DeepEar: Sound Localization with Binaural Microphones

Qiang Yang, Yuanqing Zheng
Department of Computing, *The Hong Kong Polytechnic University*, Hong Kong, China
{csqyang, csyqzheng}@comp.polyu.edu.hk

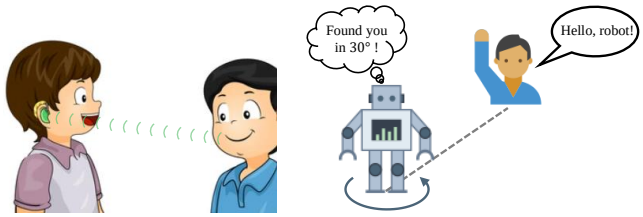
***Abstract*—Binaural microphones, referring to two microphones with artificial human-shaped ears, are pervasively used in hearing aids and humanoid robots to improve sound quality. In many applications, it is crucial for such devices to interact with humans by finding the voice direction. However, sound source localization with binaural microphones remains challenging, especially in multi-source scenarios. Prior works utilize microphone arrays to deal with the multi-source localization problem. Extra arrays yet have more space constraints for deployment in many scenarios (*e.g.*, hearing aids). However, human brains have evolved to locate multiple sound sources with only two ears. Inspired by this fact, we propose DeepEar, a binaural microphone-based localization system that can locate multiple sounds. To this end, we develop a neural network to mimic the acoustic signal processing pipeline of the human auditory system. Different from hand-crafted features used in prior works, DeepEar can automatically extract useful features for localization. More importantly, the trained neural networks can be extended and adapt to new environments with a minimum amount of extra training data. Experiment results show that DeepEar can substantially outperform the state-of-the-art deep learning approach, with a sound detection accuracy of 93.3% and an azimuth estimation error of 7.4 degrees in multi-source scenarios.**

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

I. Introduction

Sound localization can provide context information to improve user experience and enable a variety of innovative applications such as gaming, smart environment, and human-computer interaction. As shown in Fig. 1, people with hearing difficulties could also benefit from sound localization. If the hearing aids they wear can distinguish the sound location, then the binaural microphones in the ears can amplify the sound from this direction and substantially improve their quality of life when talking with others as well as their safety when walking outside. Moreover, humanoid service robots with binaural microphones and speakers can interact with users to promote products, give directions, and take care of kids and elders. When a user talks to a service robot, it would be great if the robot can figure out the voice direction, turn to the user, and provide customized location-aware services.

Currently, many microphone array-based sound localization technologies have been proposed, such as beamformer-based SRP-PHAT [1], spectral estimation-based MUSIC [2], triangulation based approach [3], and deep learning based methods [4, 5]. However, an extra microphone array brings about additional deployment costs, making hearing impaired users inconvenient to wear hearing aids. Moreover, the above



(a) Hearing aids. (b) Humanoid robots.

Fig. 1. Application scenarios. (a) Binaural microphones in hearing aids can localize the sound location and amplify the sound for hearing impaired people and improve their life quality. (b) Humanoid robots are equipped with artificial ears. When a user calls the robot, it should be able to locate the voice and turn around to the user.

microphone array-based solutions cannot be directly applied to binaural microphones due to very few microphones. For example, the correlation-based time difference technique can estimate the AoA of a sound with multiple (*e.g.*, 4) microphones. However, using only two microphones will lead to the cone of confusion problem [6], which means the sound source can be located in multiple locations with the same time difference. In horizontal 2D space, this problem causes front-back confusion. Moreover, when more than one sound source is present, they will interfere with each other, raising more challenges to separate multiple sound sources.

Existing binaural microphone-based solutions train machine learning models on the raw audio data directly [7] or hand-crafted features (*e.g.*, interaural time difference (ITD) or interaural level difference (ILD) [8, 9]). However, these works can only locate one source, or they assume the number of sound sources is known beforehand. In real usage scenarios, such assumptions are hard to guarantee and their performance degrades since they cannot handle the interference of multiple sources. On the other hand, the human auditory system has naturally evolved to locate multiple sounds simultaneously. In this paper, we aim to imitate the human auditory system and achieve multiple-sound localization with binaural microphones. To enable such human-like sound localization, we identify the following key objectives and design requirements:

- **Full-field localization.** Different from the existing methods (*e.g.*, correlation-based methods) which suffer the cone of confusion problem, human beings can normally differentiate whether the sound is from the front or from the back. Accordingly, we expect that our target system should be able to avoid such a confusion problem and

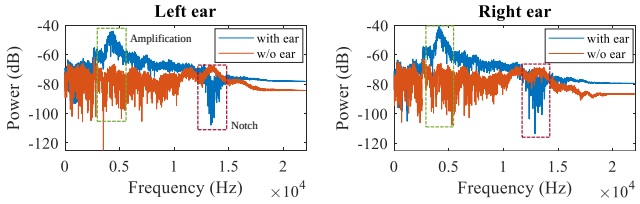
support full-field localization.

- **Multi-source localization.** Previous works typically formulate the single sound localization task as a single-label classification problem [10]. They estimate the most likely one direction among several pre-defined degrees. However, it is nontrivial to extend such a single source localization method to multi-source scenarios, especially when the number of sound sources is not known.
- **New environment adaptation.** We observe a substantial performance decrease of previous works when they work in a new environment. For example, neural network-based source localization methods suffer dramatic performance degradation when estimating new data collected in unseen environments (*i.e.*, unseen data). Ideally, our system should evolve and adapt to new working environments with minimum extra training.

