Machine learning in spatial audio

Tantep Sinjanakhom
Aalto Universtiy
Master's Programme CCIS / AAT

tantep.sinjanakhom@aalto.fi

Abstract

This paper reviews the application of machine-learning techniques in the field of spatial audio. Different aspects of spatial audio tasks, including capture, processing, and reproduction, are covered. The review begins with a comprehensive overview of spatial audio, followed by an investigation of various machine-learning techniques and their relevance to this domain. Different architectures of neural networks are employed for distinct spatial audio tasks. Sound source localization is used as an example for spatial audio-capturing tasks. For processing tasks, sound fields reconstructed based on neural networks are reviewed. As an example for spatial audio reproduction tasks, head-related transfer function modeling is covered. Trade-offs between each method can include accuracy and efficiency in terms of real-time processing. The findings from numerous studies have consistently demonstrated that machine learning models can surpass traditional techniques in performance and achieve tasks that were previously beyond the capabilities of traditional methods.

1 Introduction

Spatial audio refers to the technology that involves audio signal processing and acoustics to create the sensation of perceiving sound in environments [15]. This concept originated with stereophony [2] and has evolved into numerous modern applications, including ambisonics [26], reverberation synthesis [9], binaural audio [6], and 3-D sound systems used in virtual reality [27].

Many research studies have been carried out to improve the quality of different aspects of spatial audio processing. Recently, attention has shifted towards machine learning techniques across various fields, including spatial audio. This is mainly

because some traditional methods are not accurate or efficient enough. Thus, formulating these tasks into machine learning problems is convincing, and it can learn to perform different analysis and reconstruction tasks from the collected data. In the context of spatial audio tasks, machine learning is applied to tasks such as sound source localization, spatial audio synthesis, room impulse response estimation, acoustic scene analysis, and sound event detection. Machine learning algorithms can learn from large datasets of audio recordings and corresponding spatial information to improve the accuracy and efficiency of spatial audio processing techniques.

This paper is structured as follows: Section 2 provides an overview of various machine learning model architectures and discusses their application in spatial audio. Section 3 describes machine learning methods for spatial audio capture. Section 4 reviews neural network models designed for sound field reconstruction, which falls under the context of spatial audio processing. Section 5 discusses approaches to synthesizing head-related transfer functions. Section 6 concludes the paper.

2 Machine learning in different aspects of spatial audio

The typical pipeline of spatial audio tasks consists of spatial capturing, spatial processing, and spatial reproduction [4]. Spatial capture is the first stage of the pipeline. This procedure includes gathering spatial information, which can be in the form of raw recorded signals [10], spherical harmonics [18], and feature representations [22]. This information is then passed to the next step, which is spatial processing. At this stage, the collected spatial data can go through various transformations, depending on the tasks. The signal could be analyzed and decomposed to estimate parameters. Examples of these parameters include room dimensions, source locations, and reverberation times. Other transformations that are done during this stage are the reconstruction of the sound field [1], resynthesis of sound scene [5], and spatial audio coding [17]. After the spatial information is transformed, the final step is reproduction. Two main ways to reproduce spatial audio are based on loudspeaker [16, 3] and binaural sound [14]. For loudspeaker-based reproduction, it can be as simple as surround sound or it could be more intricate, such as wave field synthesis [5]. For binaural reproduction, the main tool to use is head-related transfer functions (HRTFs). This helps create the sensation of different sound directions and distances according to how humans perceive sound.

Most of the processes mentioned above can be carried out by using a multitude of traditional techniques based on digital signal processing algorithms and acoustics theory. However, some of those might not be able to produce the most accurate results or sometimes require high computational power. Therefore, many researchers have now conducted studies where conventional methods are being replaced by machine learning models. In this context, when referring to the machine learning (ML)

or deep learning (DL) approach, the neural network is the main tool. there are many different topologies of neural networks. A generic fully connected feedforward network is usually referred to as a deep neural network (DNN) or a multilayer perceptron (MLP). This type of network can learn to approximate nonlinear functions in order to estimate the complicated relationship between input and output data. A more intricate type of feedforward network is the convolutional neural network (CNN). In this architecture, the network has multiple kernels striding over the input data performing convolution, hence the name. This type of network can perform feature selection as the kernels can filter and extract meaningful features making it suitable for many different tasks, especially when the input data has more than 1 dimension.

