Acoustic Source Localisation in Constrained Environments

Elizabeth Vargas

OVERVIEW

What is Acoustic Source Localisation?

Answer **where** is sound coming from

What is Acoustic Source Localisation?

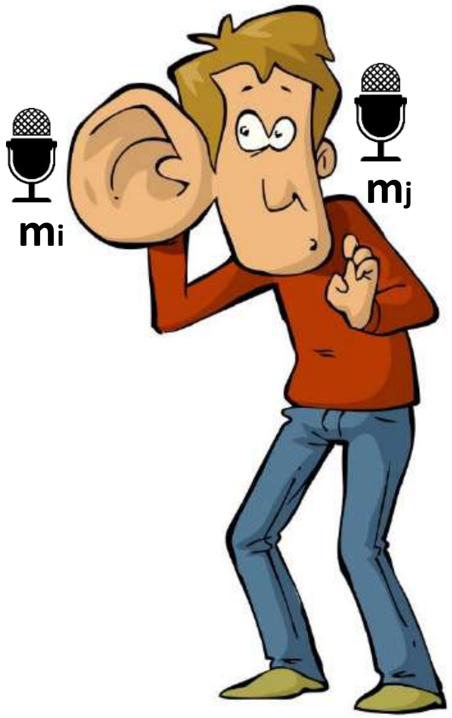
Answer **where** is sound coming from

HUMANS

What is Acoustic Source Localisation?

Answer **where** is sound coming from

HUMANS

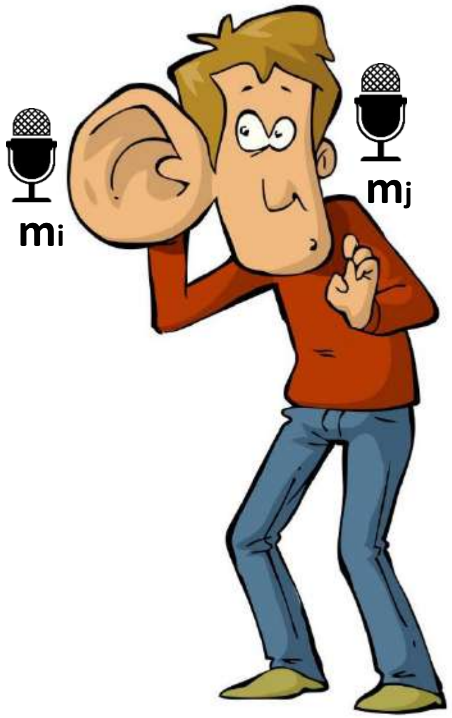


We use audio cues, such as time and intensity differences between both ears.

What is Acoustic Source Localisation?

Answer **where** is sound coming from

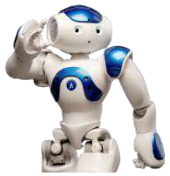
HUMANS



We use audio cues, such as time and intensity differences between both ears.

ARTIFICIAL

INPUT



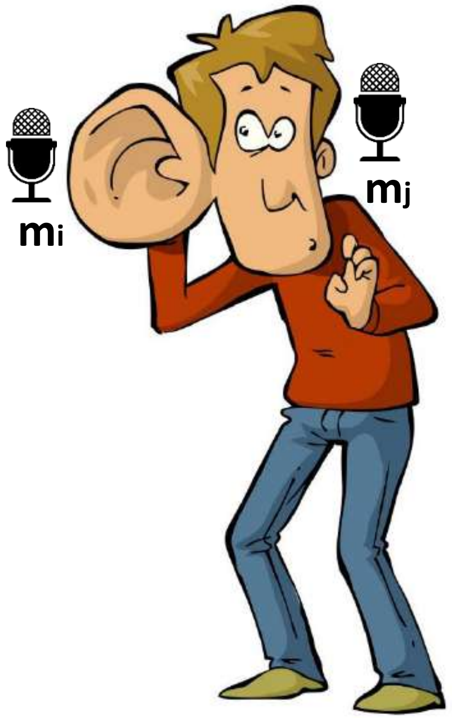
Binaural

Microphone Array

What is Acoustic Source Localisation?

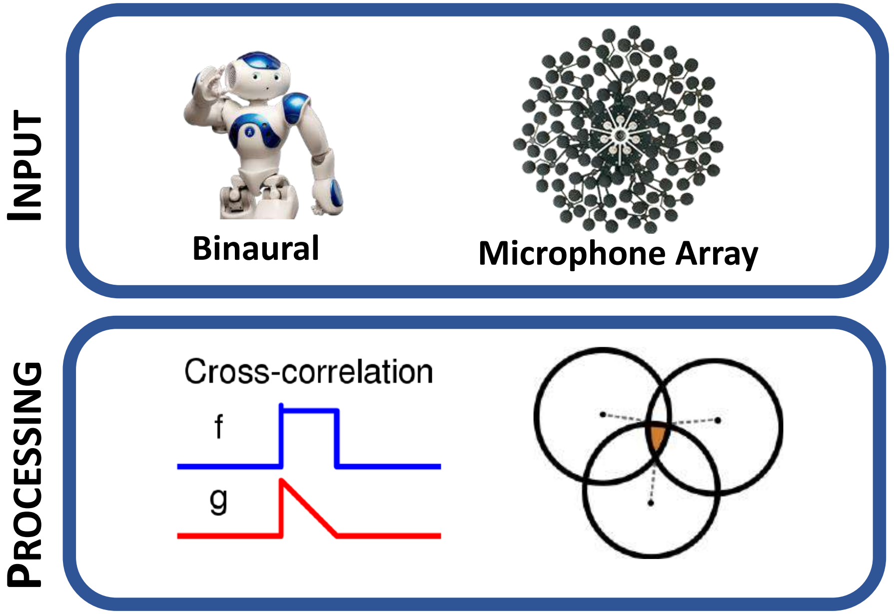
Answer **where** is sound coming from

HUMANS



We use audio cues, such as time and intensity differences between both ears.

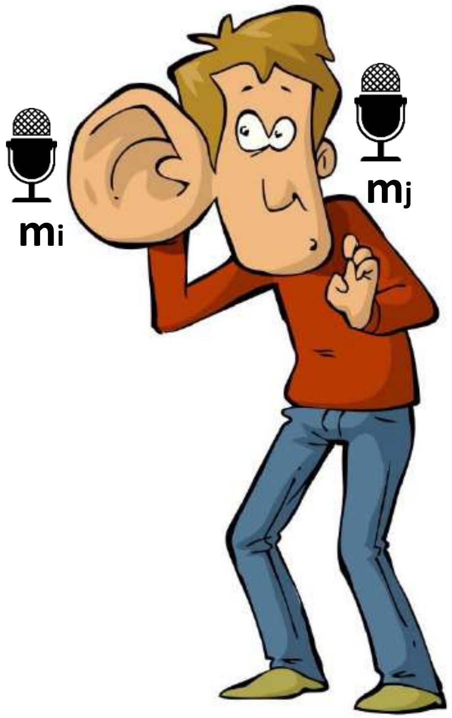
ARTIFICIAL



What is Acoustic Source Localisation?

Answer **where** is sound coming from

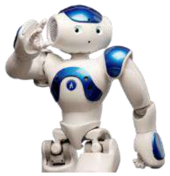
HUMANS



We use audio cues, such as time and intensity differences between both ears.

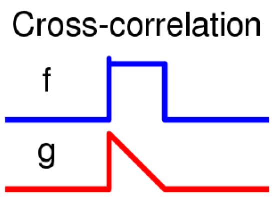
ARTIFICIAL

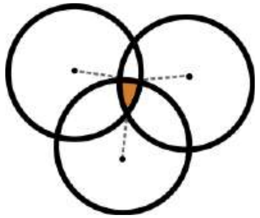
INPUT


Binaural


Microphone Array

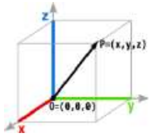
PROCESSING

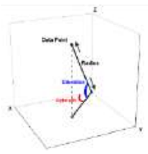




OUTPUT

Angle: Direction of Arrival (DOA)

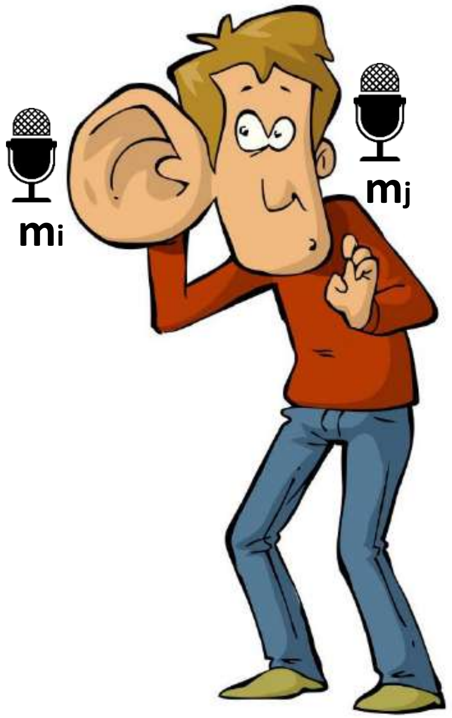

3D Coordinate: Exact location


Azimuth, Elevation and Range

What is Acoustic Source Localisation?

Answer **where** is sound coming from

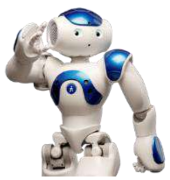
HUMANS



We use audio cues, such as time and intensity differences between both ears.

ARTIFICIAL

INPUT

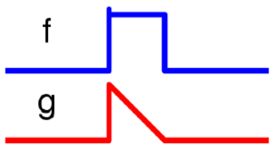


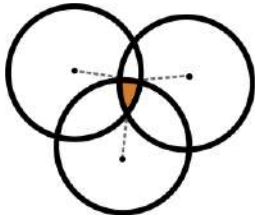
Binaural

Microphone Array

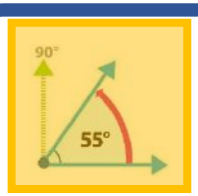
PROCESSING

Cross-correlation

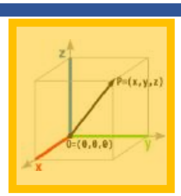




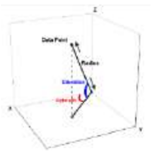
OUTPUT



Angle: Direction of Arrival (DOA)



3D Coordinate: Exact location



Azimuth, Elevation and Range

Application Scenarios

Robot waitress

Rescue drones

Speech

Constrained Environments

Atypical scenarios, in which the conditions to solve ASL are restricted, and therefore conventional methods do not work.

Constrained Environments

Atypical scenarios, in which the conditions to solve ASL are restricted, and therefore conventional methods do not work.

Number and configuration of microphones (CHAPTER 4)

- Limited number of microphones available
- TDOA available for some microphone pairs.
- Microphone arrangement

Constrained Environments

Atypical scenarios, in which the conditions to solve ASL are restricted, and therefore conventional methods do not work.

Number and configuration of microphones (CHAPTER 4)

- Limited number of microphones available
- TDOA available for some microphone pairs.
- Microphone arrangement

Signal samples (CHAPTER 5)

- The use of the full length of an acoustic signal is unavailable
- Limited communication
  - Low bandwidth

Constrained Environments

Atypical scenarios, in which the conditions to solve ASL are restricted, and therefore conventional methods do not work.

Number and configuration of microphones (CHAPTER 4)

- Limited number of microphones available
- TDOA available for some microphone pairs.
- Microphone arrangement

Signal samples (CHAPTER 5)

- The use of the full length of an acoustic signal is unavailable
- Limited communication
  - Low bandwidth

Data available for training (CHAPTER 6)

- Machine/deep learning approaches
- Insufficient training data
- Test data differs from training data

Hypothesis

It is possible to accurately localise sound sources, even in constrained scenarios.

Publications

1. **E. Vargas**, K. Brown, K. Subr, "Impact of Microphone Array Configurations on Robust Indirect 3D Acoustic Source Localization", in *International Conference on Acoustics, Speech and Signal Processing* **(ICASSP)**, Calgary, Canada, April 2018. ***(Oral Presentation)***

2. **E. Vargas**, J. R. Hopgood, K. Brown, K. Subr, "A Compressed Encoding Scheme for Approximate TDOA Estimation", in *European Signal Processing Conference*, **(EUSIPCO)**, Rome, Italy, September 2018. ***(Oral Presentation)***

Contributions

This thesis presented work on **Acoustic Source Localisation (ASL) in constrained environments**. The three constraints studied were the number and configuration of sensors; the signal samples; and training data.

