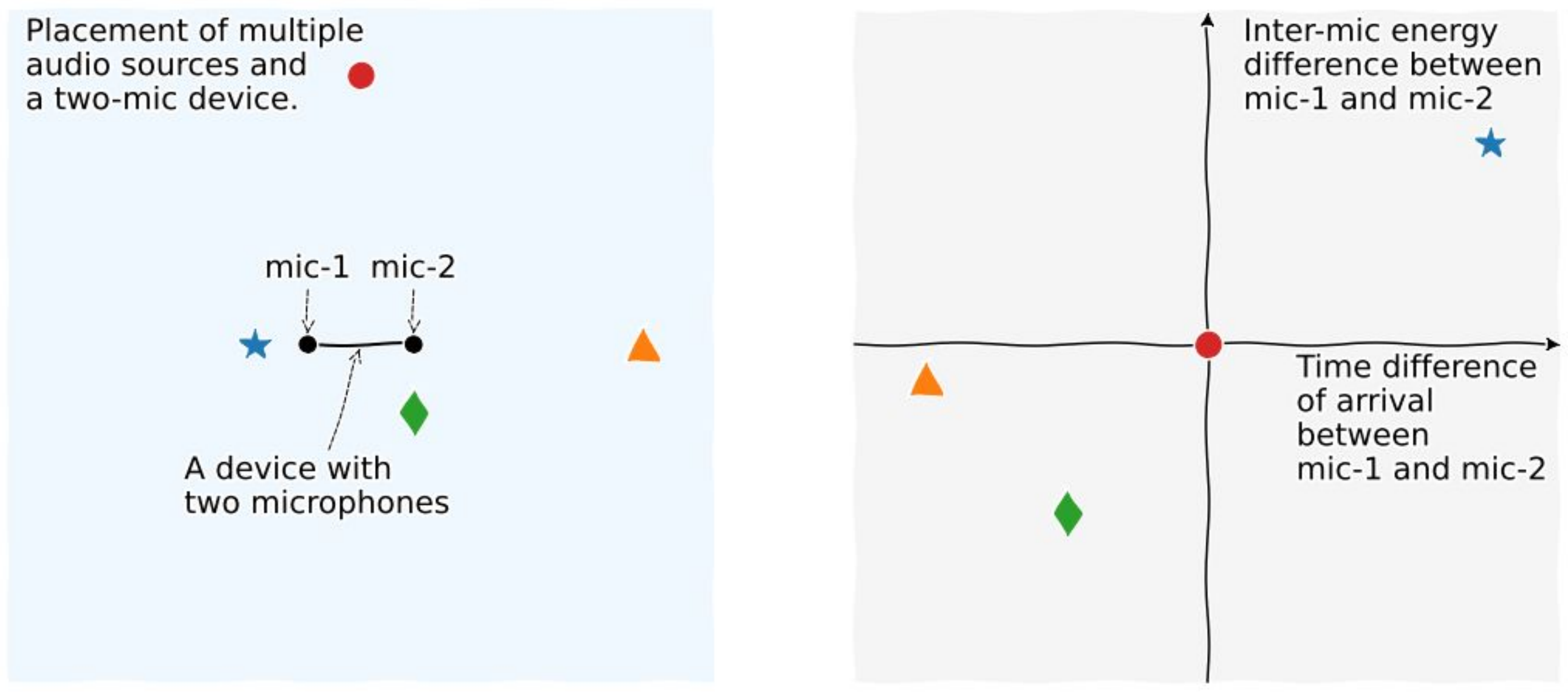
BASNet: Binaural Angular Separation Network

Yang Yang, George Sung, Shao-Fu Shih, Hakan Erdogan, Kevin Lee, Matthias Grundmann
{yanghm, gsung, shaofu, hakanerdogan, chehunglee, grundman}@google.com

Introduction


Placement of multiple audio sources and a two-mic device.
mic-1
mic-2
A device with two microphones
Inter-mic energy difference between mic-1 and mic-2
Time difference of arrival between mic-1 and mic-2

Definition

- **Delay contrast**: difference of TDoA (time difference of arrival) across multiple audio sources.
- **Gain contrast**: difference of inter-mic level difference across multiple audio sources.

Background

- Most devices are equipped with at least two microphones.
- The availability of a two-channel input provides spatial diversity of audio signals in the form of **delay and gain contrast**.