Apart from feedforward topologies, neural networks with feedback paths are generally referred to as recurrent neural networks (RNNs). There are many variants of RNNs, with one of the most common being the long short-term memory (LSTM) unit. This type of network is appropriate for modeling a stateful system.

3 Capture: Feature-based representations

This section reviews the representation of spatial information that is suitable for data-driven approaches. This is essentially a regression task where acoustic features are used as input to the neural network in order for it to predict meaningful values such as the locations of sound sources. On the other hand, the neural network can be trained to transform the input features into different representations that can later be used to construct a subsequent tool, such as a spectral mask. The main topic to be discussed is the sound source localization techniques with convolutional neural networks [7] and recurrent neural networks [24, 23].

3.1 Localization with CNNs

In [7], a neural network model consisting of two CNNs of identical architecture is proposed to perform sound localization in a shallow water environment. The CNNs take different acoustic features as inputs, namely, the generalized cross-correlation (GCC) and the power cepstrum. The model diagram is shown in Fig. 1. The network was trained to predict both the range and the bearing of the source. The power cepstrum is used as it contains echo information from multi-path reflections. On the other hand, the GCC helps estimate the time-of-arrival differences in noisy and uncorrelated signals. Experimental results showed that the neural network outperform the baseline algorithmic method based on time-of-arrival differences in [19]. Despite the fact that shallow water environments have many reflections coming from sea surfaces, the trained CNNs were shown to be robust and the predictions

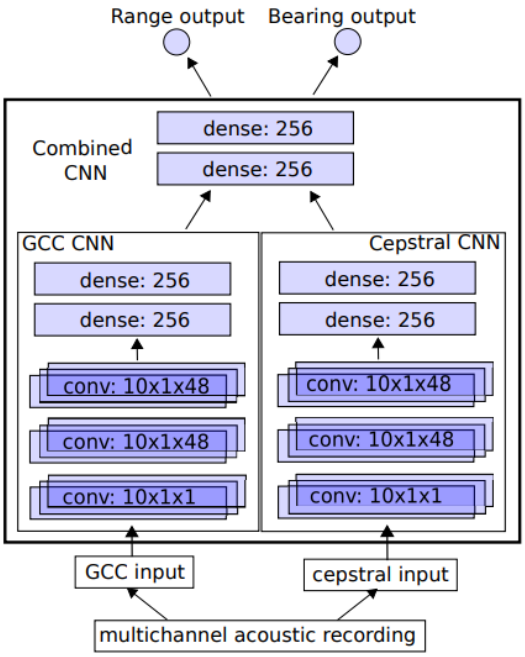

Figure 1: *CNN architecture for the acoustic source localization. Adopted from [7].*

were not degraded as much as algorithmic methods. The only possible drawback of this method is the computational complexity. The convolution operation is known to be a computationally heavy process. If there are many hidden layers with each layer being large in terms of kernel size, then the calculation speed can be low and it would be difficult to track the sound source in real-time.

3.2 Speech separation with LSTMs

Besides environmental source localization, neural networks are also used for localizing speakers, which can then be used for speech separation in noisy or reverberant environments. In [23], a bidirectional-LSTM (BLSTM), which is a non-causal RNN model, was used for clustering. The model topology is shown in Fig. 2. This paradigm is also known as unsupervised learning. In this type of learning, the neural networks are not exposed to target data as they could be unknown. However, the network will be trained to optimize some objective functions. In this case, the BLSTM model takes the input features consisting of spectral information and spatial features such as the GCC, along with additional phase difference features from the microphones. The features are clustered and encoded into an embedding representation. The embeddings are learned such that the time-frequency units that belong to the same speaker are brought closer together, while those from different speakers are pushed further apart.

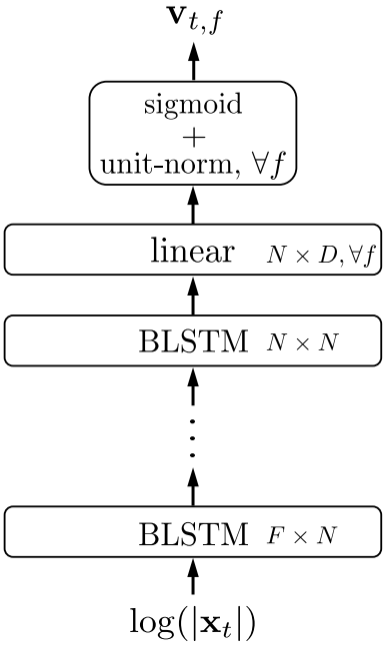

Figure 2: *BLSTN architecture for clustering, where* $\boldsymbol{v}$ *is the embedding vector. Adopted from [23].*

During inference, time-frequency features on each speaker are clustered via k-means using the trained embeddings. Then, a spectral mask can be constructed to be used for speech separation. Results demonstrate that the separated speech from this method has higher signal-to-distortion ratios than old techniques such as the Wiener filter and models based on the expectation-maximization algorithm.