Number and configuration of microphones (CHAPTER 4)

- Limited number of microphones available
- TDOA from some microphone pairs
- Microphone arrangement

Accuracy can be maintained at state-of-the-art levels (SRP) while **reducing computation six fold,** while **circular arrays** leads to **highest errors.**

Contributions

This thesis presented work on **Acoustic Source Localisation (ASL) in constrained environments**. The three constraints studied were the number and configuration of sensors; the signal samples; and training data.

Number and configuration of microphones (CHAPTER 4)

- Limited number of microphones available
- TDOA from some microphone pairs
- Microphone arrangement

Signal samples (CHAPTER 5)

- The use of the full length of an acoustic signal is unavailable
- Limited communication
- Low bandwidth

Accuracy can be maintained at state-of-the-art levels (SRP) while **reducing computation six fold,** while **circular arrays** leads to **highest errors.**

The algorithm presented in this work outperforms an audio fingerprinting baseline while maintaining a **compression ratio of 40:1**.

Contributions

This thesis presented work on **Acoustic Source Localisation (ASL) in constrained environments**. The three constraints studied were the number and configuration of sensors; the signal samples; and training data.

Number and configuration of microphones (Chapter 4)

- Limited number of microphones available
- TDOA from some microphone pairs
- Microphone arrangement

Signal samples (Chapter 5)

- The use of the full length of an acoustic signal is unavailable
- Limited communication
- Low bandwidth

Data available for training (Chapter 6)

- Machine/deep learning approaches
- Insufficient training data
- Test data differs from training data

Accuracy can be maintained at state-of-the-art levels (SRP) while **reducing computation six fold,** while **circular arrays** leads to **highest errors.**

The algorithm presented in this work outperforms an audio fingerprinting baseline while maintaining a **compression ratio of 40:1**.

Music training data is used to record an **improvement of 19%** against a noise training baseline **using only 25% of the training data.**

CONTRIBUTIONS

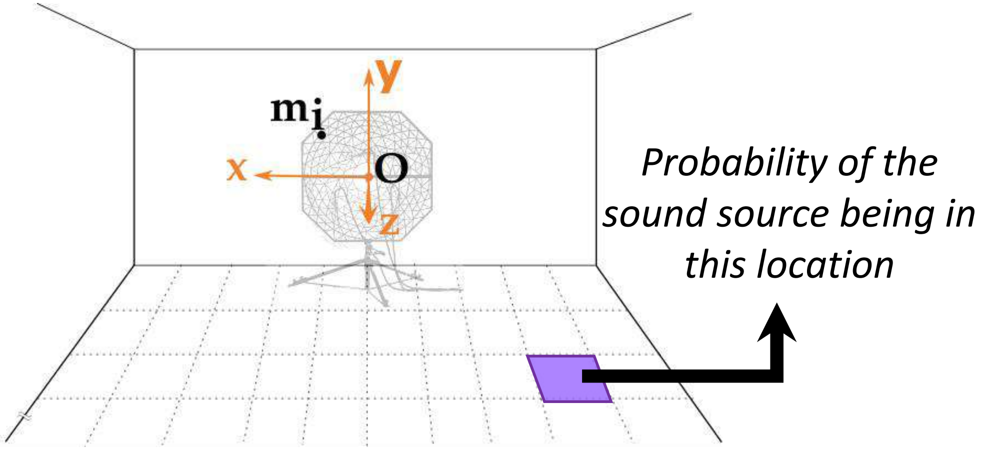
CONTRIBUTION I

Number and Configuration of Microphones (CHAPTER 4)

E. Vargas, K. Brown, K. Subr, "Impact of Microphone Array Configurations on Robust Indirect 3D Acoustic Source Localization", in *International Conference on Acoustics, Speech and Signal Processing* **(ICASSP)**, Calgary, Canada, April 2018. ***(Oral Presentation)***

Locating The Source In 3D

Steered Response Power (SRP)



Most likely position amongst a grid of candidate locations

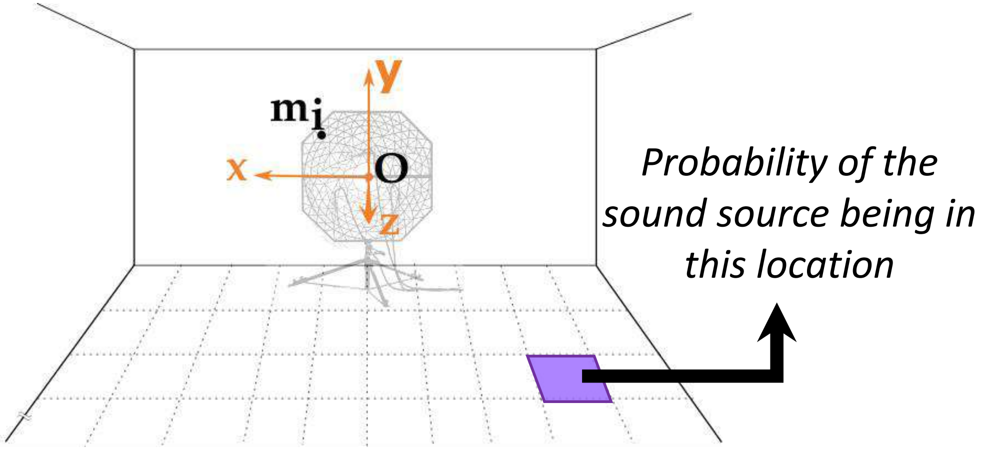
✓Accurate

✘ Slow

Lima et al., 2015

Locating The Source In 3D

Steered Response Power (SRP)



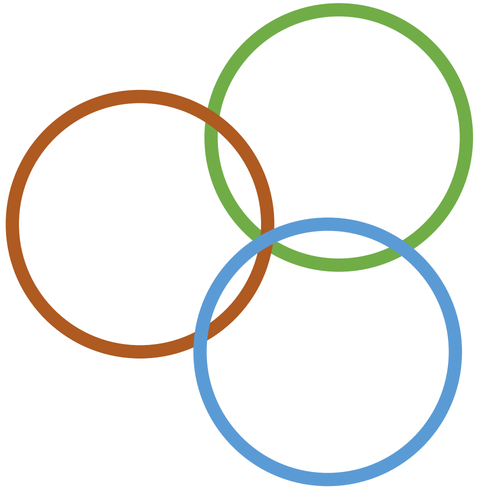
Most likely position amongst a grid of candidate locations

✓Accurate

✘ Slow

Lima et al., 2015

Multilateration



Infer the source position via least squares optimisation

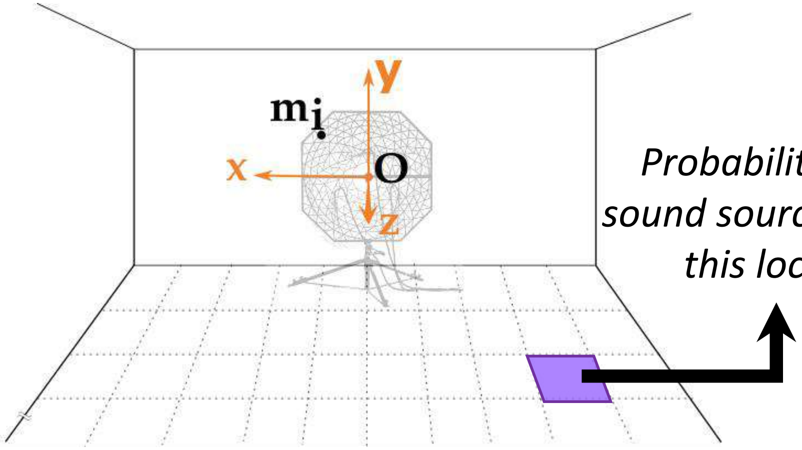
Qu et al., 2016

✓ Fast

✘ Non-convex function, local minima

Locating The Source In 3D

Steered Response Power (SRP)

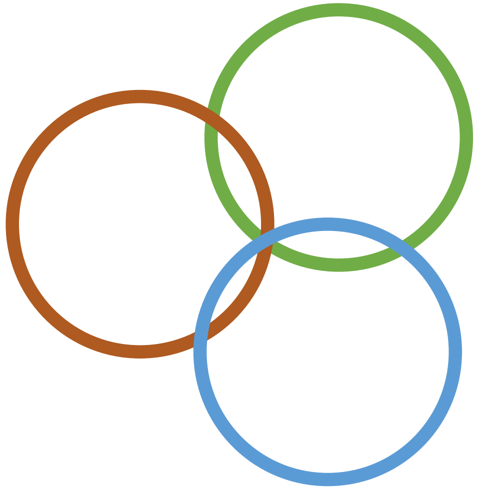


Probability of the sound source being in this location

Most likely position amongst a grid of candidate locations
✓Accurate
✘ Slow

Lima et al., 2015

Multilateration

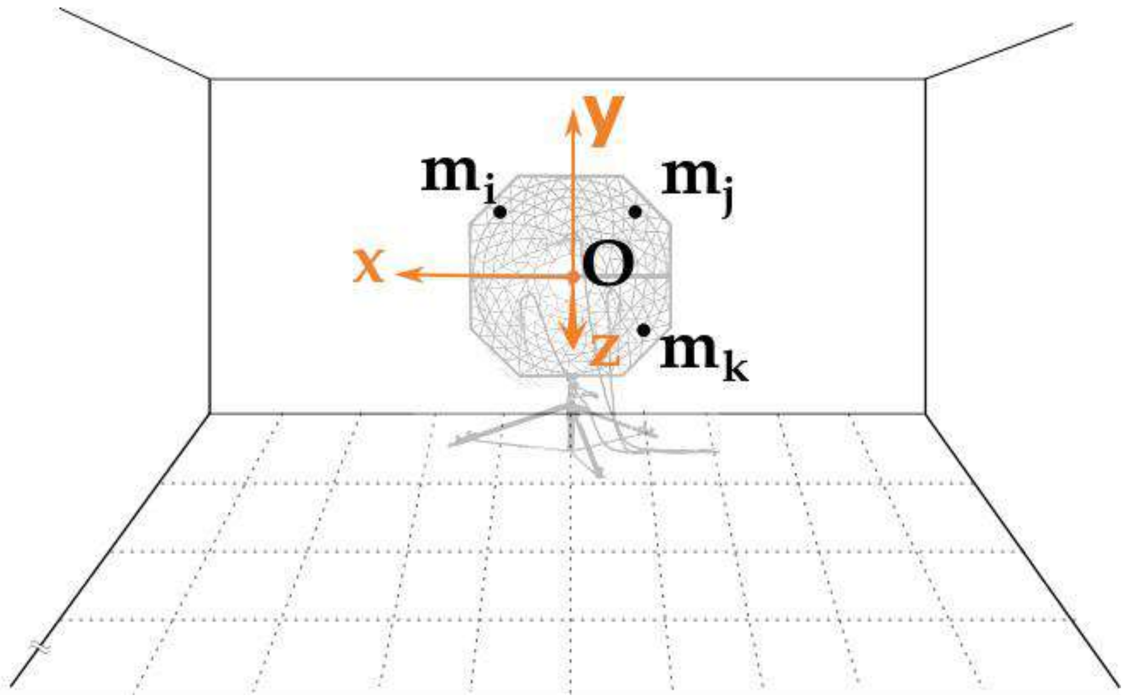


Infer the source position via least squares optimization
✓ Fast
✘ Non-convex function, local minima

