SOUND SOURCE DISTANCE ESTIMATION IN DIVERSE AND DYNAMIC ACOUSTIC CONDITIONS

Saksham S. Kushwaha[1,2], Iran R. Roman[2], Magdalena Fuentes[2,3], Juan P. 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

IEEE WASPAA 2023

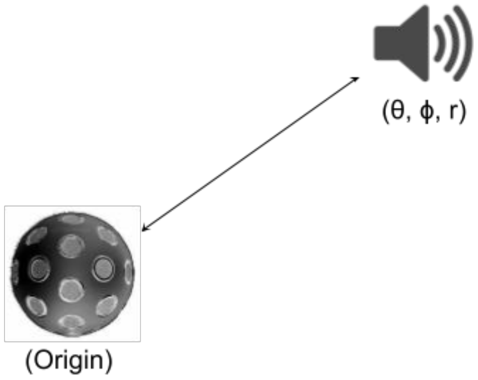
NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU | TANDON SCHOOL OF ENGINEERING

Integrated Digital Media

MARL

Introduction



- Sound Source Localization
  - Direction of arrival (DOA) estimation (θ,ϕ)
  - Source distance (r)
- Several applications. For eg.
  - Audio-based navigation
  - Sound source separation
- Recent research development has focused on DOA estimation
- Sound distance estimation remains understudied
  - Difficult task: reverberation, reflections, noise etc.
  - Lack of annotated data
- Previous attempts are not generalizable:
  - Often reformulated as classification problem / range estimation
  - Existing methods test on synthetic data OR handful of real data

Contributions

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

Related Work

- DOA estimation:
  - Primarily focusing on direction rather than distance [Shimada et.al]
  - Many studies on DOA estimation have employed CRNNs [Adavanne et.al]
  - Open source real data and Room impulse responses (RIR) based data generators [Politis et.al]

- Distance estimation:
  - Given the sound onset(t_0) and arrival times(t_a)
    - Distance, $d = c \times (t_r - t_o)$ where c: speed of sound
  - Previous approaches:
    - Human listening inspired approaches. Eg. direct-to-reverberant ratio [Zahorik et.al]
    - Data driven approaches. Eg. FNN, CNN [Yiwere et. al, Takeda et. al]
  - Most of work assumes non-coincident microphones
  - (Baseline) Generalized Time-domain Velocity Vector based method [Kitic et. al]

Zahorik et.al Direct-to-reverberant energy ratio sensitivity, Yiwere et al Distance estimation and localization of sound sources in reverberant conditions using deep neural networks, Takeda et al Sound source localization based on deep neural networks with directional activate function exploiting phase information, Kitic et al Generalized time domain velocity vector, Adavanne et. al: A multi-room reverberant dataset for SELD, Shimada et. al Multi-accdoa: Localizing and detecting overlapping sounds from the same class , Politis et al: STARSS23: SonyTAu Realistic Spatial Soundscapes 2023

Our study: Datasets

- **DCASE**: Synthetic data + Room impulse responses(RIRs)
- **STARSS**: Real-world recordings + masked overlapping sounds
- **LOCATA**: Single room real speech recordings
- **3D-MARCo**: Musical recordings + reverberant church
- **METU-SPARG**: RIRs + 3D grid + office setup

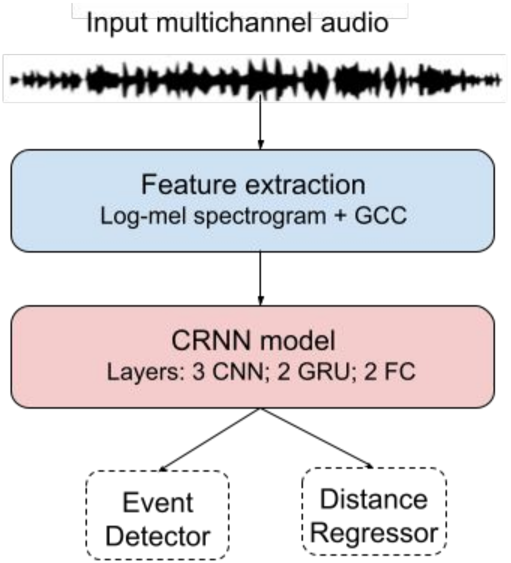
Dataset	Range(m)	Avg dist(m)	#train	#test	Avg. dur(s)	#Room	Moving Sources?
DCASE	1.35-7.15	3.34	900	300	60.0	9	Y
STARSS	0.42-7.02	1.83	87	74	162.2	16	Y
LOCATA	0.50-3.49	1.78	27	5	18.9	1	Y
MARCo	2.6-12	4.01	5	7	78.6	1	N
METU	0.3-2.2	1.41	146	98	2.0	1	N

Model

- Aim: Non-simultaneous + moving sources
- Adapt the CRNN model from Adavane et al.
- Distance output is ignored if sound is not present

$$L = \underbrace{\frac{1}{N}\frac{1}{T}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1} d_{n,t}\mathcal{E}(y_{n,t}, \hat{y}_{n,t})}_{\text{Masked distance estimator}} + \underbrace{\mathrm{BCE}(d_{n,t}, \hat{d}_{n,t})}_{\text{Event detector}}$$

- Here, ε can be different regression losses.



