Acuity: Creating Realistic Digital Twins

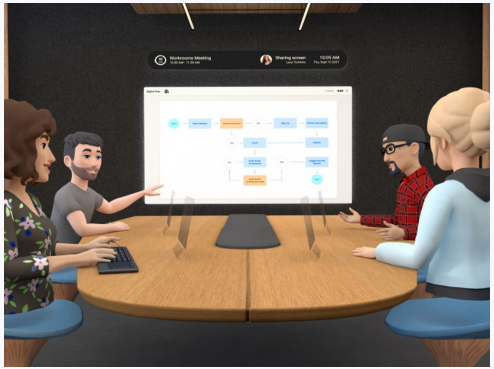
Through Multi-resolution Pointcloud Processing and Audiovisual Sensor Fusion

Jason Wu, Ziqi Wang, Ankur Sarker, Mani Srivastava (UCLA)

01

Introduction

How to represent human subjects audiovisually at high fidelity in real time?



Visual Domain Goals

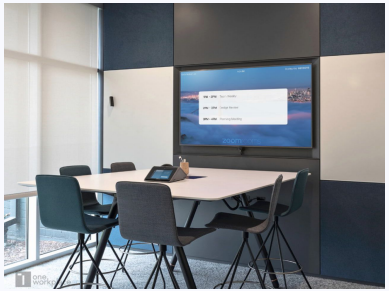
Point Clouds

Avatars

Visual Domain Goals

Obtain High Resolution Point Clouds

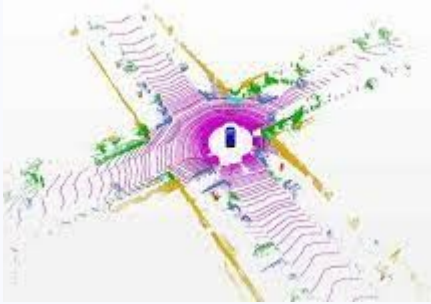
Isolate point clouds of human subjects



Place within virtual scene

Visual: Current Challenges

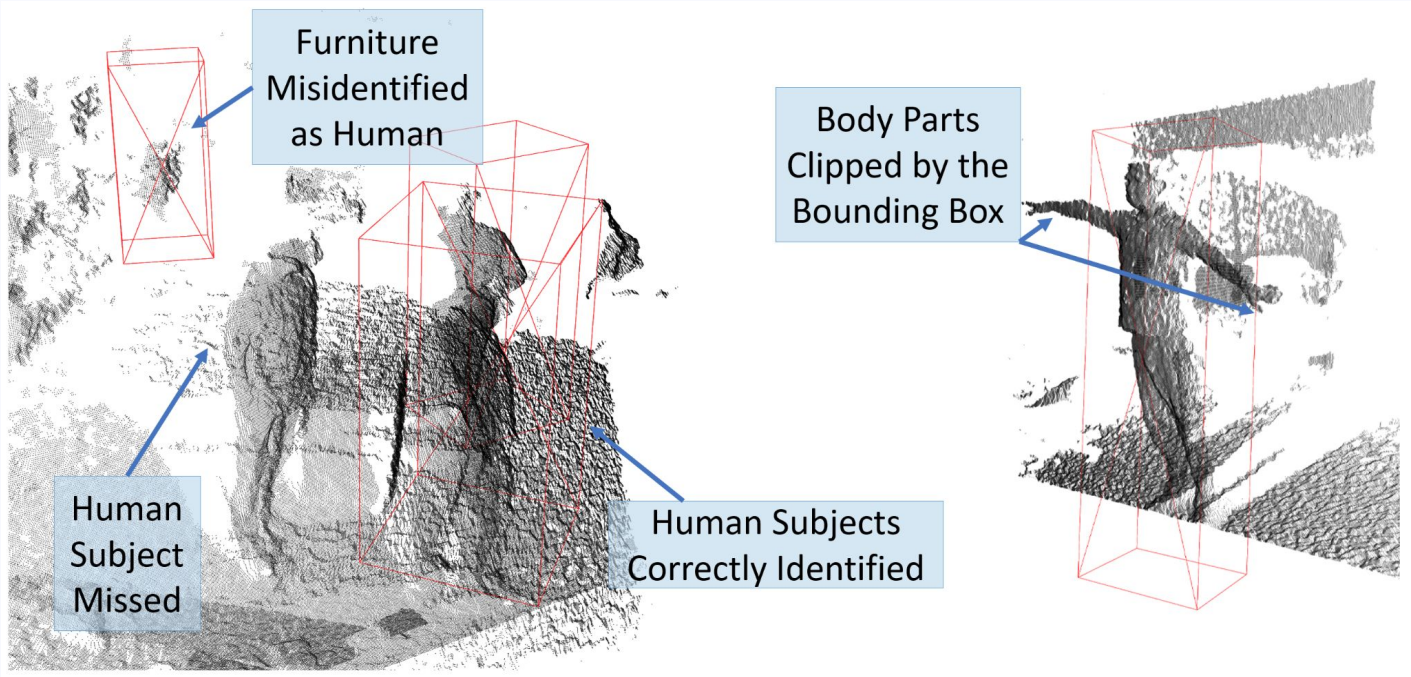
- Existing neural networks trained on sparse point clouds input from KITTI Dataset
- Suffer from runtime issues
- Generalize poorly to dense point cloud input