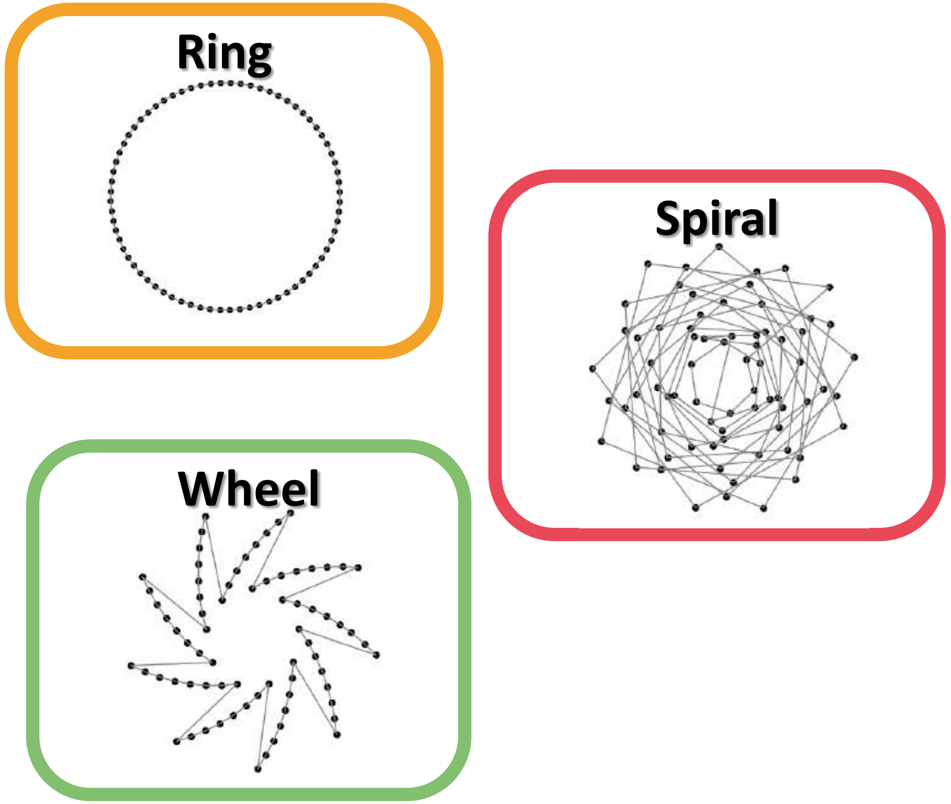
Qu et al., 2016

PROBLEM 1: Localisation needs to be accurate and fast at the same time

Experimental Set Up

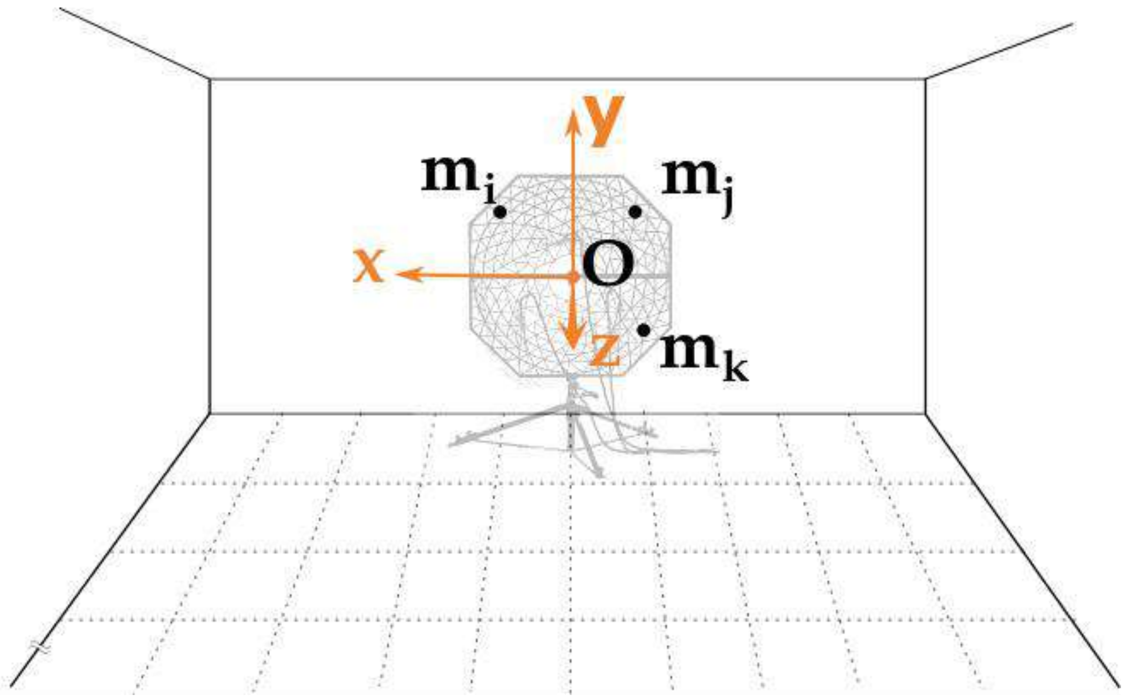


- **72** microphones
- Sampled at **192 kHz**

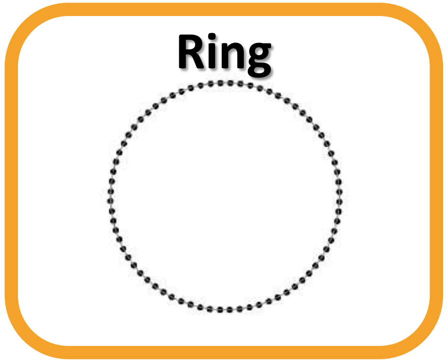


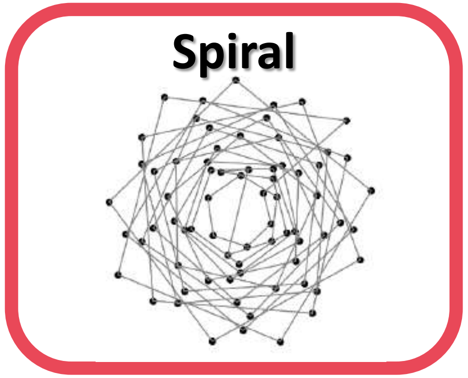
- **Three configurations** spanning the same area

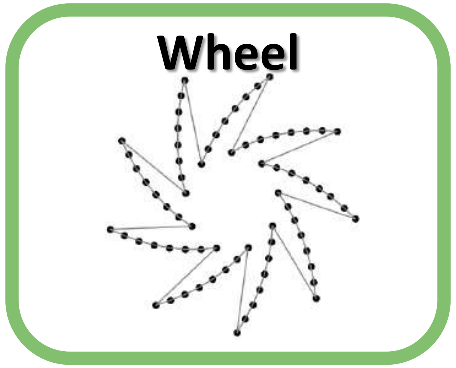
Experimental Set Up



- **72** microphones
- Sampled at **192 kHz**







- **Three configurations** spanning the same area

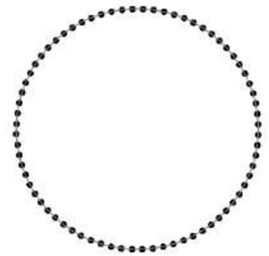
Problem 2: Which configuration to use in order to obtain the most accurate localisation?

HYPOTHESIS I

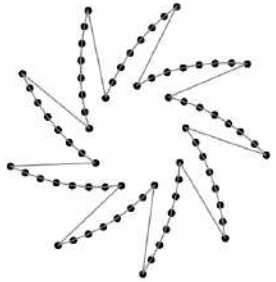
Using one particular **microphone configuration** over another could bring **more accuracy** to the estimation of sound source localisation. Moreover, **multilateration could be fast and reliable** when used in combination with the right amount of microphone pairs.

Simulated Room Impulse Response (RIR)