4 Processing: DL-driven sound field reconstruction

This section reviews different deep learning models that can be optimized for sound field analysis and reconstruction, such as CNNs [13] and RNNs [21]. Sound field reconstructions relate to the simulation of sounds that match the real environment. CNNs are suitable for this task since they are typically employed with 2-D data. They can be used for both feature learning and generating tensors of two or more dimensions. In this case, they can be used to reproduce a 2-D representation of a sound pressure field from 3-D data. The RNNs on the other hand can also predict the sound field, but they require less computational resources during inference due to their recursive nature.

4.1 Sound field reconstruction with CNNs

Research work in [13] showed that a CNN model can be designed and trained to reconstruct the sound field of a room in the form of sound pressure magnitude. The

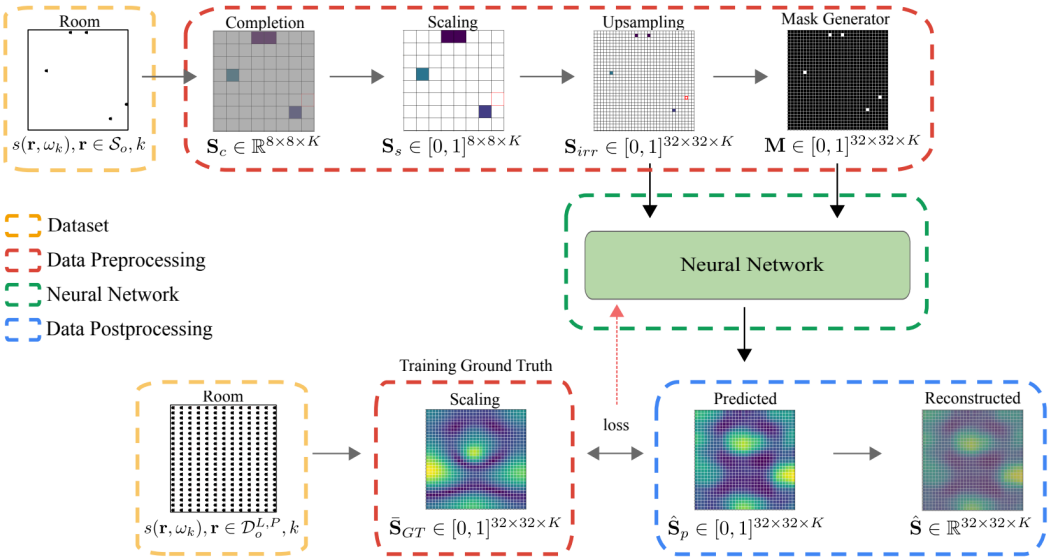

Figure 3: *Diagram of the sound field reconstruction model. Adopted from [13].*

neural network was trained on simulated sound fields from common rectangular rooms to predict sound field structures in the frequency range of 30 Hz to 300 Hz. This model aims to accurately extrapolate and interpolate the sound field into a higher-resolution representation. The network takes the magnitude of the Fourier transform of the spatial sound field in three-dimensional space as input and predicts the pressure magnitude of spatial sample points in a two-dimensional rectangular plane.

The input data are preprocessed by completion, scaling, upsampling, and the generation of an encoding mask that stores the location information. The completion step assigns a constant arbitrary value to positions in the data where the microphone is absent. The diagram of the overall process is illustrated in Fig. 3. The neural network architecture is a U-Net which is a variant of autoencoder models based on CNNs. This topology is suitable for reconstruction tasks. There are skip connections between layers which help retain previous meaningful latent spaces for the latter layers. This was shown to improve the model performance.

This method is said to require a low number of training samples, and the microphone can also be unevenly distributed. This approach is an improvement of traditional methods for sound field reconstruction, such as the model-based method [8] since it does not require a high density of microphones to achieve low reconstruction errors. Additionally, it was suggested that it is feasible to employ this model for reconstructing the sound field in rooms of different shapes as well as using a three-dimensional representation of the spatial sample points, which has the same dimensionality as its input.