KITTI Point Cloud (Sparse)

Realsense Point Cloud (Dense)

Visual: Current Challenges



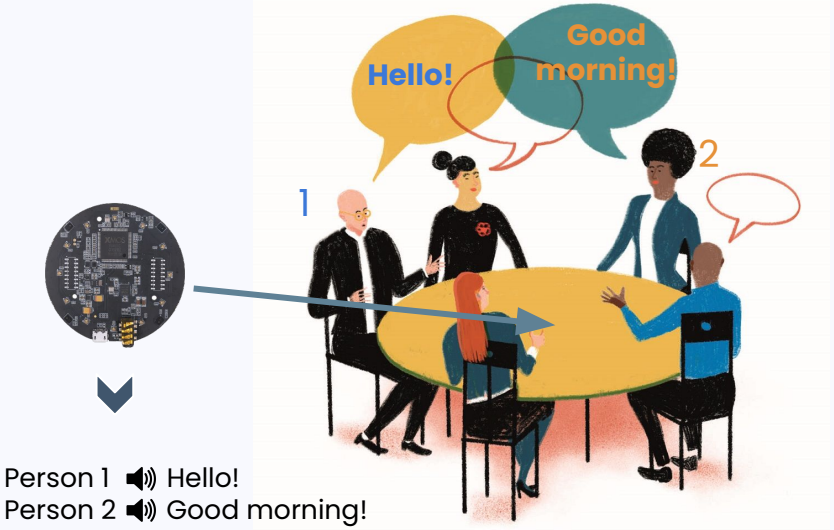
Example: Failure Cases of NN-based Human Subject Bounding Box Detection

Challenge

How do we isolated human subjects from dense point clouds in real time?

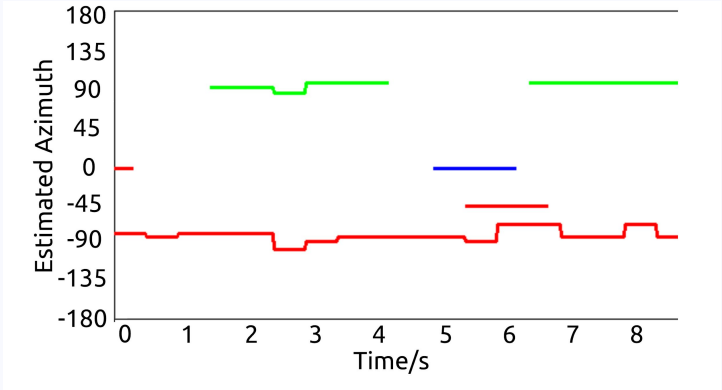
Audio Domain Goals

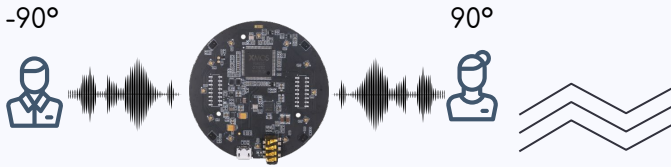
- Audio quality degrades with high background noise or background speech
- **Sound Source Separation:** Leverage beamforming with microphone array to isolate sound from particular direction
- Requires the angle of arrival (AoA) to be known