Our previous work

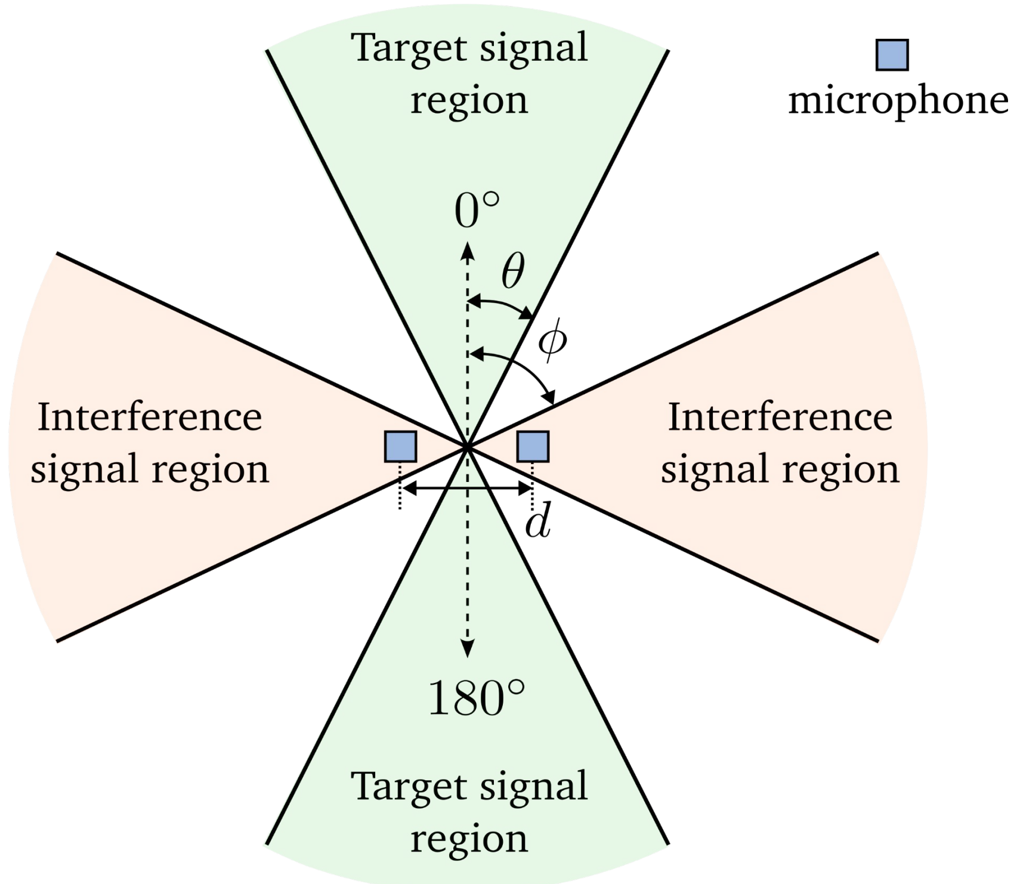
- We explored using gain contrast for audio separation in our previous work: **Guided Speech Enhancement Net**.

Motivating question:

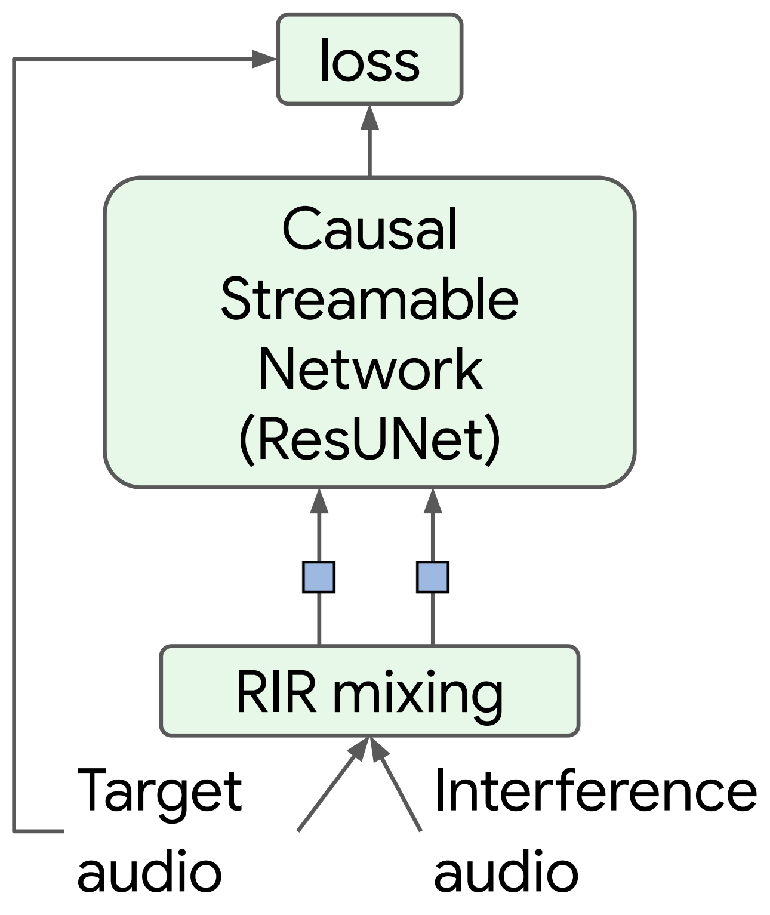
- Can we train a ML model that exploits the **delay contrast** information implicitly to achieve **spatial separation of audio**?