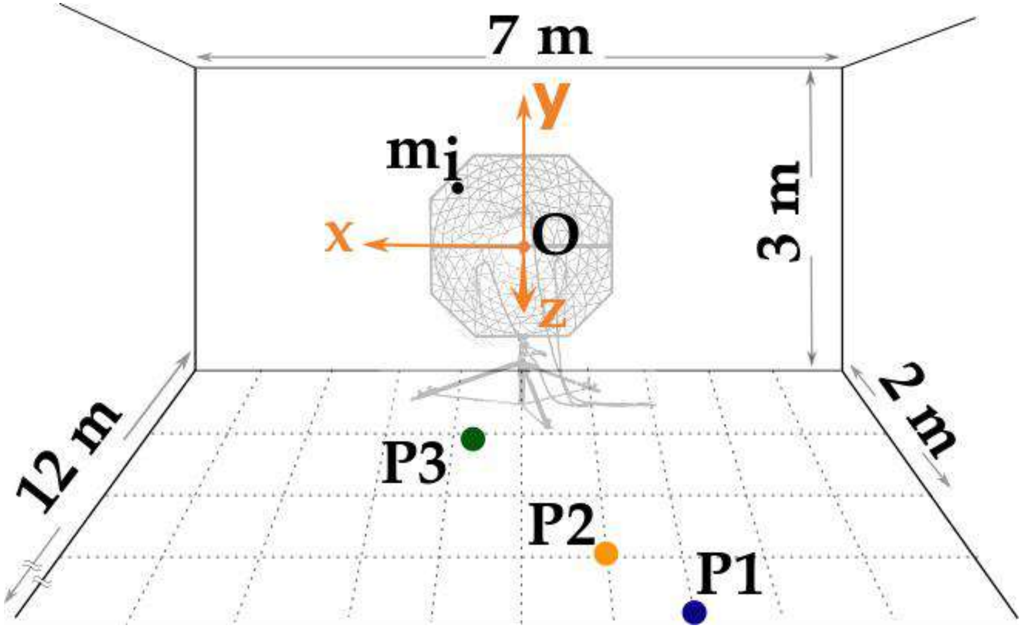
Ring



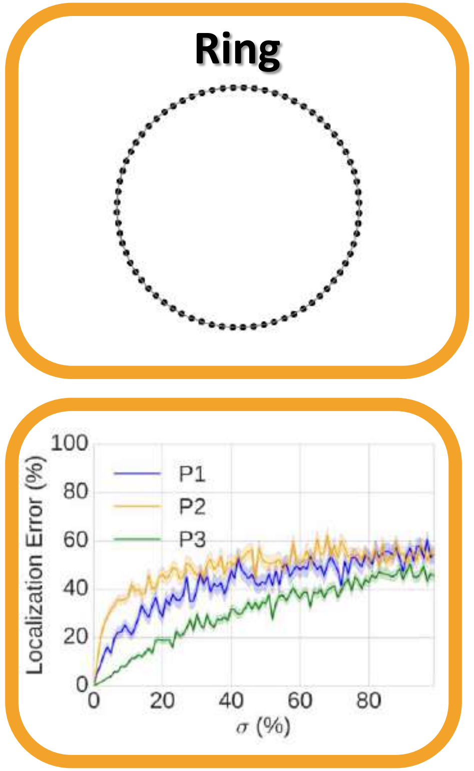
Wheel

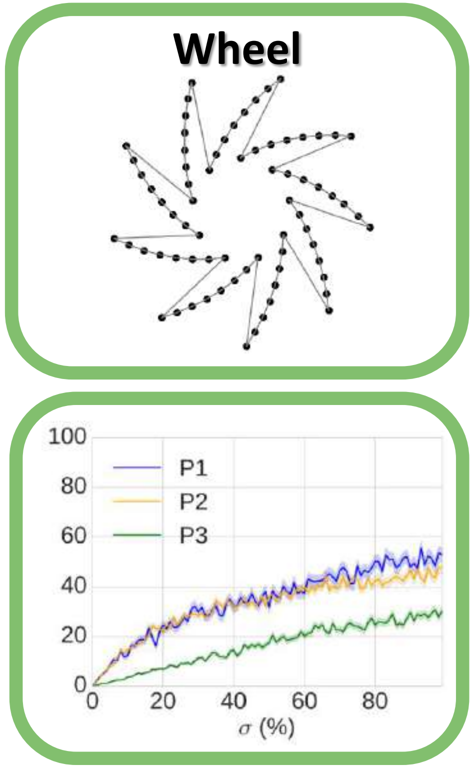


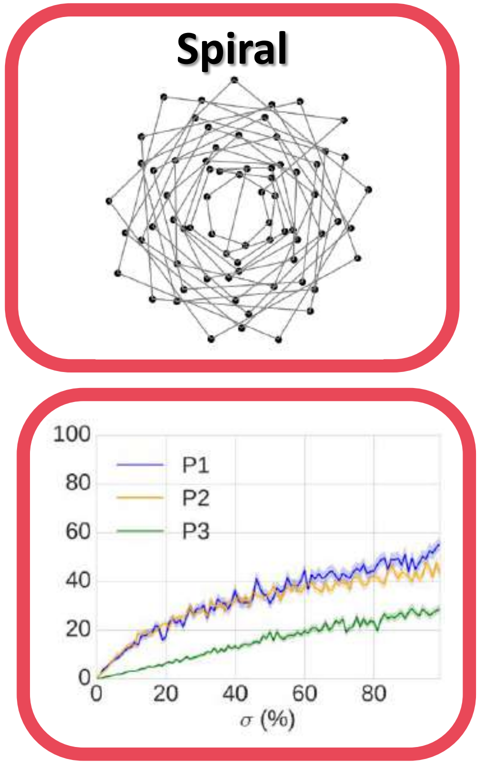
Spiral

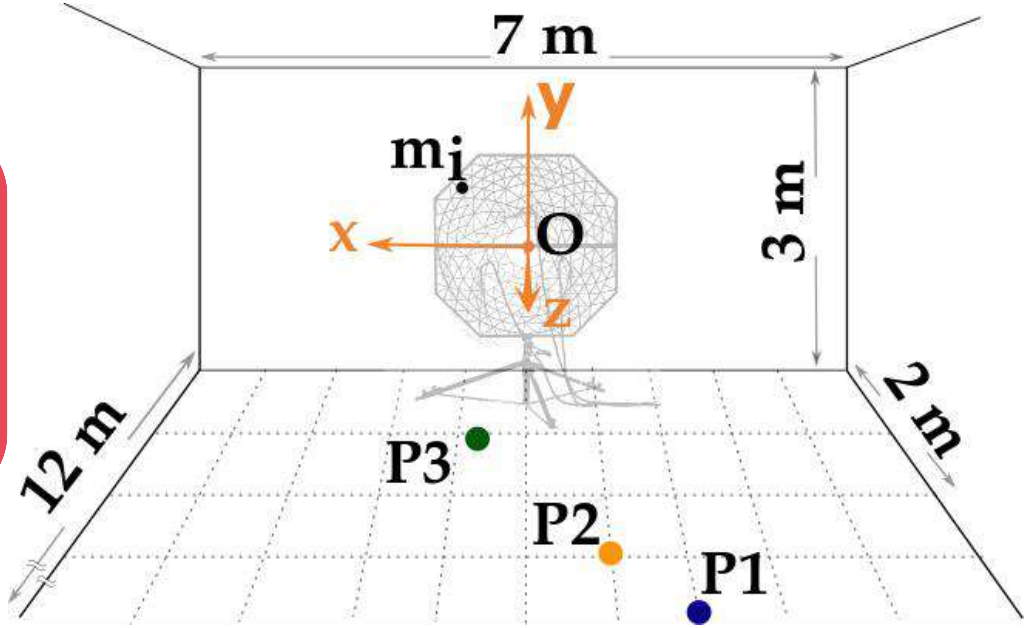


Simulated Room Impulse Response (RIR)

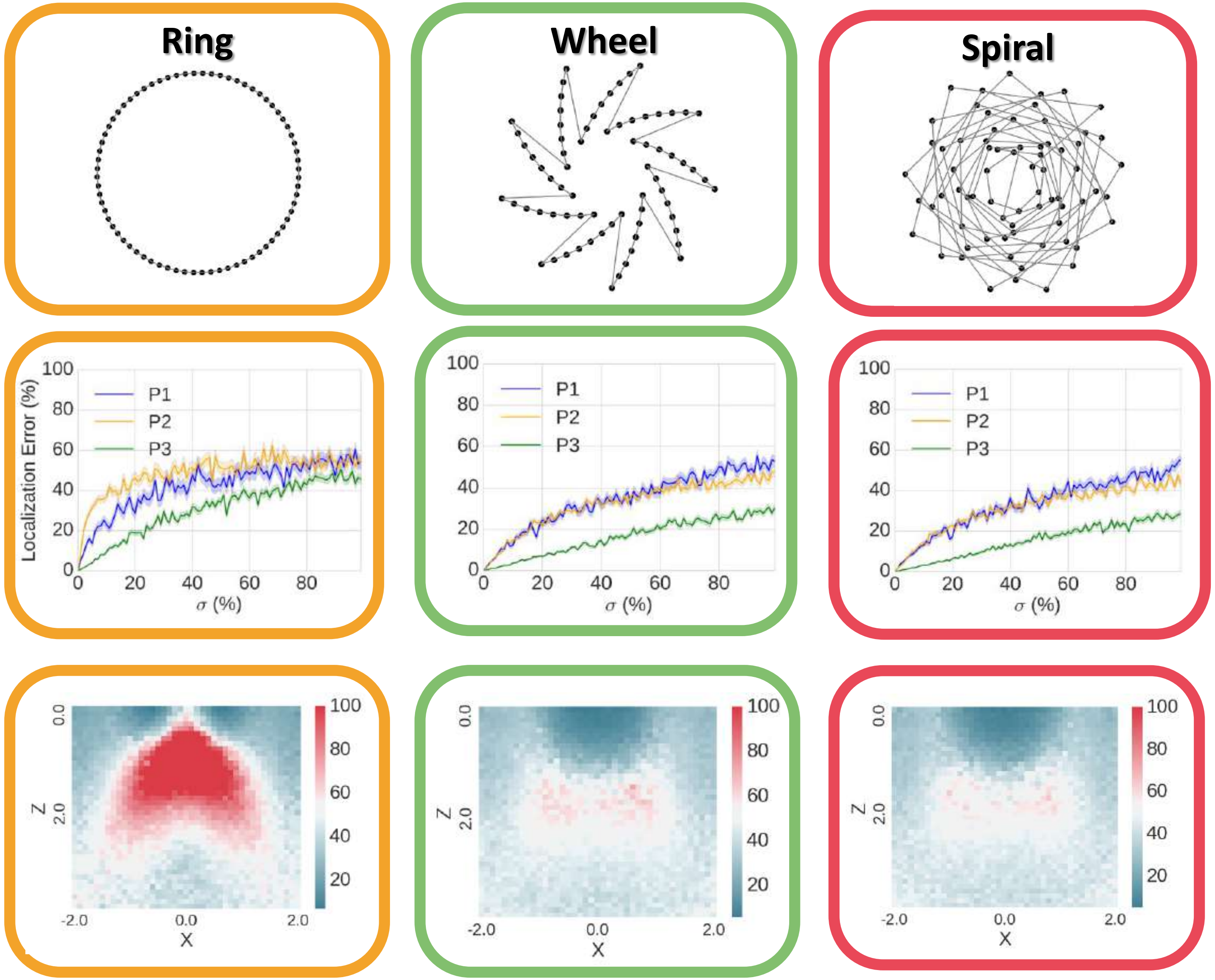


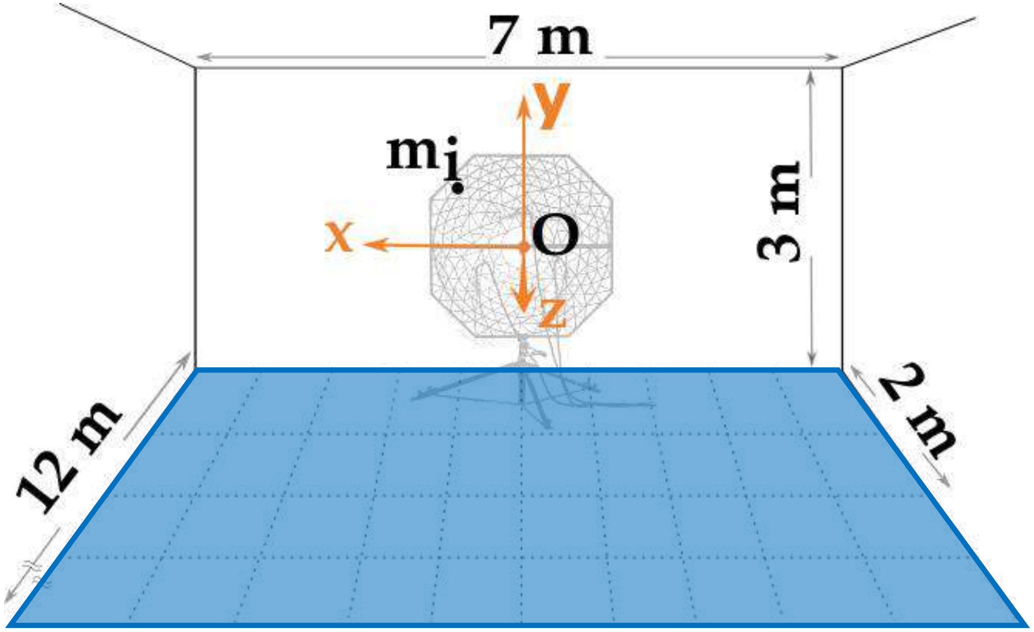




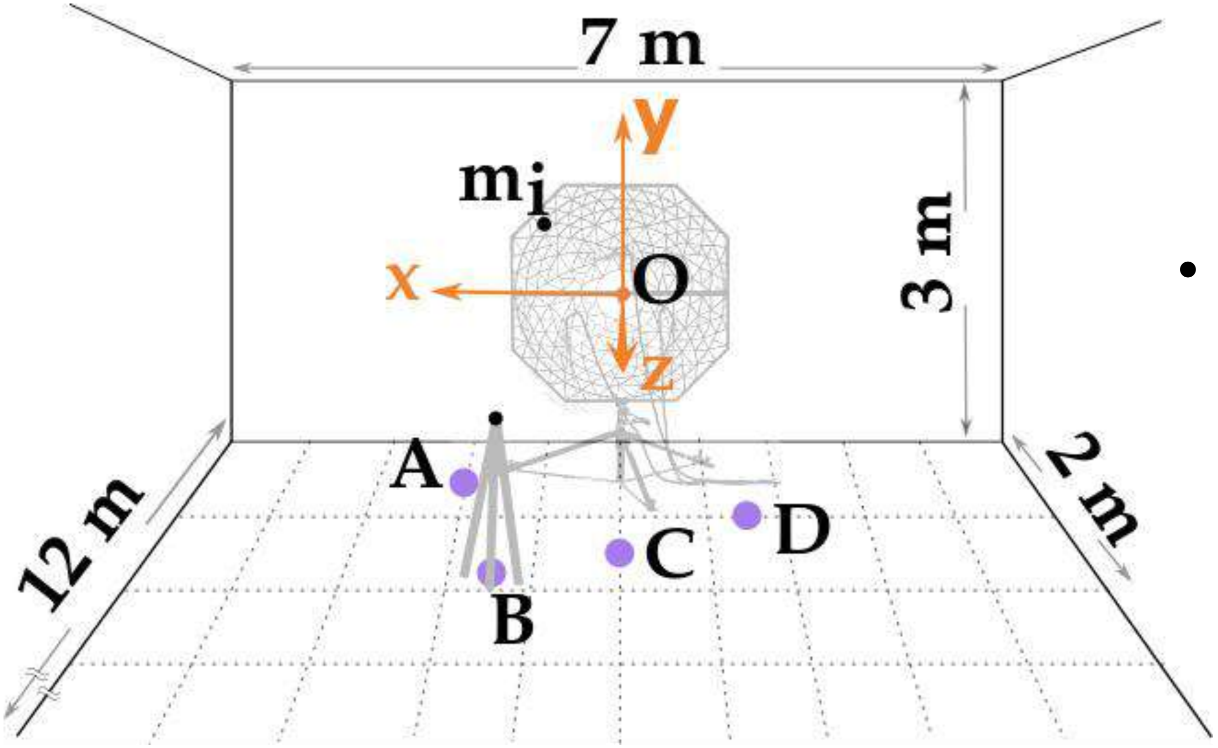


Simulated Room Impulse Response (RIR)