Audio: Current Challenges

- AoA Estimation Methods: SRP-PHAT and MUSIC
- Learning based methods: We require real-time, mobile subjects, and variable number of subjects

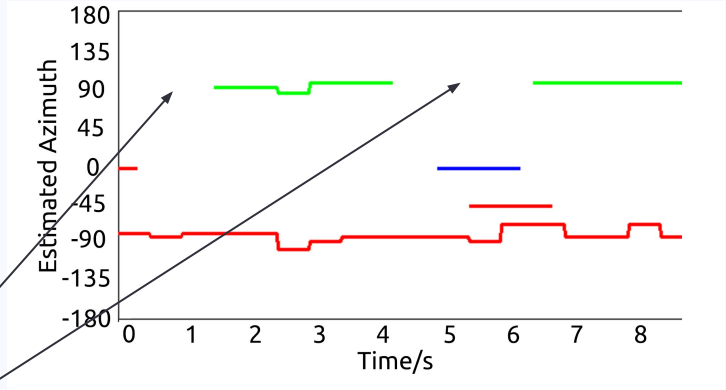




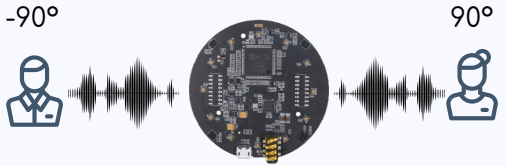
MUSIC Localization: 2 Subjects

Audio: Current Challenges

- AoA Estimation Methods: SRP-PHAT and MUSIC
- Learning based methods: We require real-time, mobile subjects, and variable number of subjects



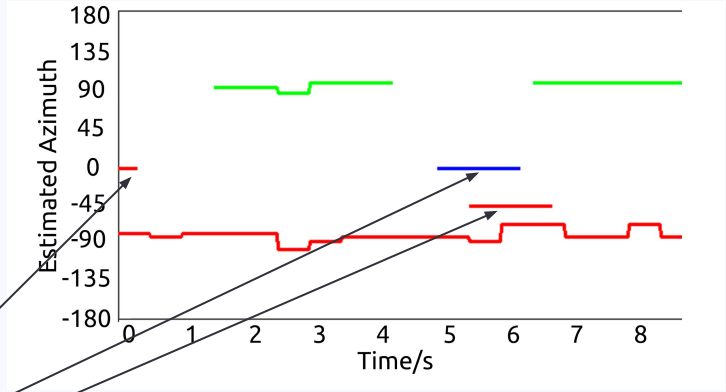
Missing Subject



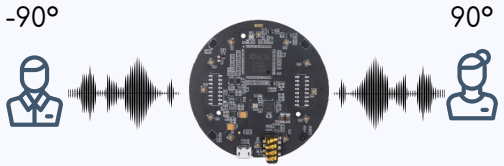
MUSIC Localization: 2 Subjects

Audio: Current Challenges

- AoA Estimation Methods: SRP-PHAT and MUSIC
- Learning based methods: We require real-time, mobile subjects, and variable number of subjects



“Ghost” Subjects



MUSIC Localization: 2 Subjects

Challenge

How do we obtain the AoA in real time?

Acuity

Double Background Subtraction Pipeline

- Leverage **background subtraction** at two resolutions to efficiently isolate point clouds

- Runs at 30 fps with resolution of 640x480 and three subjects in the scene

Multimodal Fusion With Visual Localization

- Leverages localization information from the visual pipeline to obtain AoA (**audiovisual fusion**)