Method

An example config of our RIR simulator:



Training scheme:

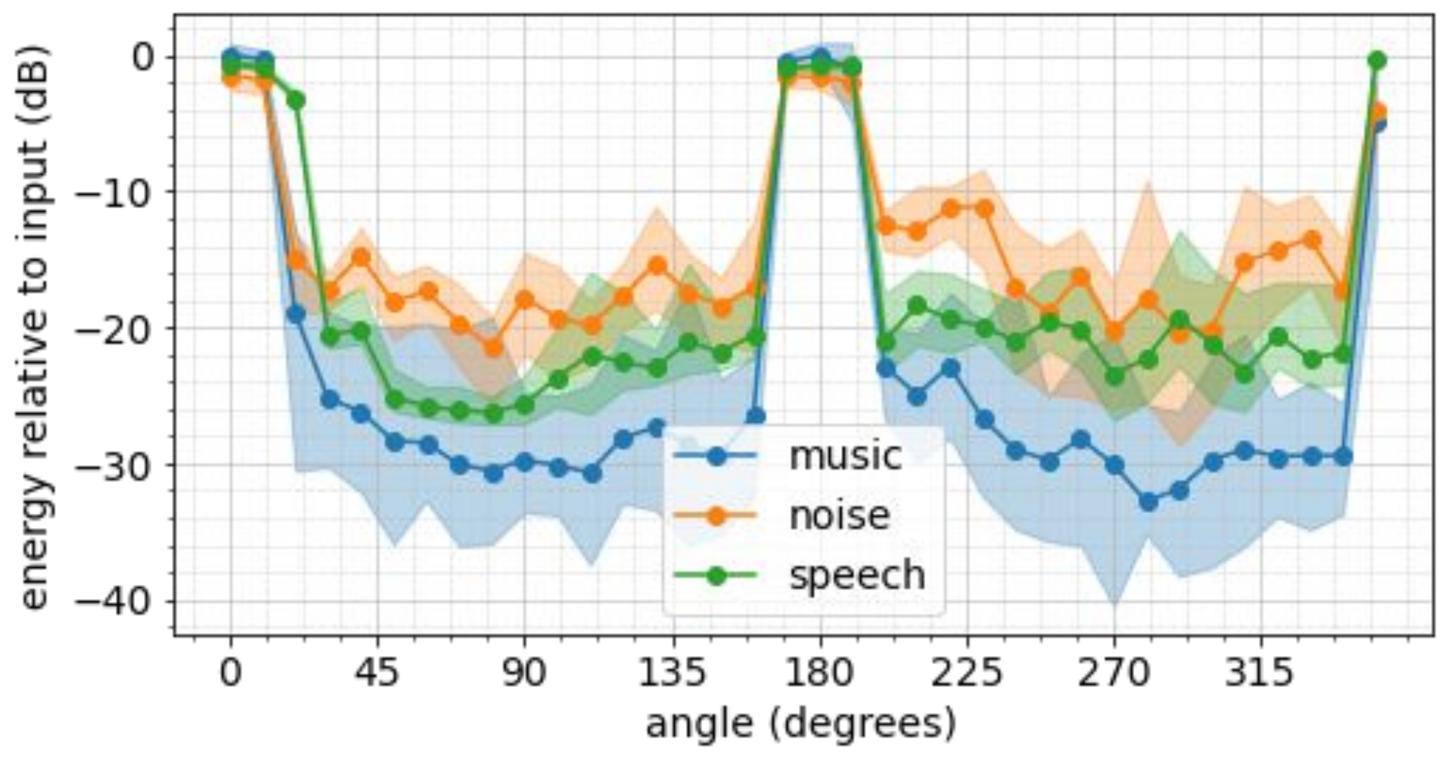


Key idea:

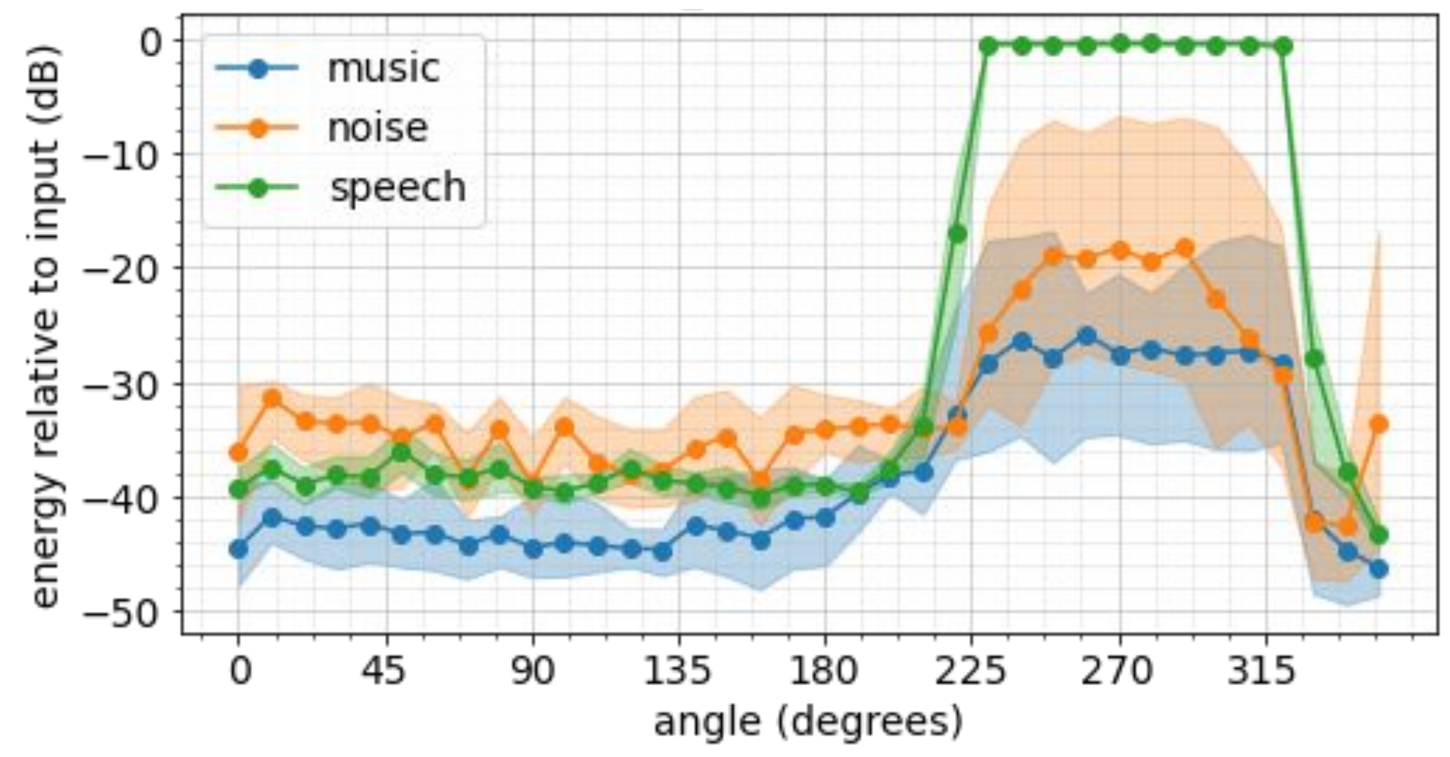
- Use a Room Impulse Response (RIR) simulator to generate two-channel audio inputs, where the target signal and interference signal come from different angular regions.
  - Which ensures that there is a delay contrast between the target and interference sources that can be exploited by the model.
  - Inter-mic distance needs to fit the device of interest.
- NN is trained to preserve the target signals and reject interference ones, implicitly exploiting the delay contrast between the two.

