RAW WAVEFORM BASED END-TO-END DEEP CONVOLUTIONAL NETWORK FOR SPATIAL LOCALIZATION OF MULTIPLE ACOUSTIC SOURCES

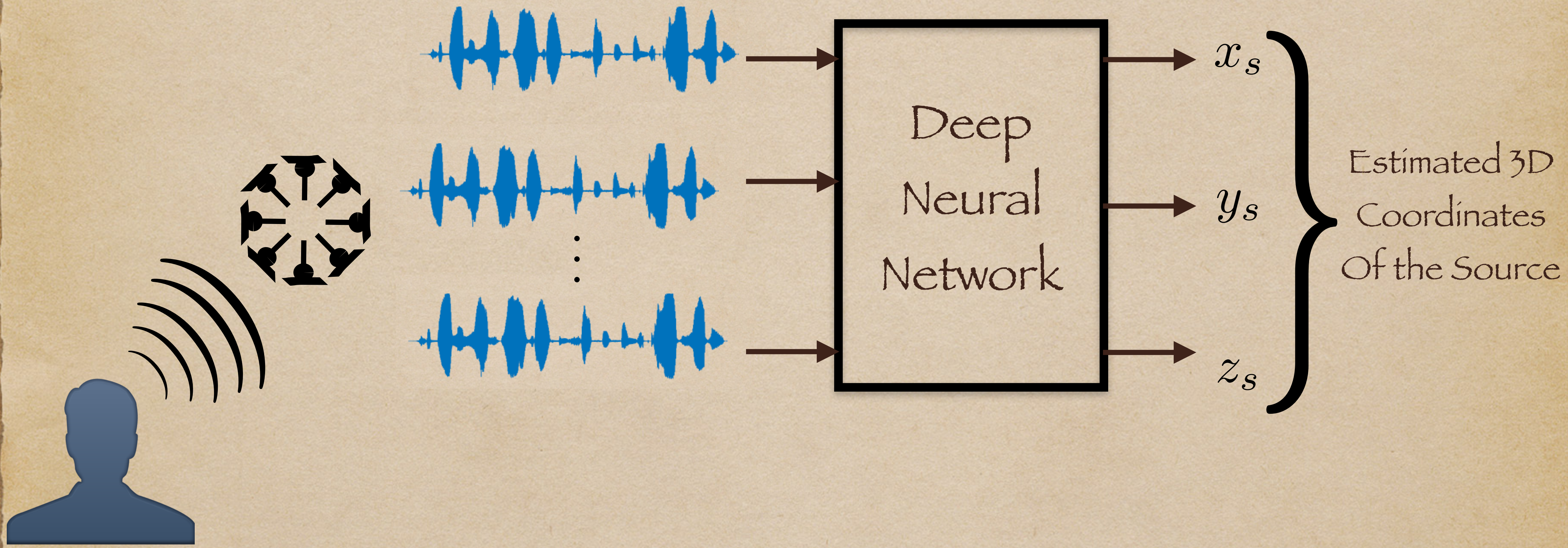
Harshavardhan Sundar

Weiran Wang

Ming Sun

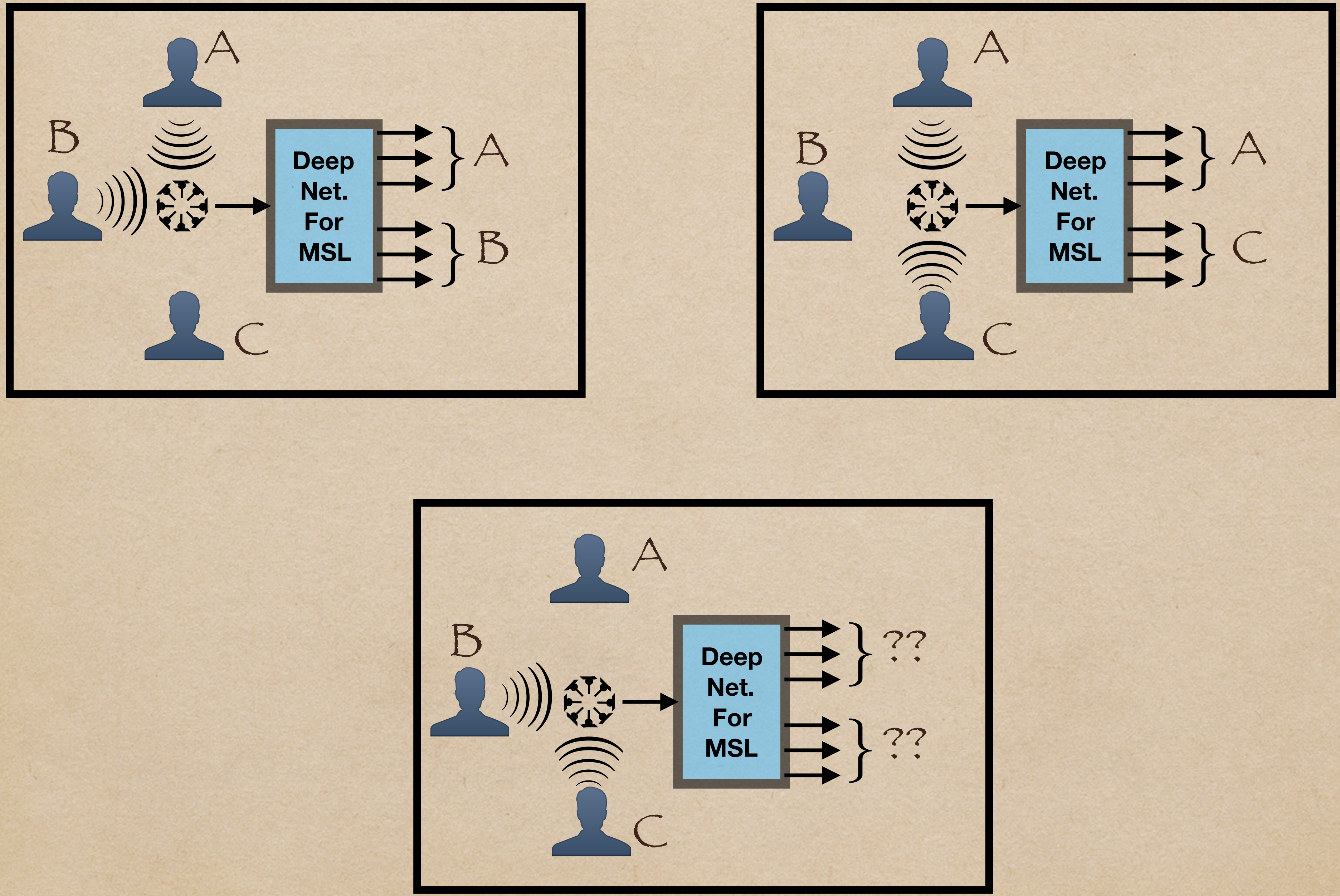
Chao Wang

Raw Audio based Acoustic Source Localization Using Deep Learning

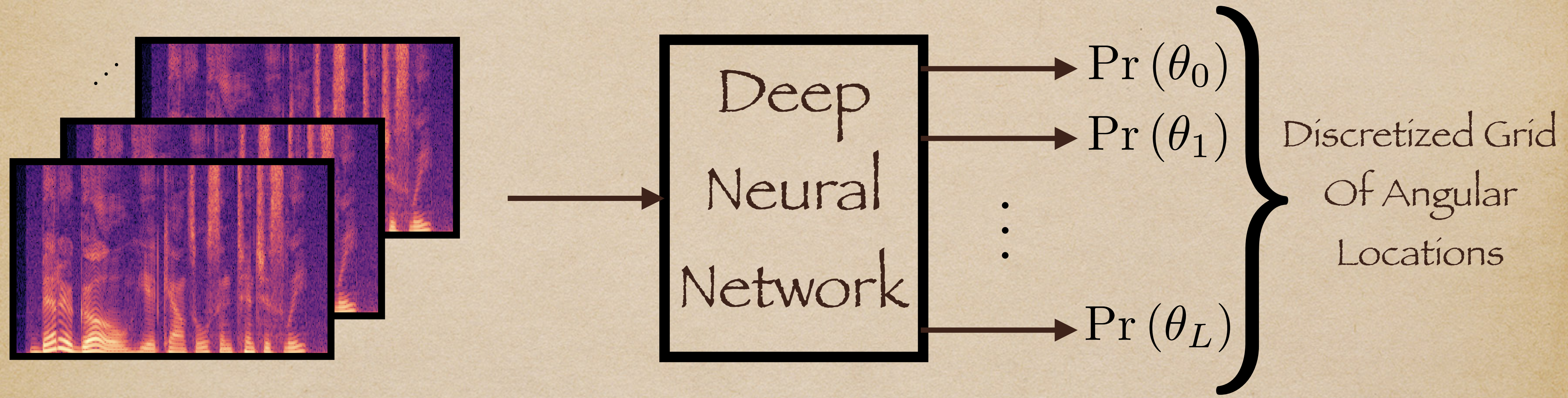


Ref: J. M. Vera-Diaz, D.Pizarro, and J. M. Guarasa, “Towards end-to-end acoustic localization using deep learning: From audio signal to source position coordinates,” CoRR, vol. abs/1807.11094, 2018.