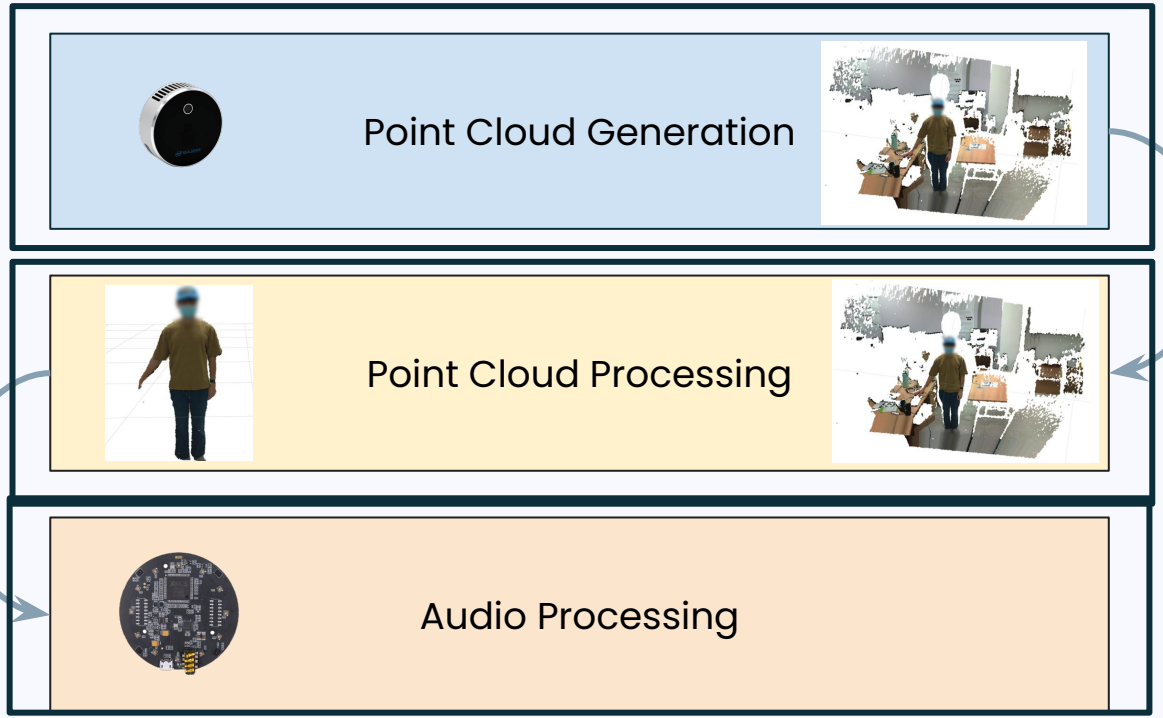
- Uses the centroid of the isolated point cloud

- Latency of 30 ms (< 45 ms)

02

System Design And Implementation

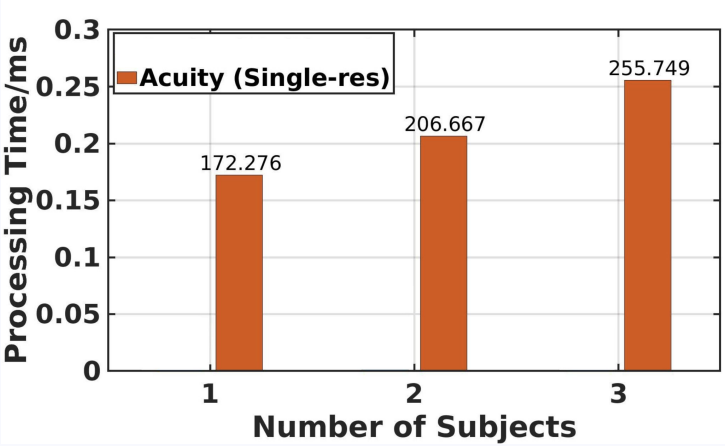
Acuity Pipeline



(c_x, c_y, c_z)

Point Cloud Processing

- **Background subtraction:** Compare each new frame with reference frame without subjects. Removes all the *voxels* (points) that are identical
- Apply clustering to remove noise