Results

Spatial separation of audio signals under different settings

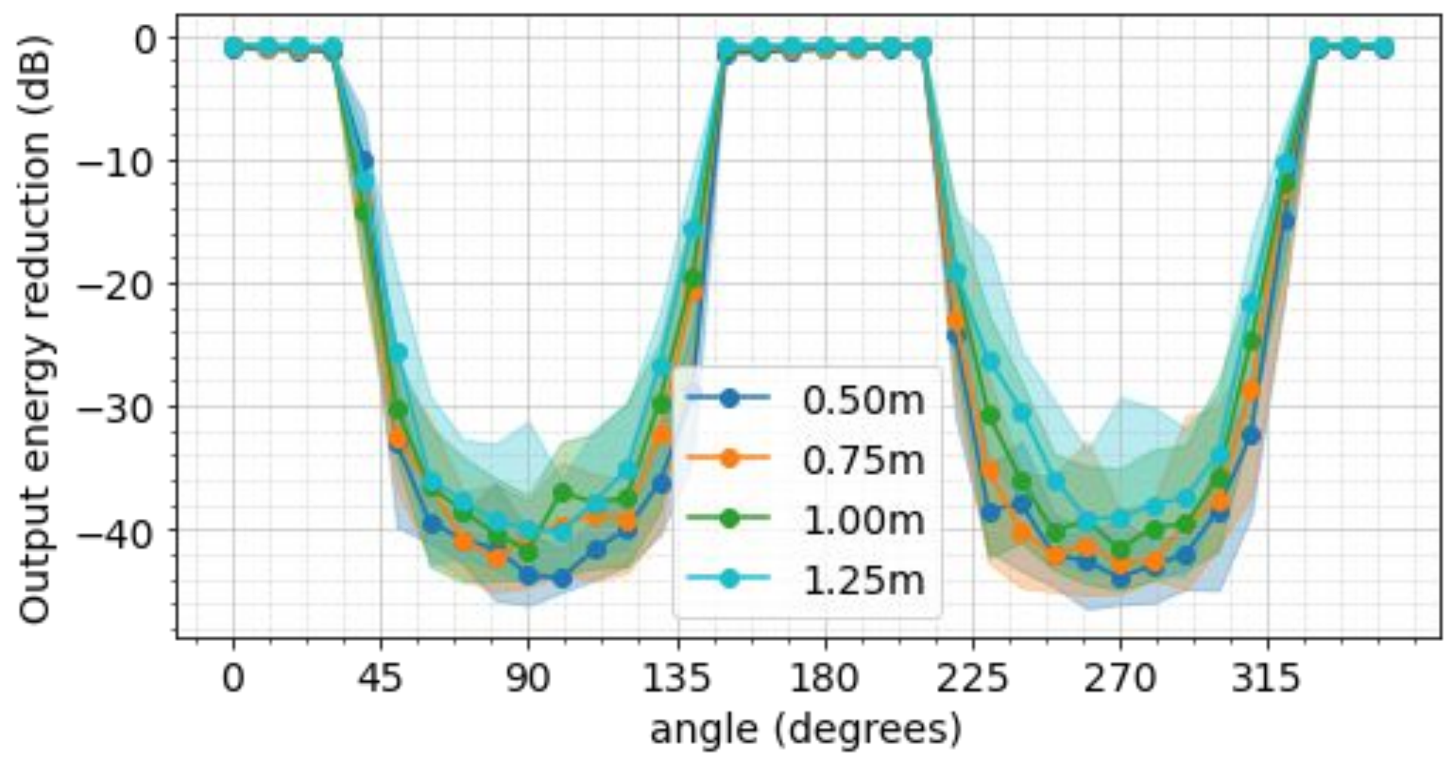


- Audio separation with target region = [-10°,+10°]