Multiple Source Localization: Permutation Problem



Multiple Source Localization as Multi-Label Classification



- No Permutation Problem
- Regression —> Classification

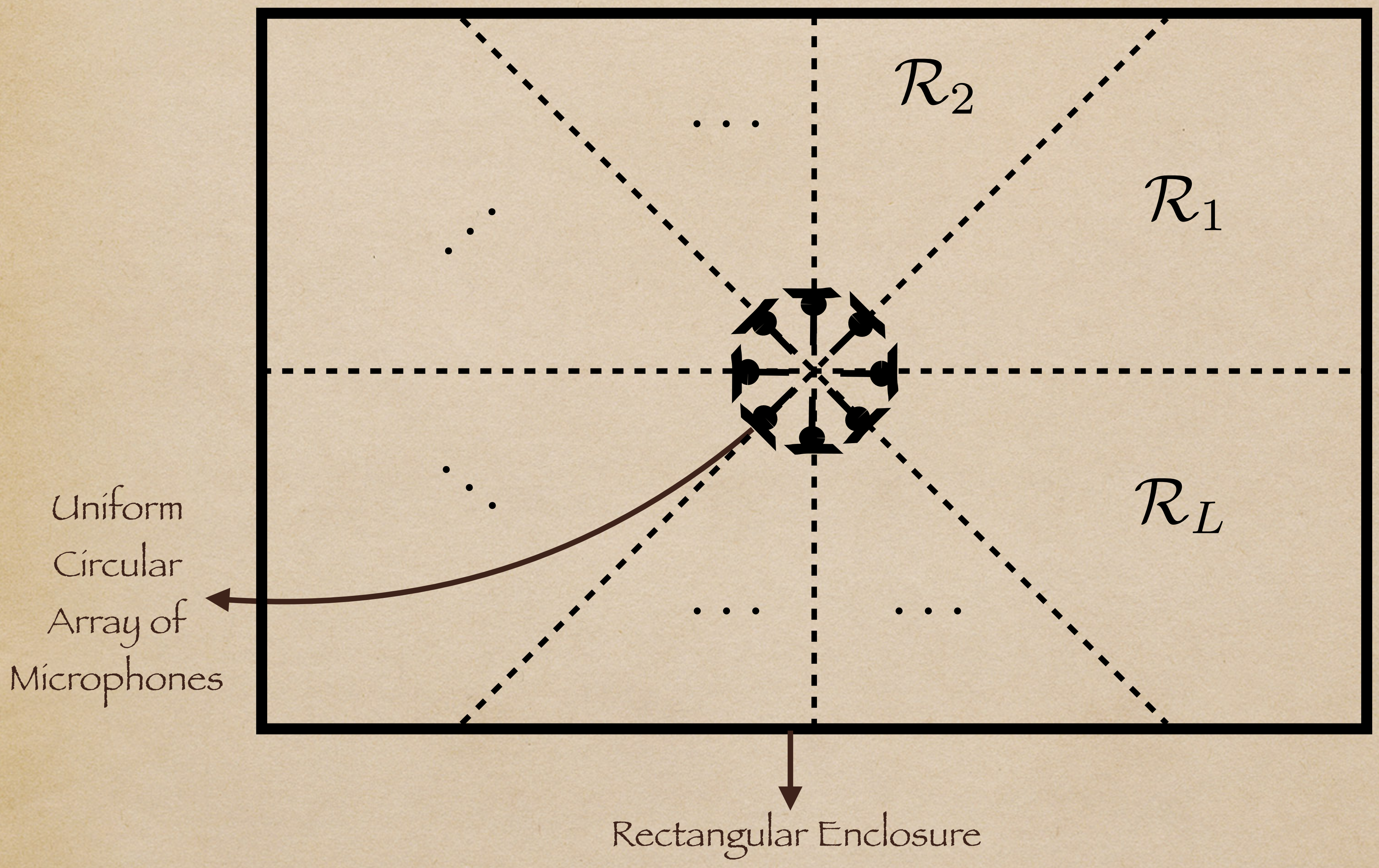
- Spatial resolution is limited by grid size
- Off grid?
- Requires training from all combination of grid points chosen up to 3 at a time.

Ref: S. Adavanne, A. Politis, and T. Virtanen, “Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network,” in 2018 26th European Signal Processing Conference (EUSIPCO), Sep. 2018, pp. 1462–1466.

