DeepEar: Sound Localization with Binaural Microphones

Qiang Yang, Yuanqing Zheng

The Hong Kong Polytechnic University

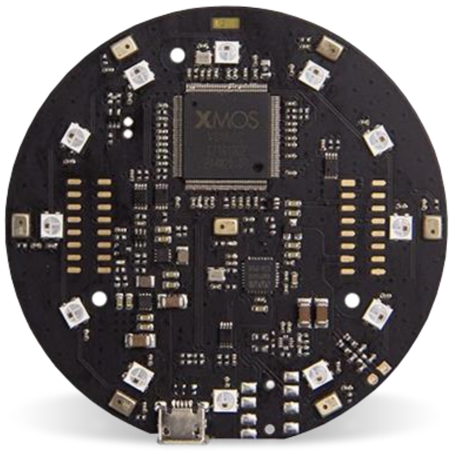
Hong Kong

4 May 2022

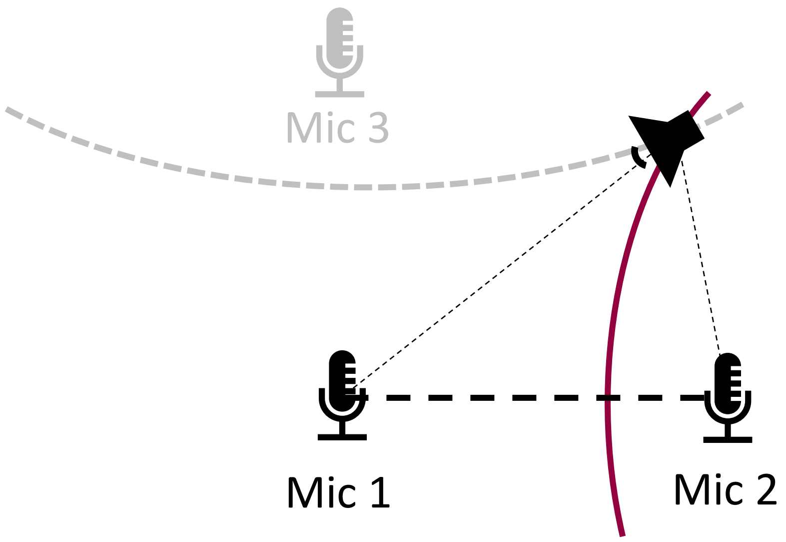
Motivation

Sound Localization: microphone array

- Conventional approaches fail with two microphones due to ambiguity
- Any locations on the hyperbola have the same TDoA.



Microphone array



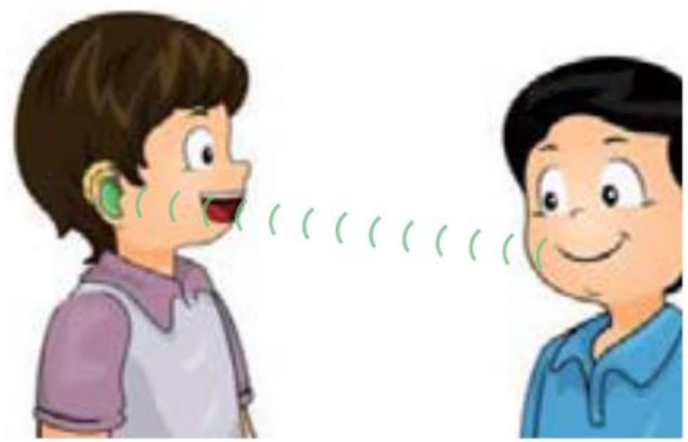
Two-microphone case

Motivation

Application scenarios of binaural localization

- Localization with two microphones

Humanoid robots



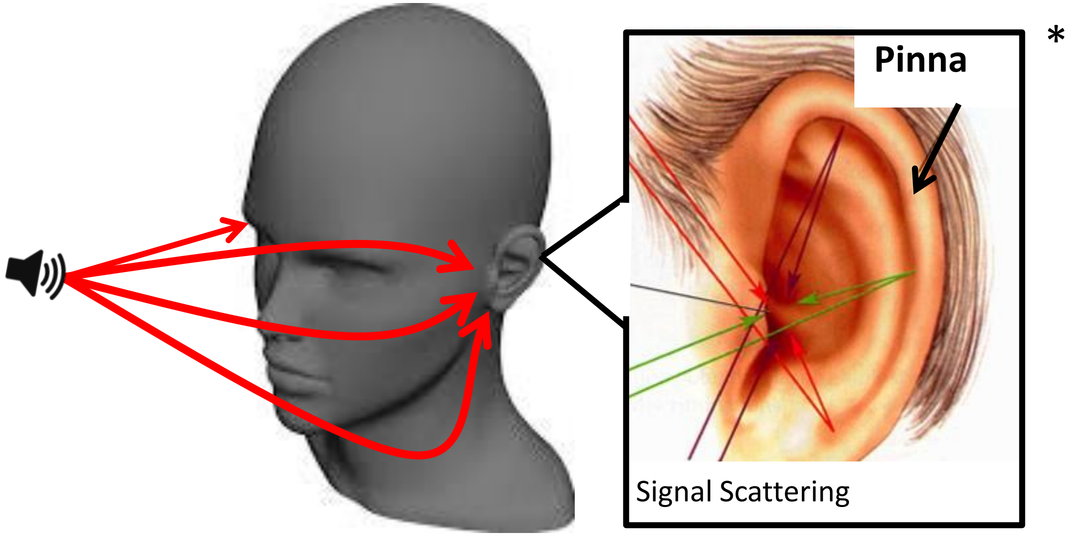
Hearing aids

How to locate sound sources with only two microphones?