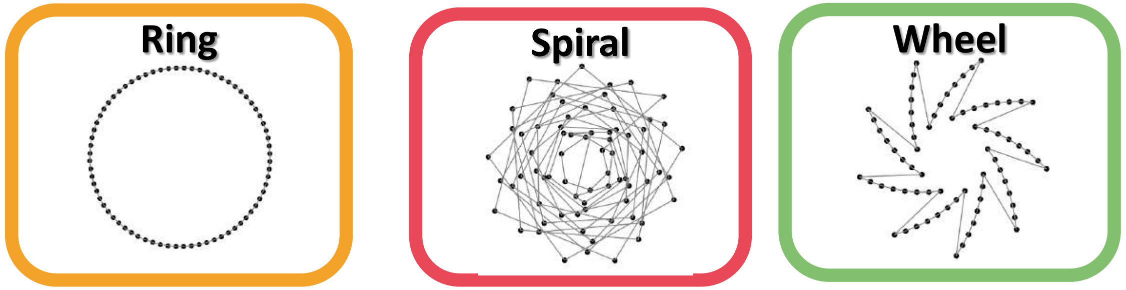




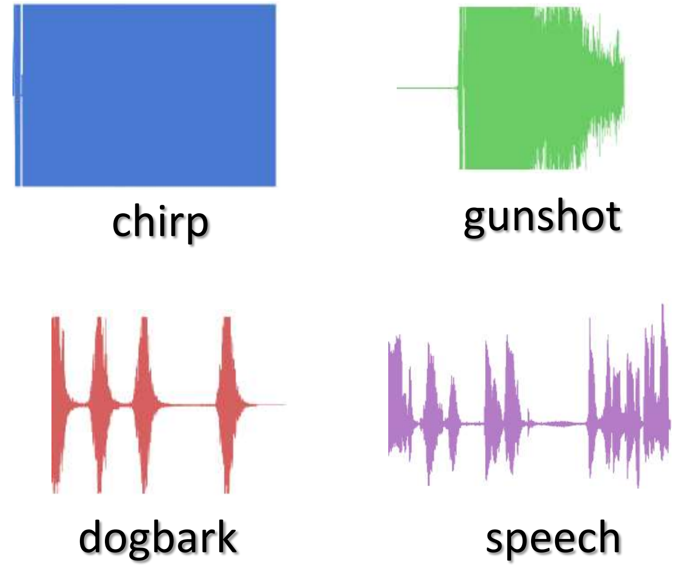
Using Real Data



- **72** microphones
- Sampled at **192 kHz**

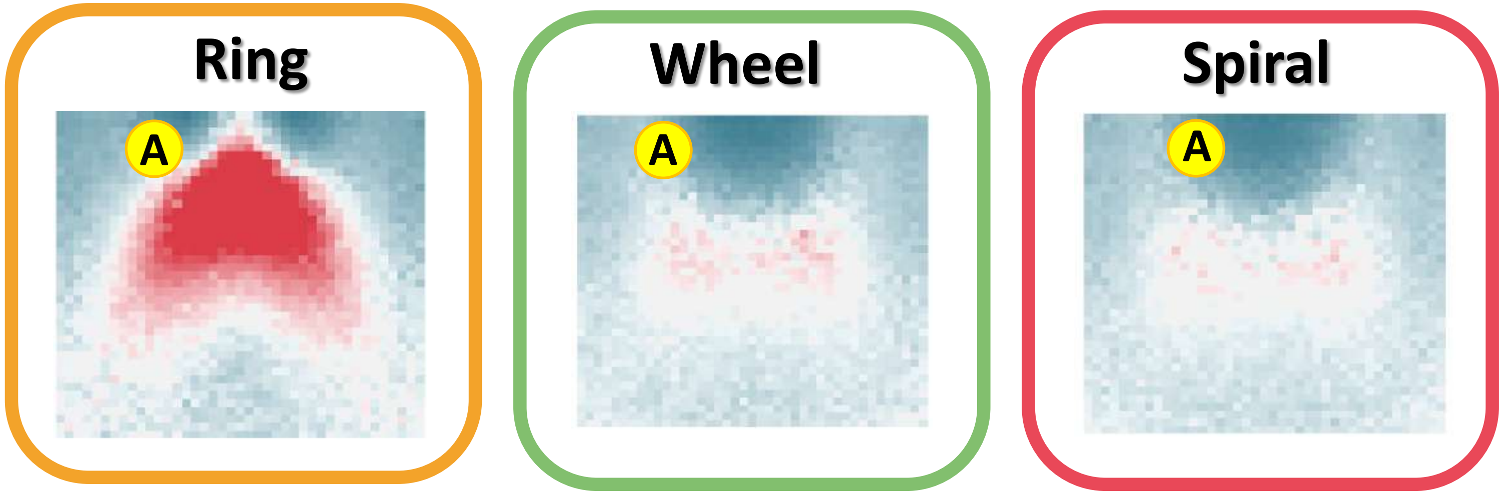


- **Three configurations** spanning the same area

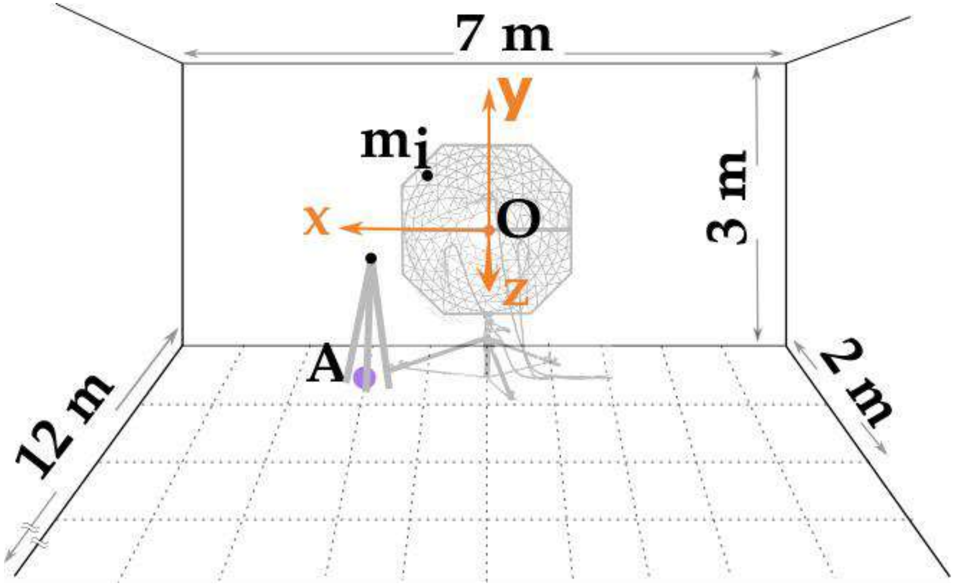


- **Four** different **signals**

Short Range Localisation

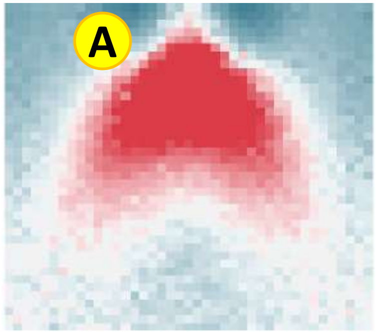


A: (2.0,-0.32,0.5)

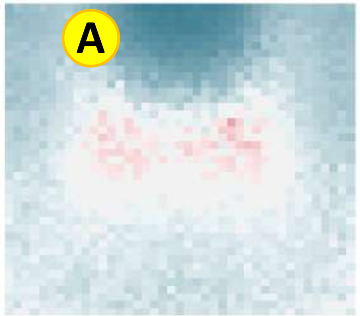


Short Range Localisation

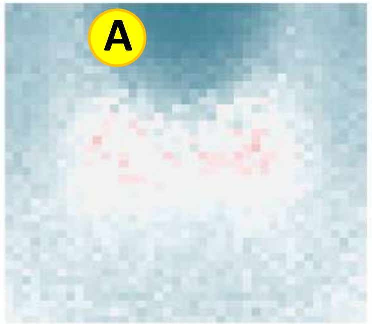
Ring



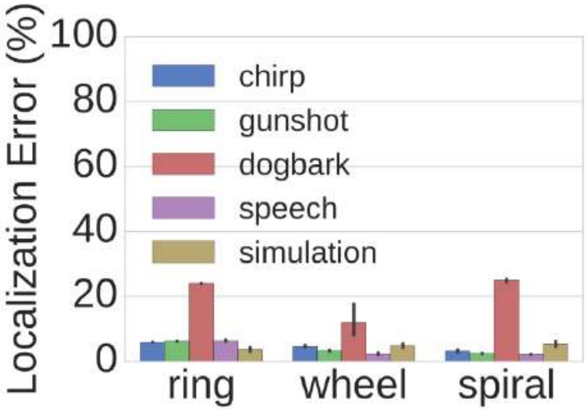
Wheel



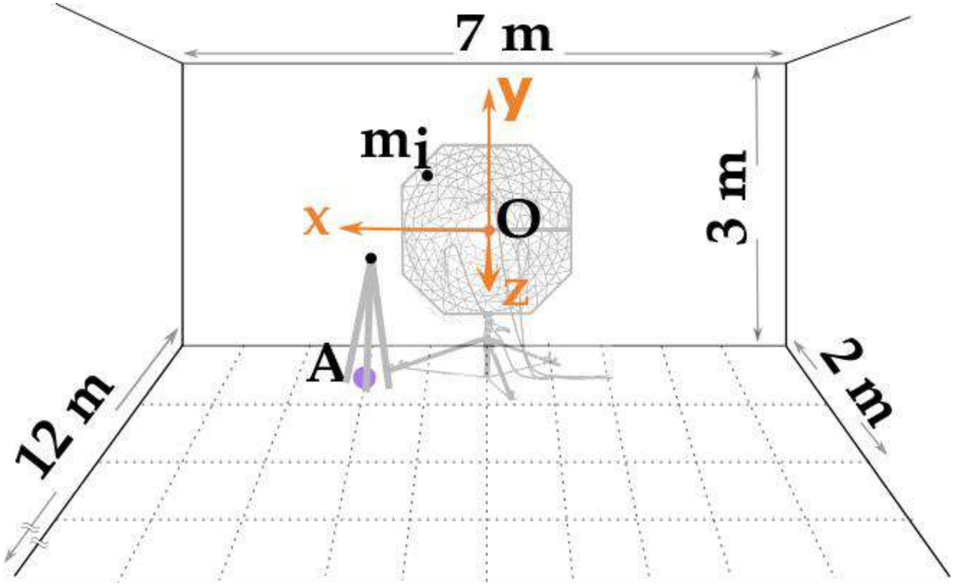
Spiral



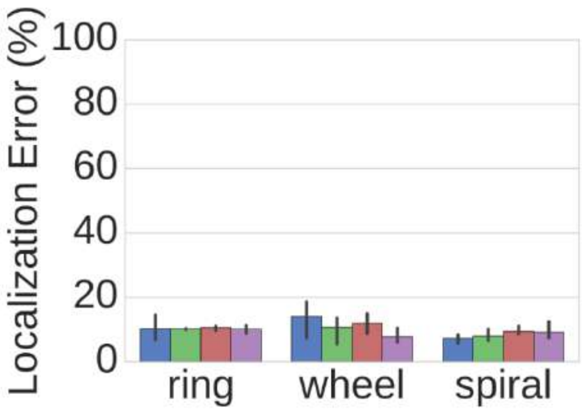
Multilateration



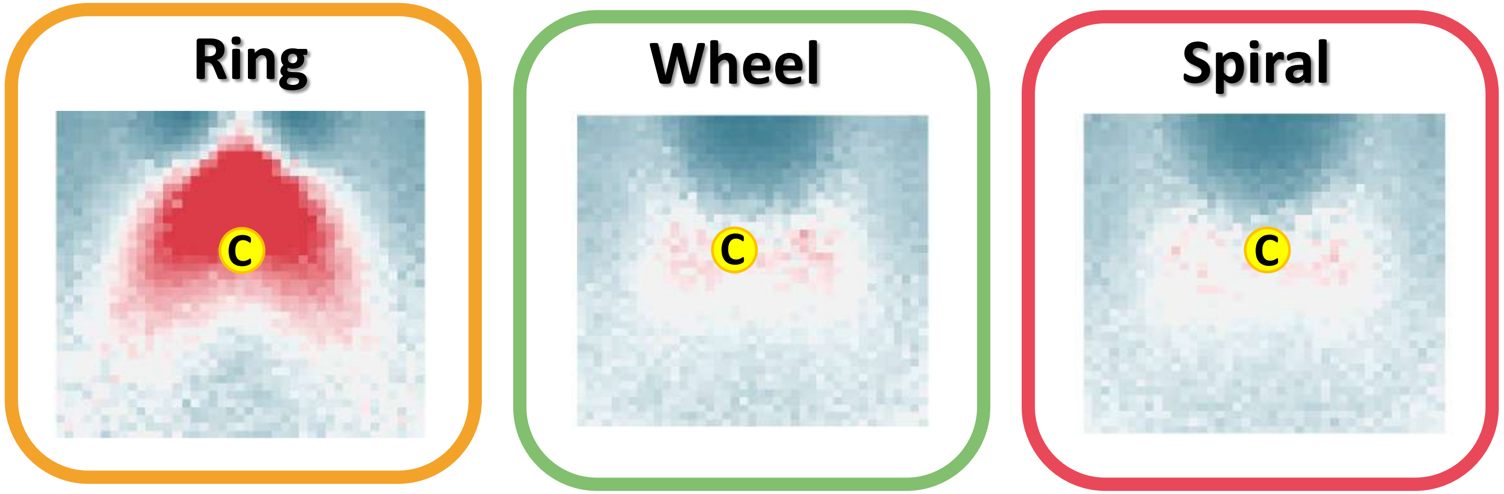
A: (2.0,-0.32,0.5)