To this end, we propose DeepEar, a multi-source localization system with binaural microphones. DeepEar mimics the signal processing pipeline of the human auditory system. First, the audio data is transformed into the time-frequency domain on the equivalent rectangular bandwidth (ERB) scale. Then, a temporal encoder network is designed to extract the latent representation of sounds. To enable multi-source fullfield localization, we partition the 2D horizontal space into a number of sectors, and model the multiple sound localization as a multi-label classification problem. The number of subsectors can be configured and changed according to application requirements. To adapt the model to new environments, we first train a global model on a large amount of readily available public data sets. Note that during the training of the global model, we do not need to collect any data from end-users or their working environments such as their homes or offices, which dramatically simplifies data collection and global model training. To bootstrap the training process, DeepEar harnesses a transfer learning strategy and fine-tunes the global model with a small amount of new data collected in the target environments during the usage of end-users. In this way, our method can reduce the data collection overhead involved in training a global model, as well as cope with the heterogeneity of working environments with the minimum effort of end-users. The contributions of this paper can be summarized as follows.

To highlight the contribution of this paper, we propose DeepEar, the first bionic sound localization system for binaural microphones that can locate multiple sources without a priori knowledge of the number of sources. Comprehensive experiments are conducted in both anechoic and reverberant environments. The results show a 93% sound detection accuracy and 7.4° azimuth estimation error in the multiple-source scenario, which outperforms the deep learning-based state-of-the-art in various experiment settings. A real-world case study also illustrates that the ears of binaural microphones play a pivotal role in disambiguation, which can improve sound localization performance significantly.

The paper is organized as follows. In Sec. II, we briefly introduce the background and present the empirical results of our feasibility study. We elaborate the detailed design in Sec. III. Then, Sec. IV and Sec. V describe the implementation and evaluation results. Related works are summarized in Sec. VI. Finally, Sec. VII concludes this paper.



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

Fig. 2. Frequency response with and w/o ears.

II. Background and Empirical Study

Human-shaped outer ears are an important part of the human auditory system, which helps in locating sound sources. We first conducted a feasibility study to evaluate the influence of artificial human-shaped ears on acoustic signals. As shown in Fig. 3, we placed a miniDSP EARS binaural microphone at the center of a meeting room. Then, we used a portable speaker to play an exponential sweep sine as the excitation signal $1m$ away in front of the binaural microphone. This excitation signal was recorded with the microphones to calculate the Binaural Room Impulse Response (BRIR), which describes the acoustic channel from the speaker to the microphone in this room. After that, we kept all settings unchanged but only detached the two artificial human-shaped ears from the microphones and repeated the measurement. As shown in Fig. 2, the frequency responses with the artificial ears substantially differ from those without ears. Specifically, we can see an amplification with ears at the voice frequency region ($<$ 10 kHz), since ear canals act as tubes and amplify the frequency band where human voices mainly reside. Besides, there is a noticeable frequency notch in a high-frequency band (10 kHz $\sim$ 20 kHz). That is because the ears with many wrinkles can cause special multipath reflection and destructive interference as reported in the literature [11]. This result validates that ears can significantly distort and filter the sound in certain frequencies.

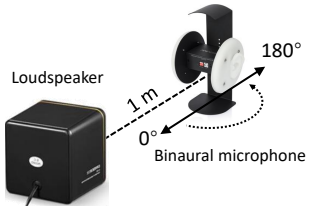
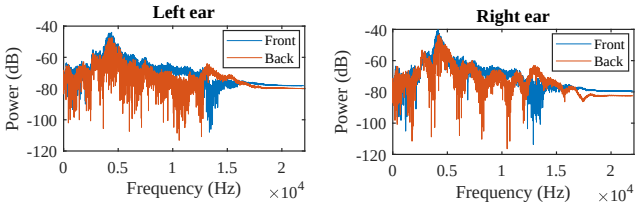


Fig. 3. Preliminary experiment setting.

As shown in Fig. 3, we conducted another experiment where we measured the BRIR before and after rotating the binaural microphones with ears around by 180°. We note that in the two measurements, the distances between the sound source to the



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

Fig. 4. Frequency response in the front/back.

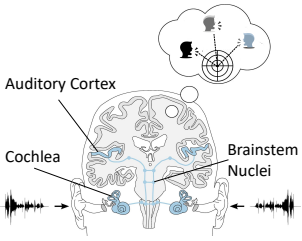


Fig. 5. Illustration of the human auditory system [16].

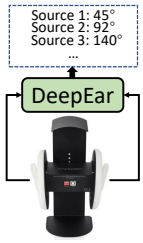


Fig. 6. Sound localization with binaural microphones.

two microphones remain the same. Intuitively, we expect that these two responses will be similar since all settings are kept fixed but only with the small ears orientation rotated. However, we see that the frequency responses significantly differ from each other in Fig 4. The reason is that, when the sound wave travels to a user, it will be scattered, reflected, and diffracted by the body, head, and especially the ears of the user (which can be described by the Head-Related Transfer Function, HRTF). The ears and head shape the acoustic signals by filtering and absorbing different frequency bands, thus the HRTF is both frequency and direction-dependent [12]. Therefore, our brain can learn to associate these subtle difference patterns with certain spatial locations, which helps resolve the ambiguity and perform source localization, even in multi-source scenarios [13].