Motivation

Human can naturally locate multiple sounds simultaneously.

- With only two ears, why?



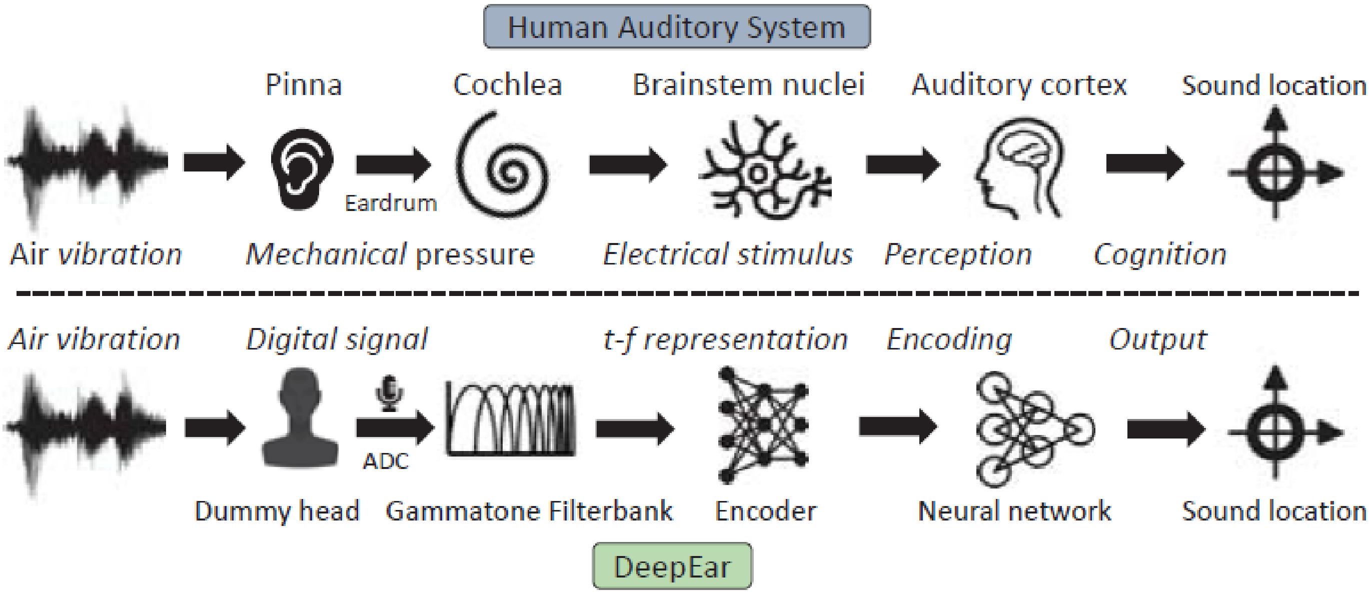
Human beings have ears and brain!

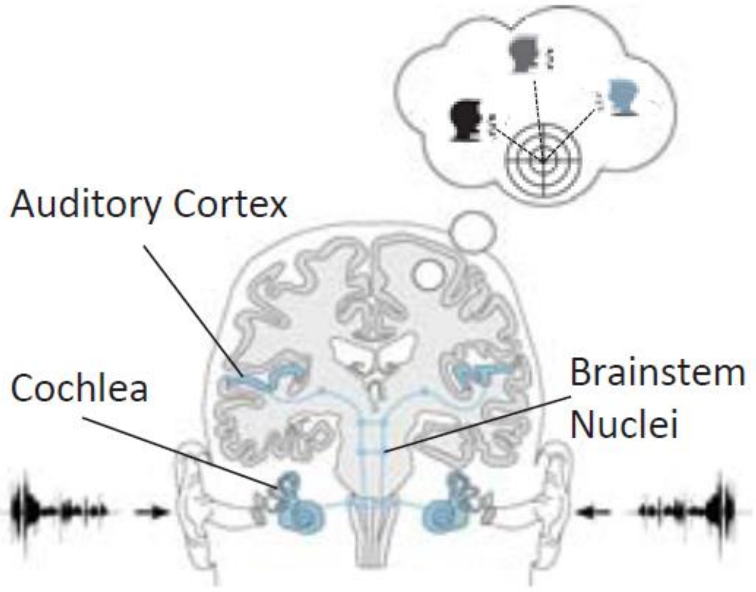
* Yang, Z., & Choudhury, R. R. (2021, August). Personalizing head related transfer functions for earables. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference* (pp. 137-150).