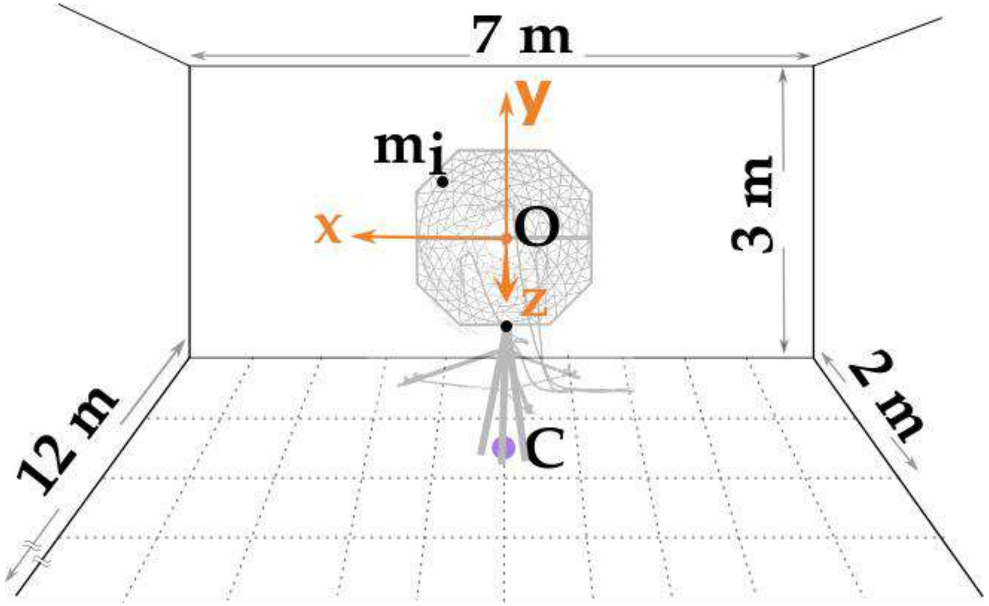
SRP



Facing The Microphone Array

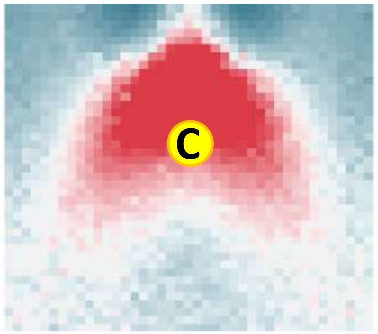


C: (0.0,-0.32,1.5)

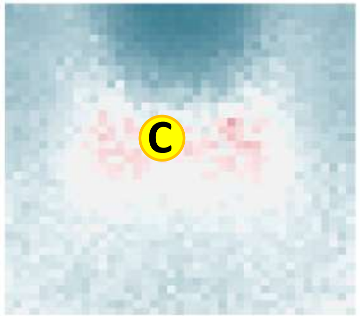


Facing The Microphone Array

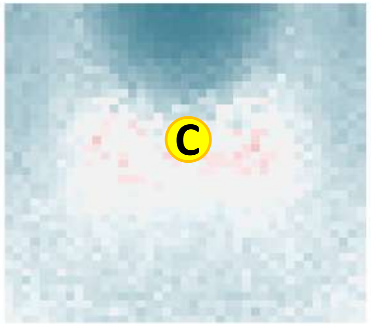
Ring



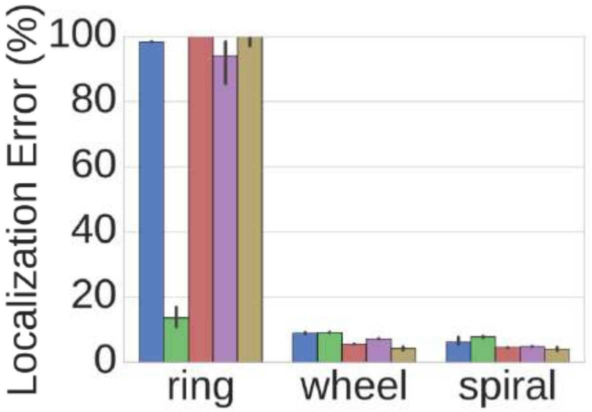
Wheel



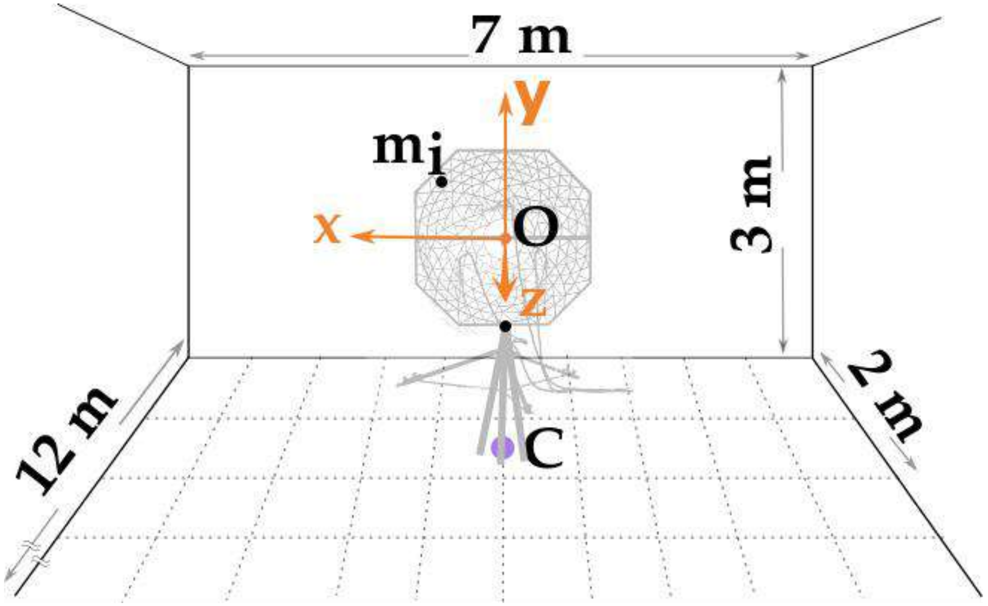
Spiral



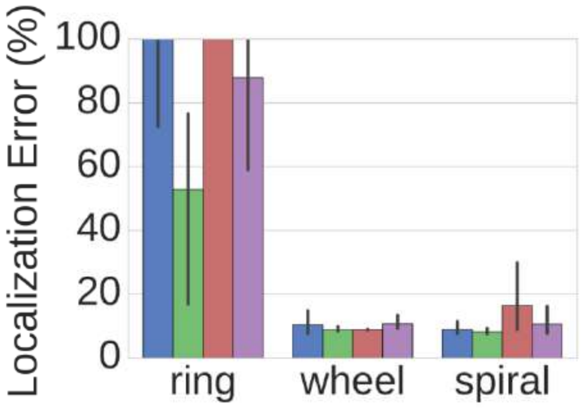
Multilateration



C: (0.0,-0.32,1.5)



SRP



Using 2556 Microphone Pairs

Accuracy (%)

Signal	SRP	Multilateration
Chirp	14.7 (25.9)	12.1 (23.2)
Gunshot	11.0 (13.3)	6.4 (3.5)
Dogbark	16.0 (28.5)	48.5 (44.6)
Speech	13.2 (21.1)	12.9 (22.5)

Using 2556 Microphone Pairs

Accuracy (%)

Signal	SRP	Multilateration
Chirp	14.7 (25.9)	12.1 (23.2)
Gunshot	11.0 (13.3)	6.4 (3.5)
Dogbark	16.0 (28.5)	48.5 (44.6)
Speech	13.2 (21.1)	12.9 (22.5)

Time (minutes)

Signal	SRP	Multilateration
Chirp	3 (0.2)	4.5 (0.03)
Gunshot	2.58 (0.2)	2.4 (0.02)
Dogbark	2.49 (0.1)	2.4 (0.02)
Speech	2.63 (0.1)	2.5 (0.02)

Using 100 Microphone Pairs

Accuracy (%)

Signal	SRP	Multilateration
Chirp	14.7 (25.9)	14.2 (25.9)
Gunshot	11.0 (13.3)	9.6 (12.8)
Dogbark	16.0 (28.5)	58.9 (38.8)
Speech	13.2 (21.1)	15.2 (23.5)

Using 100 Microphone Pairs

Accuracy (%)

Signal	SRP	Multilateration
Chirp	14.7 (25.9)	14.2 (25.9)
Gunshot	11.0 (13.3)	9.6 (12.8)
Dogbark	16.0 (28.5)	58.9 (38.8)
Speech	13.2 (21.1)	15.2 (23.5)

Time (minutes)

Signal	SRP	Multilateration
Chirp	3 (0.2)	0.5(0.01)
Gunshot	2.58 (0.2)	0.4 (0.02)
Dogbark	2.49 (0.1)	0.4 (0.02)
Speech	2.63 (0.1)	0.4 (0.02)

6 TIMES FASTER

Summary of Contributions

1. Multilateration produces localisation errors comparable to the state of the art, Steered Response Power (SRP), with 6 times less computation.

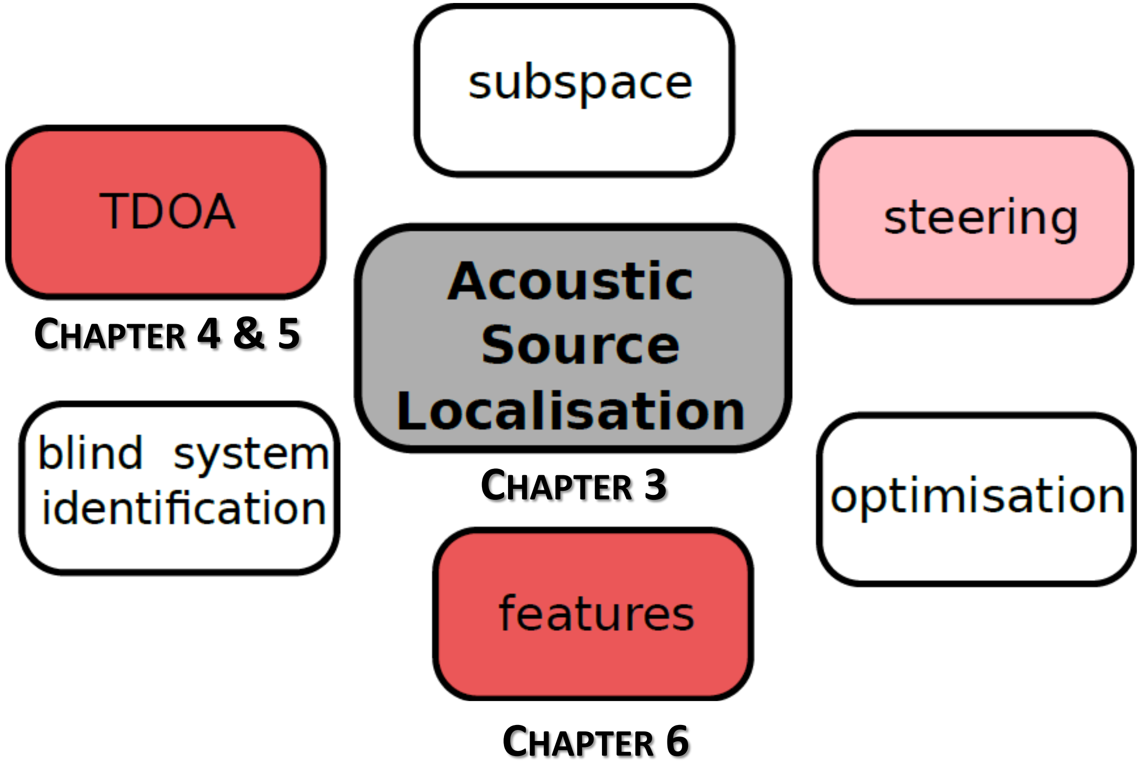
2. Circular arrays lead to higher localisation error than spiral and wheel configurations.