With the help of ears, human beings can perform accurate sound localization. Fig. 5 shows a basic human auditory system. Two ears capture and filter the sound, and then the sound wave strikes the eardrum, leading the vibration in the spiral-shaped cochlea, which transduces the sound wave to neural stimulus signals [14]. As neural activity moves along the pipeline, several brainstem nuclei encode the stimulus to perception [14, 15]. Finally, the auditory cortex in the brain interprets sound spatial information. We refer interested readers to the literature [14] for more detailed human auditory mechanisms. Inspired by this fact, we utilize binaural microphones with human-shaped ears to capture sounds, and develop a deep learning model to mimic the functions of the human auditory system and locate sound sources as illustrated in Fig. 6. In the following, we describe the design consideration and implementation detail of DeepEar.

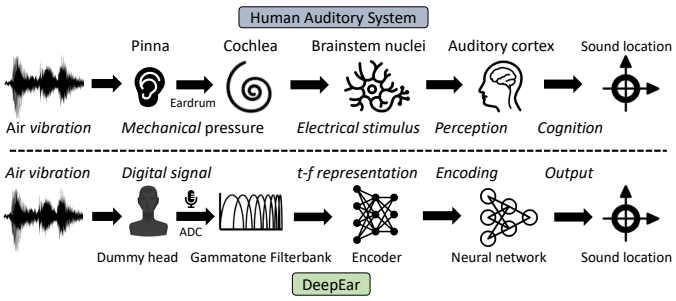


Fig. 7. System overview: an analogy between the human auditory system and DeepEar.

III. DeepEar Design

In this section, we first give a system overview of DeepEar, and then introduce the detailed components of the human-like sound processing pipeline.

A. System Overview

Fig. 7 presents a system overview of DeepEar. The upper part depicts the pipeline of the human auditory system. DeepEar is inspired by the human auditory system and we design and implement components to mimic the key functions to locate sound positions. 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 mimics the function of a cochlea in the human auditory system. After that, we train an encoder to extract the high-level representation. Finally, these sound features are input to a neural network to estimate sound locations. In the following, we introduce each component in detail.

B. Data Collection and Preprocessing

Human beings perform sound localization by learning the sound spatial patterns caused by the head, torso, and ears. Inspired by this fact, we utilize binaural microphones with human-shaped ears to capture acoustic signals. A dummy head can also be used to better capture the acoustic signals.

In the human auditory system, the cochlea is a spiral structure that is essential for frequency analysis. Along with this spiral, its different parts vibrate in response to different frequencies and convert sound waves into electrical stimuli. During this process, sounds are decomposed into many constituent frequency components. Such a frequency-selective vibration varies exponentially along the cochlea [17]. DeepEar imitates the function of a cochlea with a Gammatone filterbank, which is widely used in the literature of auditory system modeling [18]. We empirically set the number of filters as 100 to strike a balance between computational efficiency and representative sufficiency. To preserve sound temporal context, we frame the audio signals using a 100 ms Hamming window with 50 ms overlap. In this way, the output of preprocessing, Gammatone spectrogram coefficients, is a 2D matrix with size [filter size $\times$ frame number].

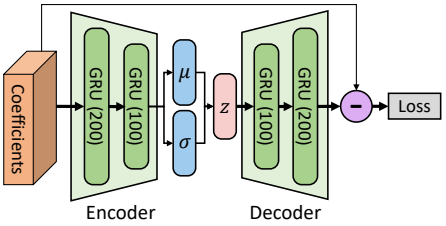
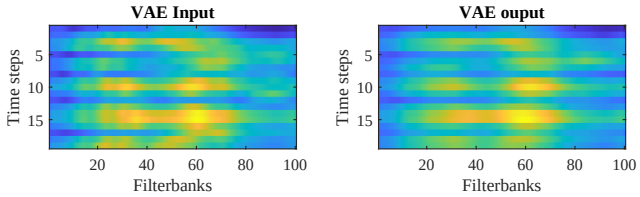


Fig. 8. Illustration of GRU VAE.



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

Fig. 9. GRU VAE can effectively extract the latent features from original data and reconstruct back with it.

C. Feature Extraction

Before a neural stimulus reaches the auditory cortex in the brain, it passes through many stages of processing by several brainstem nuclei as depicted in Fig. 5. Although the understanding of the specific processing accomplished in this stage remains not totally clear yet [19], it is commonly believed that these nuclei perform a function similar to feature mapping and encoding for sound localization and recognition [15]. Such a compressing and extraction process is able to prevent the overload of information in a short time [20].

This neural coding procedure inspires us to exploit an autoencoder to automatically extract compact sound representations. An autoencoder is trained to compress or encode data to a high-level latent feature space, which can be reconstructed back into the original input data without much information loss. An autoencoder consists of two parts: an encoder and a reversed structure named decoder. As the preprocessed result is a 2D time series, we use the seq2seq framework [21] to encode the data. As shown in Fig. 8, we build a Gated Recurrent Unit (GRU) variational autoencoder (VAE), which reads the Gammatone spectrogram coefficients and maps them to a fixed-length feature vector z. Two GRU layers are used to form an encoder. Similar to LSTM (Long Short-Term Memory), it can also learn the long and short-term temporal context, while it has fewer parameters and better generalization capability. Moreover, instead of coding the latent features from the input independently, we use a variation autoencoder to map the data into a multivariate normal distribution. This constrains the encoder to learn a smoother representation, which is more generalizable to reconstruct unseen data. After the training process, the decoder part can be cut off and only the encoder is used in DeepEar. Fig. 9 illustrates the original and reconstructed Gammatone coefficients of one sample. We can see that GRU VAE can extract representative high-level features from the original input.