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

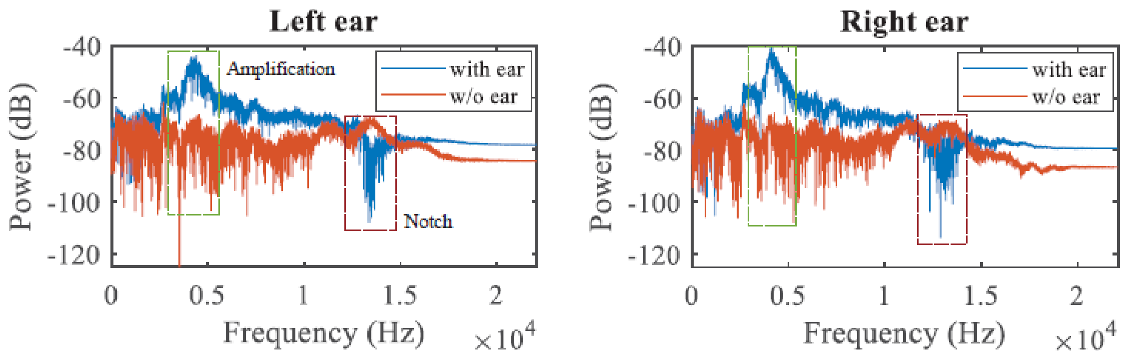




Signal Collection

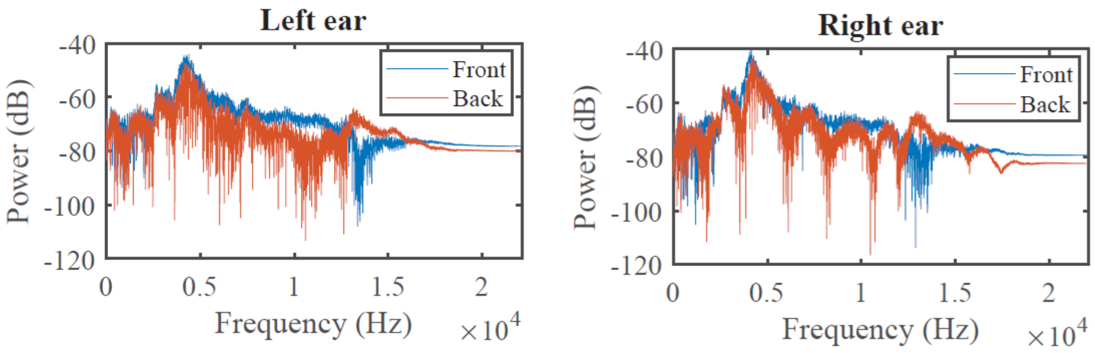
Binaural Microphones

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



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

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



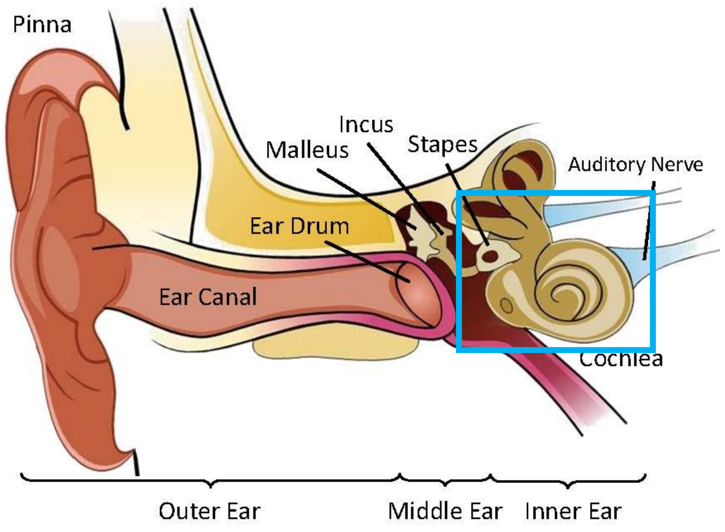
(a) Left ear. (b) Right ear.

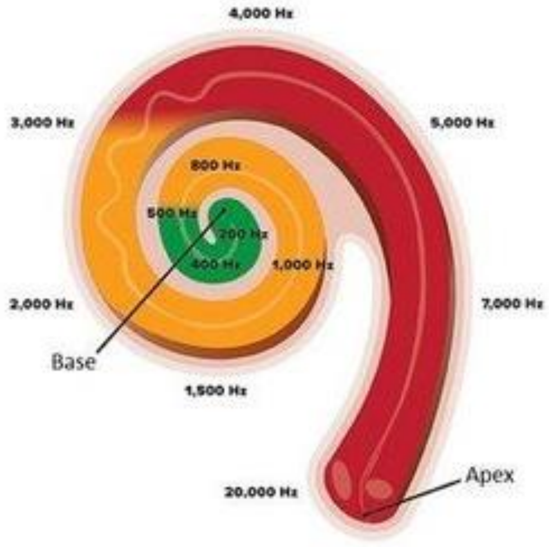
Fig. 4. Frequency response in the front/back.