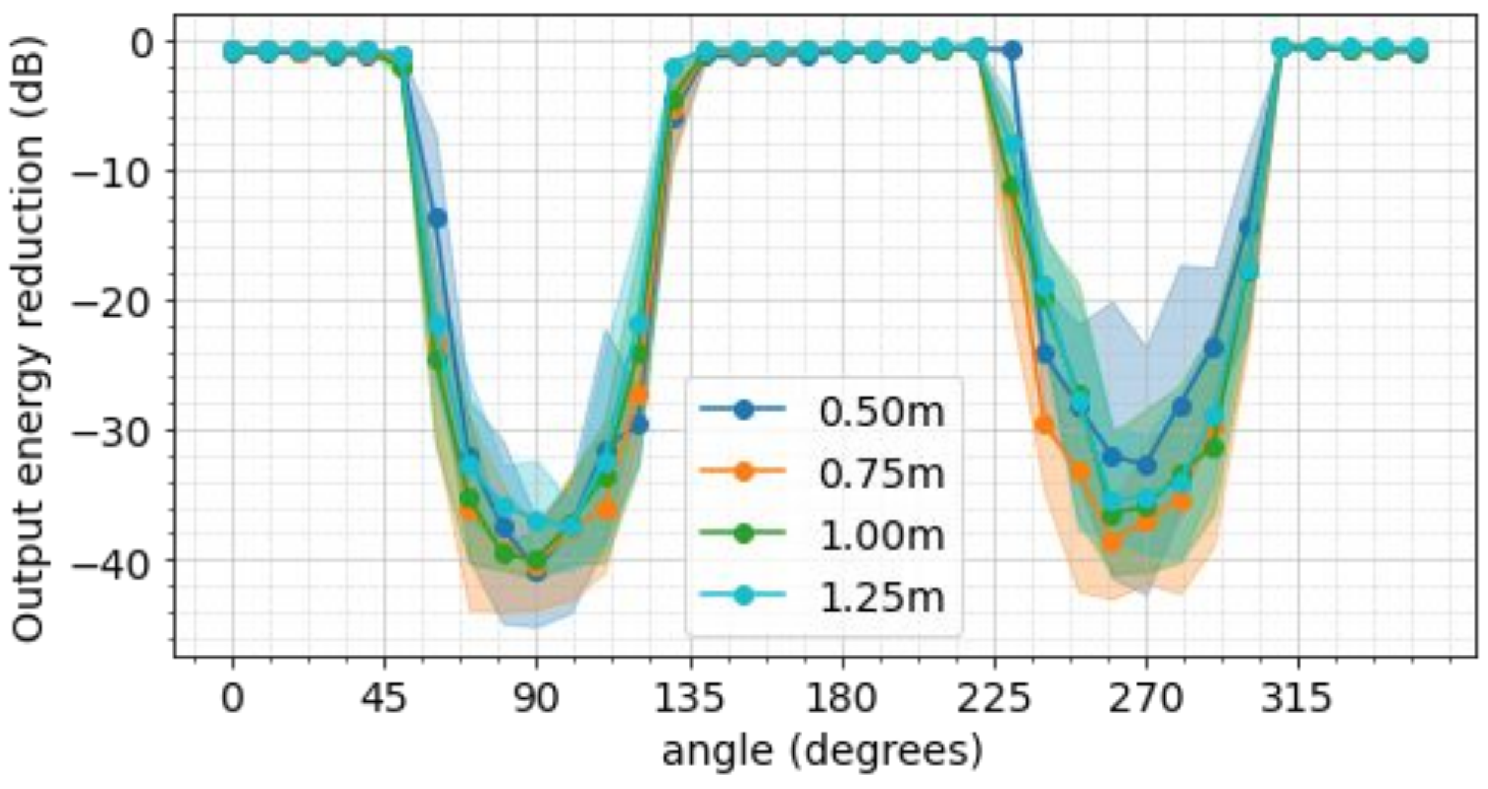


- Audio separation + speech denoising with target region = [-225°,+315°]

Beamwidth is consistent for sources with varying distances away from the device.