As we mentioned before, the human brain perceives the spatial patterns in sound to perform localization. On the one hand, different propagation paths cause subtle sound differences between the two ears [22]. For example, the interaural time differences (ITD) can help us to infer the sound azimuth. As such, we perform cross-correlation GCC-PHAT [23] between the signals of two ears. The distance between two ears determines the maximum time difference from a sound source. Considering extra multipath caused by the head and body, we take the middle 100 coefficients instead of all correlation results as a part of features. However, there is no one-to-one mapping between ITD to sound direction or sound location because of ambiguities as we discussed. On the other hand, the ears produce micro-echoes to the arriving sound, leading to spectral distortion associated with certain spatial locations. These two patterns jointly help humans to locate sound signals. Therefore, along with the encoded features from the left and right ears, we also subtract two outputs and measure the feature differences between the two ears. Finally, all of these features are concatenated to form the final feature representation.

D. Sound Localization

DeepEar first detects whether a sound is present in a specific sector, and then estimates its AoA and distance if a target is present. We introduce the neural network design as follows.

1) Network Structure Design: With the extracted features, we construct a neural network to perform multiple sound localization. A subsector-based output is used to facilitate simultaneous multiple source localization with arbitrary spatial resolution. In this paper, we set the number of sectors to 8, and we release the assumption of previous work that the number of sound sources is known beforehand. Instead, we assume that there is at most only one source in a sector. This also means that DeepEar supports up to 8 simultaneous acoustic sources localization. We can increase the number of subsectors to increase the spatial resolution and the maximum supporting number of concurrent sources according to application requirements.

Fig. 10 shows the network design of DeepEar. The extracted features of binaural channels are subtracted in the subtract layer to obtain the difference between the two ears. After that, all features are concatenated to a feature vector and input to the sound localization network. We only use several dense layers to construct this network. To prevent overfitting, dropout layers are attached after each dense layer with a drop rate of 0.2. We formulate the full-field localization as a multi-task learning problem. The first three layers learn a general shared spatial pattern of the sound, followed by eight subnets that are responsible for each subsector. In each subnet, three task sub-networks share a common dense layer. The first task subnet is *SoundNet*, which detects whether an acoustic source is present in this sector and outputs a binary result indicating the presence of a target. The second task subnet *AoANet* predicts the AoA of the target. AoANet is a regression net,

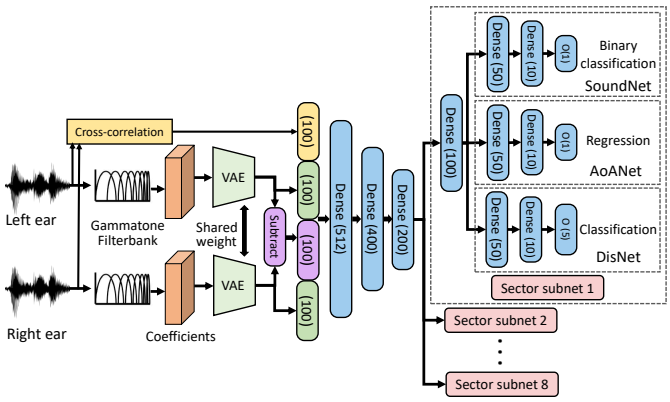


Fig. 10. DeepEar network design.

whose output is a normalized value in (0,1] indicating the minimal and maximal degree in the sector. For example, 0 and 1 represent 1° and 45° in sector 1, respectively. If there is no sound source in the sector, this value is set to 0 in the corresponding target label. *DisNet* is the third task subnet, which estimates the distance between ears and the target source. Note that humans perform distance estimation with sound loudness and the ratio of direct to reverberant sound, which is much worse compared with AoA estimation [24]. Therefore, we classify the sound distance into five classes, and among them the last category represents the no-source case.

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. All SoundNets can be regarded as a multilabel classification problem, so the activation function is sigmoid and binary cross-entropy are used as the loss function:

$$\mathcal{L}_s = -y \cdot \log(\hat{y}) - (1-y) \cdot \log(1-\hat{y}) \tag{1}$$

where y is the ground truth, and $\hat{y}$ is the prediction probability. As for AoANets, the mean squared error is used to qualify this regression task:

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

where $\hat{y}$ is the regression output of AoANet. Since DisNet is designed for a multiclass classification problem, we use the softmax activate function and formulate its loss function as the cross-entropy:

$$\mathcal{L}_d = -\frac{1}{C}\sum_{i=1}^{C} w_i \cdot y_i \cdot \log \hat{y}_i \tag{3}$$

where C is the number of categories (*i.e.*, 5), and w_i is the weight for each category. y_i is the $i\text{-}th$ one-hot encoding ground truth bit of this instance.

As a result, the loss of one sector subnet is constructed as a weighted sum of the losses of three task subnets:

$$\mathcal{L}_{sector} = \alpha\mathcal{L}_s + \beta\mathcal{L}_a + \gamma\mathcal{L}_d \tag{4}$$

where α, β, and γ are weights for different task subnets. The most important requirement for DeepEar is detecting the concurrent sound sources, while we also expect better AoA estimation than distance estimation. Thus, we empirically set these weights to 0.4, 0.35, and 0.25 respectively. Then we can average the losses of all sector subnets and obtain the overall loss of the DeepEar network:

$$\mathcal{L} = \frac{1}{N}\frac{1}{M}\sum_{n=1}^{N}\sum_{m=1}^{M}\mathcal{L}_{sector(m)} \tag{5}$$

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

E. Adaptation to New Environments

Humans have the ability that locates sound in various environments by continuous learning from childhood [25]. This ability indicates that humans can transfer the knowledge learned from a previous environment to new contexts. Therefore, we first build a global model for DeepEar, then we can apply transfer learning [26] to make DeepEar adaptive to new environments with a small number of new data.