Adavane et al. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks.

Training procedure and loss

- Two steps:
  - Train the model only for sound detection over DCASE (PSED)
  - Multi-task training: Sound detection + distance estimation
- Model name:
  - TWx : PSED + Train with data x
  - FWx-y: PSED + Pretrain with y + Finetune with x
- Loss functions:
  - Percentage losses:
    - Uniform error weighing with distance

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

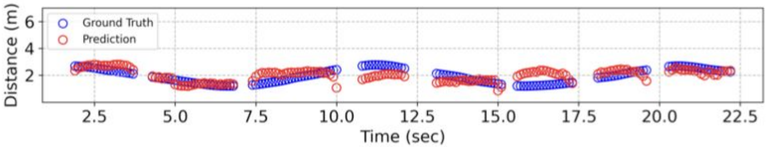
- Exp1: comparing with a baseline with
- Exp2: Effect of model pretraining over a larger dataset
- Exp3: Effect of different regression losses.

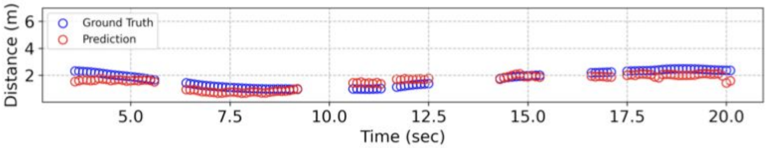
Exp1: Comparing with recent baseline

- LOCATA dataset
- Baselines:
    - Avg prediction
    - [20] Daniel et al.
- Model name:
    - TWx : PSED + Train with data x
    - FWx-y: PSED + Pretrain with y + Finetune with x
- Pretrained models perform better
- Percent error loss perform better
- Qualitative results

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
[20]			0.448	0.326	0.416

D=DCASE. A=All data,.T=METU-SPARG, L=LOCATA, S=STARSS





Daniel et al. Echo-enabled Direction-of-Arrival and range estimation of a mobile source in Ambisonic domain

Exp2: Effect of model pre-training

- Model name:
  - TWx : PSED + Train with data x
  - FWx-y: PSED + Pretrain with y + Finetune with x
- Pretrained models perform better
- Percent error loss perform better

Model	Exp	Best $\mathcal{E}$	Mean ↓	Median ↓	Std ↓
CRNN	TWD	SE	1.032	0.903	0.838
CRNN	FWD-S	AE	**0.952**	**0.731**	0.834
avg pred			1.014	0.866	**0.596**
CRNN	TWM	SE	1.346	0.417	2.158
CRNN	FWM-S	SPE	**0.811**	**0.405**	0.508
avg pred			1.183	1.611	**0.494**
CRNN	TWT	APE	**0.148**	0.122	**0.126**
CRNN	FWT-S	TAPE*	0.167	**0.114**	0.150
avg pred			0.378	0.289	0.234

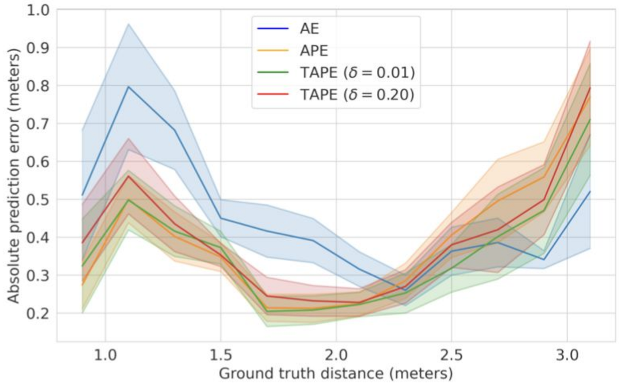
D=DCASE. M=MARCo. T=METU-SPARG

Exp3: Effect of loss functions

- Performance comparison on LOCATA
- Thresholded percent error performs best

- Error vs. ground truth distance
  - Percentage error performs better
  - Due to uniformly reduces error w.r.t distance

$\mathcal{E}$	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 and Future Work

- Extended DOA datasets and model for distance estimation
- Proposed a generalizable training schema for distance prediction
- Improved distance estimation using percent error-based loss and pretraining
- In the future:
  - Extend our approach to multiple simultaneous occurring sounds
  - Joint model for DOA + classification + distance