Sound source distance estimation in diverse and dynamic acoustic conditions

Saksham Singh Kushwaha[1,2], Iran R. Roman[2], Magdalena Fuentes[2,3], Juan Pablo Bello[2]

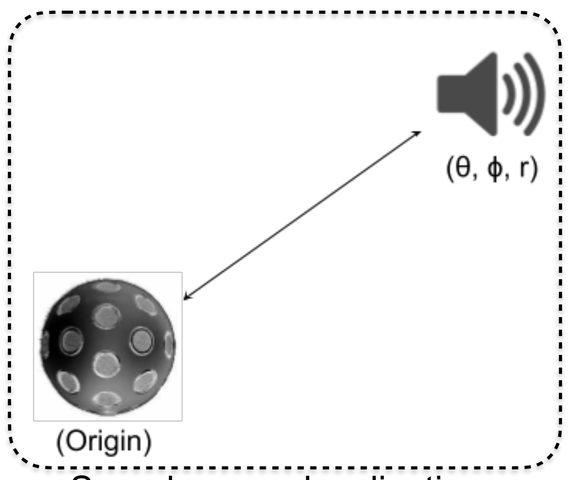
[1]Courant Institute of Mathematical Sciences, New York University, NY, USA

[2]Music and Audio Research Lab, New York University, NY, USA

[3]Integrated Design and Media, New York University, NY, USA

INTRODUCTION

- **Sound Source Localization**(SSL) consists of:
  - Direction of arrival (DOA) estimation
  - Distance estimation
- Sound distance estimation remains **understudied**[1]
  - Difficult task: reverberation, reflections, noise
  - Lack of annotated data



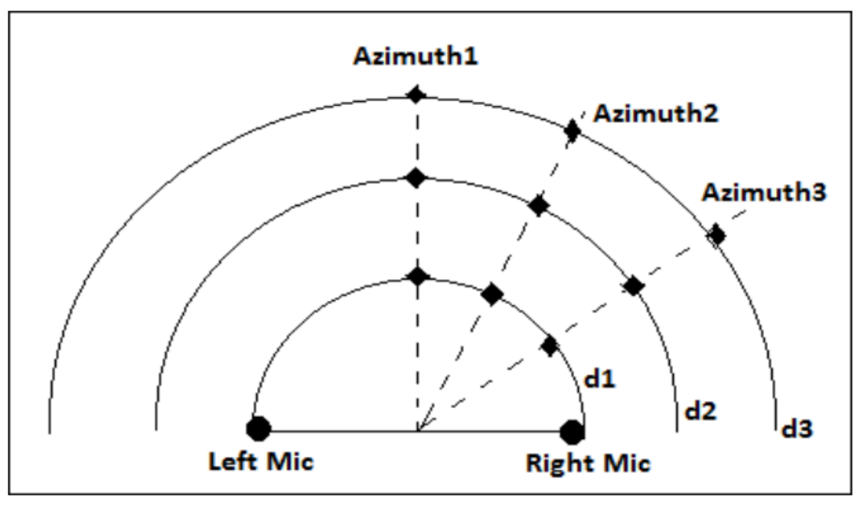
Sound source localization

Our **contributions**:

- Distance **annotations** for a collection of open-source DOA datasets
- **CRNN**-based model for non-simultaneous distance estimation
- Comprehensive testing on **diverse environments** and acoustic conditions
- Model performance analysis over different **loss functions**

Previous research in SSL:

- DOA research has been the primary focus
- Recent DL approaches in DOA use CRNN and open-source datasets[2,3]
- Previous attempts at distance estimation is not generalizable



Credits: Yiwere et. al

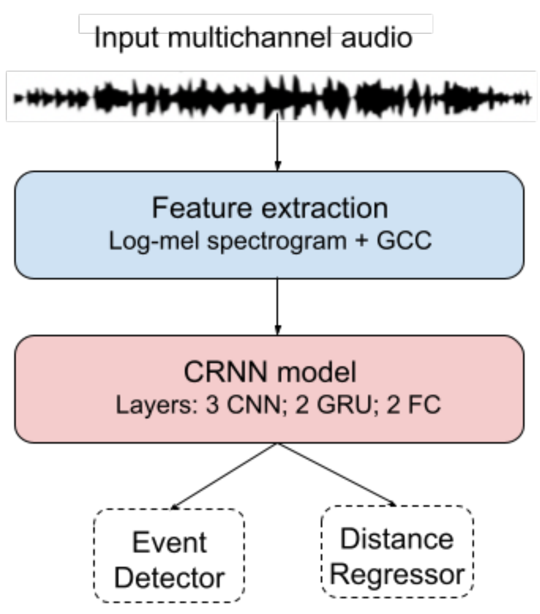
OUR APPROACH

- **Datasets:**
  - Multichannel Eigenmike recordings
  - Channel swapping[8] for small datasets (LOCATA, MARCo, METU)

Dataset	Range(m)	Avg dist(m)	#train	#test	Avg. dur(s)	#Room	Moving Sources?
DCASE	1.35-7.15	3.34	900	300	60.0	9	Y
STARSS	0.42-7.02	1.83	87	74	162.2	16	Y
LOCATA	0.50-3.49	1.78	27	5	18.9	1	Y
MARCo	2.6-12	4.01	5	7	78.6	1	N
METU	0.3-2.2	1.41	146	98	2.0	1	N

- **Model:**
  - Input: CRNN model
  - Multi-task loss

$$L_{total} = \frac{1}{N}\frac{1}{T}\sum_{n}\sum_{t}\overbrace{d_{n,t}\cdot \epsilon(y_{n,t}, \hat{y}_{n,t})}^{\text{Masked distance estimator}} + \overbrace{BCE(d_{n,t}, \hat{d}_{n,t})}^{\text{Event detector}}$$