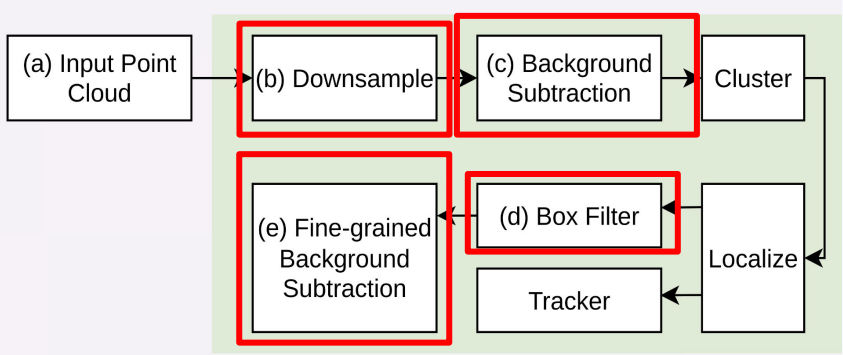
Challenge

Background Subtraction and Clustering do not run in real time

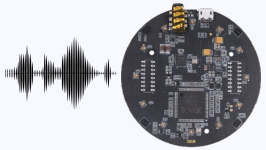
Solution

Utilize **multi-resolution** processing!