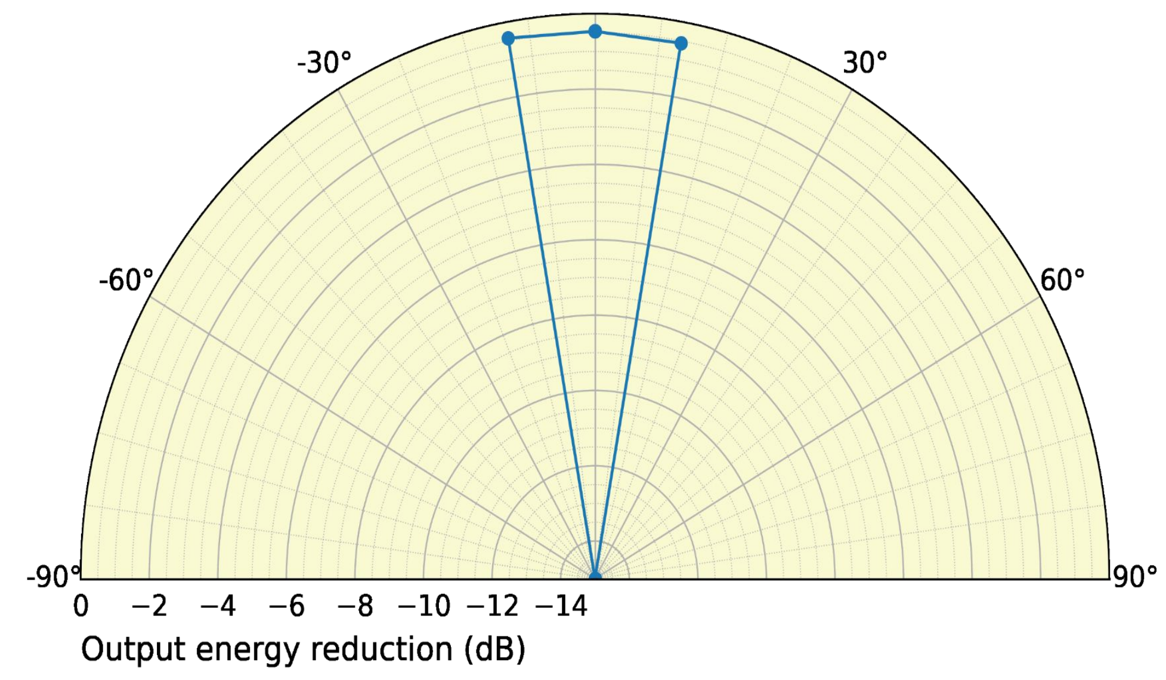


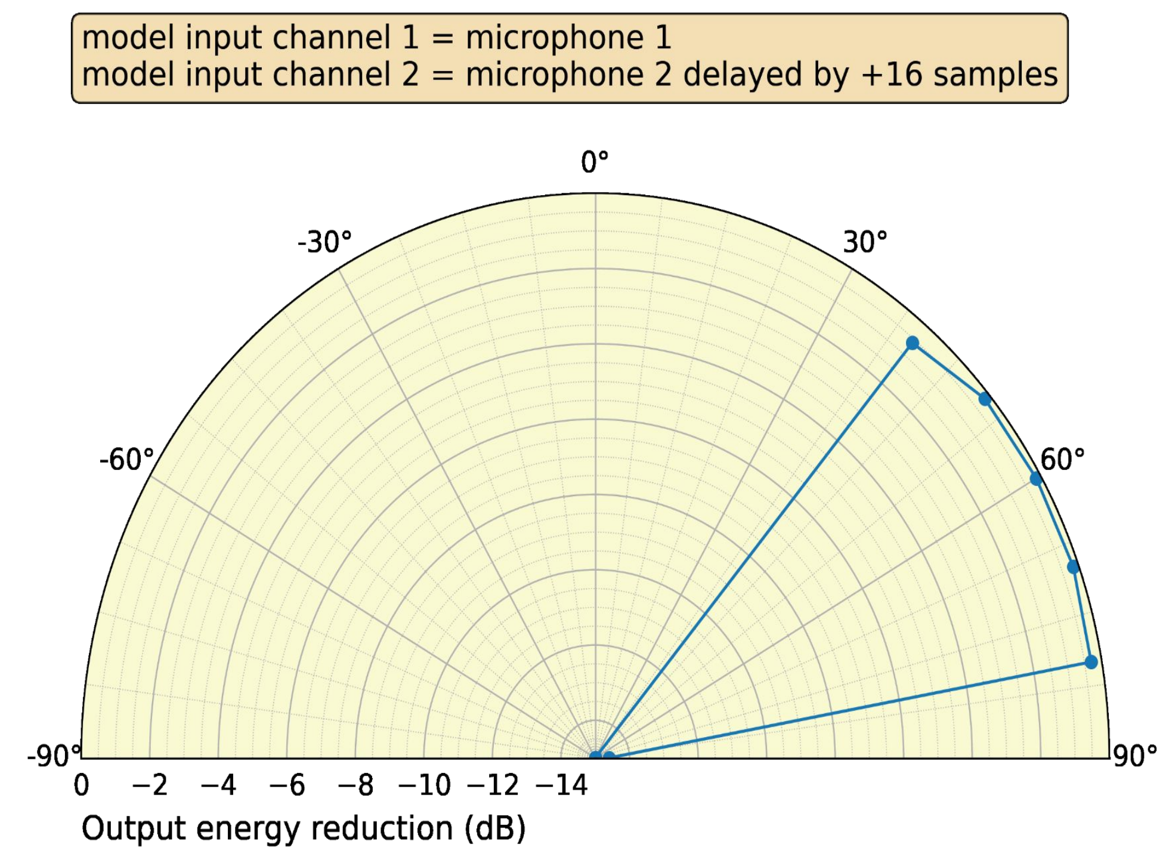
- Target region = [-30°,+30°]