DeepEar network can be divided into three components. The first one is the feature extraction module, including VAE and feature concatenation layer. Then three dense layers are used for learning the general spatial pattern knowledge. Finally, eight subnets are responsible for learning specific context information and performing localization tasks. Thus, based on the pre-trained global model, we freeze the first two parts and fine-tune 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.

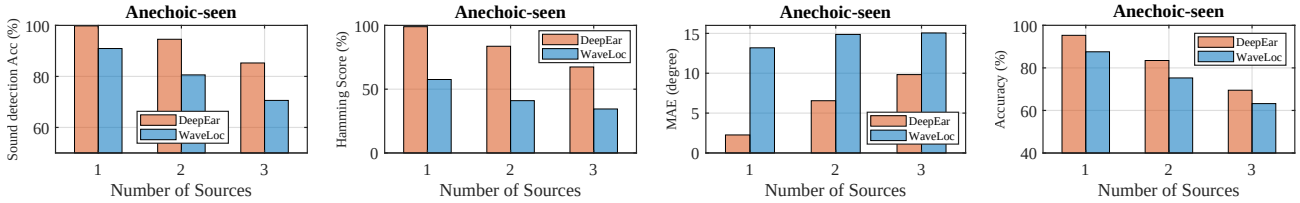
IV. Implementation

System 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. Early-stopping was applied to prevent overfitting if no performance improvement on the validation set was observed for more than 5 epochs. The loss of VAE is the mean square error between input Gammatone coefficients and reconstruct output.

Data Synthesis. Same as the previous binaural localization work [7–9], we synthesized binaural spatial sounds by convolving clean speech audio recordings with the binaural room impulse responses (BRIR) of different locations. The clean speeches were randomly chosen from a corpus named TIMIT [27], containing the recordings of 630 speakers with eight major dialects of American English, each reading ten sentences. We choose a publicly available BRIR dataset named TU Berlin [28]. This dataset was measured with a KEMAR dummy head (*i.e.*, binaural microphones) in three different rooms, including an anechoic chamber, a small meeting room named Spirit, and a mid-size lecture room called Auditorium3. Considering that the maximal number of concurrent sound

Table I. Dataset summary.

Dataset	Anechoic-training	Anechoic-validation	Anechoic-testing-seen	Anechoic-testing-unseen	Spirit-testing	Auditorium3-testing
BRIR convolved	Anechoic	Anechoic	Anechoic	Anechoic	Spirit	Auditorium3
Sample size	72000	9000	9000	9000	9000	9000
Usage	Training	Validation	Testing	Testing	Testing	Testing



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

Fig. 11. Performance comparison between DeepEar and WaveLoc on seen data in the anechoic environment.

Metircs	Sound detection (%)				Hamming score (%)				AoA (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 II. Performance comparison between DeepEar and WaveLoc on unseen data in the anechoic environment.

sources is typically small in the real world, we set it as 3. The number of sources, AoA, and distance are all randomly generated but with a constraint that only one source presents in a sub-sector. All synthesized data were sampled at 16 kHz and sliced to 1-second instances.

V. Evaluation

A. Experiment Setup

We first train a global model for DeepEar only with the publicly available data. After that, DeepEar can be customized and adapted to the real-world application environments by transfer learning with a minimum amount of new data collected in target working environments.

The clean speech recording corpus consists of two portions, TRAIN, and TEST. We randomly selected speeches in the TRAIN portion, convolved with Anechoic BRIR to get the anechoic synthesized data for training a global model. These data were divided into training-validation-testing three parts with the ratio 8:1:1. Given that these training data and testing data are split from the same dataset, the evaluation result will be overestimated since the trained model may have seen the test data. Therefore, we then separately took random clean speech recordings in the TEST portion and synthesized a new testing dataset, denoted Anechoic-testing-unseen. Moreover, we similarly convolved clean speeches in the TEST portion with the real-world BRIRs of the meeting room (Spirit) and lecture room (Auditorium) to generate other two testing datasets. Overall, we have six datasets: one for training, one for validation, and four for model testing. We summarize the names, sizes, usages of all datasets in Tab. I.

For comparison, we implemented a binaural localization state-of-the-art WaveLoc [7]. WaveLoc decomposes binaural signals into 32 frequency bands, and then employs CNN (convolutional neural network) on the raw waveform in each band to classify the AoA. Noted that WaveLoc only supports one source azimuth classification, so we replaced the last layer of WaveLoc with the sector-subNets of DeepEar to enable multiple sound localization. To illustrate the importance of ears, we also conduct a real-world case study with a binaural microphone to locate the sound with and without the presence of ears.