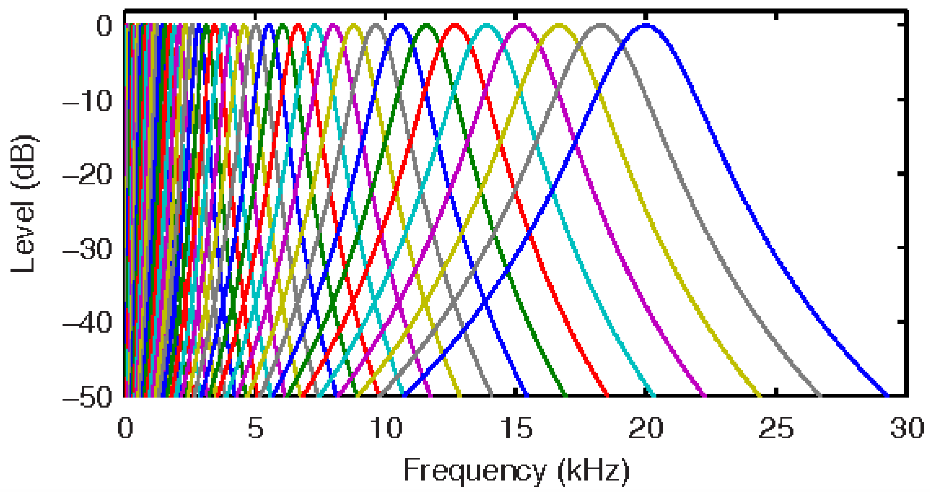
Signal Processing

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.





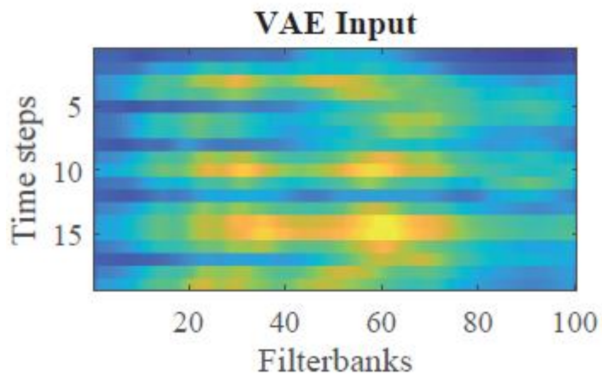


İlik, Bedirhan. *MEMS thin film piezoelectric acoustic transducer for cochlear implant applications*. MS thesis. Middle East Technical University, 2018.

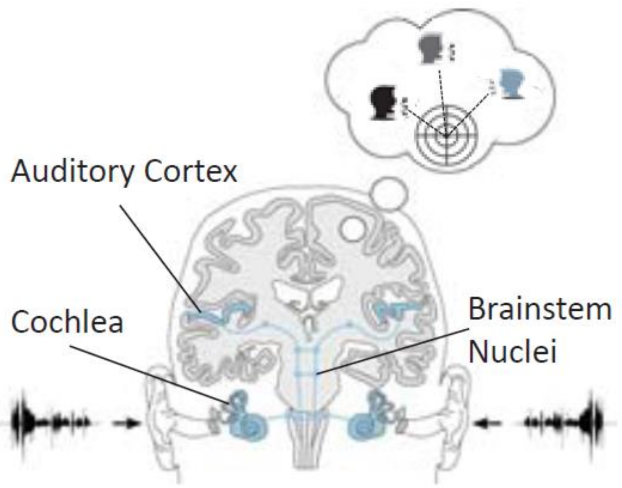
Feature Extraction

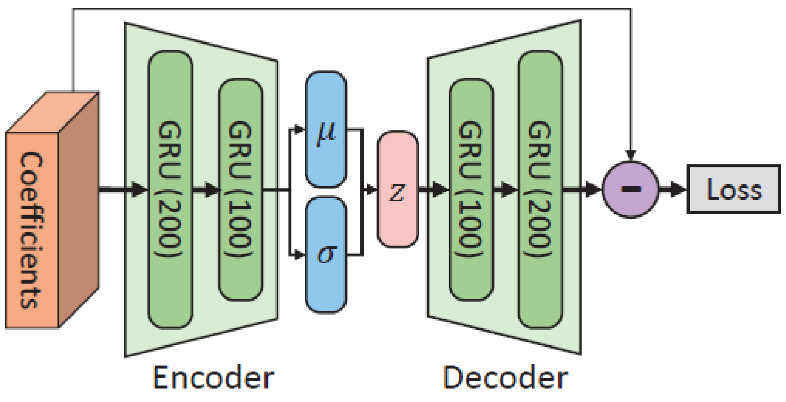
GRU-based Variational Autoencoder (VAE)

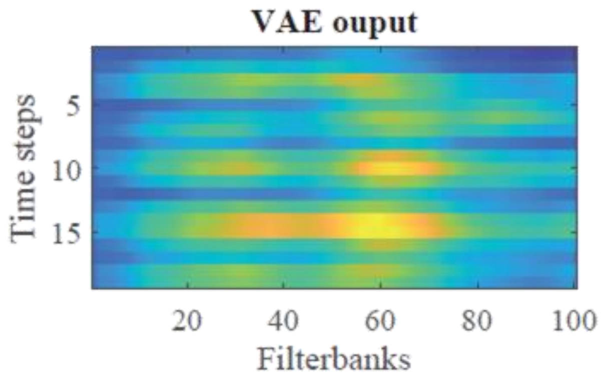
- Brainstem nuclei will compress and encode the signal to prevent the overload of information in a short time.
- Gammatone Coefficients is a 2D matrix with the time information.
- We use a GRU-based VAE to map the data into a multivariate normal feature vector.