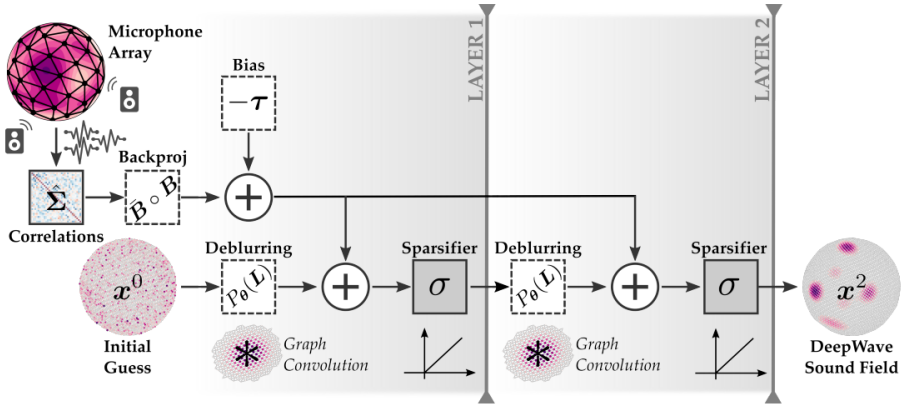

Figure 4: *DeepWave RNN architecture. Adopted from [21].*

4.2 Acoustic imaging with RNNs

In [21], a recurrent model is designed to perform real-time acoustic imaging called DeepWave. For real-time applications, using RNNs is more suitable than CNN. This is because it requires less computational power. We think of RNNs as a set of nonlinear IIR filters, while CNN is a model that consists of many nonlinear FIR filters. The diagram of the DeepWave model is illustrated in Fig. 4. This model takes the recordings from the microphone array as input and learns to map the complex correlation with microphones into a spherical map. The deblurring matrix learns to reduce artifacts that can cause errors. The output of the model is a sound field of pressure that changes in real time.

The trade-off of the RNN models is accuracy. They can be less accurate when compared to CNNs since recurrent models tend to struggle with long-term dependencies. If previous data from many past time frames needs to be computed with the current data frame, then RNN might not be the best choice as that information might have vanished through time already. Thus, this is a clear example where we have to choose between high-accuracy or high-speed calculation for real-time systems.

5 Reproduction: HRTF personalization and generalization

This section focuses on spatial reproduction tasks driven by deep learning techniques. Specifically, the modeling of HRTFs, which can be used for binaural sound reproduction, is reviewed. The HRTF is distinguished by the unique shapes of individuals' heads and ears, aiding listeners in localizing sound sources. This characteristic is crucial because, in addition to interaural time differences (ITD) and interaural level differences (ILD), frequency response also plays a significant role in distinguishing the direction of the sound.

5.1 Principal component analysis

HRTF has been extensively studied in previous research, revealing that traditional methods like principal component analysis (PCA) can approximate HRTFs by modeling them with multiple orthogonal linear basis functions [11, 25]. This helps reduce the need to store HRTFs for different azimuth and elevation values, thereby reducing dimensionality. In this previous study, it was shown that HRTFs can be approximately represented by a linear combination of five basic spectral shapes. Thus, five basis functions derived via the PCA method are used.

A recent study in [25] incorporated a feedforward neural network to help predict weights for the spatial PCA (SPCA) model. The neural network takes the anthropometric parameters of an individual as inputs and maps them to SPCA weights. Through the SPCA techniques, HRTFs can be decomposed into a combination of spatial principal components which is similar to the spherical harmonics basis functions. With additional ITD modeling from the desired azimuth and elevation, a new HRTF can be reconstructed for arbitrary spatial directions.

5.2 Deep neural networks

In recent years, deep neural networks have emerged as a promising approach for directly modeling HRTFs with greater accuracy. In [12], a deep learning model that consisted of three sub-networks was proposed. The deep neural network architecture is illustrated in Fig. 5. The first sub-network (A) is a fully connected feedforward neural network that takes anthropometric measurements as inputs. These measurements involve the dimensions of the head, pinna, neck, and torso. The second sub-network (B) is a convolutional neural network that takes in a 2-D image of the ear as input. The outputs of both networks are then combined and passed into the third sub-network (C) to predict the HRTF and compare it with the target data. By training on this intricate and nonlinear mapping, neural networks can achieve precise HRTF modeling.

