DeepEar: Sound Localization with Binaural Microphones

Qiang Yang, *Student Member, IEEE,* and Yuanqing Zheng, *Senior Member, IEEE*

Abstract—The binaural microphone, which refers to a pair of microphones with artificial human-shaped ears, is widely used in hearing aids and spatial audio recording to improve sound quality. It is crucial for such devices to find the voice direction in many applications such as binaural sound enhancement. However, sound localization with two microphones remains challenging, especially in multi-source scenarios. Most previous work utilized microphone arrays to deal with the multi-source localization problem. Extra microphones yet have space constraints for deployment in many scenarios (*e.g.*, hearing aids). Inspired by the fact that humans have evolved to locate multiple sound sources with only two ears, we propose DeepEar, a binaural microphone-based sound localization system. To this end, we design a multisector-based neural network to locate multiple sound sources simultaneously, where each sector is a discretized region of the space for different angle of arrivals. DeepEar fuses explicit hand-crafted features and implicit latent sound representatives to facilitate sound localization. More importantly, the trained DeepEar model can adapt to new environments with a minimum amount of extra training data. The experiment results show that DeepEar substantially outperforms the state-of-the-art binaural deep learning approach by a large margin in terms of sound detection accuracy and azimuth estimation error.

Index Terms—Binaural localization, Multi-source localization, Earable computing.

1 Introduction

Sound localization can provide context information to improve user experience and enable a variety of innovative applications such as human-computer interaction, smart home, and helping disadvantaged groups. As shown in Fig. 1, people with hearing difficulties generally wear a pair of hearing aids to help amplify sounds when listening to others. However, all ambient sounds, including noise, will be enhanced in this case. Thus, binaural beamforming algorithms have been applied to further improve speech intelligibility [1]. If hearing aids can distinguish the sound location, then the beamforming algorithms can focus on the desired direction to improve the Speech to Noise Ratio (SNR). Furthermore, when hearing-impaired people walk on a street, it is essential to detect nearby sounds and alert them timely to avoid potential accidents. Such binaural localization would substantially improve their communication quality and life experience.

Over the years, many microphone array-based sound localization approaches have been proposed, such as cross-correlation based methods [2], [3] and subspace-based MUSIC [4]. These approaches typically require a large number of microphones and are difficult to directly apply to binaural microphones. For example, rigidly employing the cross-correlation on only two microphones leads to the front-back confusion problem [5]. MUSIC requires at least three microphones to estimate the Angle of Arrival (AoA) of two sound sources [6]. Many deep learning based methods using microphone arrays [7], [8], [9], [10] emerge in recent years. Although effective, these methods typically require multiple microphone channels of an array as input. The fairly large form factor of a microphone array makes it inconvenient for users to wear or integrate into small hearing aids.

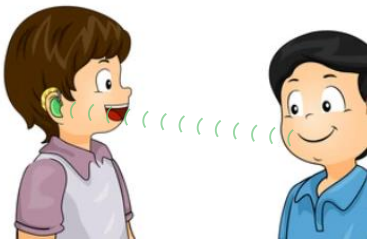


Fig. 1. Application scenario. The binaural microphones in hearing aids can localize the sound location and amplify the sound for hearing impaired wearers to improve their communication quality.

Benefiting from its good modeling capability, researchers also exploit deep learning to achieve sound localization with binaural microphones [11], [12], [13], [14]. Most works treat AoA estimation as a classification problem by dividing the space into several sub-regions. Consequently, the resolution of the AoA is limited to the region size. Moreover, some of them locate one source [11], [14] or assume that the number of active sound sources is known beforehand [12], [13]. According to our experiment results (Sec. 5), the localization performance degrades dramatically if the number of active sound sources increases.

Our work is based on the fact that the human auditory system has naturally evolved to locate multiple sounds simultaneously and accurately. Biological research found that the outer ears shape the sound waveform from different directions and provide additional spatial information which helps humans locate sounds [15]. Inspired by observation, in this paper, we investigate the mechanism of the human auditory system and propose DeepEar, a Deep Neural Net-

• *Q. Yang and Y. Zheng were with the Department of Computing, The Hong Kong Polytechnic University, Kwloon, Hong Kong. E-mail: {csqyang,csyqzheng}@comp.polyu.edu.hk*

work (DNN) based machine hearing framework to fully leverage the help of the ear-shaped binaural microphones. We identify the following key objectives and challenges to enable binaural localization for multiple sources:

i) How to characterize and exploit the ear filtering effects? Although we know that ears cause unique distortion for sound signals from different directions, how to exploit such a filtering effect is still challenging. Most previous works utilize either raw acoustic signals [11] or hand-crafted features (*e.g.*, Interaural Time Difference (ITD) or Interaural Level Difference (ILD)) [13] as the input, overstating or understating the signal in the localization procedure. To address this challenge, we adopt an analogous processing pipeline to the human auditory system and transform audio signals into the time-frequency domain. Then, a temporal autoencoder is designed to extract the latent sound representation automatically. Apart from this, we also combine the explicit ITD feature with encoded representatives to facilitate sound localization.

ii) How to achieve fine-grained multi-source localization? Intuitively, regression-based methods produce potentially higher-resolution results than classification since there is no quantization [16]. However, it is non-trivial to directly reform a classification layer in previous methods to a regression node to achieve fine-grained localization. First, for multiple sources, the number of active sound sources may not be known and can vary over time. Second, multiple regression nodes usually face the source permutation problem of associating outputs to their corresponding target sources [16]. To this end, we use a multi-task learning framework to detect the sound existence and estimate sound locations simultaneously. Specifically, we partition the 2D horizontal space into several sectors and formulate multiple sound detection as a multi-label classification problem. Each sector represents a certain range of the search space, in which we model sound localization as a regression problem. These sectors pose a spatial constraint for different sources and hereby avoid the label permutation problem. Therefore, DeepEar can detect multiple sound sources dynamically and then estimate their fine-grained positions in each sector. Moreover, the number of sectors can be configured according to the application requirement.

iii) How to adapt to new environments? Many machine learning-based methods highly depend on the data used for training, which are susceptible to new environments due to different room reverberations [16]. Our experiment (Sec. 5.4) indicates a substantial performance degradation of a baseline approach when tested in unseen rooms. In this case, training the model in new environments from sketch involves a huge data collection overhead. To ease this burden, we first train a global model on a large amount of available datasets. To bootstrap the adaption process, DeepEar then harnesses a transfer learning strategy and fine-tunes the global model with a small amount of new data collected in the target environments. By doing so, our method significantly alleviates the data collection overhead and cope with the heterogeneity of working environments with the minimum effort of end-users.

In summary, the contributions of this paper can be summarized as follows.

- We propose DeepEar, a human-inspired sound localization framework for binaural microphones that can locate multiple sources without the number of sources. We also propose two variants, namely Complex DeepEar and Monaural DeepEar. The former further improves the localization performance with the phase of the sound. The latter verifies that DeepEar can still work with only one ear.
- DeepEar fuses explicit binaural time clues and implicit sound representatives to facilitate sound localization. It features a sector-based DNN model to enable dynamic sound source detection and simultaneous multisource fine-grained localization.
- Comprehensive experiments are conducted in both anechoic and reverberant environments. The results demonstrate that DeepEar outperforms a binaural state-of-the-art in various experiment settings. A real-world case study illustrates that the ears of binaural microphones play a pivotal role in sound localization performance, especially for disambiguation.

The paper is organized as follows. Related work is summarized in Sec. 2. We elaborate on the detailed system design of DeepEar in Sec. 3. Then, Sec. 4 and Sec. 5 describe the implementation and evaluation results. We also discuss some open problems in Sec. 6. Finally, Sec. 7 concludes this paper.

2 Related Work

2.1 Sound Localization

Sound source localization has been studied for many years [16]. DeepEar is most related to binaural sound localization. We divide existing works into four categories based on two features, *i.e.*, microphone array-based/binaural microphone-based and one source/multiple source(s) in Tab. 1.

Microphone array-based methods. Many prior research work utilizes microphone arrays to estimate the AoA of a unknown sound source. [17] performs a spiking neural network (SNN) with a 4-mic array for AoA Estimation. By exploiting the sound reflections, VoLoc [18] and [19] can locate the voice position with a microphone array. When multiple sound sources are present, their interference raises practical challenges for localization. [3] explores the microphone redundancy in an array to achieve multisource localization. Many works [7], [8], [9], [10], [20] adopt Convolutional Neural Network (CNN)-based model to localize multiple sources with a microphone array.

Binaural microphone-based approaches. Unlike the microphone array, the binaural microphone consists of only two microphones with human-shaped artificial ears. Our experiment in Sec. 5.11 shows that localization with only two microphones without ears suffers from the ambiguity problem. Some researchers tried to exploit the ear filtering effect and perform binaural localization with deep learning techniques [21], [22] . WaveLoc [11] inputs raw waveforms into a CNN and classifies sounds into 37 directions. [14] utilizes CNN on audio spectrograms to perform azimuth and elevation classification. Both works use the softmax

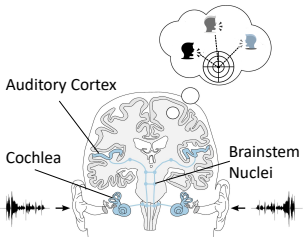


Fig. 2. Illustration of the human auditory system [24].

Fig. 3. Sound localization with binaural microphones.

Table 1. A taxonomy of related works on sound localization.

Sound localization	Mic array	**Binaural mics**	
One source	[17], [18], [19]	[11], [14], [21], [22]	
Multiple sources	[3], [7], [8], [9], [10], [20]	Known number	**Unknown number**
		[12], [13], [23]	**DeepEar**

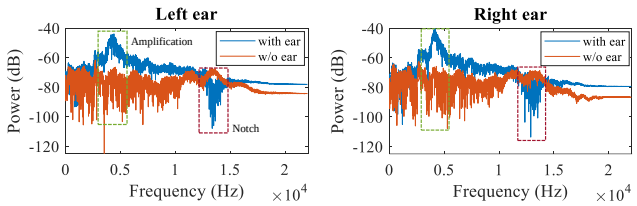
function in the classification layer, the sum of which outputs is equal to 1. These works locate one sound source with the highest probability. To achieve multiple sound localization, some works [12], [13], [23] train machine learning models and aggregate the estimates of different frequency bands or time segments. However, they assume the prior knowledge of exact number of coactive sound sources, and [12] requires an extra head rotation process. Moreover, the localization resolution of classification-based approaches is limited to the class quantization [16]. In contrast, DeepEar utilizes a sector-based network for sound detection and a regression-based network in each sector for localization, which can achieve multiple sound localization with an unknown and varying number of co-active sound sources. Here, we note that the *maximum* supporting number of sound sources is required for DeepEar.