In this Paper…

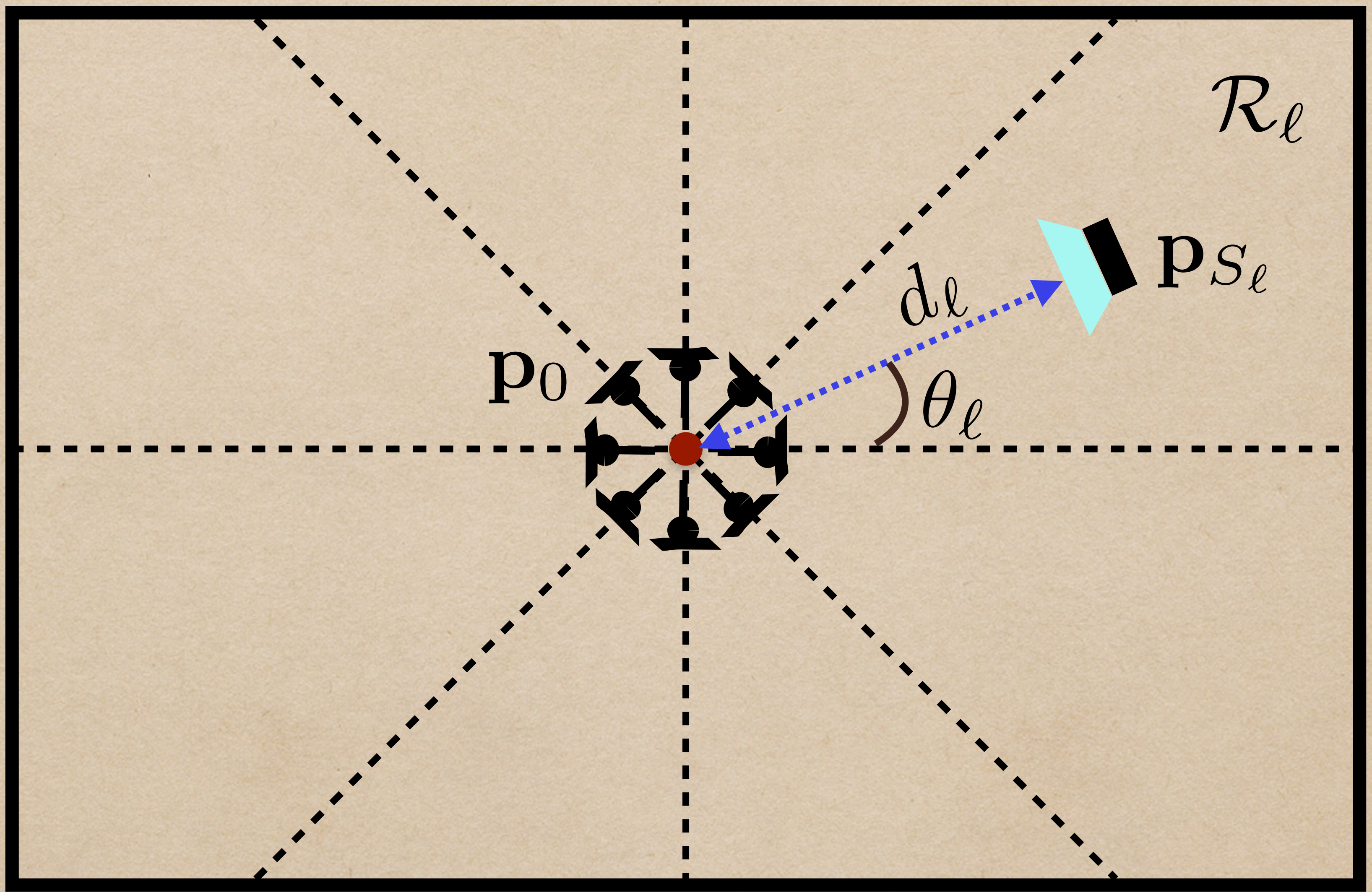
- Multiple Source Localization
  - End-to-End Starting from Raw Audio
  - Avoid Permutation Problem
  - Arbitrary Spatial Resolution

The Setup



Ref: H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "TDOA based multiple acoustic source localization without association ambiguity," IEEE/ACM Trans. on Audio, Speech, and Language Process., vol. 26, no. 11, pp. 1976–1990, Nov. 2018.

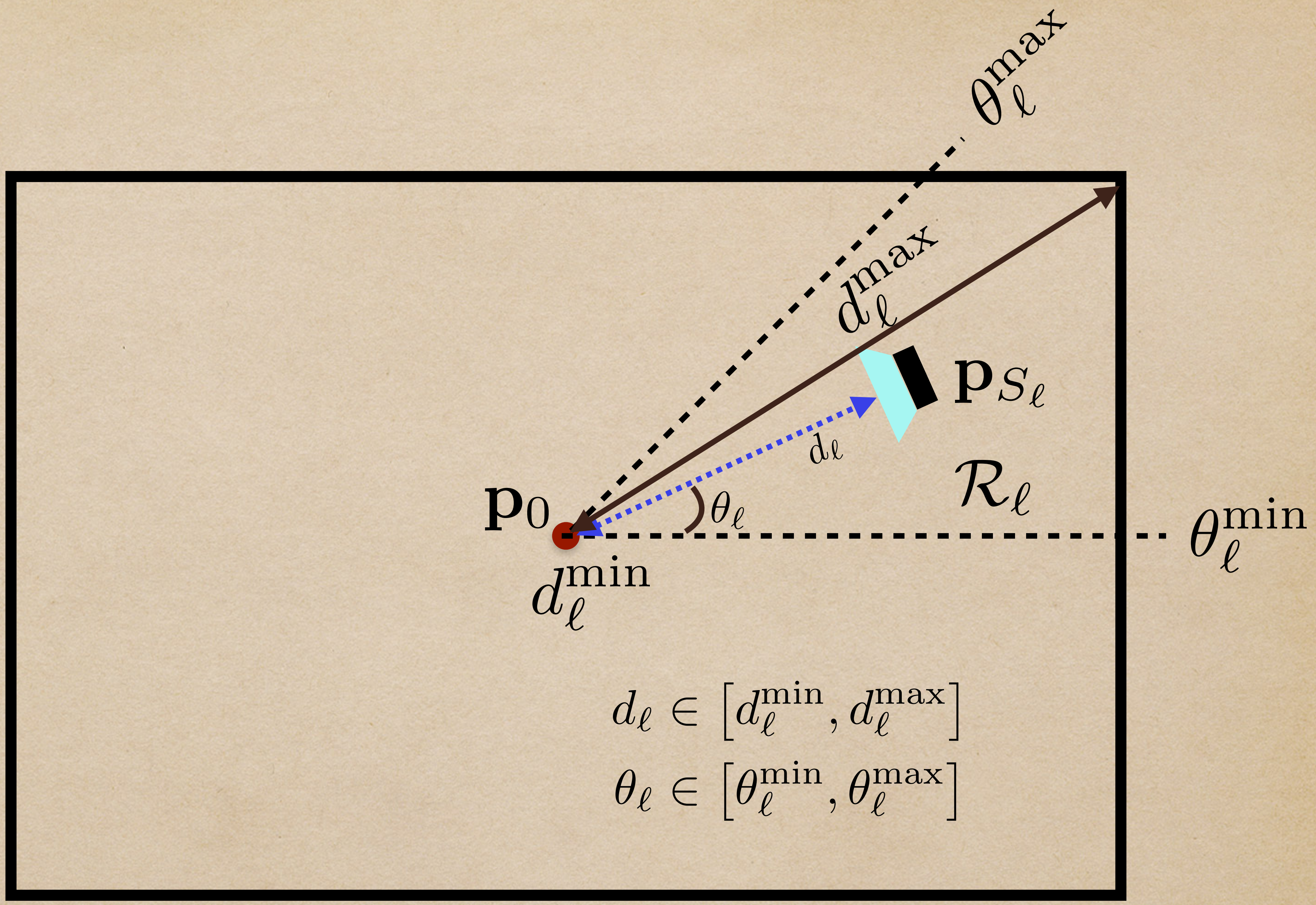
Output Encoding: Coarse -Fine Localization Strategy



$$d_\ell = \|\mathbf{p}_{S_\ell} - \mathbf{p}_0\|$$
$$\theta_\ell = \angle(\mathbf{p}_{S_\ell} - \mathbf{p}_0)$$