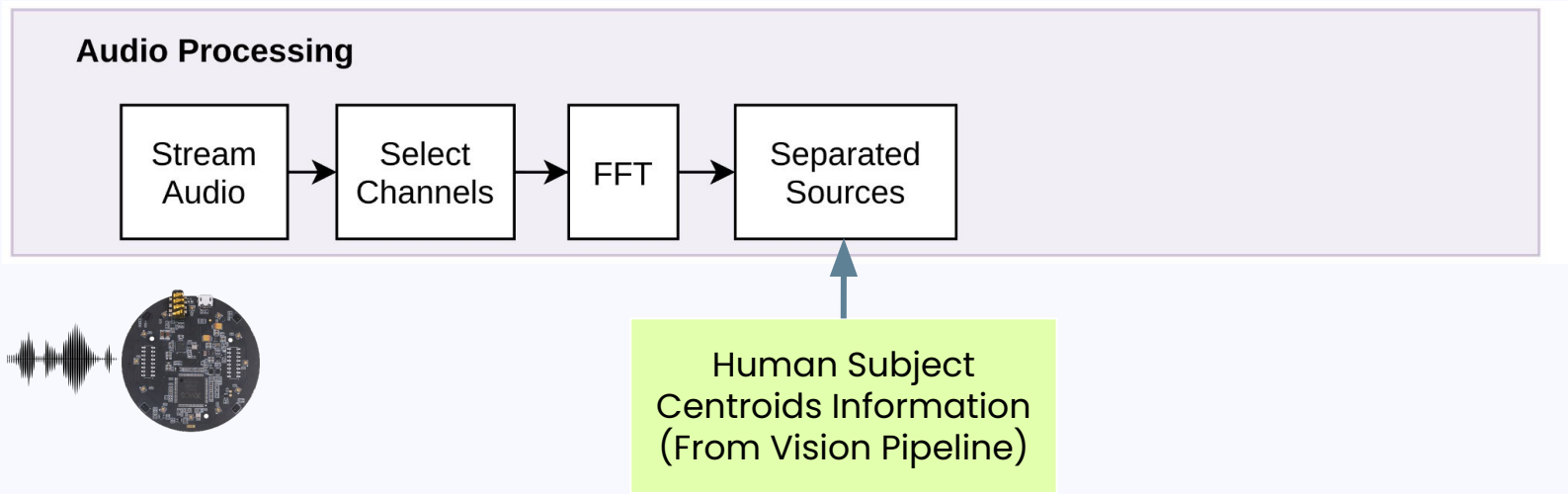
Multi Resolution Processing



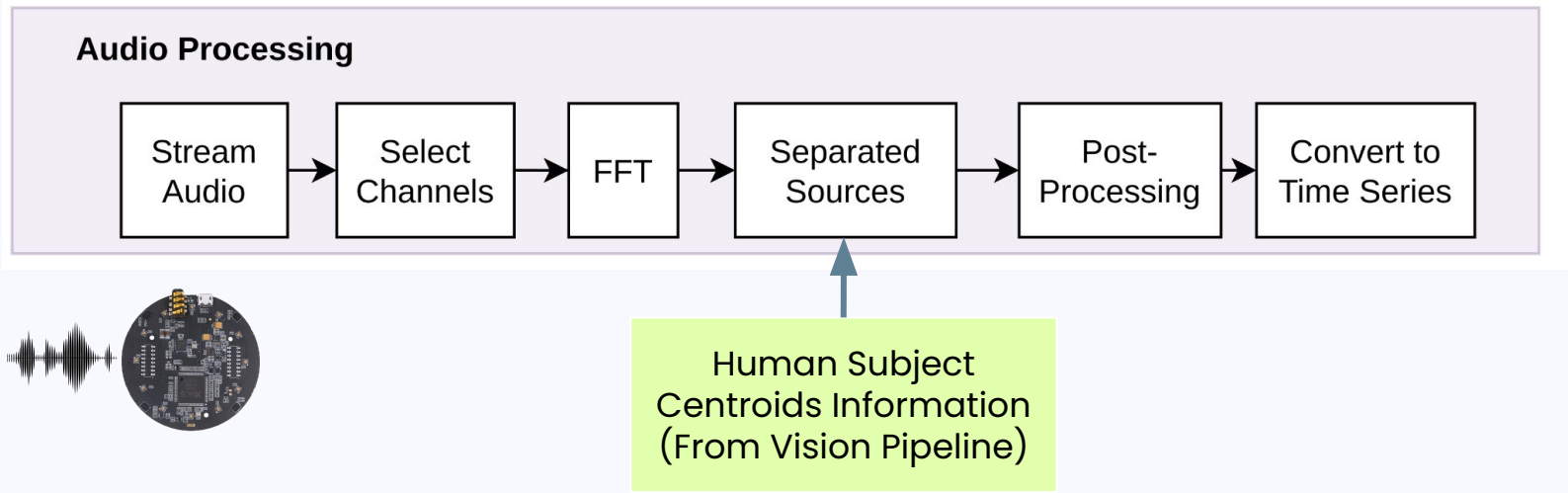
Audio Processing



Audio Processing



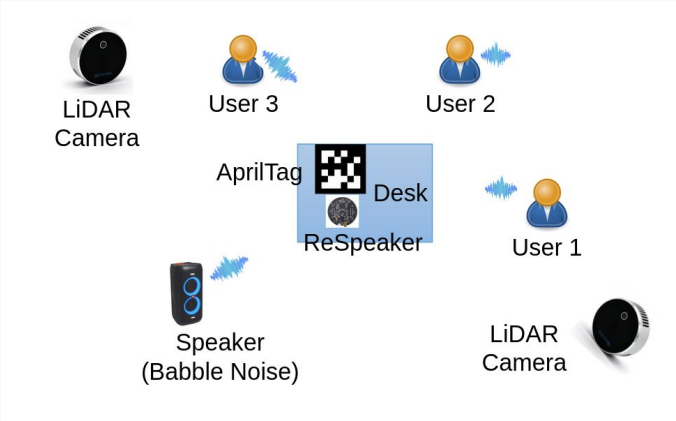
Audio Processing



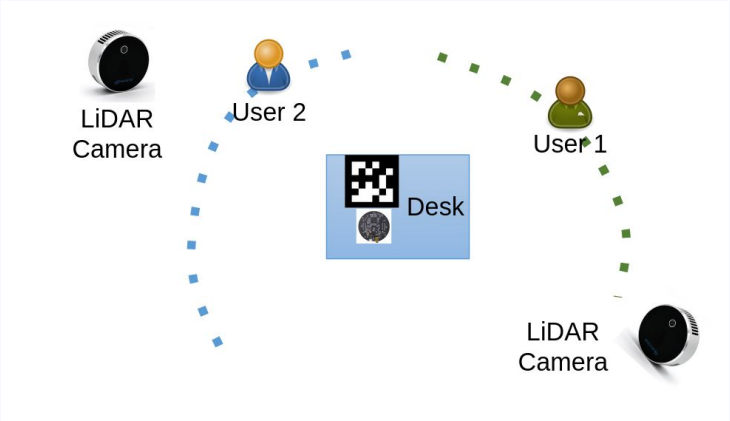
03

Experiments And Results

Experimental Setup

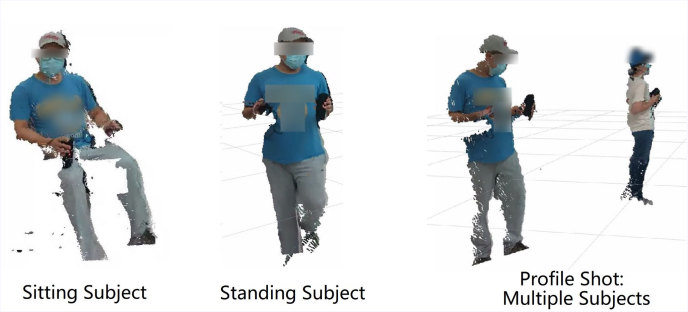


Experiment 1: Three Stationary Subjects



Experiment 2: Two Mobile Subjects

Visual Domain Results



Visual Point Cloud Output

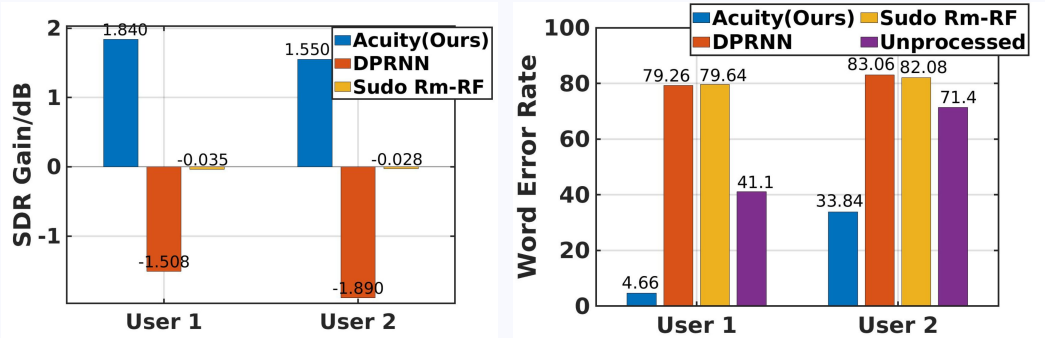
	Acuity (Ours)			PV-RCNN++			SECOND		
Number of Subjects	Latency (ms)	Accuracy	Average F1 Score	Latency (ms)	Accuracy	Average F1 Score	Latency (ms)	Accuracy	Average F1 Score
1	34	100%	1.00	1584	53.30%	0.33	33	88.50%	0.6
2	54	100%	1.00	1635	60%	0.26	32	90.80%	0.78
3	30	100%	1.00	690	76%	0.38	31	94.40%	0.83
4	44	97.40%	0.99	691	68.80%	0.33	31	71.60%	0.63
5	61	93.30%	0.96	699	75%	0.37	31	92.80%	0.77
6	69	91.70%	0.95	693	70.80%	0.41	31	63.10%	0.61

Comparison to NN

Audio Domain Results

		SDR Gain (db)			Word Error Rate (%)			