2.2 Bionic Auditory Applications

Inspired by the powerful human auditory capability, many researchers imitated the human auditory mechanism and designed several smart systems to deal with sound-related tasks such as sound classification [25], [26], speech recognition [27], and keyword spotting [28]. In addition, [29] proposed an auditory-like system to recognize the type of musical instruments, and [30] designed a machine hearing approach to predict the types of sounds. Spiking neural network [31] has been developed to closely mimic natural neural networks, which imitates the information transfer in biological neurons. It has become popular as a possible energy-efficient and neuromorphic alternative to conventional deep learning models [32]. The powerful perceptual capacity of humans is still the goal of the AI community today. Like the research on CNN and its breakthrough in computer vision tasks, we envision that modeling the human auditory system will open up a broad range of possibilities in various sound-related tasks.

3 DeepEar Design

In this section, we elaborate on the components of the human-inspired sound location pipeline. Before moving on to the details, we will first give an intuitive introduction to how humans locate sounds.



(a) Left ear. (b) Right ear.

Fig. 4. Frequency response with and without ears.

3.1 Preliminary of Human Auditory System

Figure 2 shows a basic human auditory system. When the sound waveform travels to a user, it will be scattered, reflected, and diffracted by the ears, which significantly distort and filter the sound at certain frequencies. This direction-dependent filtering effect is technically named the Binaural Room Impulse Response (BRIR) in the time domain or the Head-Related Transfer Function (HRTF) in the frequency domain [33], [34]. In Fig. 4, we illustrate HRTFs of a binaural microphone with and without artificial ears. The signal amplification (< 10 kHz) and notch (10 kHz $\sim$ 20 kHz) are observed in the HRTF with ears, which differ substantially from those without ears.

After ear filtering, the sound wave strikes the eardrum, leading to the vibration in the spiral-shaped cochlea, which transduces the sound wave to neural stimulus signals [35]. As stimulus activities move along the nerve path, several brainstem nuclei encode the stimulus to perception [35], [36]. Finally, the auditory cortex in the brain interprets perception as spatial sound information. We will elaborate more on each part in the following sections, and we refer interested readers to the literature [15], [35], [37] for more psychophysics of human sound localization.

In a nutshell, the ears distort incoming sounds, and the human brain can learn and associate these subtle difference patterns with certain spatial locations, which helps perform source localization, even in multisource scenarios [38]. Inspired by this fact, we utilize binaural microphones with human-shaped ears to capture sounds and develop a DNN-based framework to locate sound sources, as illustrated in Fig. 3.

3.2 System Overview

Figure 5 presents a system overview of DeepEar. The upper part depicts the pipeline of the human auditory system. Inspired by its powerful localization ability, we design and implement several components to mimic its key functions. We first utilize binaural microphones with human-shaped ears to capture sounds. Then, a gammatone filterbank is used to transform the audio signals into the time-frequency domain, which acts as a cochlea in the human auditory system. After that, we train an autoencoder to extract a high-level representation. Finally, these features are input to a DNN to estimate the sound locations. In the following, we introduce each component in detail.

3.3 Data Collection and Preprocessing

Human beings perform sound localization by learning the spatial patterns of sounds caused by the ears. As such,

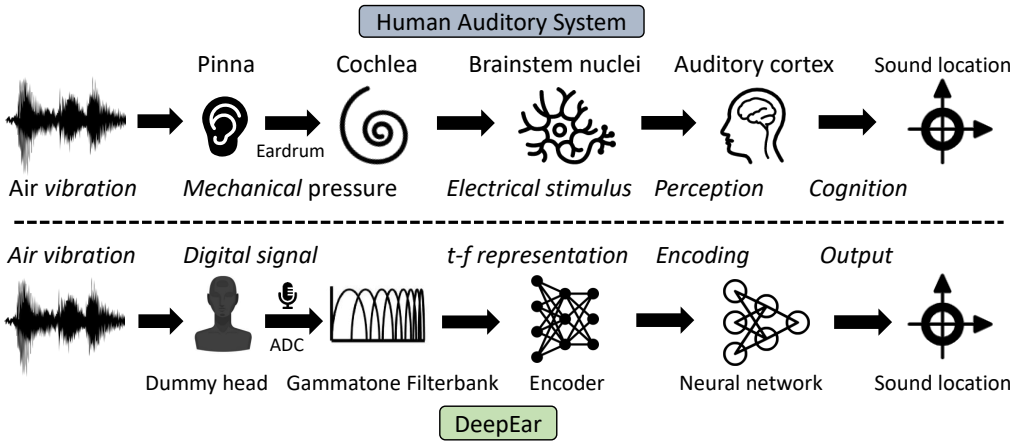


Fig. 5. System overview: an analogy between the human auditory system and DeepEar.

we use binaural microphones with human-shaped ears to capture acoustic signals. In the human auditory system, the cochlea is a spiral structure essential for frequency analysis. Along this spiral, it has a large number of inner cells that will vibrate in response to different frequencies. As a result, the sound waves are converted into electrical stimuli. During this process, the sound is decomposed into many constituent frequency components. This frequency-selective vibration varies exponentially along the cochlea [39].

DeepEar imitates the cochlea function with a gammatone filterbank. The gammatone filterbank can transform sound into multifrequency activity patterns such as those observed in the cochlea, which is widely used in the literature on auditory system modeling [40]. We employ the gammatone filterbank on the whole voice frequency band (*i.e.*, [0 Hz, 8 kHz]). The center frequencies f_c of the filterbank are equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale, where $ERB = 21.4log_{10}(0.00437f_c + 1)$ [41]. The literature shows that ears have about 3500 inner cells that decompose signals into the frequency domain with a very high resolution [35]. Although more filters provide better frequency granularity, the computational overhead increases accordingly. Hence, we empirically set the number of filters P as 100 to balance the signal representative sufficiency (*i.e.*, resolution) and the computational efficiency. Moreover, we frame the audio signals using a $100\ ms$ Hamming window with $50\ ms$ overlap to retain the frequency resolution and preserve temporal context. After filtering the audio frame in the frequency domain, we can obtain a coefficient vector with length P per frame, and the final output of the preprocessing is a 2D matrix $I \in \mathbb{R}^{P \times T}$, where T is the frame number.

3.4 Feature Extraction

A neural stimulus passes through many stages of processing by several brainstem nuclei before reaching the auditory cortex in the brain, as shown in Fig. 2. Although the understanding of the specific processing accomplished in this stage remains unclear yet [42], it is commonly believed that these nuclei perform a function similar to signal encoding for sound localization and recognition [36]. This compressing process is able to prevent an information overload in the brain in a short period of time [43].

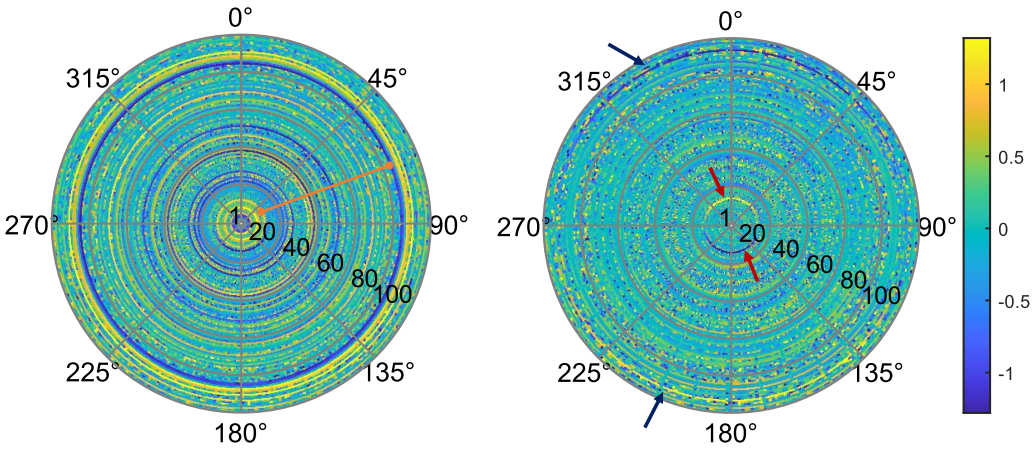
Such a neural coding procedure inspires us to automatically exploit an autoencoder to extract compact sound representations. Therefore, we train an autoencoder to compress and encode data to a high-level latent feature space. An autoencoder consists of two parts: an encoder to compress data and a reversed structure named a decoder, which can reconstruct encoded features into the original input without much information loss.

As the input is a 2D temporal series, we use the seq2seq framework [44] to build a Gated Recurrent Unit-based Variational AutoEncoder (GRU-VAE). As shown in Fig. 7, two GRU layers are used to form an encoder. Like the Long Short-Term Memory (LSTM) layer, GRU can learn the long and short-term temporal context while having fewer parameters and better generalization capability. The encoder reads the gammatone coefficients I and maps them to a feature vector z with 100 dimensions. Instead of encoding latent features for the input data independently, we use a variational autoencoder to map the data into a multivariate normal distribution $\mathcal{N}(\mu, \sigma) \in \mathbb{R}^{100}$. After that, a latent feature z is sampled from this distribution. This variational design forces the encoder to learn a smoother feature representation, which is more generalized to unseen data. As a symmetric structure, two GRU layers are used to construct a decoder to recover the latent feature z to the data domain. Specifically, the VAE module V is pre-trained using massive audio samples in a self-supervised way by minimizing the following loss:

$$\mathcal{L}_v = KL(\mathcal{N}(\mu, \sigma), \mathcal{N}(0, 1)) + \lambda \left\| V(I) - I \right\|^2 \qquad (1)$$

where KL is the Kullback-Leibler divergence that measures the difference between two probability distributions, and $\lambda \left\| V(I) - I \right\|^2$ is the Mean Square Error (MSE) loss to guarantee that z is informative enough for reconstruction. λ is a weight constant. Once the training process is completed, the decoder part is cut off; the encoder is then frozen and grafted into the DeepEar framework.

Figure 8 illustrates the original and reconstructed gammatone coefficients of an audio sample. We can see that our GRU-VAE can extract representative high-level features from the original input without much information loss. We visualize the encoded latent features of this clip as the AoA changes in Fig. 6. The radius of these polar figures



(a) Left ear. (b) Difference between the two ears.

Fig. 6. The latent feature encoded by VAE along with different directions. The radius is the feature dimension, and the angle is the sound AoA. The left figure illustrates the latent features in the left ear, and the right one is the difference between the left and right ear.

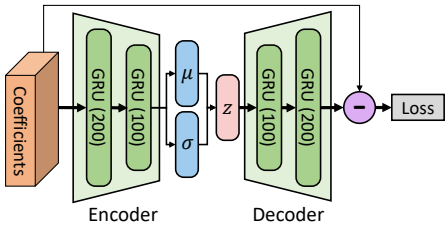
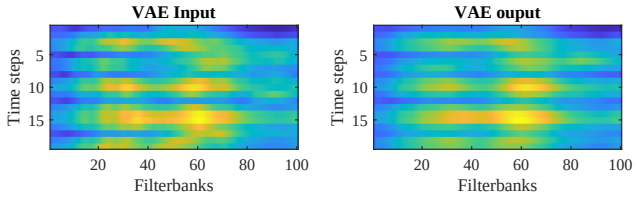


Fig. 7. Illustration of GRU VAE.