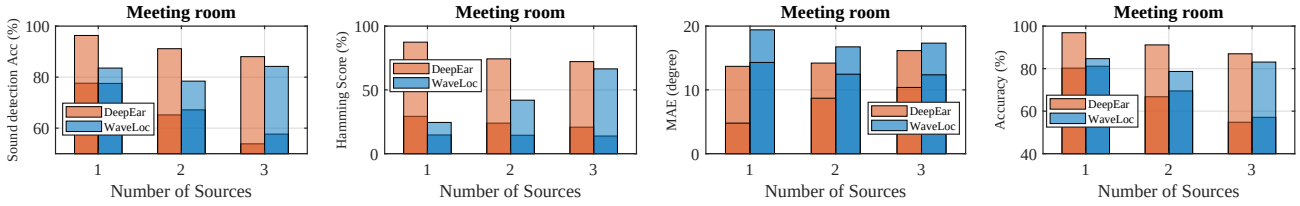
B. Evaluation Metrics

We evaluate DeepEar with the following metrics:

- Sound detection accuracy. It measures the binary classification accuracy of SoundNet in detecting whether there is a sound source or not.
- Hamming score of sound detection. Hamming score is defined as the proportion of the predicted correct labels to the total positive labels (predicted and actual) for that instance:

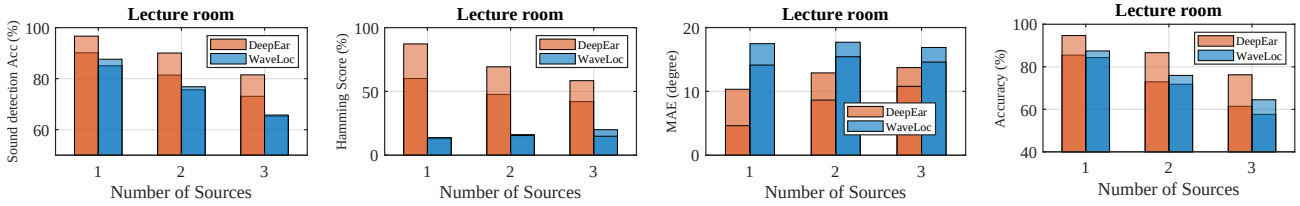
$$H = \frac{1}{N}\sum_{n=1}^{N}\frac{sum(y_n \ \& \ \hat{y_n})}{sum(y_n \mid \hat{y_n})} \tag{6}$$

  where y_n is the ground truth of eight SoundNets of the $n\text{-}th$ instance. $\hat{y_n}$ is the corresponding classification result. $\&$ and $\mid$ represent bit-wise AND and OR operation, respectively. Compared to binary accuracy, Hamming score ignores the true negative (*i.e.*, a no-source case is correctly detected) as well as penalizes false-positive cases (*i.e.*, a no-source case is detected as an active source by mistake).
- Mean absolute degree error of AoA (MAE). MAE means the absolute degree error between prediction AoA and



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

Fig. 12. Performance comparison in Spirit meeting room. The darker bars refer to Accuracy before transfer learning or MAE after transfer learning.



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

Fig. 13. Performance comparison in Auditorium lecture room. The darker bars refer to Accuracy before transfer learning or MAE after transfer learning.

ground truth. We average MAE of all AoANet as the overall MAE of DeepEar.

- Distance classification accuracy. This metric refers to the averaged accuracy of all DisNets.

C. Anechoic Environment

1) Seen Data: Fig. 11 shows the performance of the global model on the seen data in the anechoic environment. Overall, the sound detection accuracies of DeepEar and WaveLoc are 93.3% and 80.9%, respectively. Surprisingly, DeepEar has a high detection accuracy of 99.8% in the one-source scenario. In comparison, the performance of WaveLoc is a little lower with the detection accuracy of 90.9%. We can see that the performance of both models decreases with the increasing number of sounds. 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%, which is slightly lower than the binary accuracy since all no-source cases 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.

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 2.3° error. However, the MAE of WaveLoc is 13.2°, much larger than DeepEar. It is because that WaveLoc performs CNN directly on raw waveform data, missing the key time difference information between binaural channels. With the number of sources increasing, multiple sounds may interfere with each other so that time differences will be confused, leading to a higher estimation error.

The average distance accuracies of all source-cases are 82.9% and 75.6% for DeepEar and WaveLoc, respectively. There is no large gap between them like before due to narrow possible categories. Same as before, the larger number of active sources is, the lower the estimation performance is.

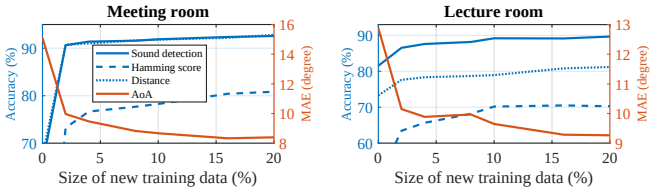
2) Unseen Data: We also evaluate DeepEar on the unseen data. This dataset is generated separately instead of splitting from the original training data. The result is listed in Tab. II. Overall, the sound detection accuracy and Hamming score of DeepEar are 91.9% and 80.4% respectively. This performance is almost the same as that on the seen data, so are AoA MAE (8°) and distance accuracy (82%). The performance of WaveLoc presents similar trends like DeepEar, indicating that both systems generalized well to anechoic unseen data. The reason is that we synthesized massive training data, which describes the data space to a great extent.

D. Reverberant Environment

We know that an anechoic environment is hardly possible in real life, so it is our turn to examine DeepEar on the data in real reverberant rooms, including a small meeting room and a larger lecture room.