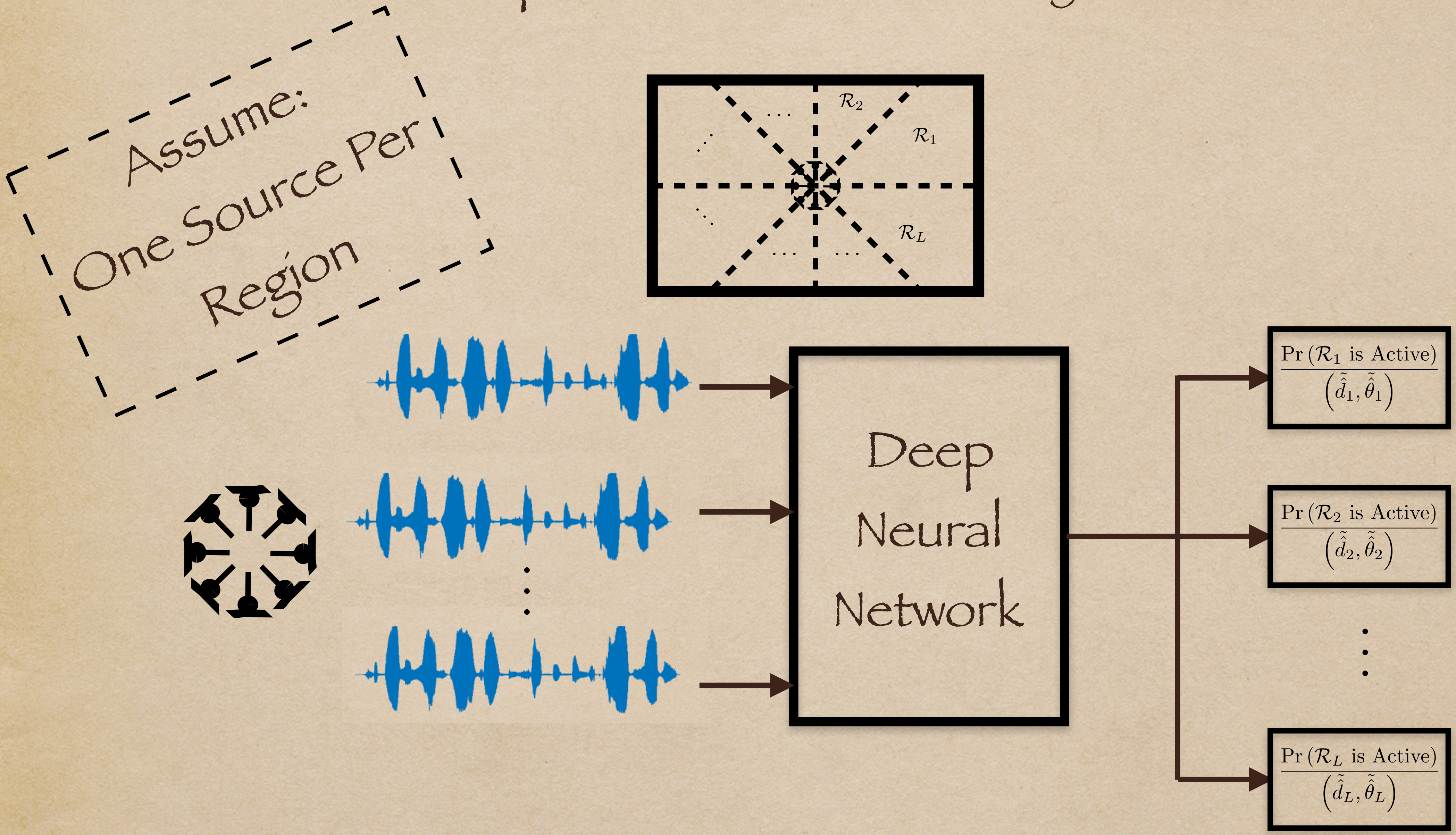
$$\boxed{\mathbf{p}_{S_\ell} = d_\ell \ \angle\theta_\ell}$$

Normalized Source Co-ordinates



$$\tilde{d}_\ell = \frac{d_\ell - d_\ell^{\min}}{d_\ell^{\max} - d_\ell^{\min}} \qquad \tilde{\theta}_\ell = \frac{\theta_\ell - \theta_\ell^{\min}}{\theta_\ell^{\max} - \theta_\ell^{\min}} \longrightarrow \begin{array}{l} \tilde{d}_\ell \in [0,1] \\ \tilde{\theta}_\ell \in [0,1] \end{array}$$

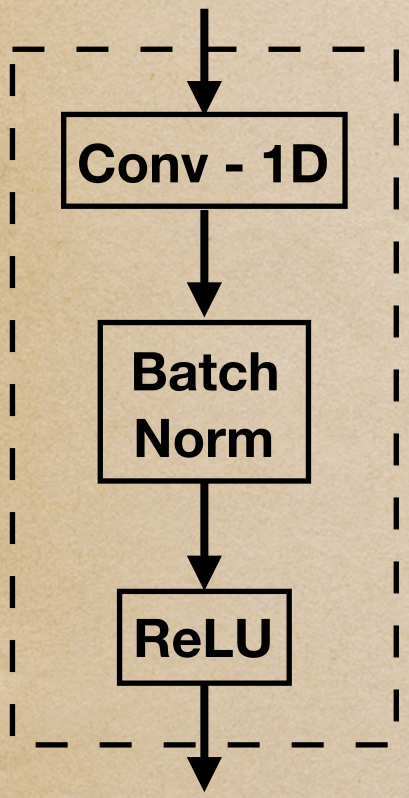
Input-Output Description of the Proposed End-to-End System



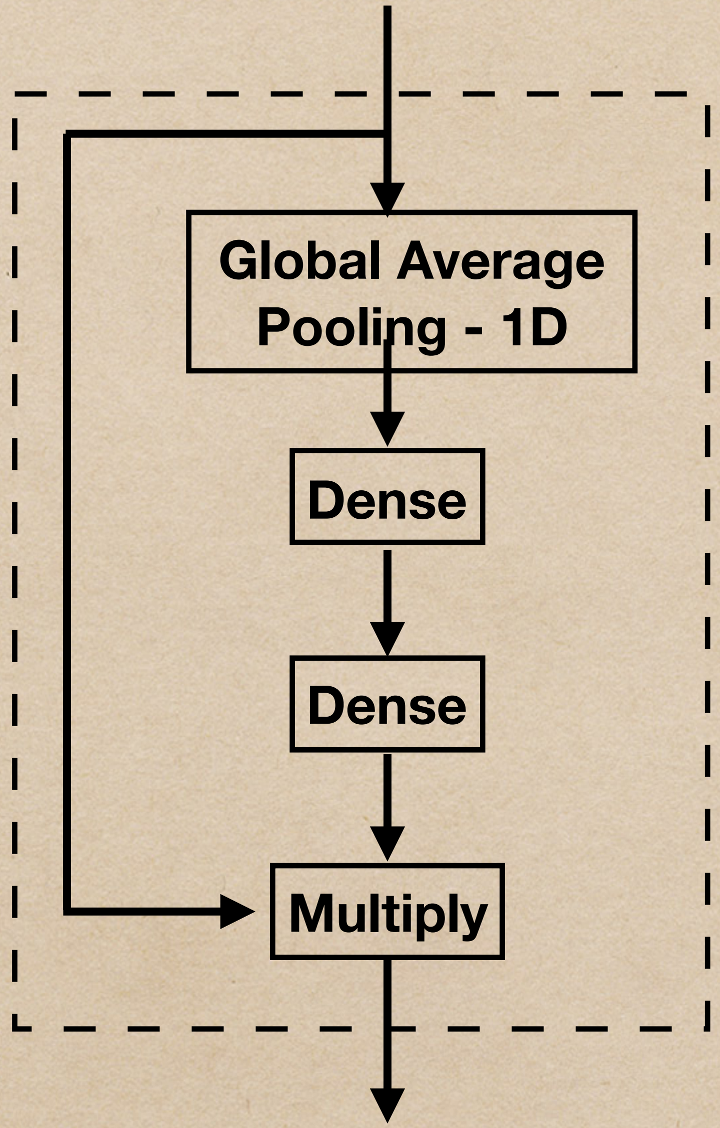
L Coarse regions $\rightarrow 3L$ Outputs