(a) VAE input. (b) VAE output.

Fig. 8. GRU VAE can effectively extract the latent features from original data and reconstruct them back with it.

is the feature dimension (*i.e.*, 1∼100). Figure 6(a) shows the latent features in the left ear. We can see that some parts of the latent features (*e.g.*, the $90th$ dimension) look similar in different directions and form several circles because the distortion effect of a single ear is not very notable. However, there are still diverse patterns in other dimensions (*e.g.*, $20th \sim 90th$), which provide direction clues for different AoAs. In Fig. 6(b), we subtract the latent features of the left ear from those of the right ear. This subtraction operation removes the signal impact and makes the difference between the two ears stand out. As a result, we observe rare apparent circles, indicating that the latent features of all directions are different from each other. More importantly, we can see clear distinguishable patterns between the front and back semi-field at $15th$ (red arrows) and $90th$ (blue arrows) dimensions. This observation confirms that binaural filtering performed by human shaped ears on sounds can effectively alleviate the front-back ambiguity problem.

As we mentioned before, the human brain uses the spatial patterns in sounds to perform localization. This spatial pattern arises from two aspects. First, different propagation paths cause subtle time differences between the two ears [45], so the ITD is associated with the sound azimuth. As such, we perform GCC-PHAT [46] between the signals of two ears as part of the features. The distance between two ears limits the maximum time difference between two ears. Hence, we only take the middle 100 coefficients ($\pm 3\ ms$) instead of all correlation results considering the extra multipath caused by the head and body. However, there is no one-to-one mapping between ITD and sound direction because of the ambiguity problem as we discussed. Then, ear filtering, as a second feature, can help. The ears produce microechoes to the arriving sound, leading to spectral distortion associated with specific spatial locations. Therefore, we fuse explicit correlation features and implicit latent sound representatives after ear filtering to jointly help DeepEar locate sound sources. Besides the separately encoded features from the left and right ears, we also subtract them and measure the feature differences between the two ears. Finally, all of these features are concatenated to form the final feature vector for sound localization.

3.5 Sound Localization

DeepEar divides 2D space into several equal sectors and detects whether a sound is present in a specific sector. If yes, then the AoA and distance of the sound source are estimated. We introduce the neural network design as follows.

3.5.1 Network Structure Design

With the extracted features, we design a neural network to perform multiple sound localization. In this research, a sector-based output is used to facilitate simultaneous multiple source localization with arbitrary spatial resolution. For example, we set the number of sectors as eight, which means that DeepEar supports up to eight co-active sources. We assume that there is at most one source in a sector. Although two sources may sometimes present in the same sector, it is sufficient for some applications such as hearing aids since

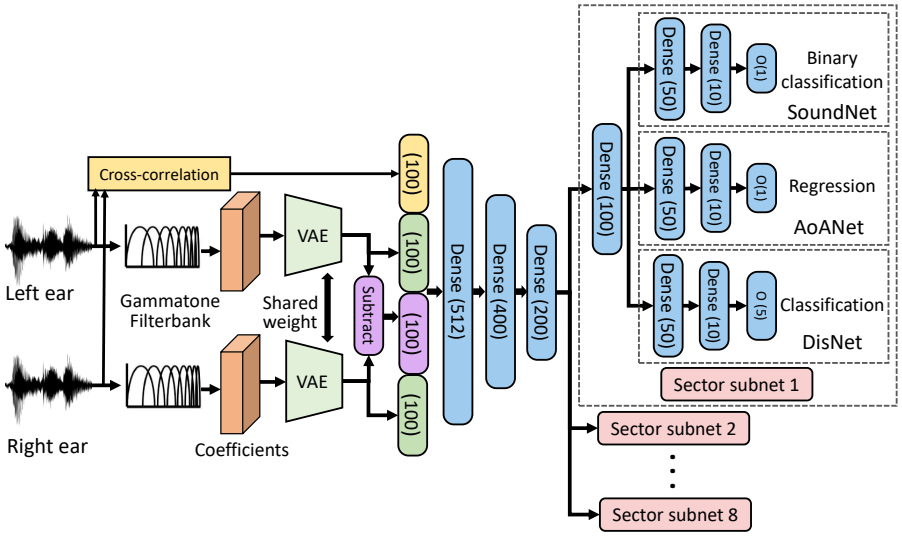


Fig. 9. DeepEar network design.

users do not need to strictly distinguish two very close sound sources. We can surely further increase the number of sectors to increase the spatial resolution according to specific application requirements (*e.g.*, an extreme case, one degree per sector). Here, we assume that the *maximum* number of concurrent sound sources is less than eight.

Figure 9 shows the DeepEar network design. The extracted features of the binaural channels are subtracted in the subtract layer to obtain the feature difference between the two ears. After that, all features are concatenated to a feature vector and input to the DNN-based sound localization network. We formulate the full-field localization as a multitask learning problem. The first three layers learn a general shared spatial pattern, followed by eight sector subnets responsible for each sector ($45°$). In each sector subnet, three task subnets share a common dense layer. The first task subnet is *SoundNet*, which detects if an acoustic source is present in this sector and produces a binary result. The second task subnet named *AoANet* predicts the AoA of the target. AoANet is a regression net whose output is a normalized value in [0,1], indicating the minimal and maximal degree in the sector. But we note that two adjacent sectors have a common degree on the boundary. For example, the degree range of sector 1 is [$0° \sim 45°$] and the scope of sector 2 is [$45° \sim 90°$]. They have an overlapped degree (*i.e.*, $45°$) at the sector boundary and so are other sectors. That is to say, for each sector, the first angle already appears in the previous sector. Therefore, the regression value 0 is meaningless. Thus, we leave it for the case where no sound source is present in the sector. Note that the loss of all sectors is counted even if the ground truth of a sector is the no-source case, since we expect that these samples could also help update AoANet to learn a complete feature space for AoA prediction. *DisNet* is the third task subnet that estimates the distance between the ears and the target source. Note that humans estimate distance by the sound loudness and the Direct to Reverberant sound Ratio (DRR). This perception result is much worse compared to the AoA estimation [47]. Therefore, we model the distance estimation as a classification problem and add an extra category for the no-present source case.

3.5.2 Loss Function

Overall, DeepEar has a 56-dimension output, and the whole network can be trained by minimizing the loss between the network output and ground truth. The SoundNets of all sectors can be regarded as a multilabel classification problem, so the activation function is sigmoid and the binary cross-entropy is used as the loss function:

$$\mathcal{L}_s = -y^s \cdot \log\left(\hat{y^s}\right) - (1 - y^s) \cdot \log\left(1 - \hat{y^s}\right) \tag{2}$$

where y^s is the ground truth of the DisNet, and $\hat{y^s}$ is the prediction probability.

As for AoANets, the mean squared error (MSE) is used to qualify this regression task:

$$\mathcal{L}_a = (y^a - \hat{y^a})^2 \tag{3}$$

where $\hat{y^a}$ is the regression output of AoANet.

Since DisNet is a multiclass classification problem, we use the softmax activate function and formulate its loss function as the cross-entropy:

$$\mathcal{L}_d = -\frac{1}{C}\sum_{i=1}^{C} w_i \cdot y_i^d \cdot \log \hat{y_i^d} \tag{4}$$

where C is the number of quantization distances, and w_i is the weight for each category. y_i^d is the $i\text{-}th$ one-hot encoding ground truth of this instance. We seldom observe many simultaneous sound sources (*e.g.*larger than 3), which leads to unbalanced data in the category without sound present. Therefore, we add weights to different categories to improve the generalization of DisNet.

In this case, the loss of one sector subnet $\mathcal{L}_{sector}$ is constructed as a weighted sum of the losses of three task subnets:

$$\mathcal{L}_{sector} = \alpha\mathcal{L}_s + \beta\mathcal{L}_a + \gamma\mathcal{L}_d \tag{5}$$

where α, β, and γ are weights for different task subnets. The most important requirement for DeepEar is successfully detecting concurrent sound sources, while we also expect a better AoA estimation than distance estimation. Thus, we empirically set these weights at 0.4, 0.35, and 0.25, respectively.

Finally, we can average the losses of all sector subnets and obtain the overall loss of the DeepEar network:

$$\mathcal{L} = \frac{1}{N}\frac{1}{M}\sum_{n=1}^{N}\sum_{m=1}^{M}\mathcal{L}_{sector(m)} \tag{6}$$

where M is the sector number and N is the number of training data in a batch.

3.6 Adaptation to New Environments

Humans have the ability to locate sound in various environments through continuous learning from childhood [48]. This ability indicates that humans can transfer knowledge from previous environments to a new context. Therefore, we first build a global model for DeepEar with the publicly anechoic dataset to learn the unalloyed ear filtering patterns for sounds from different directions. Then, we apply transfer learning [49] to make DeepEar adapt to new environments with a small number of new data.

DeepEar network can be divided into two components. The first is the general feature extraction module including the VAE, the feature concatenation layer, and three dense layers to learn the general knowledge of spatial patterns. Another part consists of eight subnets responsible for learning specific context information and performing several localization tasks. Thus, we employ transfer learning by freezing the first part of the pre-trained global model and fine-tuning the remaining subnets with a small amount of data from new environments. In this way, DeepEar can adapt to different working environments quickly, saving redundant and burdensome training overhead for users.

3.7 DeepEar Variants

To further investigate the localization capability of DeepEar, we propose two variants, namely, Complex DeepEar and Monaural DeepEar.

3.7.1 Complex DeepEar

The acoustic signals in the frequency domain include not only the amplitude but also the phase. In DeepEar, we directly input the magnitude coefficients after the gammatone filterbank, fusing the effect of magnitude and phase. Therefore, we propose Complex DeepEar, whose input consists of both magnitude and phase information. The target of Complex DeepEar is to investigate whether the phase can help further improve localization performance. Although we have used the cross-correlation to extract the most prominent time difference between two ears, other subtle time delay information at different frequencies may be neglected.

Biological literature reports that some brain stem nuclei, such as the superior olivary complex have the property of "phase locking" [50]. Consequently, they can compare the timing of stimulus spikes within the auditory nerve linking two ears and obtain the interaural time delay on different frequencies [51]. For this consideration, besides the power (*i.e.*, magnitude) of the gammatone spectrogram, we also feed phase values into Complex DeepEar, as shown in Fig. 10. Specifically, we perform Fast Fourier Transform (FFT) on each audio frame and extract the phase values at the center frequencies of gammatone filters. Then, the gammatone coefficients and phase values are fed into VAE separately. We trained a new VAE for phase encoding in the same approach as the magnitude VAE. After the feature encoding process, the magnitude and phase representatives are concatenated as the final feature vector for the localization network.