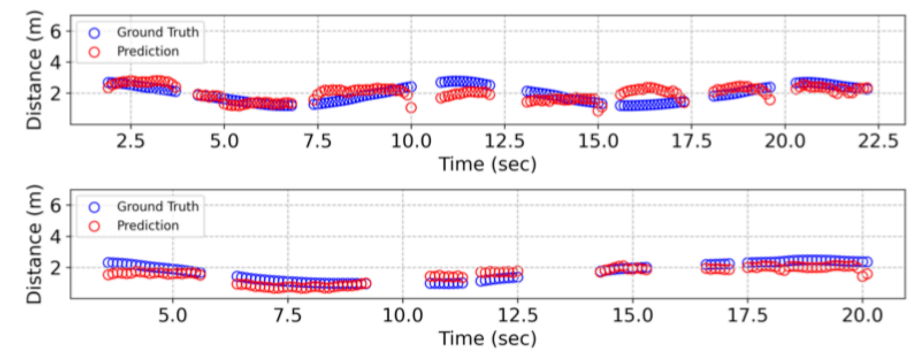
- **Training and Losses**
  - Two steps:
    - PSED: Pre-trained sound detector
    - Sound detection + distance estimation
  - Model Naming:
    - TWx : PSED + Train with data x
    - FWx-y: PSED + Pretrain with y + Finetune with x

Acronym	Full name	$\mathcal{E}$
AE	absolute error	$\lvert y-\hat{y}\rvert$
SE	squared error	$(y-\hat{y})^2$
APE	absolute percent error	$\frac{1}{y}\lvert y-\hat{y}\rvert$
SPE	squared percent error	$(\frac{1}{y}(y-\hat{y}))^2$
TAPE	thresholded APE	$\max(\delta, \frac{1}{y}\lvert y-\hat{y}\rvert)$

EXPERIMENTS

- Comparing our approach with baseline

Model	Exp	Best $\mathcal{E}$	Mean ↓	Median ↓	Std ↓
CRNN	TWL	SPE	0.413	0.330	0.347
	TWA	SE	0.368	0.340	**0.244**
	FWL-S	APE	**0.337**	0.290	0.246
	FWL-D	APE	0.352	**0.269**	0.275
avg pred			0.452	0.410	0.283
GTVV[5]			0.448	0.326	0.416

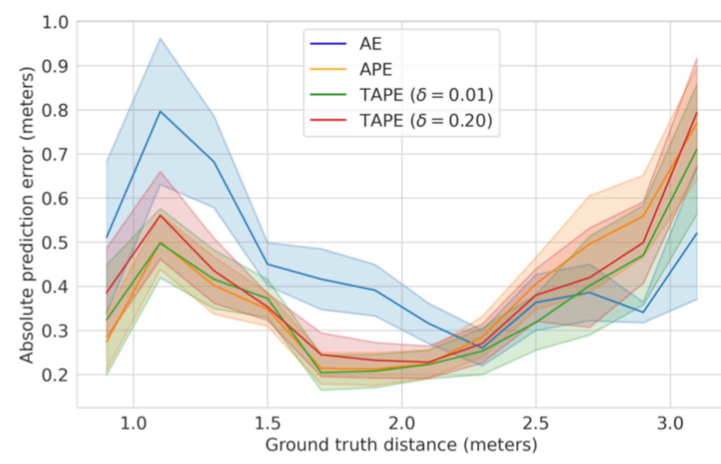


- Does pretraining help?

Model	Exp	Best $\mathcal{E}$	Mean ↓	Median ↓	Std ↓
CRNN	TW<u>D</u>	SE	1.032	0.903	0.838
CRNN	FW<u>D</u>-S	AE	**0.952**	**0.731**	0.834
avg pred			1.014	0.866	**0.596**
CRNN	TW<u>M</u>	SE	1.346	0.417	2.158
CRNN	FW<u>M</u>-S	SPE	**0.811**	**0.405**	0.508
avg pred			1.183	1.611	**0.494**
CRNN	TW<u>T</u>	APE	**0.148**	0.122	**0.126**
CRNN	FW<u>T</u>-S	TAPE*	0.167	**0.114**	0.150
avg pred			0.378	0.289	0.234

- Effect of loss functions

$\mathcal{E}$	Mean ↓	Median ↓	Std ↓
AE	0.438	0.360	0.342
SE	0.374	0.319	0.256
APE	0.337	0.290	0.246
SPE	0.334	0.292	0.259
TAPE ($\delta = 0.01$)	0.322	0.248	0.261
TAPE ($\delta = 0.10$)	0.361	0.312	0.250
TAPE ($\delta = 0.20$)	0.346	0.282	0.260



CONCLUSION

- Our approach performs better than baselines
- Qualitative results show model is adaptable to moving sound source
- Pretraining with larger dataset i.e. STARSS[8] helps
- Percentage based error gives the optimal results

- **Future Work:**
  - Extend to multiple simultaneous events
  - Joint model for sound event DOA + distance + classification

REFERENCES

[1] Grumiaux et. al, “A survey of sound source localization with deep learning methods”

[2] Adavanne et. al, “A multi-room reverberant dataset for SELD”

[3] Politis et. al, “STARSS23: SonyTAu Realistic Spatial Soundscapes 2023”

[4] Daniel et. al, “Echo-enabled direction-of-arrival and range estimation of a mobile source in ambisonic domain”

[5] Yiwere et. al, “Distance estimation and localization of sound sources in reverberant conditions using deep neural networks”

[6] Takeda et. al, “Sound source localization based on deep neural networks with directional activate function exploiting phase information”

[7] Wang et. al, “A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection”

[8] Politis et al, “STARSS23: Sony-TAu Realistic Spatial Soundscapes 2023”