Training Targets $\rightarrow$ Active Regions + Normalized Source Co-ordinates

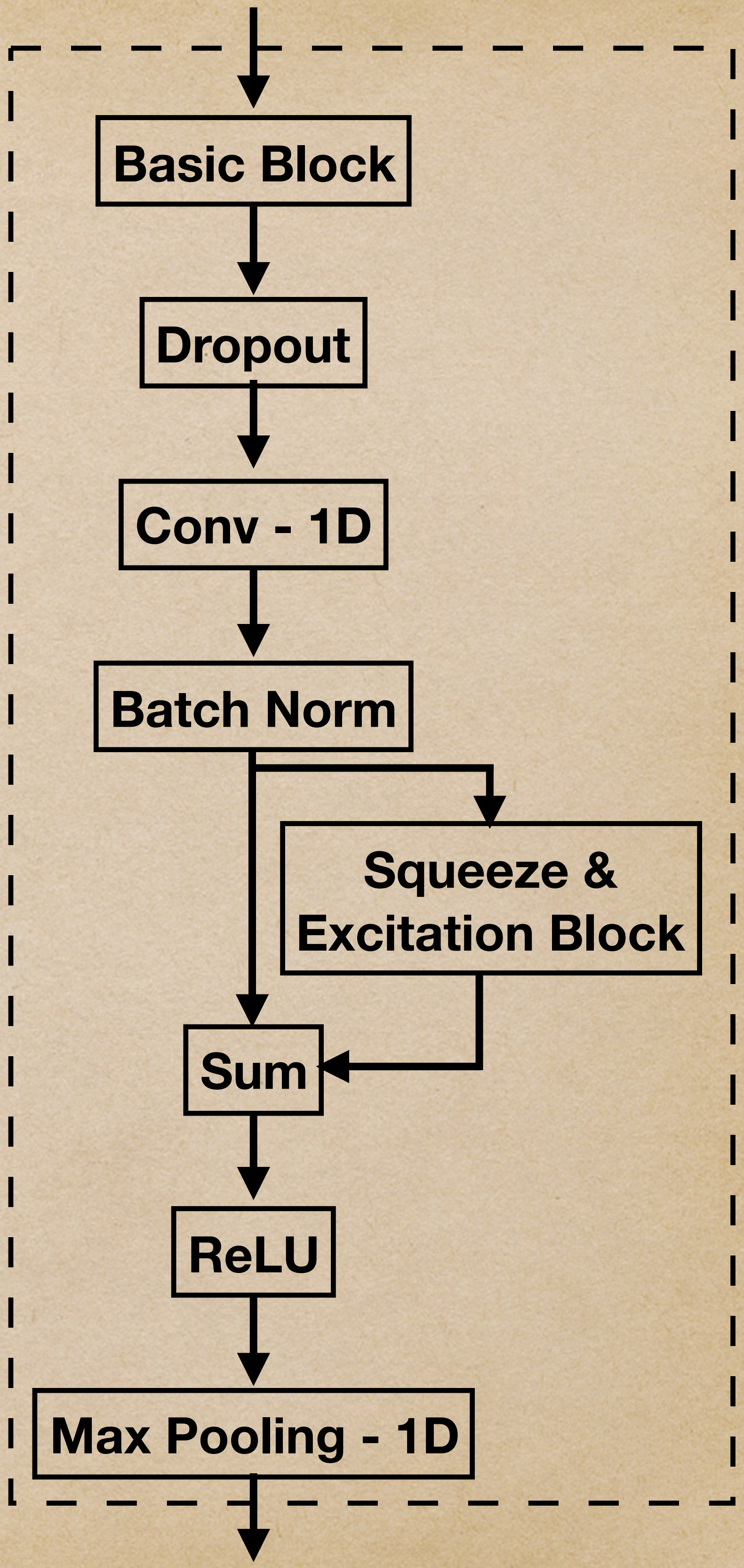
Building Blocks



Basic Block



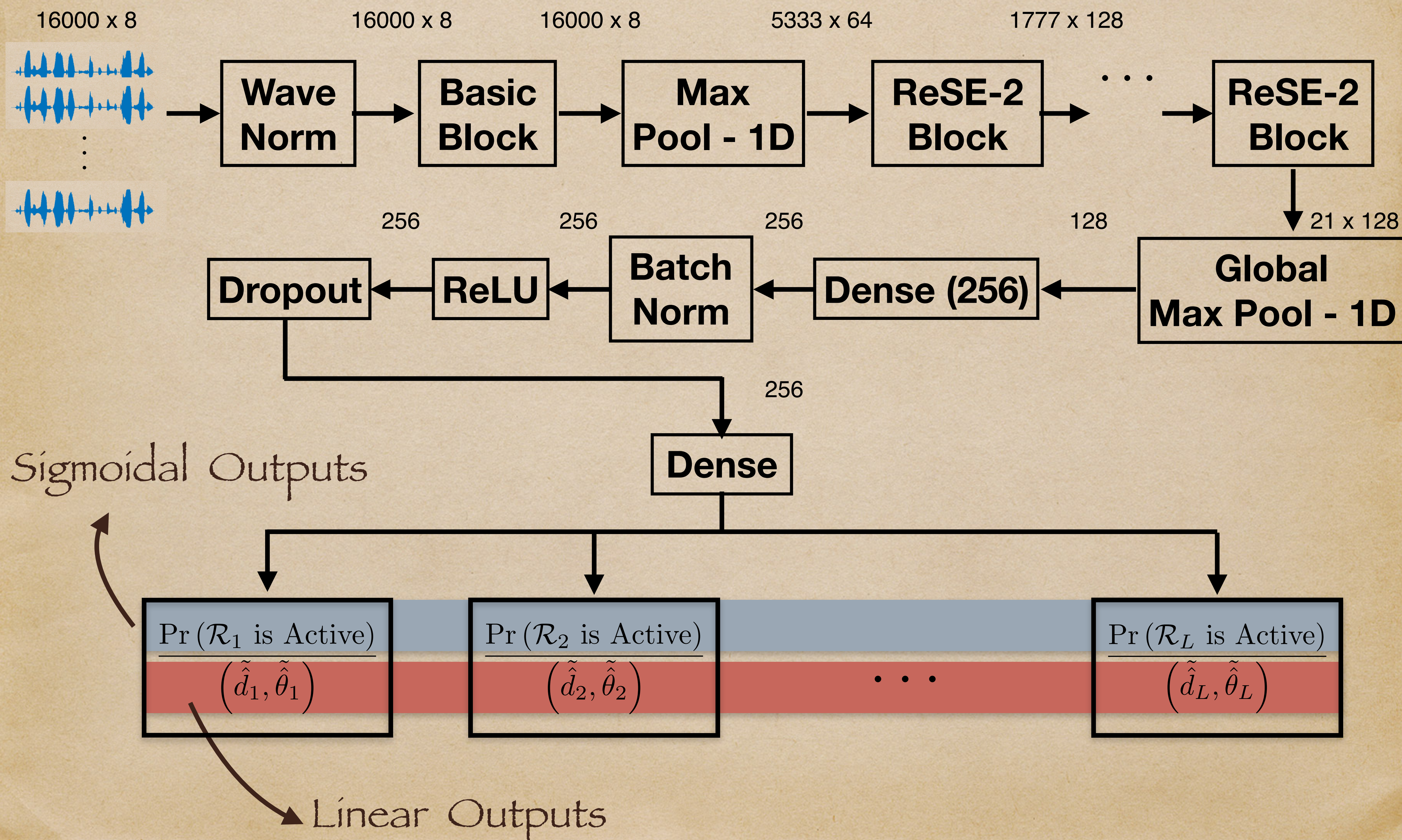
Squeeze & Excitation Block



ReSe-2 Block

Ref: T. Kim, J. Lee, and J. Nam, "Comparison and analysis of SampleCNN architectures for audio classification," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp.285–297, May 2019.