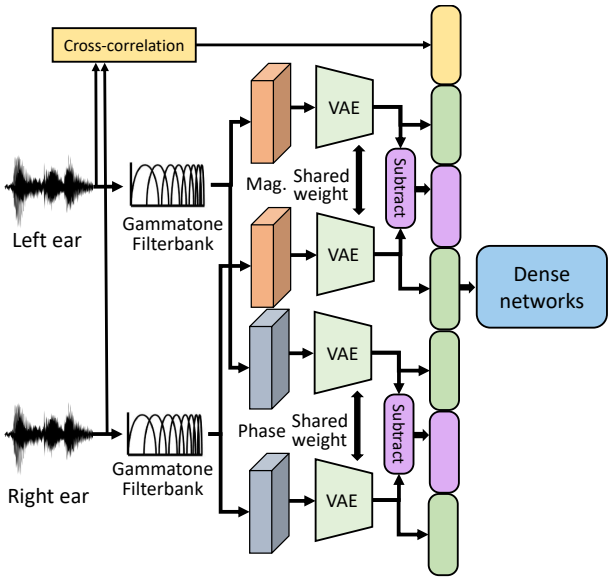


Fig. 10. Complex DeepEar network design. The Dense networks are the same as the right side of the original DeepEar (Fig. 9).

3.7.2 Monaural DeepEar

Despite the binaural localization, we notice that some hearing-impaired people only have one functional ear. As such, we also need to investigate whether we can apply the DeepEar methodology to a single ear. Previous studies show that, despite the difficulty, hearing-impaired listeners with only a single functional ear can also distinguish the sound direction to some extent [52], [53]. Such monaural localization is made possible by the external ears, which also distorts the sound depending on different angles, even with one ear. As a result, although without the binaural time clue, human beings can still learn special filtering patterns for different incident angles. Hence, we propose Monaural DeepEar, which only uses the sound of one ear as input to exploit the feasibility of monaural localization. Since the monaural clues rely primarily on the amplitude distortion rather than the phase, we only feed the gammatone coefficients of the left or right ear into Monaural DeepEar. After the encoding process, the latent features are forwarded into the localization network. This monaural localization approach loosens the binaural requirement, which benefits people who suffer from severe hearing diseases with only a single functional ear.

Table 2. Dataset summary.

Dataset	Anechoic-training	Anechoic-validation	Anechoic-testing1	Anechoic-testing2	Spirit	Auditorium	Rostock
BRIR convolved	Anechoic	Anechoic	Anechoic	Anechoic	Spirit	Auditorium3	Rostock
Sample size	72000	9000	9000	9000	9000	9000	9000

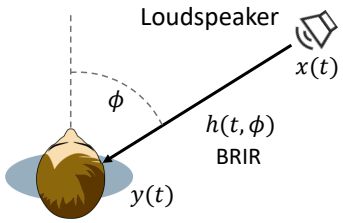


Fig. 11. Binaural spatial sound synthesis with BRIR data.

4 Implementation

We implemented DeepEar with Python and TensorFlow. The neural network and VAE were trained on a workstation with an Nvidia GeForce RTX 2080 Ti. We applied a dropout rate of 0.2 for each dense layer to prevent overfitting. The early-stopping strategy was used if no performance improvement was observed on the validation set for more than five epochs. DeepEar has $584K$ and $785K$ parameters for the VAE and the localization network. The feature extraction and model inference time for a sample is about 181.4 ms and 69.4 ms, respectively.

We follow existing binaural localization works [11], [12], [13] to generate binaural spatial sounds through synthetic recordings. As shown in Fig. 11, a loudspeaker source emits sound signals $x(t)$. It travels through the air channel and is then distorted by the ears, which can be characterized by BRIR $h(t, \phi)$, where ϕ is the incident angle of the sound. Finally, ears capture the sound $y(t) = x(t) \bigotimes h(t, \phi)$. Therefore, we can synthesize a variety of binaural sounds by convolving clean speech audio recordings $x(t)$ with BRIRs $h(t, \phi)$ of different locations. It is possible since various speech signals and BRIRs of different rooms are available in public datasets.

To this end, we randomly chose clean speeches from a corpus named TIMIT [54], which contains the speech recordings of 630 speakers with eight major dialects of American English. We choose the BRIR from the TU Berlin BRIR dataset [55]. This dataset was measured with a KEMAR dummy head (*i.e.*, binaural microphones with a head) in three different rooms, including an anechoic chamber [56], a small meeting room named Spirit [57], and a mid-size lecture room called Auditorium3 [58].

In the anechoic chamber, BRIRs were measured in the horizontal plane with a resolution of 1° for four different distances of 0.5 m, 1 m, 2 m, 3 m. There are 360 (degrees) × 4 (distances) = 1440 BRIR measurements in total accordingly. For the meeting room, BRIRs were measured for three different sources with a resolution of 1° and head movements from -90° to 90°. The distances between the three sound sources and the dummy head are 2 m. Therefore, the number of BRIRs is 181 (degrees) × 3 (sources) = 543 in this dataset. We note that three *sources* do not indicate three source *locations* instead of 543 locations because these sources also rotate relatively around the dummy head. Similarly, BRIRs in the lecture room were measured with the same resolution and rotation range but at six different loudspeaker positions. We also used the Rostock dataset [59] to evaluate DeepEar in more complicated environments, in which BRIRs were measured in an audio lab with 64 loudspeakers. The KEMAR dummy head used in Rostock rotates in a range of $\pm 80°$ with 2° steps. The reverberation time of this audio lab is 0.25 s at 1 kHz. We illustrate the measurement setup in Appendix and refer interested authors to the dataset references for more detailed descriptions.

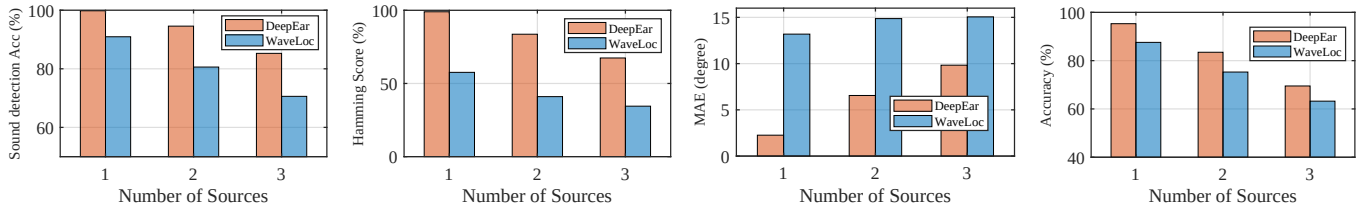
Considering that the number of concurrent primary sound sources is typically small in the real world, we set it uniformly distributed in [1, 3] but with a constraint where one source presents in a sector. The AoA is also sampled following a uniform distribution in each sector. The source index and corresponding distance are randomly selected since there are many sound sources in a dataset. In multisource cases, we add multiple sounds together as the superposition sound received by the binaural microphone. All synthesized data were sampled at 16 kHz and cut into 1-second instances for evaluation.

5 Evaluation

5.1 Experiment Setup

We first train a global model for DeepEar with anechoic data only to learn the unalloyed ear filtering patterns to the sounds from different directions. After that, DeepEar can be customized and adapted to the real-world (*i.e.*, reverberant) data by transfer learning with a minimum amount of new data collected in target working environments.

The clean speech recording corpus TIMIT consists of two portions, TRAIN and TEST. No human speakers and no speech text overlap between these two portions. We first randomly selected speeches in the TRAIN portion, and convolved them with the anechoic BRIR as anechoic data to build a global model. These data were divided into three parts with a ratio of 8:1:1. We denote them as anechoic-training, anechoic-validation, and anechoic-testing, respectively. Given that anechoic-training and anechoic-testing data are split from the same portion (*i.e.*, TIMIT TRAIN), their speech text or speakers may overlap, although their sound locations are different. Therefore, we separately took random clean speech recordings in the TEST portion and synthesized a new testing dataset to evaluate the model robustness to unseen speeches and speakers, denoted Anechoic-testing-2 (accordingly, the former testing set is renamed as Anechoic-testing-1). Moreover, we similarly convolved the clean speeches randomly selected in the TEST portion with the real-world BRIRs to generate other



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 12. Performance comparison between DeepEar and WaveLoc on the anechoic-testing1 dataset.

Metircs	Sound detection (%)				Hamming score (%)				AoA MAE (degree)				Distance (%)			
Source #	ave	1	2	3	ave	1	2	3	ave	1	2	3	ave	1	2	3
DeepEar	91.9	99.8	92.5	83.5	80.5	99.1	78.2	64.1	8.0	2.3	7.7	10.1	81.6	95.2	81.2	68.4
WaveLoc	80.4	90.9	80.0	70.3	43.2	56.7	39.3	33.7	14.5	13.2	15.2	14.5	75.0	87.5	75.0	62.6

Table 3. Performance comparison between DeepEar and WaveLoc in the anechoic-testing2 dataset.

three testing datasets, including a meeting room (Spirit), a lecture room (Auditorium), and a lab (Rostock). Overall, we obtained seven datasets: one for training, one for validation, and five for model testing. We summarize the name, size, and usage of all datasets in Tab. 2. We should note that there are only four distances in the training data (*i.e.*, the anechoic chamber, Sec. 4), and most distances in other testing rooms are inconsistent. In this case, we regard it as a correct prediction if the distance in other rooms is classified into its closest distance in the training data (*e.g.*, 2.93 m $\rightarrow$ 3 m) since the classification results are discrete values.

For comparison, we implemented a state-of-the-art binaural localization WaveLoc [11]. WaveLoc decomposes binaural signals into 32 frequency bands and then employs a CNN on the raw waveform to classify the AoA. Note that WaveLoc only supports single-source azimuth classification, so we replaced the last layer of WaveLoc with DeepEar's localization network (*i.e.*, sector subnets) to enable multiple sound localization to evaluate the effectiveness of the proposed feature. In addition, we also conducted a real-world case study with a binaural microphone to locate the sound with and without ears to further verify the importance of ears.