When compared to the SPCA approach in [25], despite the fact that both used MLP as a part of the methods. This model is more straightforward and requires fewer procedures for decomposing and reconstructing the HRTFs through PCA. Another work in [20] proposed a simpler MLP-based model that was trained to select HRTFs instead of generating a new one. The anthropometric measurements are the inputs to the model. The model will then choose existing HRTFs from a large database that are the most correlated to the ear dimension and other measurements. The drawback of the approach is the requirement of large memory space to store as many HRTFs as possible in order to increase generalizability.

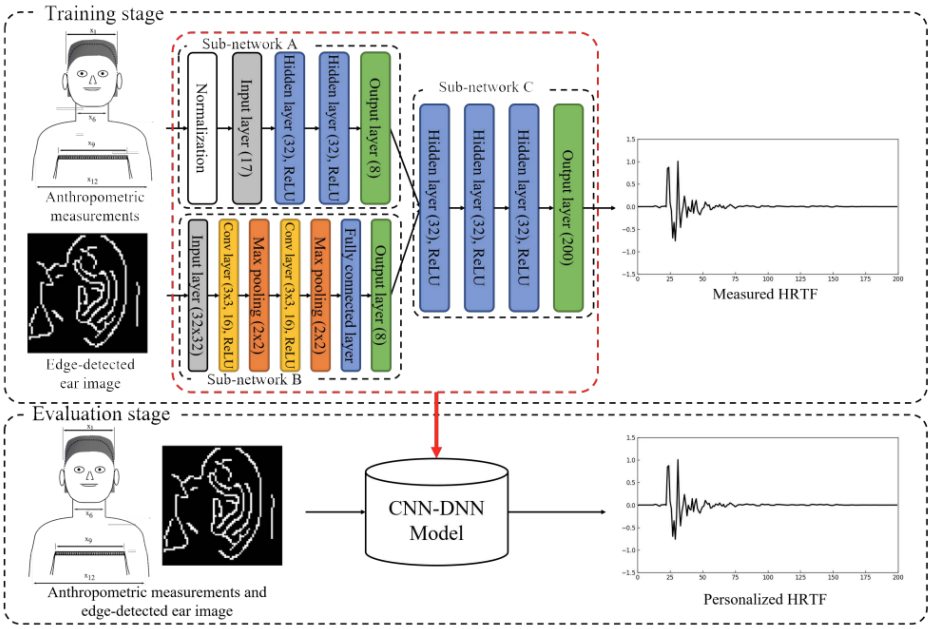

Figure 5: *Block diagram of a deep neural network (DNN)-based HRTF estimation method using anthropometric measurements and ear image. Adopted from [12].*

For the HRTF modeling task, employing a deep neural network may offer superior accuracy. This advantage stems from the network's ability to utilize multiple hidden layers to map and interpolate intricate HRTFs. While the PCA-based method can approximate HRTFs, it necessitates numerous basis functions to precisely estimate them. Additionally, the PCA-based approach lacks generalizability since the functions are constructed based on a single individual, whereas the neural network model can generate corresponding HRTFs for any individual with a picture of their ear and anthropometric measurements. This makes it more versatile. A potential avenue for future research, building on the work of Lee et al. [12], is to develop a CNN model that predicts HRTFs solely from ear images without requiring anthropometric measurements. Such an approach would be more convenient for individuals seeking to utilize their HRTFs without the need for tedious measurements.

6 Conclusions

In summary, this review paper explored different aspects of machine learning technique utilization for tasks related to spatial audio.

Various neural network architectures can be applied for different tasks within the spatial audio pipeline. For instance, in feature-based representation, neural networks such as LSTMs and CNNs are shown to be suitable for sound source localization when trained as regression models. Moreover, both CNNs and RNNs demonstrate proficiency in sound field reconstruction. Notably, RNN-based models offer the

advantage of real-time sound field display due to their low computational cost. For the reproduction side of spatial audio, HRTF modeling can be done with a more straightforward architecture of multilayer perception. It is apparent that the model does not need many analysis tasks as opposed to PCA-based methods. Despite the non-complex topology, the machine learning-based method can learn from real-world data and outperform conventional algorithmic methods. This versatility in using neural networks highlights how adaptable and effective machine learning approaches are in spatial audio processing.

