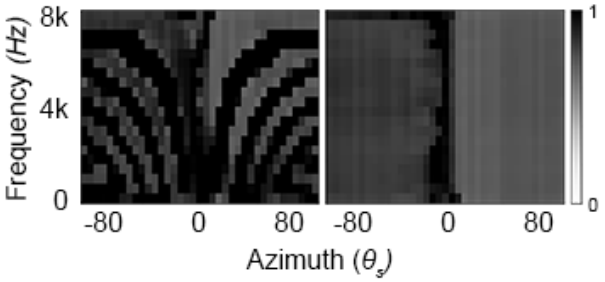

Figure 2: *The attention maps of IPD (left) and ILD (right). The high energy in these plots indicates that the model is attending to specific frequency ranges to localize sound at each azimuth.*

6. Conclusion

Despite not following the selective attention pattern predicted by the Duplex Theory, the proposed FAViT has shown a significant improvement in BSSL when dealing with unseen and noisy environments. Furthermore, the model boasts a much lower number of parameters compared to other deep learning models, especially CNNs and the original ViT, making it a valuable tool for real-time processing devices. It would be worthwhile to further explore the potential of this model for multiple SSL and separation, as well as for moving sources.

7. References

[1] D. Desai and N. Mehendale, "A review on sound source localization systems," *Archives of Computational Methods in Engineering*, vol. 29, no. 7, pp. 4631–4642, 2022.

[2] E. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raumel, and S. Argentieri, "Binaural localization of multiple sound sources by non-negative tensor factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1072–1082, 2018.

[3] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[4] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, 2018.

[5] Y. Yang, J. Xi, W. Zhang, and L. Zhang, "Full-sphere binaural sound source localization using multi-task neural network," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 432–436.

[6] 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.

[7] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint itd/ild model and head movement tracking," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 430–434.

[8] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1900–1912, 2010.

[9] R. Li, F. Zhao, D. Pan, and L. Dong, "Speech enhancement based on binaural sound source localization and cosh measure wiener filtering," *Circuits, Systems, and Signal Processing*, vol. 41, pp. 395–424, 2022.

[10] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 2927–2932.

[11] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 451–455.

[12] J. C. R. Licklider, "A duplex theory of pitch perception," *The Journal of the Acoustical Society of America*, vol. 23, no. 1, pp. 147–147, 1951.

[13] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[14] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 405–409.

[15] S. Jiang, L. Wu, P. Yuan, Y. Sun, and H. Liu, "Deep and cnn fusion method for binaural sound source localisation," *The Journal of Engineering*, vol. 2020, no. 13, pp. 511–516, 2020.

[16] H. Liu, P. Yuan, B. Yang, G. Yang, and Y. Chen, "Head-related transfer function–reserved time-frequency masking for robust binaural sound source localization," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 1, pp. 26–33, 2022.

[17] 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.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[20] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 771–775.

[21] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," *arXiv preprint arXiv:1904.12769*, 2019.

[22] B. Gardner, K. Martin *et al.*, "Hrft measurements of a kemar dummy-head microphone," 1994.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[24] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[25] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The qut-noise-timit corpus for evaluation of voice activity detection algorithms," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2010, pp. 3110–3113.

[26] L. S. Baltzell, A. Y. Cho, J. Swaminathan, and V. Best, "Spectro-temporal weighting of interaural time differences in speech," *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 3883–3894, 2020.