5.2 Evaluation Metrics

We evaluate DeepEar with the following metrics:

- Sound detection accuracy. It measures the binary classification accuracy of SoundNet for detecting whether there is a sound source in a spatial sector.
- Hamming score of sound detection. Hamming score is defined as the proportion of the correctly predicted labels to the total positive labels (predicted and actual) for a sample:

$$H = \frac{1}{N}\sum_{n=1}^{N}\frac{sum(y_n^s \;\&\; \hat{y_n^s})}{sum(y_n^s \mid \hat{y_n^s})} \tag{7}$$

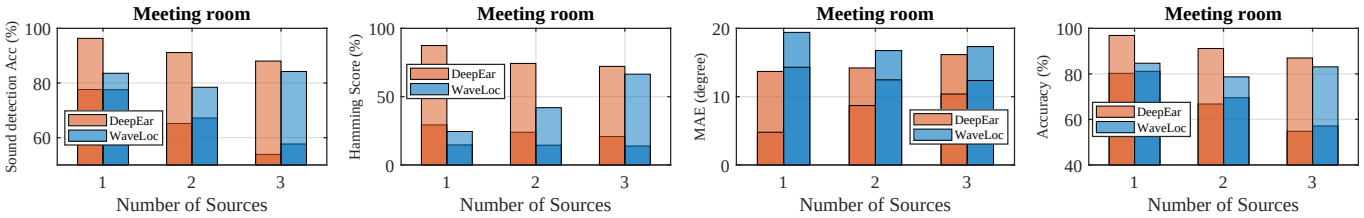
where y_n^s is the ground truth of eight SoundNets of the n-th instance. $\hat{y_n^s}$ is the corresponding classification result. $\&$ and $\mid$ represent bitwise AND and OR operations, respectively. Compared to detection accuracy, the Hamming score ignores the true negative (*i.e.*, a no-source case is correctly recognized) and penalizes false positive cases (*i.e.*, a no-source case is mistakenly detected as an active source).
- Mean Absolute degree Error (MAE) of AoA. MAE means the average absolute degree error between the predicted AoA and the ground truth. We ignore the sectors corresponding to no-source cases.
- Distance classification accuracy. This metric refers to the average accuracy of all DisNets.

5.3 Overall Performance

Figure 12 shows the performance of the global model in the anechoic-testing1 data. Overall, the sound detection accuracies of DeepEar and WaveLoc are 93.3% and 80.9%, respectively. Furthermore, DeepEar has a high detection accuracy of 99.8% in the one-source scenario. In comparison, the performance of WaveLoc is slightly lower, with a detection accuracy of 90.9% in this case. We can see that the performance of both models decreases with increasing number of sound sources. When three sources coexist, the detection accuracy of DeepEar drops to 85.3%, and WaveLoc's accuracy decreases to 70.6%.

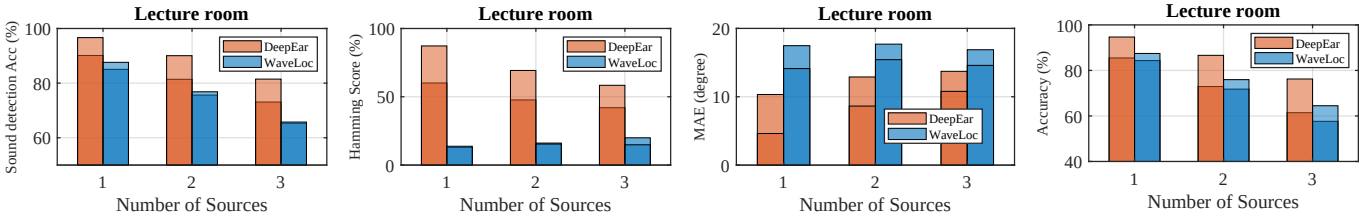
In general, the Hamming score of DeepEar is 83.5%, slightly lower than the detection accuracy, since all cases without sound sources are excluded. However, the performance of WaveLoc drops by almost a half and decreases to 44.6%. This degradation indicates that WaveLoc makes more false positive sound detection than DeepEar.

As for AoA estimation, the mean absolute degree error of DeepEar is 7.4°, which is nearly a half of WaveLoc's. In the one-source case, DeepEar can even predict AoA within an error of 2.3°. However, the MAE of WaveLoc is 13.2° in this setting, much larger than DeepEar. It is because that WaveLoc performs CNN directly on raw waveforms, missing the key time difference information between binaural channels and filtering patterns in the frequency domain. With the increasing number of sources, multiple sounds interfere with each other and their time differences are confused, leading to a higher estimation error.



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 13. Performance comparison in Spirit meeting room. The darker bars refer to the accuracy before transfer learning or MAE after transfer learning.



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 14. Performance comparison in the Auditorium lecture room. The darker bars refer to the accuracy before transfer learning or MAE after transfer learning.

The average distance accuracies of all source cases are 82.9% and 75.6% for DeepEar and WaveLoc, respectively. Same as before, the larger the number of active sources, the lower the estimation performance.

We also evaluate DeepEar on the anechoic-testing2 dataset. This dataset is generated separately rather than splitting from the original one. The result is listed in Tab. 3. Overall, the sound detection accuracy and Hamming score of DeepEar are 91.9% and 80.4%, respectively. The performance is nearly the same as that on anechoic-testing1 data, as well as AoA MAE (8°) and distance accuracy (82%)). The performance of WaveLoc is still lower than DeepEar in terms of all metrics. This result indicates that DeepEar generalizes well to unseen data. This is likely because we synthesized massive training data to train a global model, and the VAE can also learn a smooth latent feature space to adapt unseen speakers and locations.

5.4 Real Environment

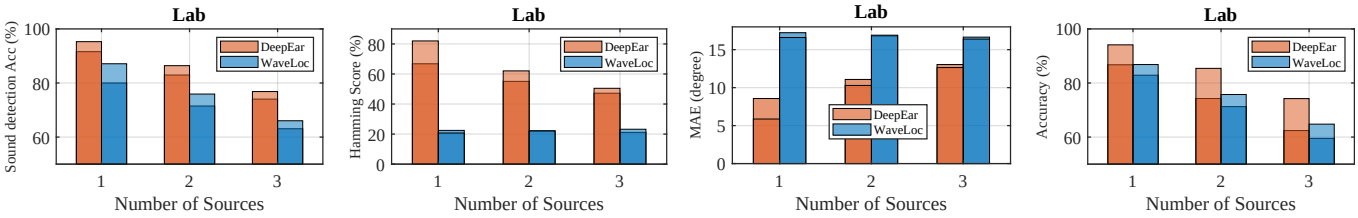
The DeepEar trained in the anechoic environment has learned the spatial filtering patterns of the ear, so it is our turn to examine DeepEar in real reverberant rooms, including a small meeting room, a larger lecture room, and a lab room.

5.4.1 Evaluation in a Small Meeting Room

Figure 13 illustrates the performance of a small meeting room (Spirit). As we expected, directly testing the global model on the reverberant data brings about a dramatic performance deterioration. The baseline WaveLoc also performs poorly in reverberant environments. The average sound detection accuracy and Hamming score of DeepEar are 65.6% and 24.7%, while WaveLoc achieves 67.3% in sound detection and 14.3% in Hamming score, respectively. Although the sound detection accuracy of WaveLoc is comparable to that of DeepEar, the Hamming score of DeepEar is much higher than WaveLoc. Similarly, the performance of AoA and distance estimation also decreases. The reason is that signals in a reverberant environment differ substantially from those in an anechoic room.

We perform transfer learning to adapt the global DeepEar model to this meeting room. Specifically, we split the dataset (Spirit) into two portions: 10% for model adaptation (Spirit-adaptation) and the remaining 90% for performance evaluation (Spirit-testing). There is no overlap between the two portions. Besides, we also conducted an end-to-end fine-tuning for WaveLoc except for its first layer used for frequency decomposition. Both models converge fast within ten epochs and exhibit much better performance than before. The sound detection accuracy of DeepEar increases to 91.9%, while WaveLoc only achieves 82.1%. The Hamming score of DeepEar increases by 53.3%, almost double that of WaveLoc. The DeepEar's AoA MAE decreases to 8.8°, which is very close to the anechoic case. Moreover, the performance increase in terms of distance estimation is 24.4% for DeepEar (to 91.9%) and 15.1% for WaveLoc (to 82.3%), respectively. This figure shows that both methods benefit from transfer learning when tested on the new reverberant data. Nevertheless, DeepEar notably outperforms WaveLoc after the same retraining procedure. The reason may be that WaveLoc uses CNN on time-domain sample series, losing the correlation between different frequencies and ears. Therefore, it is difficult to adapt WavLoc to new environments with a relatively small number of additional training data.

We also breakdown the AoA evaluation result of DeepEar for different sources in this meeting room as shown in Fig. 18. We can observe that the AoA MAEs for three sound sources are 15.8°, 14.5°, and 15.3°, respectively. In addition, they also have a comparable performance after transfer learning, decreasing by 6.4° on average. This result shows that DeepEar generalizes well to different sound sources.



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 15. Performance comparison in Rostock lab. The darker bars refer to the accuracy before transfer learning or MAE after transfer learning.

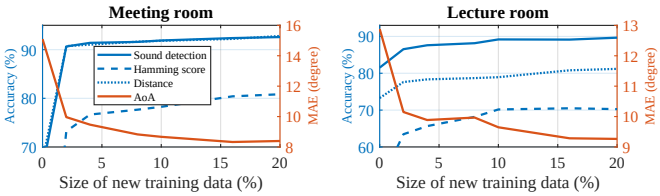
5.4.2 Evaluation in a Large Lecture Room

In this experiment, we evaluate DeepEar in a large lecture room with six different sound sources (Auditorium). As shown in Fig. 14, the overall sound detection accuracy of DeepEar is 81.5%, *i.e.*, 6.2% higher than WaveLoc. In terms of Hamming score, the performance gap is even wider. In particular, WaveLoc decreases to 16.3%, approximately one third of DeepEar (49.9%). Besides, the AoA estimation errors of these two systems are 12.9° and 17.3°, respectively. Although both methods suffer performance degradation in this reverberant environment, DeepEar still performs much better than WaveLoc. This result shows that DeepEar is more robust to the highly reverberant new environment than WavLoc.

Transfer learning is applied with 10% Auditorium data, and the remaining 90% data are used for model testing. We observe that this adaption strategy is effective in improving the performance of both models, and DeepEar benefits more than the benchmark method. Specifically, the sound detection accuracy and Hamming score of DeepEar increase to 89.4% and 71.7%, respectively. In contrast, the sound detection accuracy of WaveLoc only has an increase of 1.8%. The AoA MAEs of DeepEar and WaveLoc decrease by 3.9° and 2.5°, respectively. Furthermore, the distance accuracy of DeepEar and WaveLoc increases to 91.7% and 76.4%. Again, DeepEar still outperforms the baseline regarding distance and AoA estimation. A noteworthy aspect is that the Hamming score of WaveLoc declines from 16.3% to 14.6% after transfer learning. The main reason is that the lecture room is relatively large, which is more reverberant than the meeting room. The CNN mechanism of WaveLoc relies more on time-domain data and even hampers it from adapting to the reverberant environment. In contrast, DeepEar benefits from the variational encoding and can calibrate the feature distribution accordingly with new data, thereby achieving better performance.

5.4.3 Evaluation in a Lab with Many Sources

We also conducted an experiment in a lab with 64 loudspeakers around this room (Rostock). As shown in Fig. 15, the overall sound detection accuracy of DeepEar is 82.9%, higher than that of WavLoc by 11.4%. The Hamming scores for DeepEar and WavLoc are 56.4% and 21.9%, respectively. In addition, the DeepEar AoA MAE is 11.6°, which is less than that of WavLoc (16.8°). In terms of distance prediction accuracy, DeepEar reports 74.5% and WavLoc is 3.3% lower than it. We can see that DeepEar performs better than WavLoc in an environment with a large number of different sound sources.