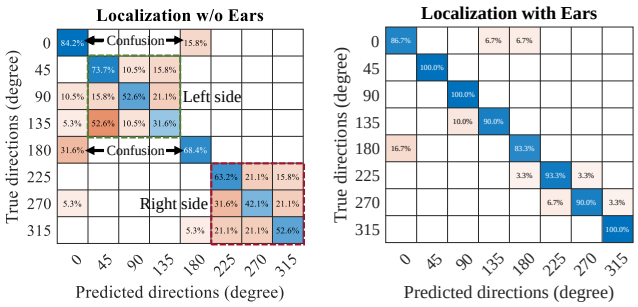
1) Evaluation in a Small Meeting Room: Fig. 12 illustrates the performance of a small meeting room. Directly testing the global model on the reverberant data brings about a dramatic performance deterioration as we expected. The benchmark WaveLoc also performs poorly. The average sound detection accuracy and Hamming score of DeepEar are 65.6% and 24.7%, while WaveLoc achieves 67.5% in sound detection and 14.3% in Hamming score, respectively. Although the sound detection accuracy of WaveLoc is slightly higher than that of DeepEar, the Hamming score of DeepEar is much higher than WaveLoc. Similarly, the performance of AoA and distance estimation also drops. The reason is that signals in a reverberant environment differ substantially from the anechoic room.

To adapt to this meeting room, we perform a transfer learning on this global model with 10% of new data. DeepEar



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

Fig. 14. The transfer learning performance of DeepEar with different sizes of new training data. Two subfigures share the same legend.



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

Fig. 15. Localization performance with and without human-shaped ears.

converges fast within 10 epochs and exhibits much better performance. The sound detection accuracy increases to 91.9%, while WaveLoc only achieves 82.1%. The Hamming score of DeepEar increases by 53.3%, which almost doubles that of WaveLoc. The DeepEar's AoA MAE decreases by half 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 and 15.1% for WaveLoc, respectively. In this figure, we can see that both methods benefit from transfer learning in testing the new reverberant data. Yet DeepEar notably outperforms WaveLoc after the same retraining procedure with the same amount of new data.

Note that the variation encoder has learned the feature distributions of different locations. Therefore, the sub-network can quickly adapt to the new data and increase the performance. In contrast, WaveLoc uses CNN to learn a discrete feature space, which is harder to adapt to new environments with a relatively small number of additional training data.

2) Evaluation in a Large Lecture Room: In this experiment, although both methods also suffer performance degradation on this unseen dataset, DeepEar performs much better than WaveLoc. As shown in Fig. 13, the overall sound detection accuracy of DeepEar is 81.5%, *i.e.*, 6.2% higher than WaveLoc. As for Hamming score, the performance gap is even wider. In particular, WaveLoc decreases to 16.3%, which is approximately one-third of DeepEar. Besides, the AoA estimation errors of these two systems are 12.9° and 17.3°, respectively. This result shows that while WaveLoc cannot deal with the reverberation interference, DeepEar can still achieve a relatively better performance because of its robustness to the highly reverberant new environment.

Transfer learning is effective in improving the performance of both models. Yet, we see that DeepEar benefits more from this 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%. 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 larger than the meeting room, meaning that the lecture room has a longer reverberation time. The CNN mechanism of WaveLoc relies more on discrete data so that it cannot adapt to the reverberant environment. In contrast, DeepEar benefits from the VAE design that enables subnets to calibrate feature distribution accordingly with new data and thereby achieves a better performance.

The AoA MAE of DeepEar and WaveLoc decreases by 3.9° and 2.5°, respectively. Furthermore, the distance accuracy of DeepEar and WaveLoc increases to 91.7% and 76.4%, respectively. Again, DeepEar still outperforms the baseline in the distance and AoA estimation.

E. Transfer Learning Performance

The experiment results above show that transfer learning effectively helps DeepEar adapt to new environments. We also test DeepEar with different sizes of new data in transfer learning. The result is illustrated in Fig. 14. We zoom into the y-axis of accuracy for clear observation. We can see that only 2% of new data can essentially boost the DeepEar performance in both the small meeting room and the large lecture room. The Accuracy (MAE) steadily increases (decreases) with the number of training data grows. In theory, if more new data is used in transfer learning, we can achieve better performance. Yet we need to balance the performance and the extra training overhead, since collecting a large number of new data in different environments could be practically challenging for ordinary users. This experiment reveals that 2% of new data (*i.e.*, 180 one-second instances) is efficient for DeepEar to yield a good adaption result in different new environments, while with 10% of new data DeepEar can achieve higher performance if needed.

F. Real-world Case Study

To further evaluate the importance of ears for sound localization in practice, we performed a real-world localization experiment. A binaural microphone (miniDSP EARS) is placed in a meeting room as the recording device. Several speech files were randomly selected from the public TIMIT corpus. Then, we used a portable loudspeaker to play the selected audio files at eight 45° evenly spaced directions $1m$ away from the microphones with and without human-shaped ears respectively. After that, audio recordings were sliced into one-second samples and 20 Gammatone coefficients were extracted from each 0.1 s frames as features.

We implemented a one-layer LSTM network consisting of 100 hidden units, stacked with a dense layer with softmax function to execute the sound localization task. Fig. 15

shows the localization confusion matrix with and without ears. Without ears, the localization accuracy is 58.6% as shown in Fig. 15(a). We can see the model suffers from front-back confusion. Moreover, although the directions on the left side and right side can be easily detected, the model can hardly identify the degrees on each side. In contrast, the overall classification accuracy increased to 92% after mounting the ears as shown in Fig. 15(b). The confusion problem was alleviated and the accuracy of almost all directions is improved, which means the human-shaped ears indeed help improve the localization accuracy significantly.