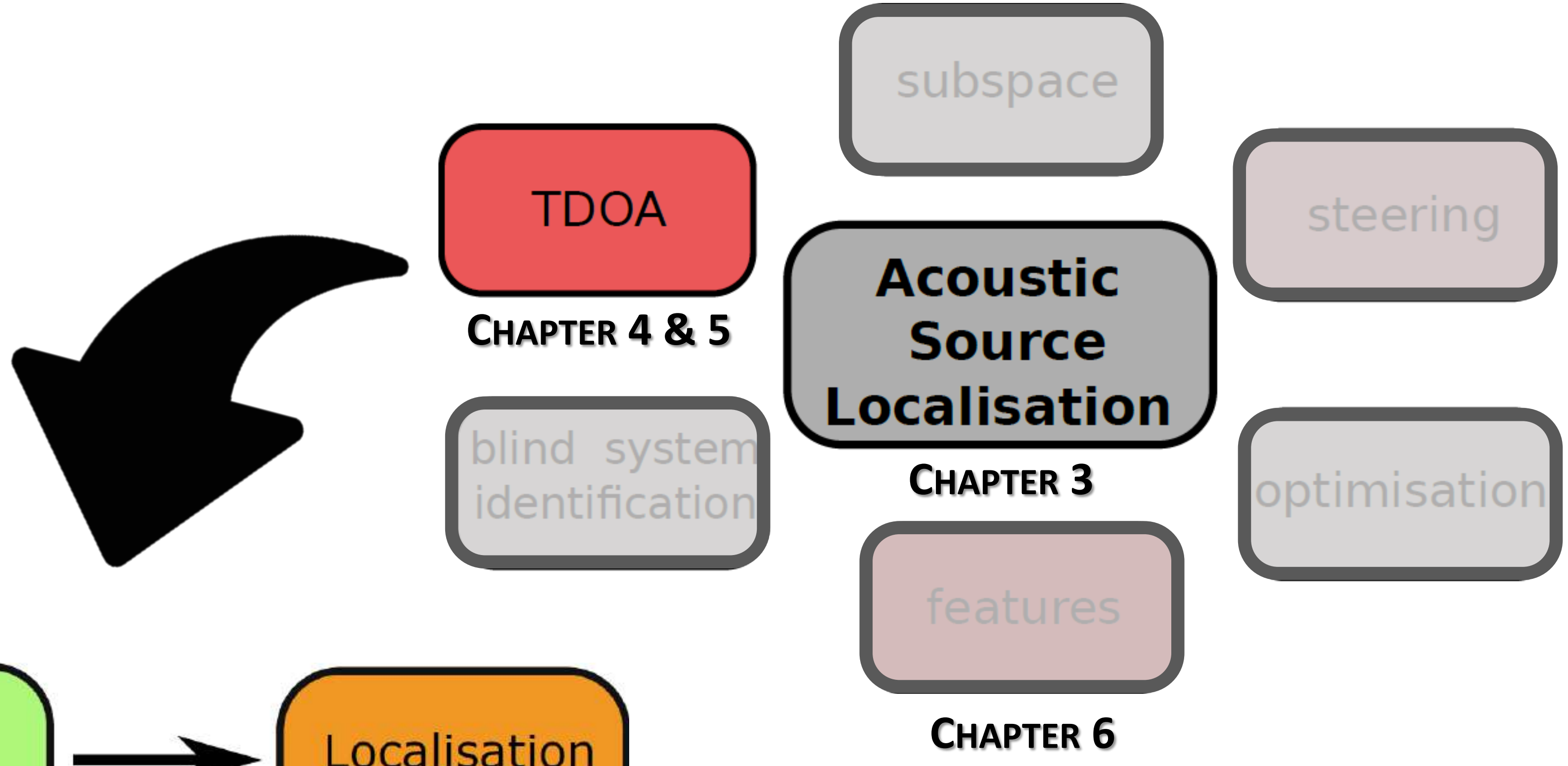
3. We confirmed our hypothesis in simulated and real scenarios.

CONTRIBUTION II

Signal Samples (CHAPTER 5)

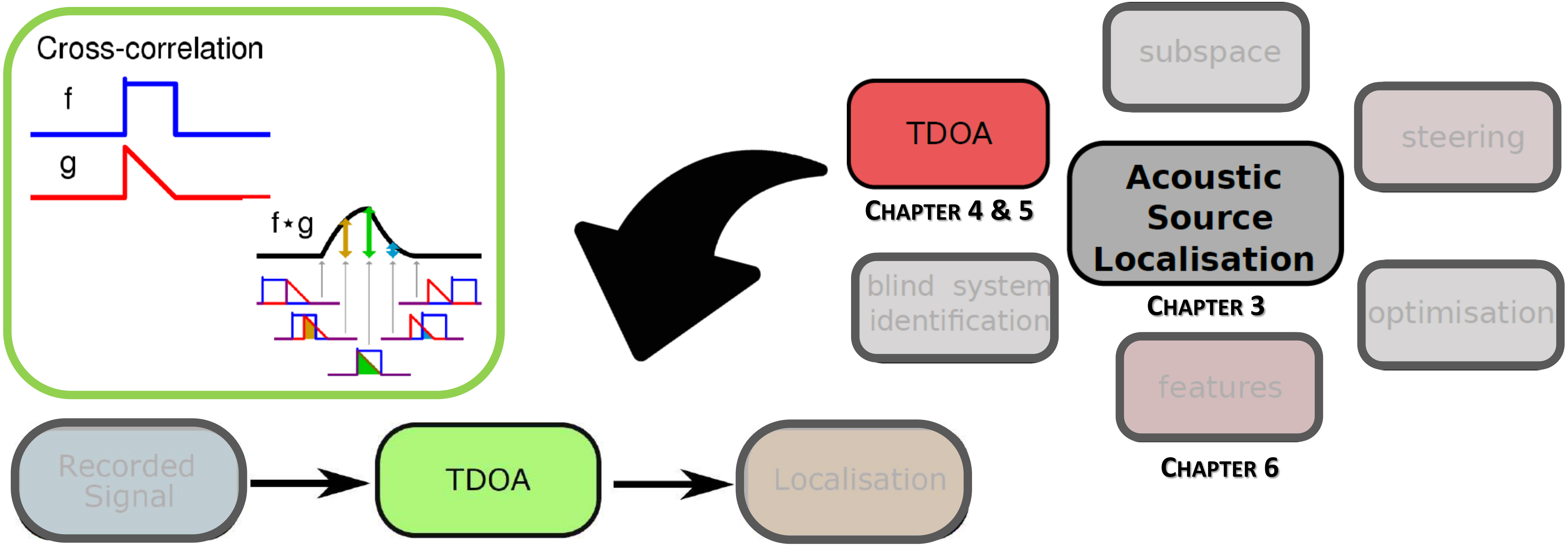
E. Vargas, J. R. Hopgood, K. Brown, K. Subr, "A Compressed Encoding Scheme for Approximate TDOA Estimation", in *European Signal Processing Conference*, **(EUSIPCO)**, Rome, Italy, September 2018. ***(Oral Presentation)***


subspace
TDOA
CHAPTER 4 & 5
steering
Acoustic Source Localisation
CHAPTER 3
blind system identification
optimisation
features
CHAPTER 6

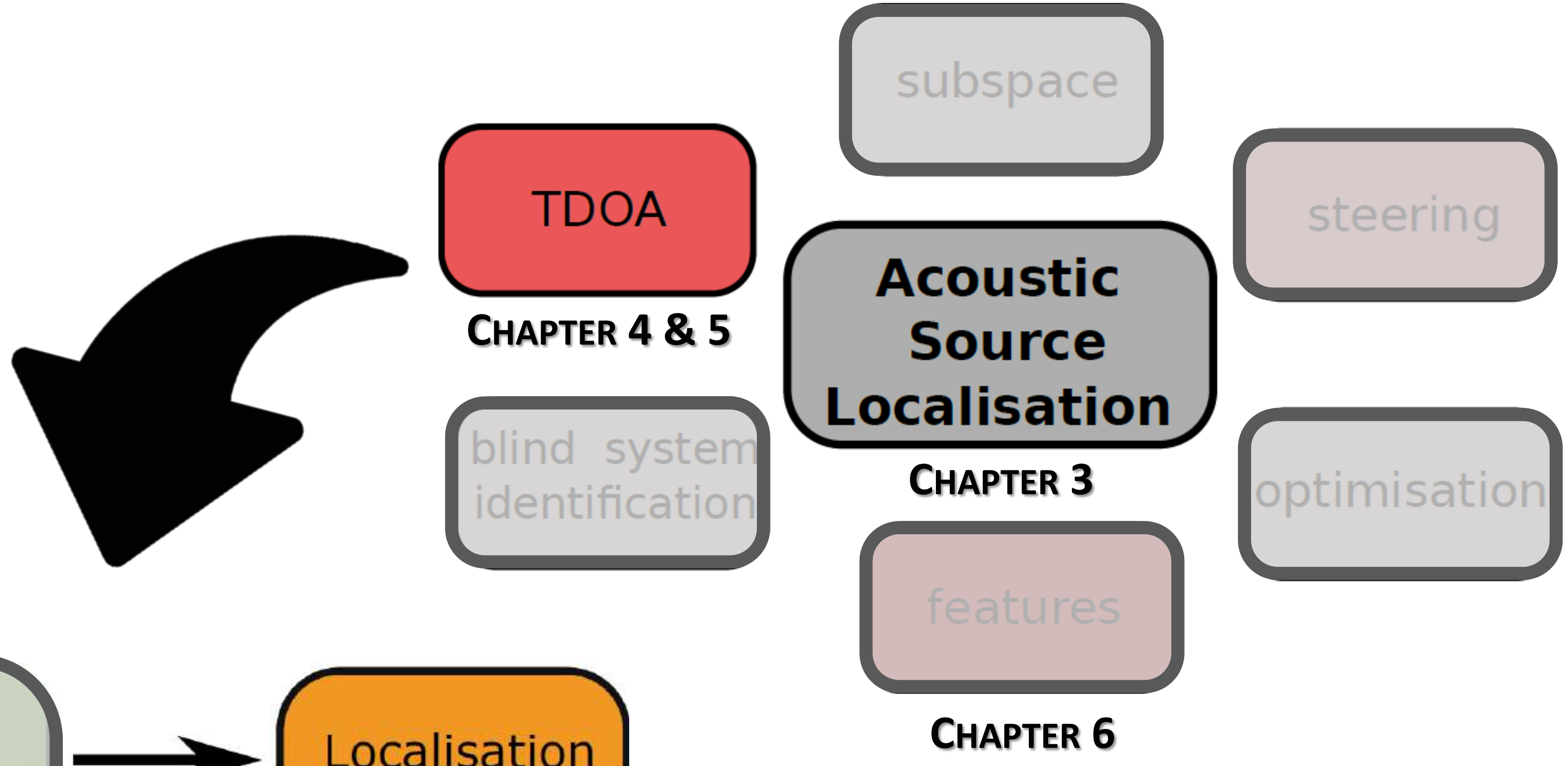


Time Difference of Arrival (TDOA)-Based Methods Pipeline

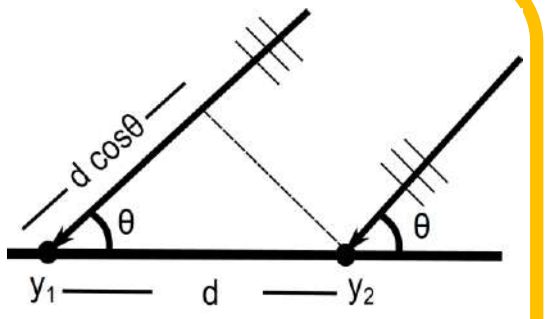
The main idea is to estimate the time delay of the signal between microphone pair(s) *(very similar to human auditory system)*