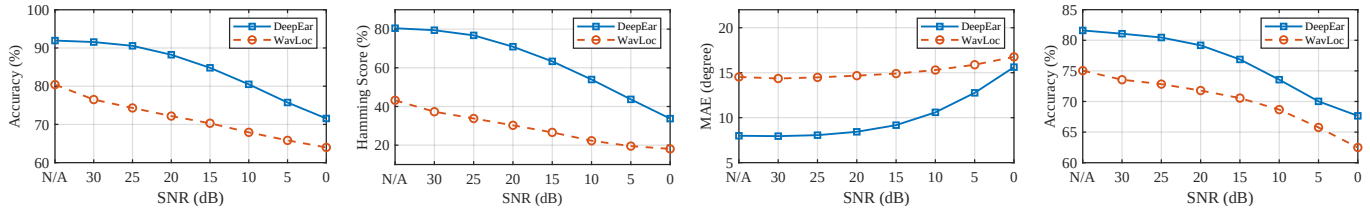
(a) Meeting room. (b) Lecture room.

Fig. 16. The transfer learning performance of DeepEar with different sizes of new training data. Two subfigures share the same legend.

After transfer learning with 10% data of the Rostock dataset, we evaluate the adapted model with remaining 90% data. As shown in Fig. 15, the sound detection accuracy of DeepEar increases to 86.2%, while WavLoc only increases to 76.4%. The distance accuracy of DeepEar increases to 84.6% while WavLoc climbs to 75.8%. Additionally, DeepEar benefits more from transfer learning than WavLoc, especially for the Hamming score and AoA estimation. Specifically, the Hamming score of DeepEar increases by 8.5%, while WavLoc only has a negligible increase (0.1%). As for AoA, DeepEar also obtains more performance gain than WavLoc, especially in one-source cases (2.7° vs. 0.6°). This result confirms that CNN-based WavLoc is difficult to adapt to a complicated environment with only a small amount of data. Although DeepEar has more performance improvement due to its human-inspired framework design, the overall performance gain from transfer learning is less than those in the meeting room and lecture room. The rationale behind this is the sophisticated reverberant environment with too many sound sources, which hinders the global DeepEar model from transferring to this new context.

5.4.4 Transfer Learning Performance

The experiment results above demonstrate that transfer learning effectively helps DeepEar adapt to new environments. We also tested DeepEar with different sizes of adaption data for transfer learning in the meeting room and lecture room because of their high performance improvement. In this experiment, we gradually increase the proportion size for adaptation data, and the rest of dataset are used for model testing. The result is illustrated in Fig. 16. We zoom in on the y-axis for clear observation. We observe that only 2% of new data can essentially boost DeepEar performance in both the small meeting room and the large lecture room. The accuracy steadily increases as the number of training data grows. Consequently, the MAE gradually decreases. In theory, the more new data used in transfer learning,



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 17. Performance comparison between DeepEar and WavLoc across different noise levels. "N/A" indicates that no noise is added to the signal.

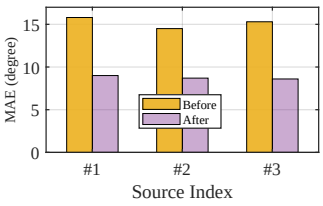


Fig. 18. DeepEar performance per source before and after transfer learning in spirit meeting room.

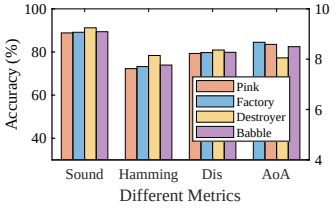


Fig. 19. DeepEar performance across different types of noise. AoA refers to the right y-axis.

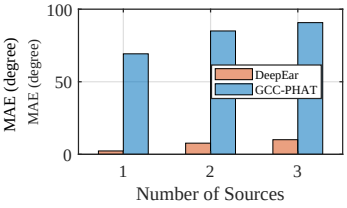


Fig. 20. AoA estimation comparison between DeepEar and GCC-PHAT.

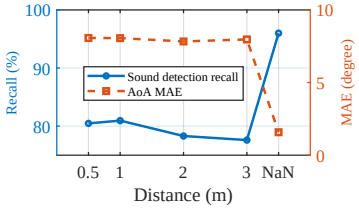


Fig. 21. DeepEar performance across different distances. The "NaN" denotes no-sound cases.

the better performance we can achieve. Nevertheless, we need to balance the performance gain and the extra training overhead introduced, since collecting a large number of new data in different environments could be practically challenging for ordinary users. This experiment reveals that 2% data for model fine-tuning (*i.e.*, 180 one-second instances) are efficient for DeepEar to produce a good adaption result, while DeepEar can achieve higher performance with 10% or more of new data if needed.

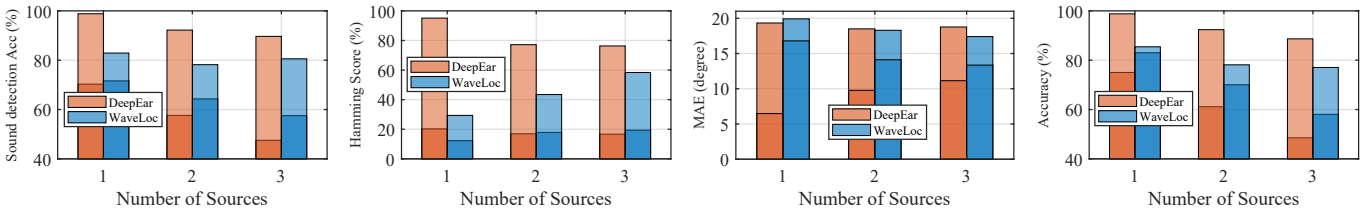
5.5 Noisy Environment

We added Gaussian noise with different signal-to-noise (SNR) levels (30 $dB \sim 0$ dB) to binaural signals in anechoic-testing2 to evaluate DeepEar in noisy environments. Figure 17 depicts the performance comparison between DeepEar and WavLoc across different SNRs. "N/A" means the result without any noise. We can observe that DeepEar keeps stable performance when the SNR is higher than 25 dB, where the sound detection accuracy, Hamming score, and distance accuracy are about 90.6%, 76.8%, and 80.5%, respectively. The corresponding AoA MAE is about 8.1°. In comparison, WavLoc suffers notable performance deterioration when encountering noise. Specifically, the sound detection accuracy and Hamming score decrease by 6.1% and 9.4% at 25 dB, respectively. As the noise level increases, the performance of both systems degrades rapidly. When SNR is 0 dB, the sound detection accuracy and Hamming score of DeepEar drop to 71.6% and 33.7%. The AoA MAE of WavLoc increases slightly slower than that of DeepEar. However, its MAE at 0 dB (16.8°) is still higher than that of DeepEar (16.8°). This result reveals that DeepEar is more robust to noise than WavLoc, but they both cannot handle relatively noisier scenarios. We note that the global DeepEar model used in this experiment is trained on anechoic data, so there is a large improvement space if we perform robust training strategies such as multi-conditional training (MCT) [12].

We also evaluate DeepEar with different kinds of noise (pink, factory, destroyer, and babble) selected in the Noise92X noise database [60]. Same as the experimental setting for Gaussian noise, we added them to binaural signals with an SNR of 25 dB. From Fig. 19, we can see that DeepEar performs slightly better under destroyer noise. But overall, the performance remains relatively stable in terms of all metrics under different types of noise.

5.6 Comparison with GCC-PHAT

Subspace-based AoA estimation methods such as MUSIC require that the number of microphones should always be larger than the sound number [6]. Since we only have two microphone channels, these approaches are not feasible in such multisource cases. Thus, we choose another typical approach GCC-PHAT for comparison. In this experiment, the dummy head can be considered a linear array with two microphones apart with head size (*i.e.*, 18 cm for the KEMAR dummy head). We used Anechoic-testing-2 as the evaluation set to exclude the noise impact, and the result is shown in Fig. 20. We can see that the AoA MAE of GCC-PHAT is 69° in the one-source case. It further increases to 85° and 91° in two-source and three-source cases, respectively, much higher than that of DeepEar. The reasons arise from many aspects. First, the signal does not travel to the ears in a straight line but diffracts due to the head curvature, leading to incorrect time delay estimation between two ears. Second, the low signal sampling rate (16 kHz) determines low spatial resolution, where one sample lag denotes 22.5° azimuth using the cross-correlation method. In addition, the cross-correlation peaks with only two microphones are easily distorted in the presence of multiple sound sources. Finally and more importantly, as we mentioned in Sec. 1, one microphone pair can only achieve the semi-field AoA estimation, which brings about a severe front-back confusion problem and significantly raises AoA MAE. And what



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 22. Performance comparison with different ear shapes (a Cortex MK2 dummy head). The darker bars refer to the accuracy before transfer learning or MAE after transfer learning.

is more, this consequence becomes worse in a multisource situation.

5.7 Impact of Distance

We conduct a detailed analysis of the evaluation result and investigate the DeepEar performance across different distances. In this setting, we assume that the distances of source sources are known as a prior, so the sound detection accuracy and Hamming score are not applicable. Instead, we adopt sound detection recall (true positive rate) as the sound detection metric, which indicates how many sound sources are correctly detected at a specific distance.

The result is shown in Fig. 21. We can see that the sound detection recall decreases with the increasing distance. However, the AoA estimation error (MAE) remains relatively stable across different distances. This is because the sound power from sources far away is very weak; hence they are hard to detect. But say, as long as DeepEar successfully detects the sound, it can extract the interaural clues and infer the corresponding sound direction. As a result, the AoA estimation performance does not suffer degradation as the distance increases. The ”NaN” on the x-axis denotes the no-source case. We use ”NaN” instead of ”0” to avoid misunderstanding. In this case, we can see that the recall is relatively high since no sound is much easier detected than sound cases. Moreover, the MAE of AoA is near 1.6°. The reason is that the ground truth label of a no-sound case is 0 (explained in Sec. 3.5.1). Thus, even though DeepEar correctly detects a no-source case, the AoA estimation value is minimal but not equal to zero. These small residuals also lead to an error.