Overall Architecture: Deep Conv. Net with Skip Connections



Training Data

$$(\boldsymbol{W}, \boldsymbol{R}, \boldsymbol{D})$$

$$\left[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \cdots, \boldsymbol{w}^{(J)}\right]$$

$J \rightarrow$ No. of Samples

$\boldsymbol{w}^{(j)} \in \mathbb{R}^{16000 \times 8}$

Multi-Channel Raw Audio Data

(Input)

$$\left[\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \cdots, \boldsymbol{r}^{(J)}\right]$$

$$\boldsymbol{r}^{(j)} = \left[r_1^{(j)}, r_2^{(j)}, \cdots, r_L^{(j)}\right]^T$$

$$r_\ell^{(j)} = \begin{cases} 1 & \text{If } \mathcal{R}_\ell \text{ is active} \\ & \text{in the } j^{th} \text{ sample} \\ 0 & \text{Otherwise} \end{cases}$$

Coarse Region Labels - Binary

(Target)

$$\left[\boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)}, \cdots, \boldsymbol{d}^{(J)}\right]$$

$$\boldsymbol{d}^{(j)} = \begin{bmatrix} \left(\tilde{d}_1^{(j)}, \tilde{\theta}_1^{(j)}\right) \\ \left(\tilde{d}_2^{(j)}, \tilde{\theta}_2^{(j)}\right) \\ \vdots \\ \left(\tilde{d}_L^{(j)}, \tilde{\theta}_L^{(j)}\right) \end{bmatrix}$$

Fine Location Labels - [0,1]

(Target)

Coarse Localization: Multi-Label Classification Loss

$$\mathcal{L}_{\text{Coarse}}^{(j)} = -\frac{1}{L}\sum_{\ell=1}^{L}\left[r_\ell^{(j)} \log\left(\hat{r}_\ell^{(j)}\right) + \left(1 - r_\ell^{(j)}\right)\log\left(1 - \hat{r}_\ell^{(j)}\right)\right]$$

$$L = \text{No. of Coarse Regions}$$

$$\hat{r}_\ell^{(j)} = \Pr\left(\mathcal{R}_\ell \text{ is Active in the } j^{\text{th}} \text{ Sample}\right)$$

Fine Localization: Regression Loss

$$\mathcal{L}^{(j)}_{\text{Fine}} = \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{1}_{\left\{r_\ell^{(j)}=1\right\}} \underbrace{\sqrt{\left(\tilde{d}_\ell^{(j)} - \hat{\tilde{d}}_\ell^{(j)}\right)^2 + \left(\tilde{\theta}_\ell^{(j)} - \hat{\tilde{\theta}}_\ell^{(j)}\right)^2}}_{\text{Squared Error}}$$

Indicates

Active Regions

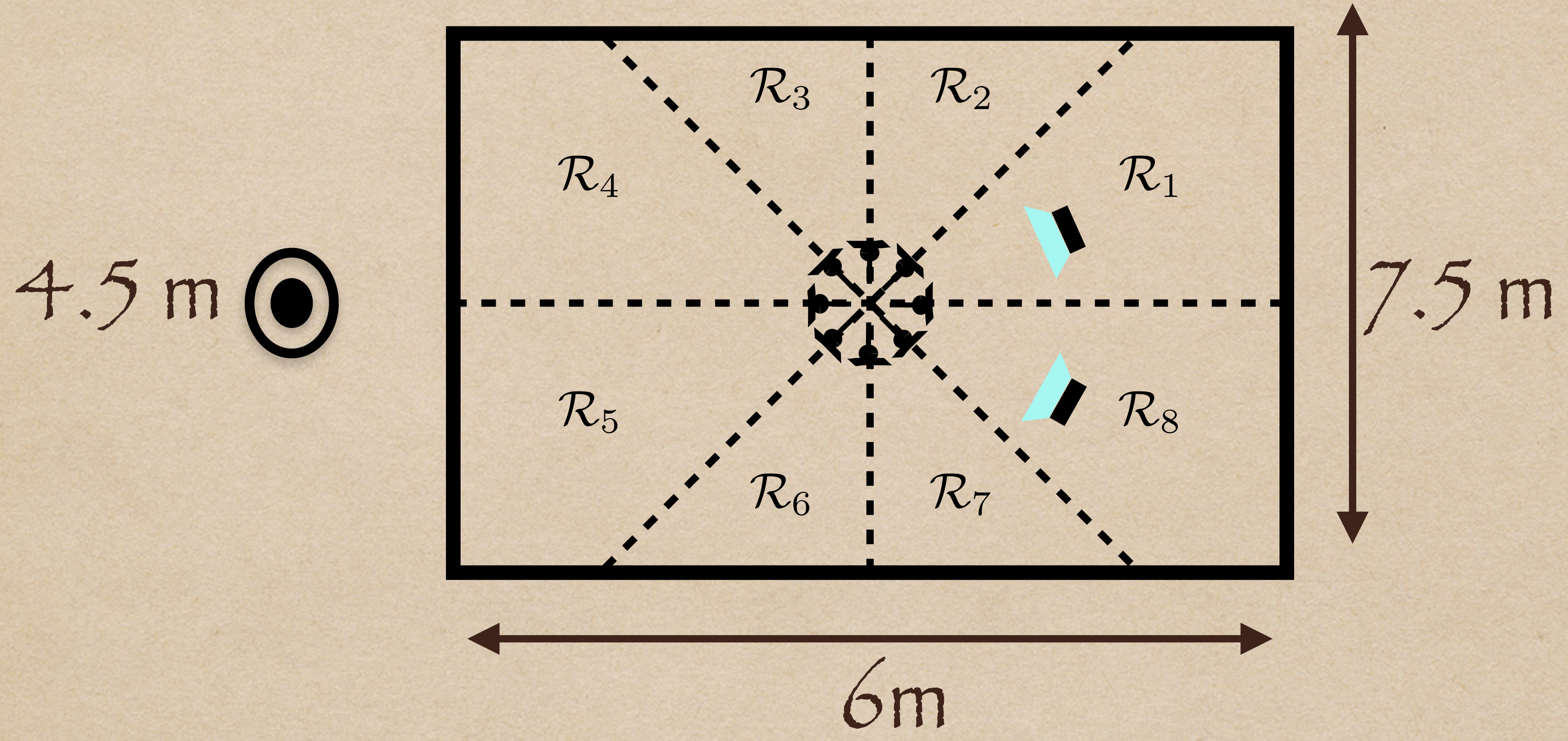
Joint Coarse - Fine Localization Loss

$$\mathcal{L}^{(j)} = \alpha \cdot \mathcal{L}^{(j)}_{\text{Coarse}} + \beta \cdot \mathcal{L}^{(j)}_{\text{Fine}}$$

$$\mathcal{L} = \frac{1}{J} \sum_{j=1}^{J} \mathcal{L}^{(j)}$$

Performance Analysis

Simulated Dataset Details



$$x_j[n] = \sum_{i=1}^{M} s_i[n] \star h_{ij}[n]$$

Microphone Signal

Clean Speech From TIMIT DR8

RIR using Image Method

Simulated Dataset Details

Acoustic Condition	Train	Validation	Test
Anechoic	33,356	443	414
Reverb (RT60 = 300 ms)	34,196	460	456

Table 1: Simulated Data set statistics. No. of 1s Audios.

Proposed Approach: SMESLP

- Sample based Multiple Encoded Source Location Predictor (SMESLP)
- Trained only on Anechoic Data: SMESLP-Anechoic
- Trained only on reverb Data: SMESLP-Reverb