References

[1] Ahrens, J. *Analytic methods of sound field synthesis.* Springer Science & Business Media, 2012.

[2] Blumlein, A. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems, 1931. *UK Patent 394325*.

[3] Borss, C. A polygon-based panning method for 3d loudspeaker setups. In *Audio Engineering Society Convention 137* (2014), Audio Engineering Society.

[4] Cobos, M., Ahrens, J., Kowalczyk, K., and Politis, A. An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction. *EURASIP Journal on Audio, Speech, and Music Processing 2022*, 1 (2022), 10.

[5] Cobos, M., and Lopez, J. J. Resynthesis of sound scenes on wave-field synthesis from stereo mixtures using sound source separation algorithms. *Journal of the Audio Engineering Society 57*, 3 (2009), 91–110.

[6] Delikaris-Manias, S., Vilkamo, J., and Pulkki, V. Parametric binaural rendering utilizing compact microphone arrays. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2015), pp. 629–633.

[7] Ferguson, E. L., Williams, S. B., and Jin, C. T. Sound source localization in a multipath environment using convolutional neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (2018), pp. 2386–2390.

[8] Grande, E. F., Rosell, A. T., and Jacobsen, F. Holographic reconstruction of sound fields based on the acousto-optic effect. In *International Congress and Exposition on Noise Control Engineering* (2013).

[9] Hold, C., McKenzie, T., Götz, G., Schlecht, S., and Pulkki, V. Resynthesis of spatial room impulse response tails with anisotropic multi-slope decays. *journal of the audio engineering society 70*, 6 (2022), 526–538.

[10] Hong, J. Y., He, J., Lam, B., Gupta, R., and Gan, W.-S. Spatial audio for soundscape design: Recording and reproduction. *Applied sciences 7*, 6 (2017), 627.

[11] Kistler, D. J., and Wightman, F. L. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America 91*, 3 (1992), 1637–1647.

[12] Lee, G. W., and Kim, H. K. Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear. *Applied Sciences 8*, 11 (2018), 2180.

[13] Lluis, F., Martinez-Nuevo, P., Bo Møller, M., and Ewan Shepstone, S. Sound field reconstruction in rooms: Inpainting meets super-resolution. *The Journal of the Acoustical Society of America 148*, 2 (2020), 649–659.

[14] Paul, S. Binaural recording technology: A historical review and possible future developments. *Acta acustica united with Acustica 95*, 5 (2009), 767–788.

[15] Politis, A., Pihlajamäki, T., and Pulkki, V. Parametric spatial audio effects. In *Proceedings of the International Conference on Digital Audio Effects* (2012).

[16] Pulkki, V. Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society 45*, 6 (1997), 456–466.

[17] Pulkki, V. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society 55*, 6 (2007), 503–516.

[18] Rafaely, B., and Avni, A. Interaural cross correlation in a sound field represented by spherical harmonics. *The Journal of the Acoustical Society of America 127*, 2 (2010), 823–828.

[19] Schau, H., and Robinson, A. Passive source localization employing intersecting spherical surfaces from time-of-arrival differences. *IEEE Transactions on Acoustics, Speech, and Signal Processing 35*, 8 (1987), 1223–1225.

[20] Shu-Nung, Y., Collins, T., and Liang, C. Head-related transfer function selection using neural networks. *Archives of Acoustics 42*, 3 (2017), 365–373.

[21] Simeoni, M., Kashani, S., Hurley, P., and Vetterli, M. Deepwave: a recurrent neural-network for real-time acoustic imaging. *Advances In Neural Information Processing Systems 32* (2019).

[22] Thuillier, E., Gamper, H., and Tashev, I. J. Spatial audio feature discovery with convolutional neural networks. In *IEEE international conference on acoustics, speech and signal processing* (2018), pp. 6797–6801.

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

[24] Wang, Z.-Q., Zhang, X., and Wang, D. Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 27*, 1 (2018), 178–188.

[25] Zhang, M., Ge, Z., Liu, T., Wu, X., and Qu, T. Modeling of individual HRTFs based on spatial principal component analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 28* (2020), 785–797.

[26] Zotter, F., and Frank, M. All-round ambisonic panning and decoding. *Journal of the audio engineering society 60*, 10 (2012), 807–820.

[27] Zotter, F., and Frank, M. *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer, 2019.