5.8 Adaption to New Ears

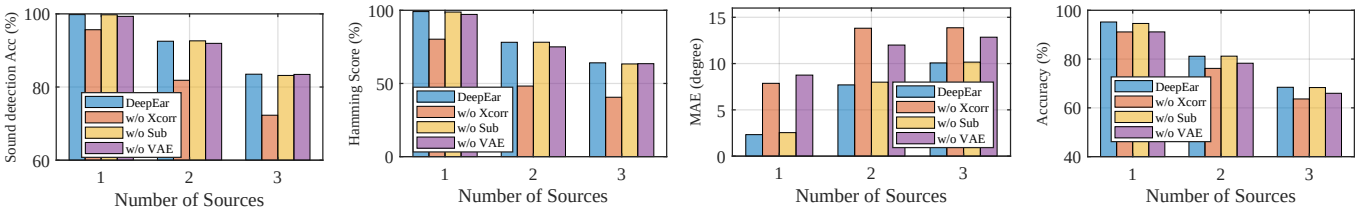
Different ear shapes may cause distinct sound distortion effects. Therefore, we synthesized a new testing set with Surrey BRIR (medium-small classroom) [61], which is recorded with a Cortex MK2 dummy head. The data synthesis setting is the same as previous datasets. Figure 22 shows the performance comparison between DeepEar and WavLoc. When we directly deploy two methods on this new dataset, the average sound detection accuracies are 58.5% for DeepEar and 64.5% for WavLoc. Although the sound detection accuracy of WavLoc is higher than DeepEar, their Hamming scores are comparable, which are 18.0% for DeepEar and 16.5% for WavLoc. Accordingly, the AoA MAEs of DeepEar and WavLoc are 18.8° and 18.1°, and their distance accuracies are 61.6% and 70.4%, respectively. The low Hamming score and the high AoA MAE denote that DeepEar can hardly locate sound sources in this context. This result indicates that, in addition to reverberation, the different ear filtering effect further degrades original models. An interesting finding is that the performance of DeepEar is much worse than that of the previous three rooms, likely because the ear-filtering features used by DeepEar are more sensitive to the ear shape (changed from a KEMAR dummy head to Cortex MK2). By contrast, the performance of WavLoc with this new ear is comparable to that of other three new rooms, although a large performance degradation is observed as well.

We also split 10% of this dataset as the adaptation set for transfer learning, and the remaining data are used for testing. As shown in Fig. 22, we observe a significant performance boost for DeepEar, especially for the cases with less number of sound sources. In particular, the average sound detection accuracy and Hamming score of DeepEar substantially increase to 93.6% and 82.9%, respectively. The same metrics for WavLoc after transfer learning are 80.6% and 43.7%. As for AoA, the MAE of DeepEar decreases almost by half to 9.9°, while that of WavLoc only decreases to 14.2%. The distance accuracy also remarkably increased for DeepEar (31.7%), higher than WavLoc (9.8%). Overall, the evaluation result shows that the transfer learning strategy can effectively help DeepEar adapt to new ears, *i.e.*, binaural microphones. The possible reason is that the human-inspired features used by DeepEar can quickly adapt the feature space to new ears, as long as with a few numbers of data.

5.9 Ablation Study

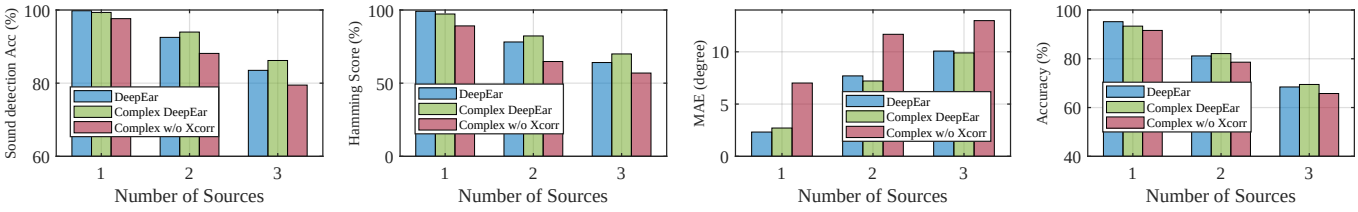
We conducted an ablation study to evaluate the importance of different components in DeepEar. Specifically, the cross-correlation and subtraction features were removed successively, and then we replaced the VAE with two general GRU layers. Anechoic-training and Anechoic-testing2 were used as the training and testing dataset, respectively.

The results are shown in Fig. 23. We can see that the sound detection accuracy decreases from 91.9% to 83.3%, and Hamming score drops by 24% without cross-correlation features. Accordingly, the AoA estimation error increases by 5.9°. This is because the cross-correlation feature apparently provides the time difference between two ears, indicating the sound direction. Thus, DeepEar is hard to accurately distinguish the sound direction without this feature. The distance accuracy, however, decreases a little (5.9%), since the direct to reverberant ratio mainly used for distance estimation is kept in the extracted gammatone coefficients.



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 23. Performance of DeepEar ablated with cross-correlation (Xcorr), subtraction (Sub), and VAE.



(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

Fig. 24. Performance of DeepEar, complex DeepEar, and complex DeepEar without cross-correlation.

After ablating the subtraction part, we observe almost the same performance as before. This result is not surprising. Subtraction is a simple task; hence the network can easily learn this operation within hidden layers. As we illustrated in Sec. 3.4, the feature difference between two ears is an essential factor helping us break the AoA front-back ambiguity. Therefore, despite a slight performance gain, adding subtraction as a part of features can reduce the learning burden and accelerate model convergence.

Replacing VAE brings about a performance decrease, especially for AoA estimation. Specifically, the Hamming score and the distance accuracy drop by 2% and 4%, and AoA estimation error increases from 8° to 11.9°. The reason is that VAE has a generalization ability to unseen data due to continuous representation distribution. Therefore, the system performance degrades without the VAE, although we have used massive data to train a global model.

5.10 Performance of DeepEar Variants

5.10.1 Complex DeepEar

We train Complex DeepEar with the anechoic-training dataset and test it on the anechoic-testing2 dataset. The result is shown in Fig. 10. Compared to the original DeepEar, sound detection accuracy and Hamming score increase by 2% and 2.3%, respectively. The AoA MAE decreases to 7.7°, and the distance accuracy increases to 82%. These results are in accordance with our expectations, since phase information provides richer time differences between two ears that can help with sound localization. The performance boost is especially notable in 2-source and 3-source cases.

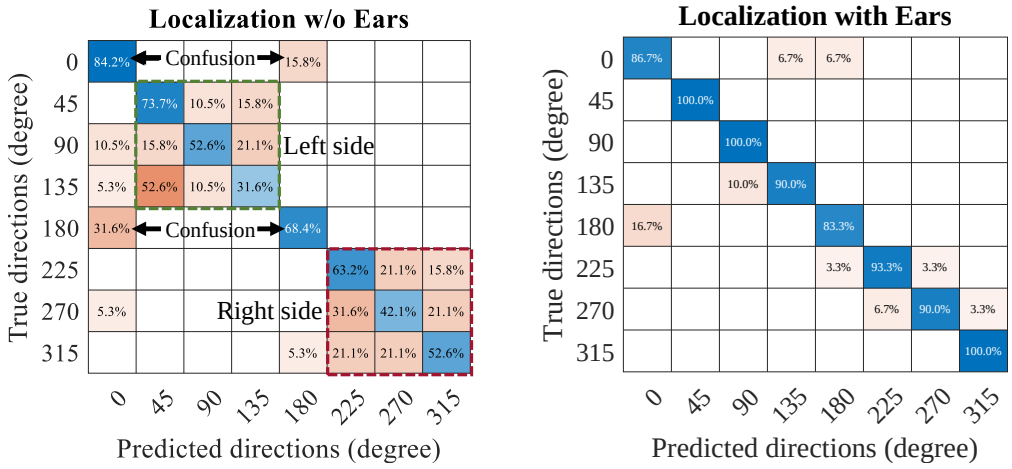
We also repeated the experiment but removed cross-correlation features from complex DeepEar. Like the result of the ablation study we have done in Sec. 5.9, the performance of Complex DeepEar decreases in terms of all metrics accordingly. We found that the performance gain of the phase is not as great as the cross-correlation. We suspect that the interaural time delay estimated with phase suffers from a "phase wrapping" problem if the phase change between two ears is greater than 2π. Compared with phase-inferred time delay information, cross-correlation can provide a more prominent time difference estimate.

However, what we want to point out here is, although Complex DeepEar experiences a large degradation without cross-correlation, its performance is still better than the original DeepEar without cross-correlation (Sec. 5.9). For example, sound detection accuracy and Hamming score are 88.5% and 70.4%, but still higher than the original DeepEar by 5.2% and 14%, respectively. The AoA MAE is 11.5°, less than the original DeepEar by 1.4°. A possible reason is that DeepEar can partly unwrap the phase and infer the interaural time differences with the redundant information of multiple frequency bands [62].

5.10.2 Monaural DeepEar

We also evaluate Monaural DeepEar in the same experimental setting as Complex DeepEar. The overall result of different metrics is illustrated in Tab. 4. We can see that Monaural DeepEar can achieve promising sound localization performance, although there is a large space to improve. The reason is that the unique pinna structure can still distort the sound and produce angle-dependent monaural clues even with one ear [52]. Furthermore, the performance of the two ears is almost the same.

However, without the help of another ear, human beings cannot cancel the sound contents between two ears and extract the binaural difference patterns. It means that listeners with one functional ear can only locate the sounds with which they are familiar [53]. Some researchers also reveal that people with hearing diseases often turn around their heads slightly and can locate a rough sound direction [63]. In this case, the head rotation leads to a different propagation path between the sound source and the ear, yielding new reference information to help achieve monaural localization. This promising result shows that Monaural DeepEar can potentially benefit people who suffer from severe hearing diseases with only a single functional ear.



(a) Without human-shaped ears. (b) With human-shaped ears.

Fig. 25. Localization performance with and without human-shaped ears.

Table 4. Performance of Monaural DeepEar of the left and right ear.

Metrics	Sound detection (%)	Hamming score (%)	AoA MAE (degree)	Distance (%)
Left Ear	82.8	50.7	13.6	77.1
Right Ear	83.3	53.1	13.1	77.5

5.11 Real-world Case Study

We conducted a real-world localization experiment to further evaluate the importance of ears for sound localization. A binaural microphone (miniDSP EARS) is placed in a meeting room as a recording device. Several speech files were randomly selected from the public TIMIT corpus to form long audio with 120 seconds. Then we used a portable loudspeaker to play the selected audio files in eight $45°$ evenly spaced directions $1m$ away from the microphones. We first recorded the binaural audio with ears and then repeated this process but detaching the human-shaped ears from the binaural microphone. After that, each long audio recording was sliced into many one-second samples. Twenty gammatone coefficients were extracted from each 0.1 s frame in a sample as a feature.

We implemented a one-layer LSTM network consisting of 100 hidden units stacked with a dense layer to execute the sound localization task. Figure 25 shows the confusion matrices of localization with and without ears. Without ears, the localization accuracy is 58.6% as shown in Fig. 25(a). We can observe that the model suffers from front-back confusion. Although directions on the left or right side (*e.g.*, 90° and 270°) can be easily detected, the model can hardly identify the degrees on each side (*e.g.*, 45°, 90°, and 135°). For comparison, the overall classification accuracy increased to 92% after mounting the ears, as shown in Fig. 25(b). The confusion problem was alleviated to a great extent and accuracy in almost all directions was improved. This result confirms that human-shaped ears indeed help to significantly improve localization accuracy, especially for AoA disambiguation.