Performance Metrics

Task	Performance Metric
Coarse Localization Accuracy	Hamming Score (Jacard Index)
Fine Localization Accuracy	Absolute Direction of Arrival Error

$$T = \text{Set of True Active Regions}$$

$$P = \text{Set of Predicted Active Regions}$$

$$\text{Hamming Score} = \frac{|T \bigcap P|}{|T \bigcup P|}$$

Baseline for Comparison

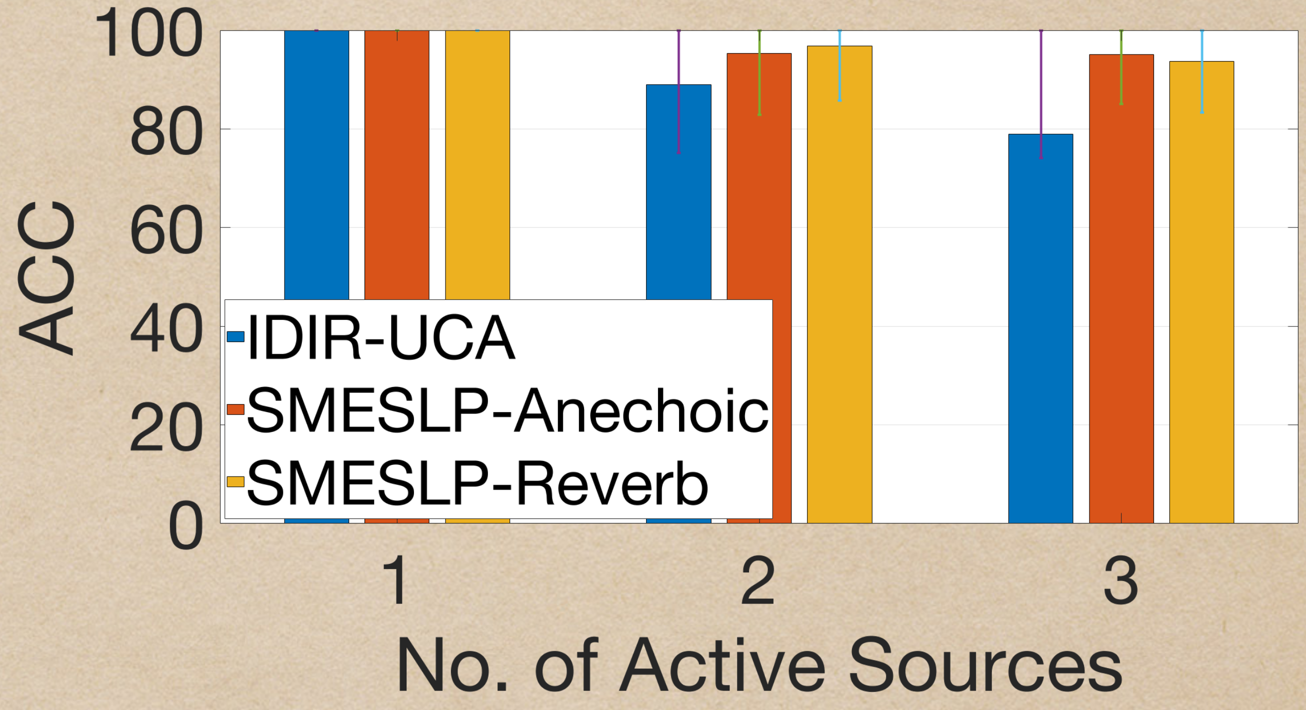
- A Signal Processing Approach based on Time Difference of Arrival (TDoA) avoiding the permutation problem.
- Also uses Uniform Circular Array (UCA)
- Referred to as IIDIR-UCA (Intersection of Inverse Delay-Interval Region)

Ref: H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "TDOA based multiple acoustic source localization without association ambiguity," IEEE/ACM Trans. on Audio, Speech, and Language Process., vol. 26, no. 11, pp. 1976–1990, Nov. 2018.