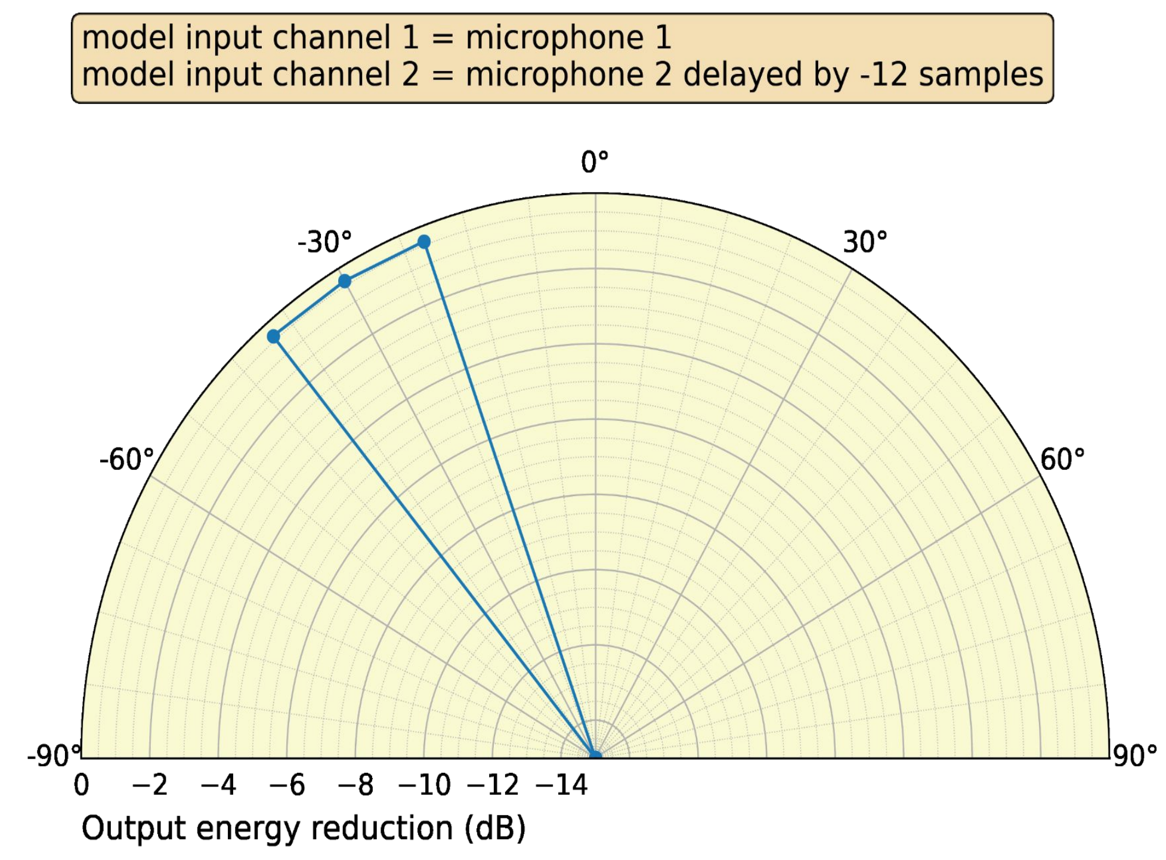


- Target region = [-45°,+45°]

Beam can be steered by adding artificial sample delay to one of the two input channels.







Conclusion

- We propose a neural network model that can separate target audio sources from interfering ones at different angular regions using two microphones.
- The model, even though trained on simulated room acoustics, performs effectively with commercial devices in real-world recording environments.
  - For optimal results, the inter-mic distance in the RIR config should closely approximate that of the device.
- **Beam steering** can be achieved by adding artificial sample delay to one of the two microphones
- Speech denoising and spatial separation can be achieved in a single model
- Please visit Google booth for a real-time demo on Pixel 8.