6 Discussion and Open Problems

6.1 HRTF Calibration

Although the ear-caused HRTF is unique and cannot be directly applied to different ears, our experiment result shows that transfer learning can help DeepEar adapt to new binaural microphones. However, the precondition is that we must collect a certain amount of data with new ears. Recent work UNIQ [64] personalizes HRTF for different users with a smartphone and a pair of in-ear microphones. [65] proposed a regression approach to estimate the HRTF based on the ear's 3D shape. These HRTF personalization approaches provide an opportunity to apply our model to different binaural microphones with only an online calibration process. Moreover, recent research found that humans can get used to new mold ears in a few weeks [48], which indicates that we may perform incremental learning strategies to facilitate HRTF generalization among different ears.

6.2 3D Localization

We focus on horizontal sound localization in this research. In fact, humans can locate full 3D sound directions with reasonably high accuracy, including both azimuth and elevation. While the primary cues for azimuth localization are binaural, the primary cues for elevation localization are often regarded monaural [66]. This is mainly due to the fact that the pinna can distort the sound in a direction-dependent manner [67]. Furthermore, the head, shoulder, and torso also produce distinct filtering patterns in different elevation angles. [68], [69] provide 3D HRTF databases that can be used for sound elevation localization. Some works also reveal that people often turn their heads slightly, and thus they can locate a sound direction more accurately [12]. We leave this for future work.

7 Conclusion

In this paper, we propose DeepEar, a sound localization framework for binaural microphones that can locate multiple sources without the number of sources. Inspired by the human auditory system, we design a machine hearing

framework to fuse binaural time differences and latent sound representatives to estimate the locations of multiple sources. To cope with the heterogeneity of working environments, a global DeepEar model is trained on available anechoic datasets. Then we take advantage of the transfer learning strategy to adapt DeepEar in real working scenarios. DeepEar investigates the significance of the ears on binaural microphones in sound localization. Experiment results demonstrate that DeepEar substantially outperforms the state-of-the-art work in terms of sound detection and localization accuracy. We believe that DeepEar could not only benefit hearing-impaired people with smart hearing aids but also fuel more binaural applications in the future.

ACKNOWLEDGMENTS

This work is supported by the Hong Kong GRF under grant PolyU 152165/19E. Yuanqing Zheng is the corresponding author.

REFERENCES

[1] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: Technologies and algorithms," *Advances in digital speech transmission*, vol. 14, pp. 401–429, 2008.

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

[3] W. Wang, J. Li, Y. He, and Y. Liu, "Symphony: localizing multiple acoustic sources with a single microphone array," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 82–94.

[4] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[5] 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.

[6] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 18–21.

[7] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4642–4646.

[8] 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.

[9] S. Chakrabarty and E. A. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[10] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[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] 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.

[13] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.

[14] 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.

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

[16] 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.

[17] Z. Pan, M. Zhang, J. Wu, J. Wang, and H. Li, "Multi-tone phase coding of interaural time difference for sound source localization with spiking neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2656–2670, 2021.

[18] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury, "Voice localization using nearby wall reflections," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[19] I. An, M. Son, D. Manocha, and S.-E. Yoon, "Reflection-aware sound source localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 66–73.

[20] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Computer Speech & Language*, vol. 75, p. 101360, Sep 2022.

[21] 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*. Tokyo: IEEE, Nov 2013, p. 2927–2932. [Online]. Available: http://ieeexplore.ieee.org/document/6696771/

[22] Q. Nguyen, L. Girin, G. Bailly, F. Elisei, and D.-C. Nguyen, "Autonomous sensorimotor learning for sound source localization by a humanoid robot," in *IROS 2018 - Workshop on Crossmodal Learning for Intelligent Robotics in conjunction with IEEE/RSJ IROS*, Madrid, Spain, 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01921882

[23] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, p. 1335–1345, Aug 2019.

[24] Two!Ears, "Media," 2021, http://twoears.eu/media/ Accessed Jul 7, 2021.

[25] J. Wu, Y. Chua, M. Zhang, H. Li, and K. C. Tan, "A spiking neural network framework for robust sound classification," *Frontiers in neuroscience*, vol. 12, p. 836, 2018.

[26] J. Wu, Z. Pan, M. Zhang, R. K. Das, Y. Chua, and H. Li, "Robust sound recognition: A neuromorphic approach." in *INTERSPEECH*, 2019, pp. 3667–3668.

[27] J. Wu, Y. Chua, and H. Li, "A biologically plausible speech recognition framework based on spiking neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[28] E. Yılmaz, O. B. Gevrek, J. Wu, Y. Chen, X. Meng, and H. Li, "Deep convolutional spiking neural networks for keyword spotting," in *Proceedings of Interspeech*, 2020, pp. 2557–2561.

[29] L. Zhang, S. Wang, L. Wang, and Y. Zhang, "Musical instrument recognition based on the bionic auditory model," in *2013 International Conference on Information Science and Cloud Computing Companion*. IEEE, 2013, pp. 646–652.

[30] D. Rothmann, "Human-like machine hearing with ai," 2021, https://towardsdatascience.com/human-like-machine-hearing-with-ai-1-3-a5713af6e2f8 Accessed Jul 29, 2021.

[31] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks," *International journal of neural systems*, vol. 19, no. 04, pp. 295–308, 2009.

[32] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: Vgg and residual architectures," *Frontiers in neuroscience*, vol. 13, p. 95, 2019.

[33] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *The Journal of the Acoustical Society of America*, vol. 61, no. 6, pp. 1567–1576, 1977.

[34] N. A. Gumerov, R. Duraiswami, and Z. Tang, "Numerical study of the influence of the torso on the hrtf," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2002, pp. II–1965.

[35] C. J. Plack, *The sense of hearing*. Psychology Press, 2013.

[36] J. J. Eggermont, "Between sound and perception: reviewing the search for a neural code," *Hearing research*, vol. 157, no. 1-2, pp. 1–42, 2001.
[37] H.-L. Han, "Measuring a dummy head in search of pinna cues," *Journal of the Audio Engineering Society*, vol. 42, no. 1/2, pp. 15–37, 1994.
[38] N. S. Harper and D. McAlpine, "Optimal neural population coding of an auditory spatial cue," *Nature*, vol. 430, no. 7000, pp. 682–686, 2004.
[39] S. J. Elliott and C. A. Shera, "The cochlea as a smart structure," *Smart Materials and Structures*, vol. 21, no. 6, p. 064001, 2012.
[40] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The journal of the acoustical society of America*, vol. 74, no. 3, pp. 750–753, 1983.
[41] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.
[42] K. M. Walker, J. K. Bizley, A. J. King, and J. W. Schnupp, "Multiplexed and robust representations of sound features in auditory cortex," *Journal of Neuroscience*, vol. 31, no. 41, pp. 14 565–14 576, 2011.
[43] A. Wingfield, "Evolution of models of working memory and cognitive resources," *Ear and hearing*, vol. 37, pp. 35S–43S, 2016.
[44] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
[45] L. A. Jeffress, "A place theory of sound localization." *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.
[46] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 280–285, 1984.
[47] I. J. Tashev, *Sound capture and processing: practical approaches*. John Wiley & Sons, 2009.
[48] P. M. Hofman, J. G. Van Riswick, and A. J. Van Opstal, "Relearning sound localization with new ears," *Nature neuroscience*, vol. 1, no. 5, pp. 417–421, 1998.
[49] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
[50] D. Tollin and T. Yin, "Sound localization: Neural mechanisms," in *Encyclopedia of Neuroscience*, L. R. Squire, Ed. Oxford: Academic Press, 2009, pp. 137–144. [Online]. Available: https://www.sciencedirect.com/ science/article /pii/ B9780080450469002679
[51] J. K. Moore, "Organization of the human superior olivary complex," *Microscopy research and technique*, vol. 51, no. 4, pp. 403–412, 2000.
[52] R. Viehweg and R. A. Campbell, "Xlix localization difficulty in monaurally impaired listeners," *Annals of Otology, Rhinology & Laryngology*, vol. 69, no. 2, pp. 622–634, 1960.
[53] A. Saxena and A. Y. Ng, "Learning sound location from a single microphone," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 1737–1742.
[54] 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.
[55] H. Wierstorf, M. Geier, and S. Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," in *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.
[56] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances," Jun. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.55418
[57] H. Wierstorf and M. Geier, "Binaural room impulse responses recorded with KEMAR in a small meeting room," Oct. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.160751
[58] ——, "Binaural room impulse responses recorded with KEMAR in a mid-size lecture hall," Oct. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.160749
[59] V. Erbes, M. Geier, S. Weinzierl, and S. Spors, "Database of single-channel and binaural room impulse responses of a 64-channel loudspeaker array," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.
[60] 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.
[61] "Iosr-surrey/realroombrirs: Binaural impulse responses captured in real rooms." https://github.com/IoSR-Surrey/RealRoomBRIRs, (Accessed on 05/30/2022).
[62] W. Xu, E. C. Chang, L. K. Kwoh, H. Lim, W. Cheng, and A. Heng, "Phase-unwrapping of sar interferogram with multi-frequency or multi-baseline," in *Proceedings of IGARSS'94-1994 IEEE International Geoscience and Remote Sensing Symposium*, vol. 2. IEEE, 1994, pp. 730–732.
[63] N. Ma, T. May, H. Wierstorf, and G. J. Brown, "A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2015-Augus, pp. 2699–2703, 2015.
[64] Z. Yang and R. R. Choudhury, "Personalizing head related transfer functions for earables," in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, 2021, pp. 137–150.
[65] M. G. Onofrei, R. Miccini, R. Unnthorsson, S. Serafin, and S. Spagnol, "3d ear shape as an estimator of hrtf notch frequency," in *17th Sound and Music Computing Conference*. Sound and Music Computing Network, 2020, pp. 131–137.
[66] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1050–1063, 1997.
[67] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual review of psychology*, vol. 42, no. 1, pp. 135–159, 1991.
[68] B. Xie, *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013.
[69] G. Bill, "Hrtf measurements of a kemar dummy-head microphone," *MIT Media Lab. Perceptual Computing-Technical Report*, vol. 280, pp. 1–7, 1994.

Qiang Yang is currently a Ph.D. student with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR. Prior to this, he received the M.E. and B.E. degrees in Computer Science and Technology from Shenzhen University, China and Henan University, China, respectively. His research interest includes Acoustic Sensing, Ubiquitous Computing, and Internet of Things (IoT). He is a student member of IEEE.

Yuanqing Zheng is an associate professor in the Department of Computing, the Hong Kong Polytechnic University. He received the Ph.D. degree in Computer Science from Nanyang Technological University, Singapore. He received the B.S. degree in Electrical Engineering and the M.E. degree in Communication and Information System both from Beijing Normal University, Beijing, China. His research interest includes wireless networking and mobile computing, acoustic and wireless sensing, and internet of things. He is on the editorial board of IEEE Transactions on Wireless Communications. He is a senior member of IEEE.