(a) VAE input.





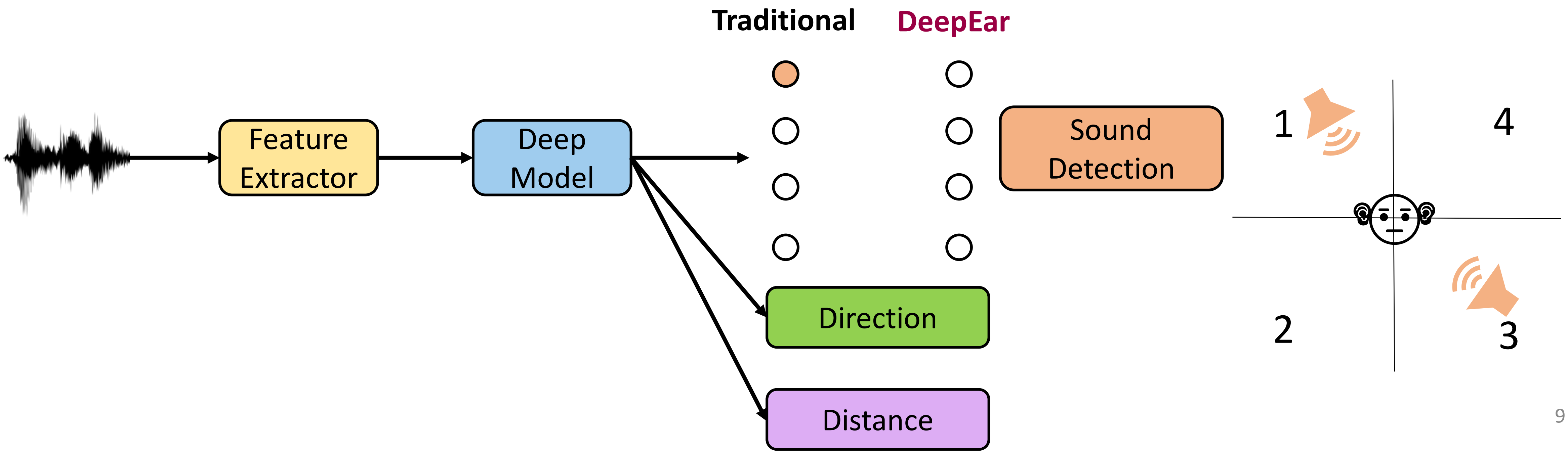


(b) VAE output.