VI. Related Work

A. Sound Localization

DeepEar is most related to sound localization, especially binaural localization. We summarise the related works in literature in Tab. III and highlight the novelty of DeepEar.

The microphone array and distributed microphone arrays have been widely used for sound localization. These works mainly target a sound source emitting pre-designed signals [29–31]. The human voice is generally unknown to microphones, which brings about challenges for localization. Rich bodies of research works utilize microphone arrays to estimate the AoA of a sound, such as SRP-PHAT [1] or MUSIC [2]. VoLoc [32] and [33] locate the voice by nearby reflections with only one microphone array. However, these methods are not suitable for binaural microphones since they either suffer the ambiguity problem, or require three microphones at least. Previous works tried to tackle this problem via deep learning techniques. WaveLoc [7] exploits a CNN on the raw waveform and classifies sound into 37 directions. [34] also employs CNN on interaural spectrograms to perform azimuth and elevation classification. These works simply treat sound localization as a classification problem, which cannot be generalized to multi-source localization and different environments.

The interference from different sound sources raises practical challenges for locating multiple sources simultaneously. [3] explores the microphone redundancy relationship to achieve multi-source localization with a single microphone array. Similarly, it requires several microphones to find the redundant spatial pattern for each source, which cannot be applied to binaural microphones. SMESLP [4] and [5] adopt a CNN to localize multiple sources also with a microphone array. [8] and [9] train a deep neural network for binaural multiple sound localization. However, they assume that the number of sources is known in advance. In contrast, DeepEar can achieve multiple sound localization with binaural microphones with an unknown number of sound sources.

B. Bionic Auditory Applications

Inspired by the powerful human auditory capability, many researchers imitated the human auditory mechanism and designed smart systems to deal with sound-related tasks. For example, [35] proposed an auditory-like system to recognize the type of musical instruments, and [36] designed a machine hearing approach to predict the types of sounds. The powerful perceptual capacity of humans is still the goal of AI technology today. Same as the research on CNN and its breakthrough in computer vision tasks, we believe modeling the human auditory system will open a broad range of possibilities in sound-related tasks.

Sound localization	Mic array	**Binaural mics**	
One source	[32, 33]	[7, 34]	
Multiple sources	[3–5]	Known number	**Unknown number**
		[8, 9]	**DeepEar**

Table III. Comparison with related works. DeepEar is the first sound localization method for binaural microphones that can locate multiple sources without a priori knowledge of the number of sources.

C. HRTF Calibration

One might be concerned that ear-caused HRTF is unique and cannot be applied to a different ear-shaped binaural microphone. Recent research found that humans can get used to new mold ears accurately within a few weeks [25], which indicates that we may perform incremental learning strategies to adapt the HRTF among different ears. Recent work UNIQ [37] personalizes the HRTF for different users with a smartphone and in-ear microphones. [38] also proposed a regression approach to estimate the HRTF based on the ear's 3D shape. We plan to study whether the DeepEar model can adapt to different ear-shaped binaural microphones in the future. DeepEar is more suitable for service robot manufacturing. This problem certainly calls for more research in the future.

VII. Conclusion

In this paper, we propose DeepEar, the first sound localization system for binaural microphones that can locate multiple sources without a priori knowledge of the number of sources. DeepEar imitates the human auditory system to transform acoustic signals and extract latent representatives. A multi-sector deep learning neural network is designed to estimate the locations of multiple sources. By leveraging a large amount of readily available datasets, we train a global model without collecting any data from end-users. To cope with the heterogeneity of working environments, DeepEar further exploits the transfer learning strategy and re-trains the global model with a small number of new data collected in real usage scenarios. Thanks to the variational encoder and novel neural network architecture design, DeepEar can generalize to unseen data and quickly adapt to new environments with minimum extra training data. Experiment results show that DeepEar substantially outperforms the state-of-the-art works in terms of sound detection as well as localization accuracy. The authors have provided public access to their code and data at https://github.com/Qiangest/DeepEar.

VIII. Acknowledgments

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

REFERENCES

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

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

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] N. S. Harper and D. McAlpine, "Optimal neural population coding of an auditory spatial cue," *Nature*, vol. 430, no. 7000, pp. 682–686, 2004.

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

[15] 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.

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

[17] S. J. Elliott and C. A. Shera, "The cochlea as a smart structure," *Smart Materials and Structures*, vol. 21, no. 6, p. 064001, 2012.

[18] 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.

[19] 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.

[20] A. Wingfield, "Evolution of models of working memory and cognitive resources," *Ear and hearing*, vol. 37, pp. 35S–43S, 2016.

[21] 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.

[22] L. A. Jeffress, "A place theory of sound localization." *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.

[23] 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.

[24] I. J. Tashev, *Sound capture and processing: practical approaches*. John Wiley & Sons, 2009.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] P. Lazik, N. Rajagopal, O. Shih, B. Sinopoli, and A. Rowe, "Alps–the acoustic location processing system," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, 2015, pp. 491–492.

[30] I. Constandache, S. Agarwal, I. Tashev, and R. R. Choudhury, "Daredevil: indoor location using sound," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 18, no. 2, pp. 9–19, 2014.

[31] L. Cheng, Z. Wang, Y. Zhang, W. Wang, W. Xu, and J. Wang, "Acouradar: Towards single source based acoustic localization," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1848–1856.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Z. Yang and R. R. Choudhury, "Personalizing head related transfer functions for earables," 2021.

[38] 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.