		Acuity	DPRNN	SuDoRMRF	Raw	**Acuity**	DPRNN	SuDoRMRF
Single Source	Source 1	**1.92**	-1.803	-0.249	6.02	**1.95**	60.43	47.075
Two Sources	Source 1	**3.194**	-1.094	0.296	79.02	**5.08**	95.5	95.95
	Source 2	**10.822**	1.107	0.104	87.82	**5.34**	78.13	75.4
Three Sources	Source 1	**5.309**	0.2899	-	100	**15.49**	100	-
	Source 2	**7.025**	-0.192	-	100	**3.13**	92.1	-
	Source 3	**4.835**	-0.145	-	100	**28.425**	95.5	-

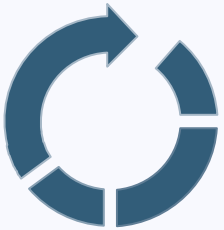
Experiment 1: Static Subject Audio



Experiment 2: Mobile Subject Audio

04

Conclusion



Limitations and Future Work

- **Real-time Point Cloud Streaming and Rendering:** Acuity does not address issues of streaming point clouds to the end user or rendering point clouds for viewing
- **Scaling up Acuity:** Acuity currently utilizes a two camera + one microphone setup, and may benefit from the introduction of additional sensors
- **Environmental Conditions:** The LiDAR camera performs poorly in low light situations, and saturates in the presence of direct sunlight

Acknowledgements

JUMP
Joint University Microelectronics Program