Deep Model

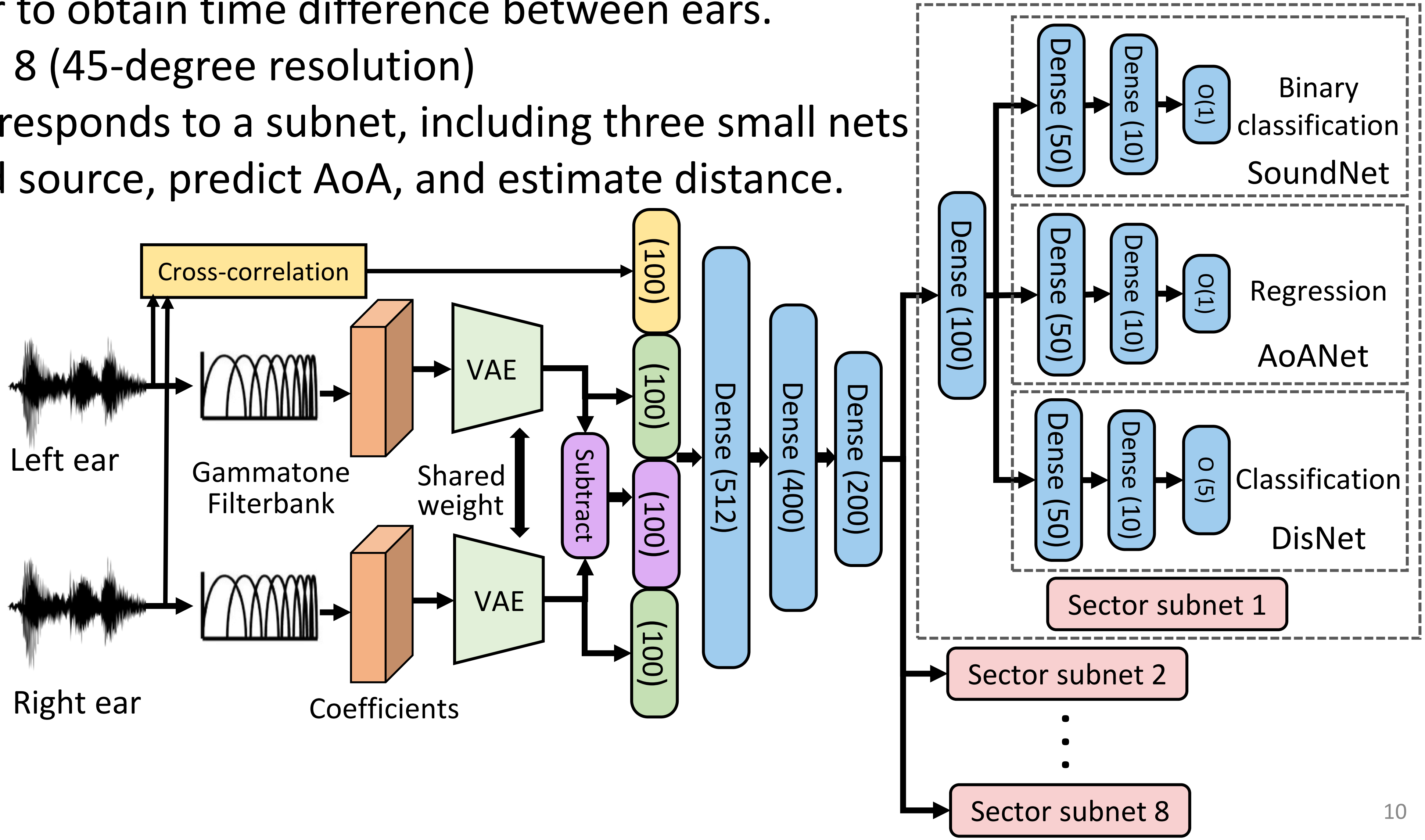
Multilabel Multitask Learning Framework

- 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.



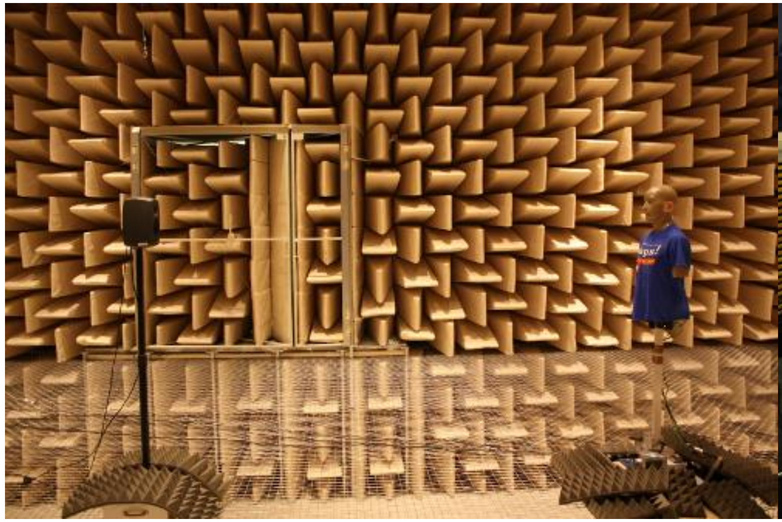
DeepEar Structure

- 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



Anechoic Chamber

Meeting Room: Spirit

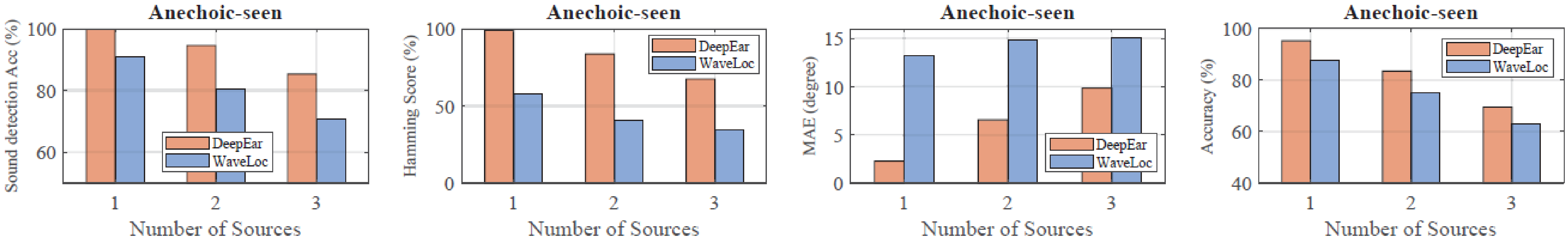
Lecture Room: Auditorium3

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

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%



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

Evaluation

Reverberant environment: meeting room

93.3%

65.6%

91.9%

Detection Accuracy

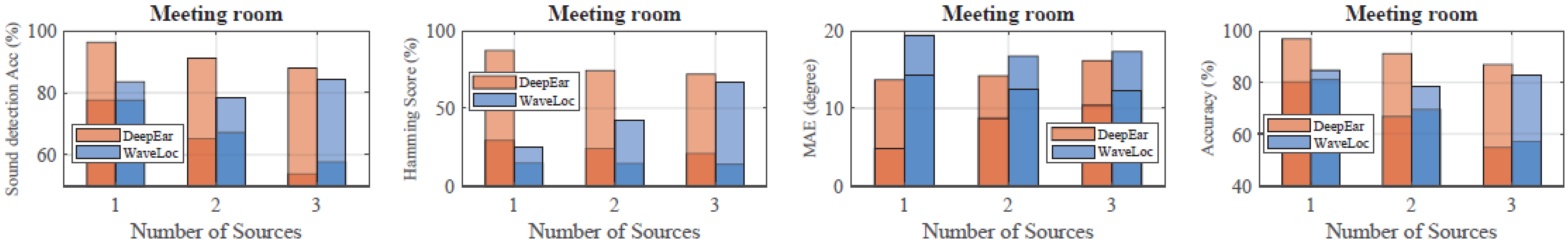
7.4°

16°

8.8°

AoA Error

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



(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.

