DeepEar: Sound Localization with Binaural Microphones

Qiang Yang, Yuanqing Zheng

The Hong Kong Polytechnic University
Hong Kong

4 May 2022
Motivation

Sound Localization: microphone array

- Conventional approaches fail with two microphones due to ambiguity
- Any locations on the hyperbola have the same TDoA.
Application scenarios of binaural localization

• Localization with two microphones

How to locate sound sources with only two microphones?
Motivation

Human can naturally locate multiple sounds simultaneously.

- With only two ears, why?

Human beings have ears and brain!

DeepEar: a bionic design

Human Auditory System

- Ears: bring unique reflection to sounds from different directions
- Cochlea: transform sounds into frequency domain
- Brainstem Nuclei: compress and encoder signals
- Cortex: interpret nerve signals to direction
Binaural Microphones

- Ears cause unique sound spatial patterns (frequency response).
- This pattern is direction-dependent.
- Human beings learn this pattern to perform localization.

![Frequency response graphs](image)

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

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

Gammatone Filterbank

• Cochlea transforms sounds into electrical signals.
• Along with this spiral shape, its different parts vibrate in response to different frequencies.
• Gammatone Filterbank is used to approximate human hearing.
Feature Extraction

GRU-based Variational Autoencoder (VAE)

• Brainstem nuclei will **compress and encode** the signal to prevent the overload of information in a short time.
• Gammatone Coefficients is a 2D matrix with the **time information**.
• We use a GRU-based VAE to map the data into a multivariate normal feature vector.
Multilabel Multitask Learning Framework

- Conventional single-label classification: how many output nodes? multiple sources?
- Our solution: sector-based multi-label classification: partition the 2D space into several subsectors.
- We can increase the number of subsectors to adjust the spatial resolution.
- Multitask learning: sound detection, direction prediction, and distance estimation.
DeepEar Structure

- Introduce xCorr to obtain time difference between ears.
- Sector number: 8 (45-degree resolution)
- Each sector corresponds to a subnet, including three small nets to detect sound source, predict AoA, and estimate distance.
Evaluation

- Dataset: TU Berlin spatial sounds
- Maximum number of co-emitting sound sources: 3
- Sound sources are sampled uniformly in arbitrary locations.
- Training: 80% anechoic data
- Test: remaining anechoic data, meeting room data, lecture room data
- Baseline: WavLoc#, an end-to-end raw waveform based CNN approach

Evaluation

Anechoic environment

• Detection Accuracy: 93.3%
• Detection Hamming Score: 83.5% (Accuracy but focusing more on the positive cases.)
• AoA Estimation Error: 7.4 degrees
• Distance Accuracy: 82.9%
Reverberant environment: meeting room

• Transfer learning: fine-tune subnets with new data and keep previous layers frozen.
• Transfer global model to new environments

Detection Accuracy

Evaluation

93.3% 91.9% 65.6%
93.3% 81.9%
7.4° 16° 8.8°
7.4° 8.8°

AoA Error

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

Reverberant environment: lecture room

Detection Accuracy

81.5% (transfer learning)
89.4%

AoA Error

9° (transfer learning)
12.9°

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

Transfer learning Performance

- Train a global model with massive anechoic data
- Transfer global model to new environment with a small number of data
- 2% of new data (180 seconds) can essentially boost the DeepEar performance.

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

- A loudspeaker is placed at different locations around a binaural microphone.
- 58% (without ears) → 92% (with ears)
- Ambiguity is remarkably reduced after mounting human-shaped ears.

Ears play a significant role in sound localization and disambiguation.
Conclusion

• We propose DeepEar, the first sound localization system for binaural microphones without a priori knowledge of the number of sources.

• DeepEar is a bionic machine hearing framework inspired by the human auditory system.

• DeepEar can quickly adapt to new environments with a small number of extra training data with the transfer learning strategy in real scenarios.
THANKS

Qiang Yang
qiang.yang@connect.polyu.hk