Coarse Localization Performance

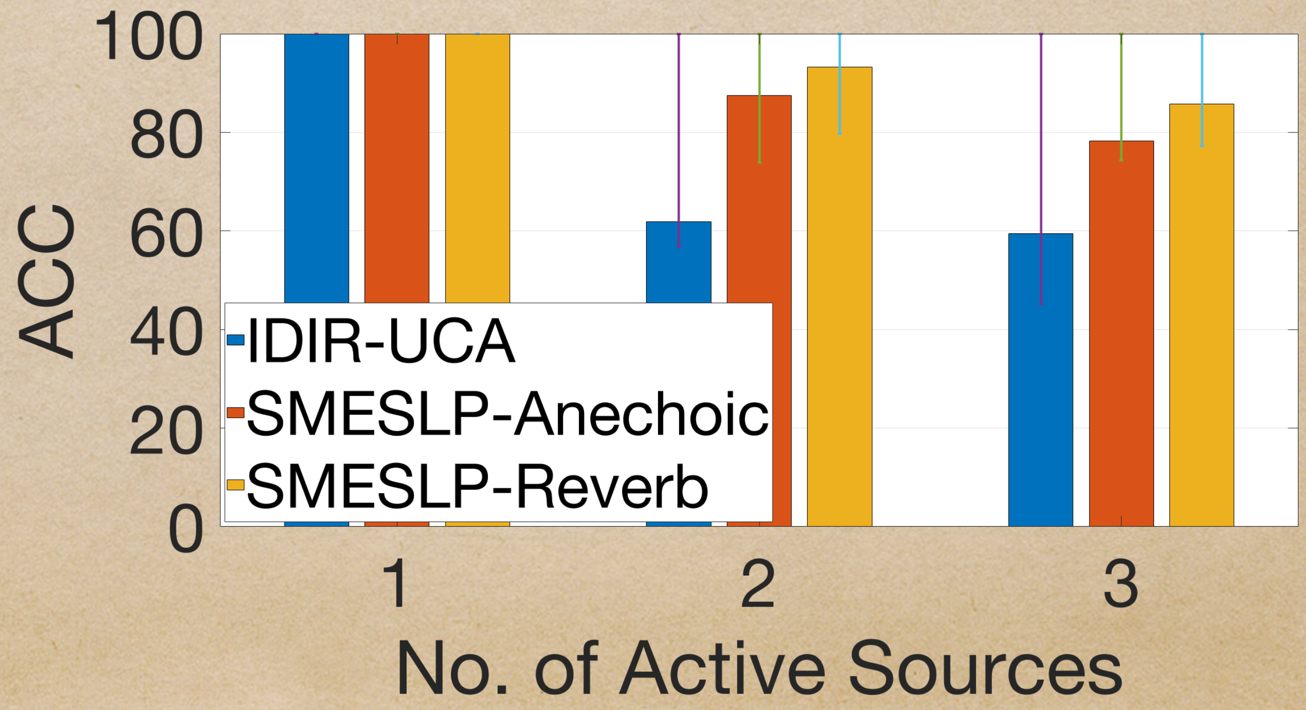
Anechoic

Test Set

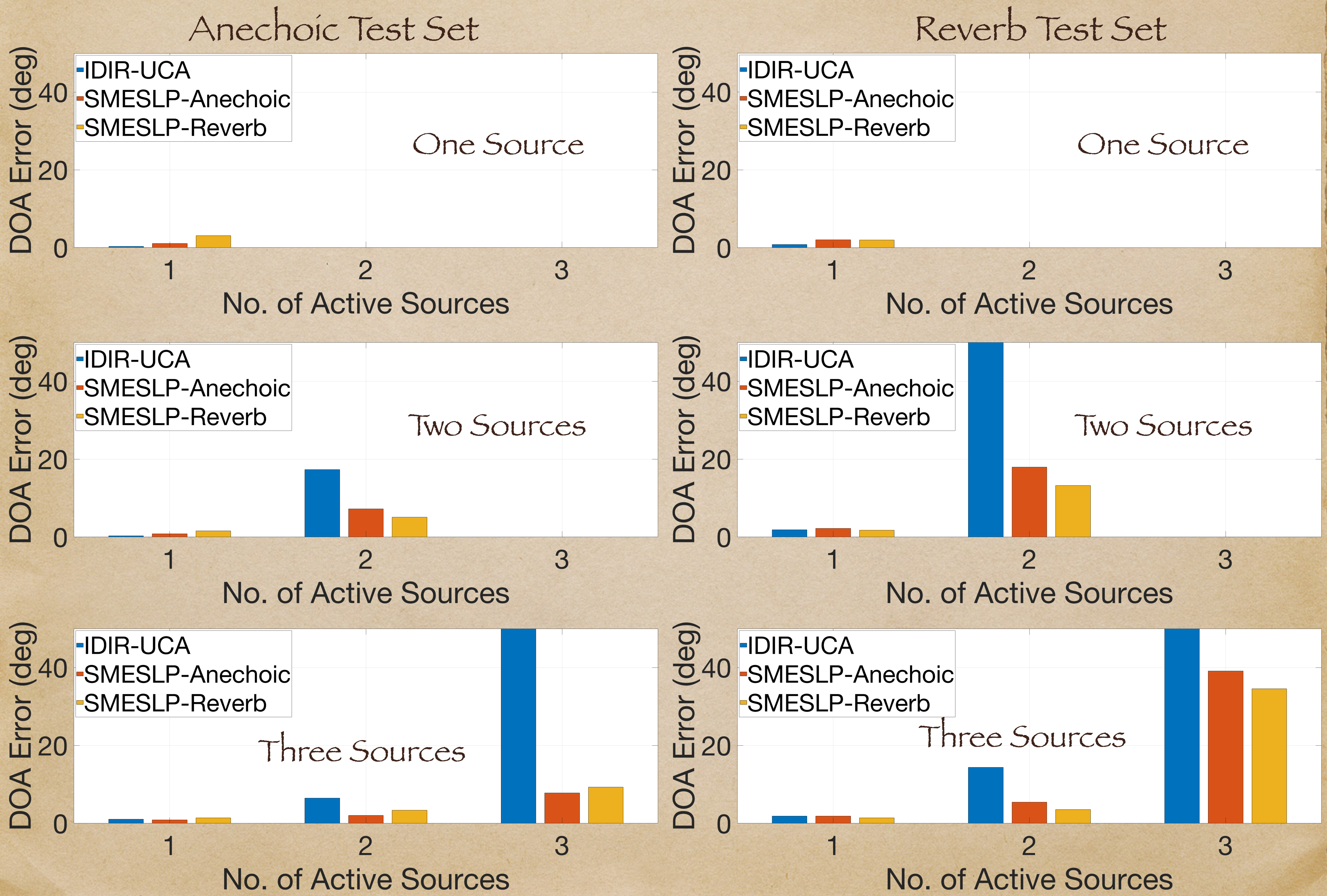


Reverb

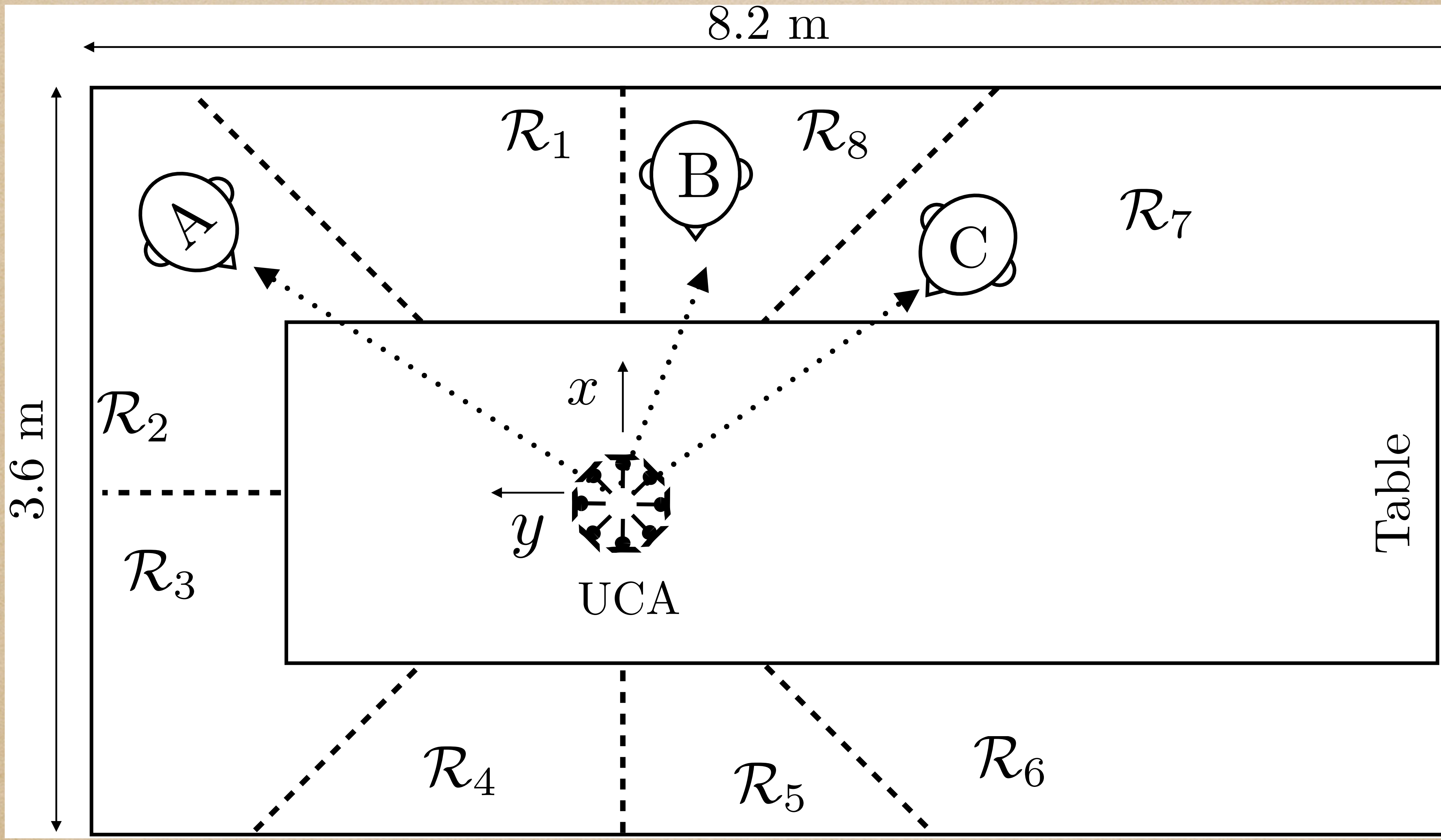
Test Set



Fine Localization Performance



Real Dataset: AV16.3 Corpus



Ref: G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in Machine Learning for Multimodal Interaction, S. Bengio and H. Bourlard, Eds., Berlin, Heidelberg, 2005, pp. 182–195, Springer.

Performance on Real Data

Fine Tuning: 110 Samples each Real Data +
100 samples each of Anechoic and Reverb Data

	Sp. B	Sp. B, C	Sp. A, B, C
	Absolute DOA Error		
SMESLP	**1.13°(100%)**	**1.96°(97.95%)**	**2.05°(100%)**
	RMSE DOA Error		
SMESLP	**1.45°(100%)**	**2.33°(97.95%)**	**2.33°(100%)**
I-IDIR-UCA [1]	1.00° (92%)	1.83° (79 %)	4.1° (60%)
CHB [2]	1.18 °	2.00 °	2.98 °

Table 1: DOA Error and Percentage of non-anomalous frames (indicated within parentheses) in real recordings for the three approaches being compared.

[1] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, “TDOAbased multiple acoustic source localization without association ambiguity,” IEEE/ACM Trans. on Audio, Speech, and Language Process., vol. 26, no. 11, pp. 1976–1990, Nov. 2018.

[2] A. M. Torres, M. Cobos, B. Pueo, and J. J. Lopez, “Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays,” J. Acoust. Soc. Amer., vol. 132, no. 3, pp. 1511–1520, 2012.

Outlook

- First End-to-End Deep Network for Localizing Multiple Sources from Raw Audio.
    - Easily Deployable with Existing DL Frameworks;
    - Easier for Model maintenance and updates.
- A novel Output Encoding Scheme based on Coarse-Fine Localization Strategy allowed for circumventing the Permutation Problem.
- Limitation: In case of multiple source in the same region (violation of assumption)
    - Active regions are still correctly detected.

Thank You!

Please Reach me at:

sundarhs@amazon.com

harshas123@gmail.com