Evaluation

Reverberant environment: lecture room

89.4%

81.5%

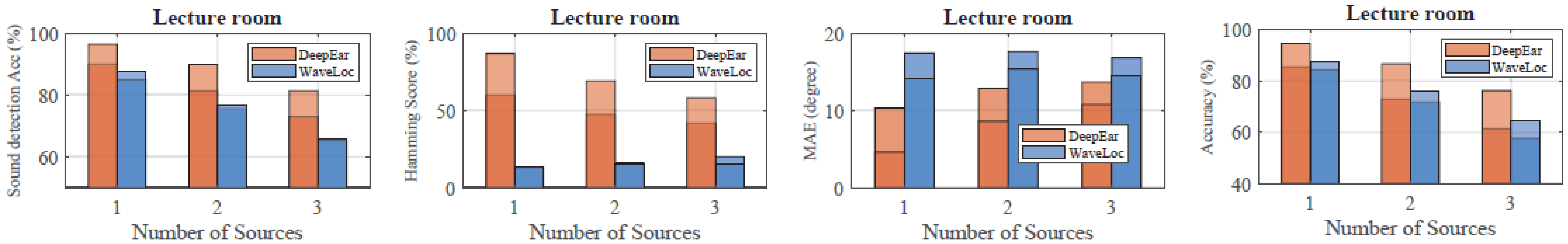
Transfer learning

12.9°

9°

Detection Accuracy

AoA Error



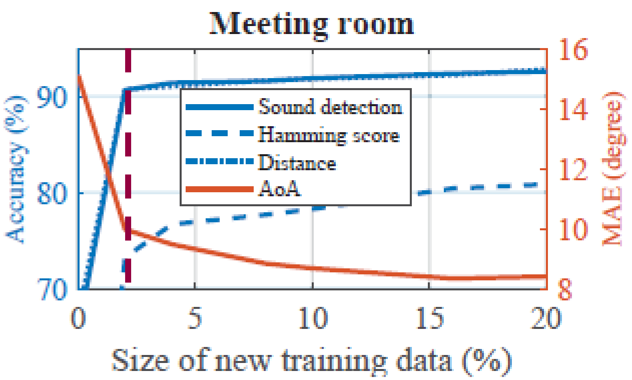
(a) Sound detection. (b) Hamming Score. (c) AoA. (d) Distance.

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

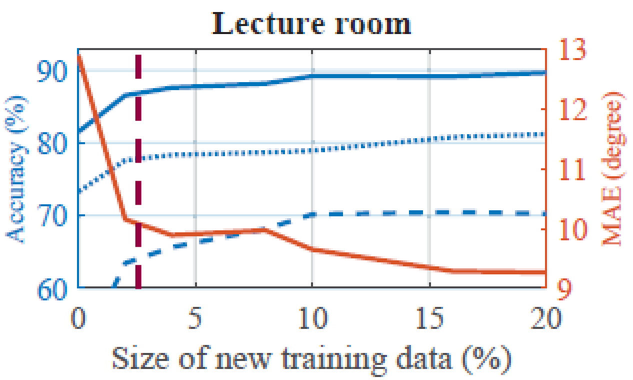
Evaluation

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.



(a) Meeting room.



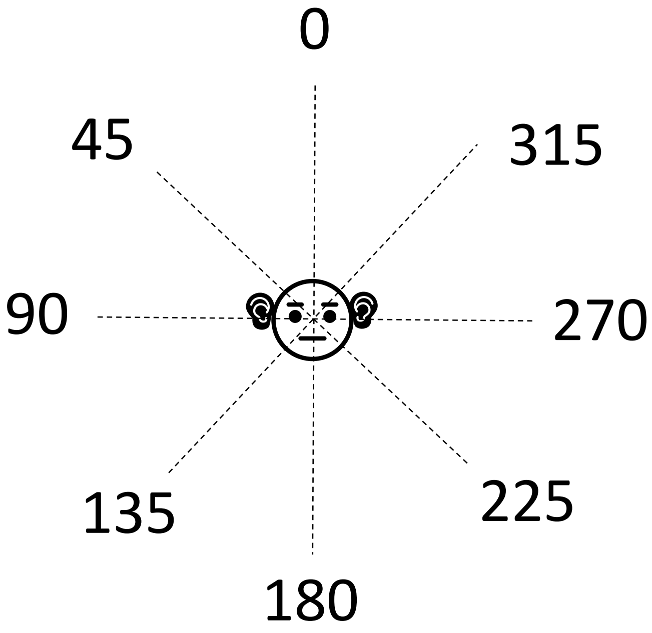
(b) Lecture room.

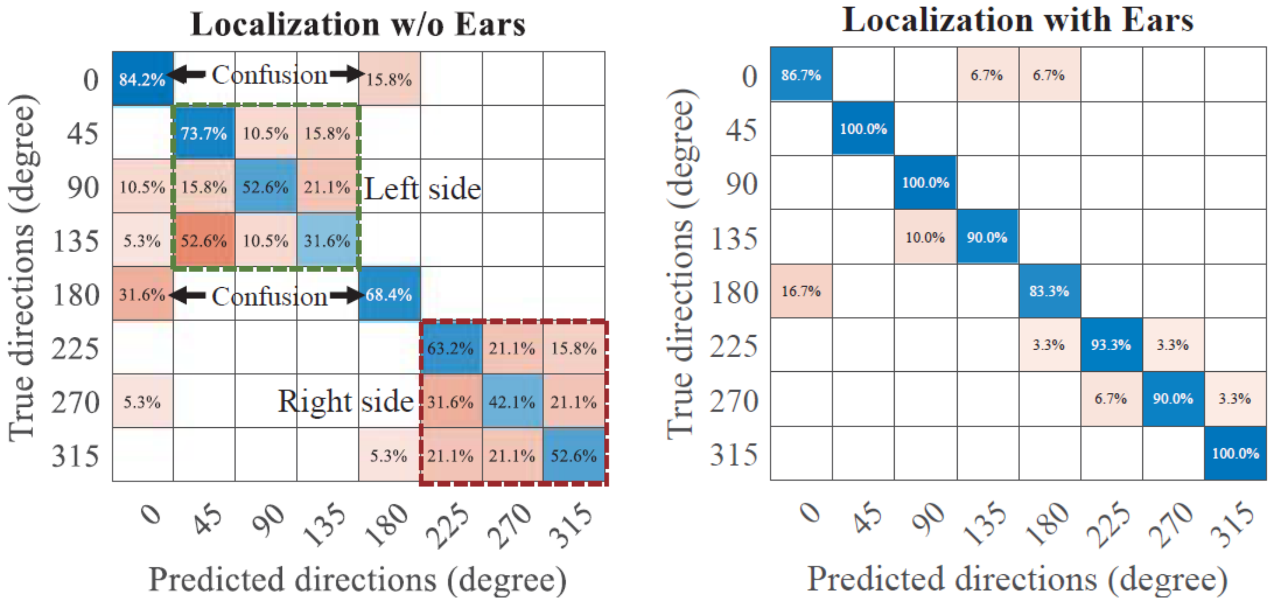
Fig. 14. The transfer learning performance of DeepEar with different sizes of new training data. Two subfigures share the same legend.

Evaluation

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.



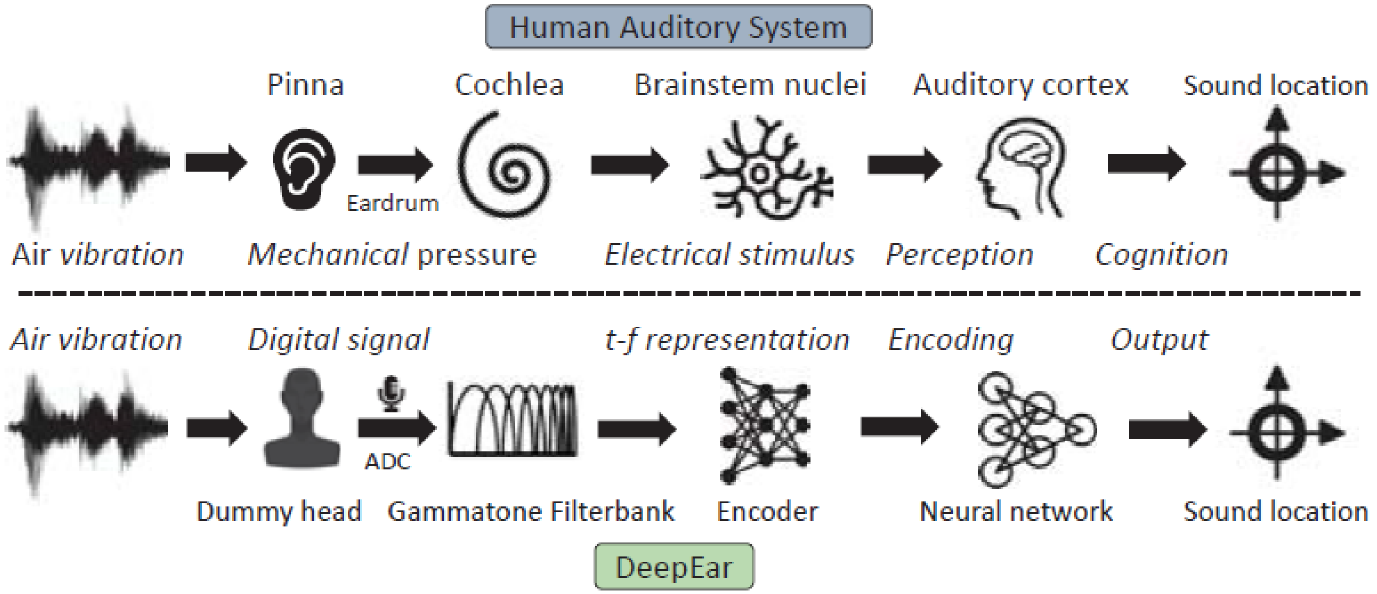


(a) Without human-shaped ears. (b) With human-shaped ears.

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