Time Difference of Arrival (TDOA)-Based Methods Pipeline

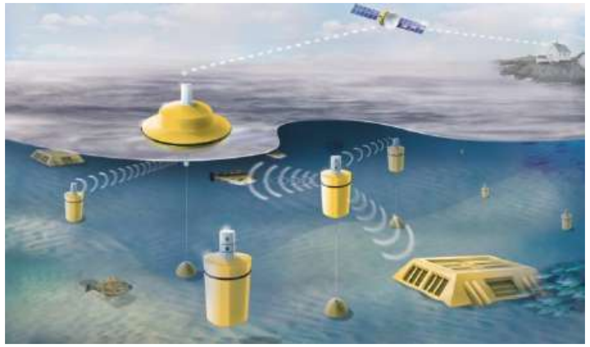


Time Difference of Arrival (TDOA)-Based Methods Pipeline

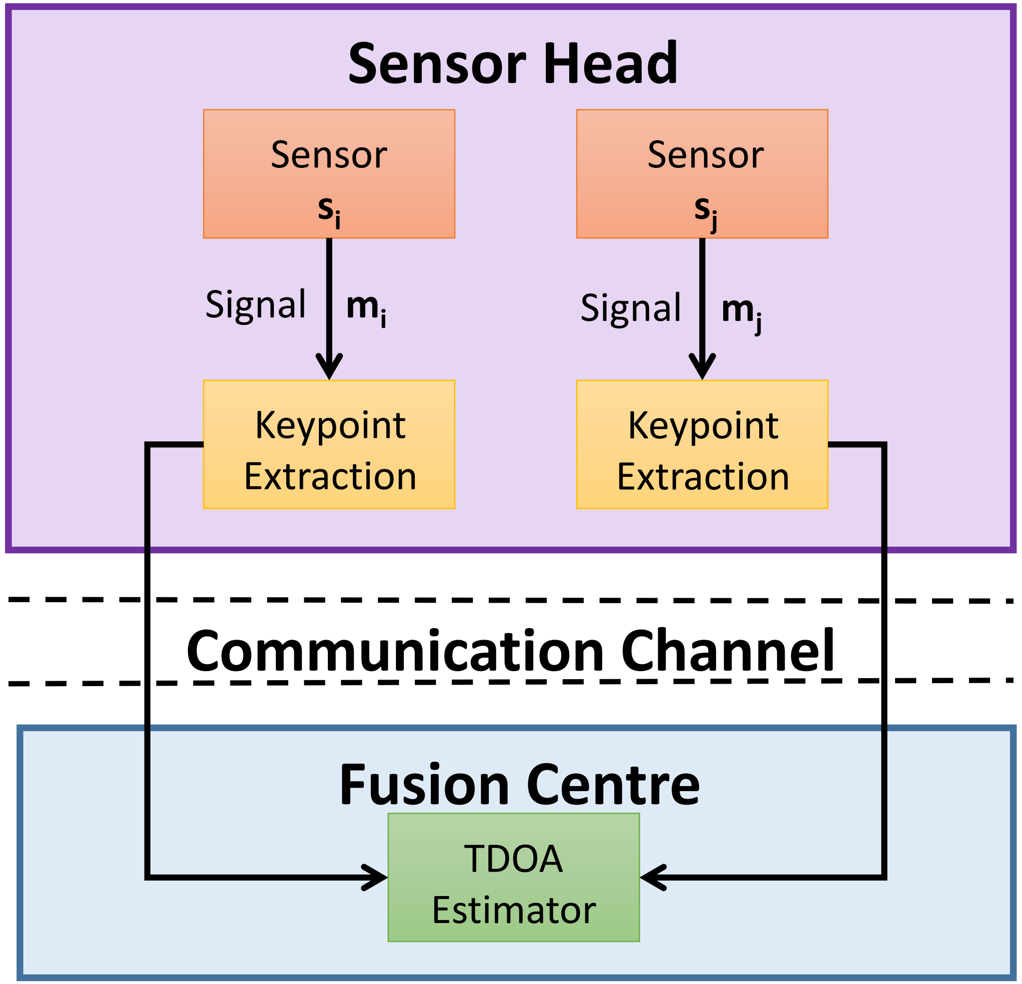


Direction of Arrival (DOA)

Experimental Set Up

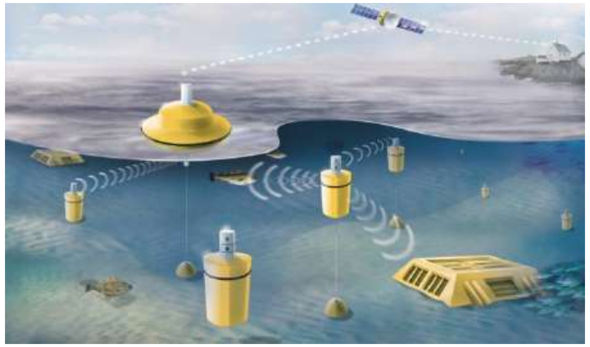


Underwater Sensor Networks

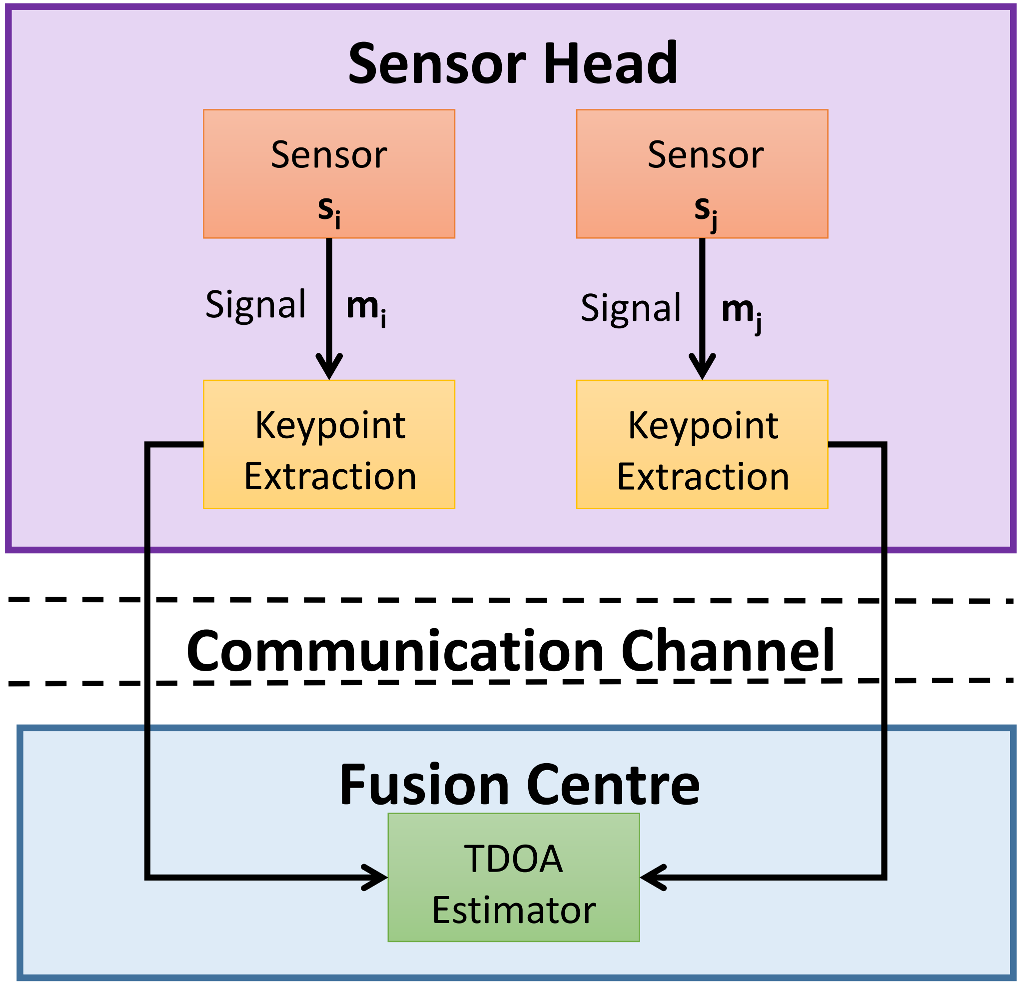


Disaster Zones

Experimental Set Up



Underwater Sensor Networks



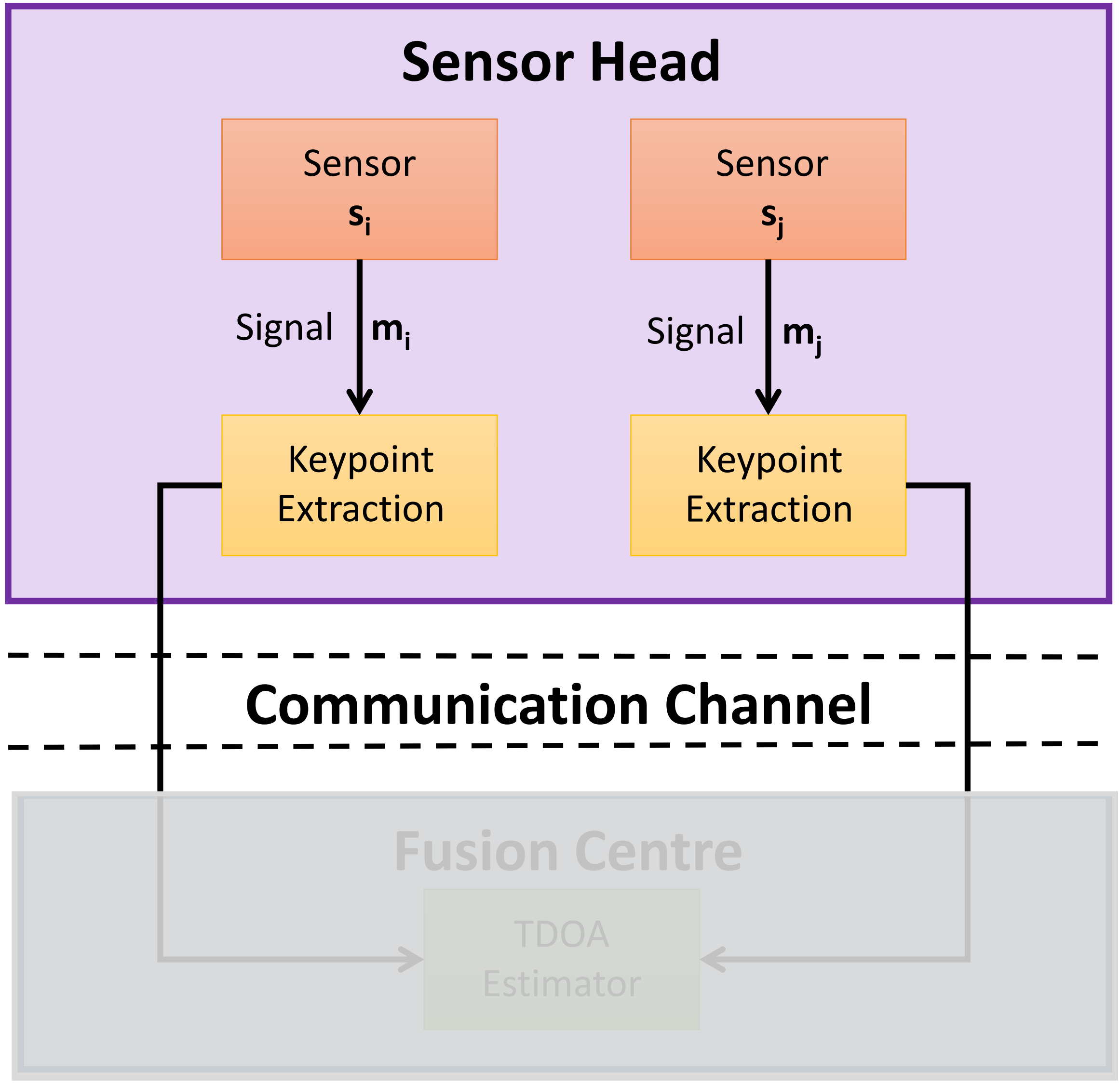
Disaster Zones

PROBLEM: There are limitations in the amount of data that could be transmitted through the communication channel.

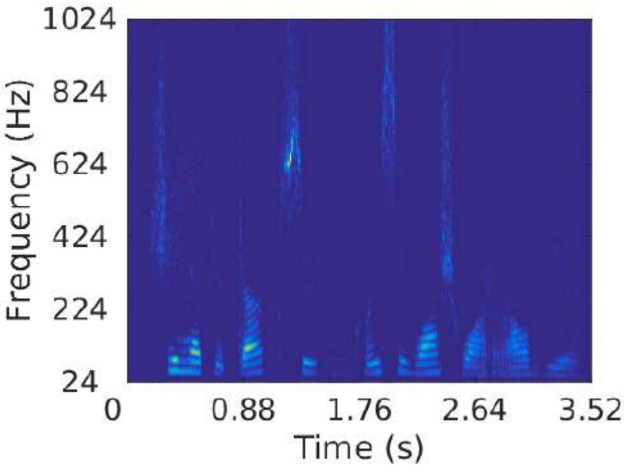
HYPOTHESIS II

It is **not necessary** to use the entire signal to accurately estimate TDOA.

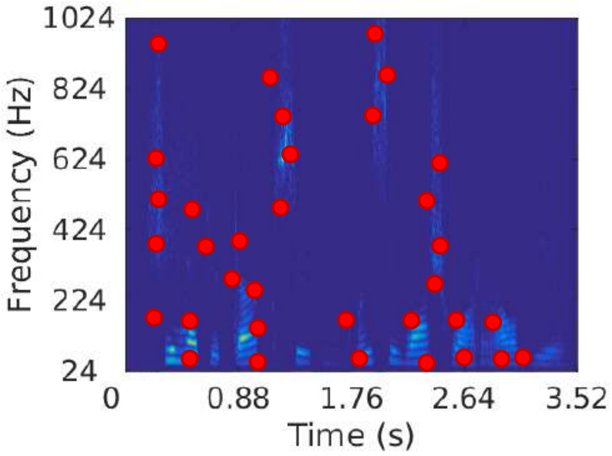
Constraint Transmission



Proposed Method



spectrogram

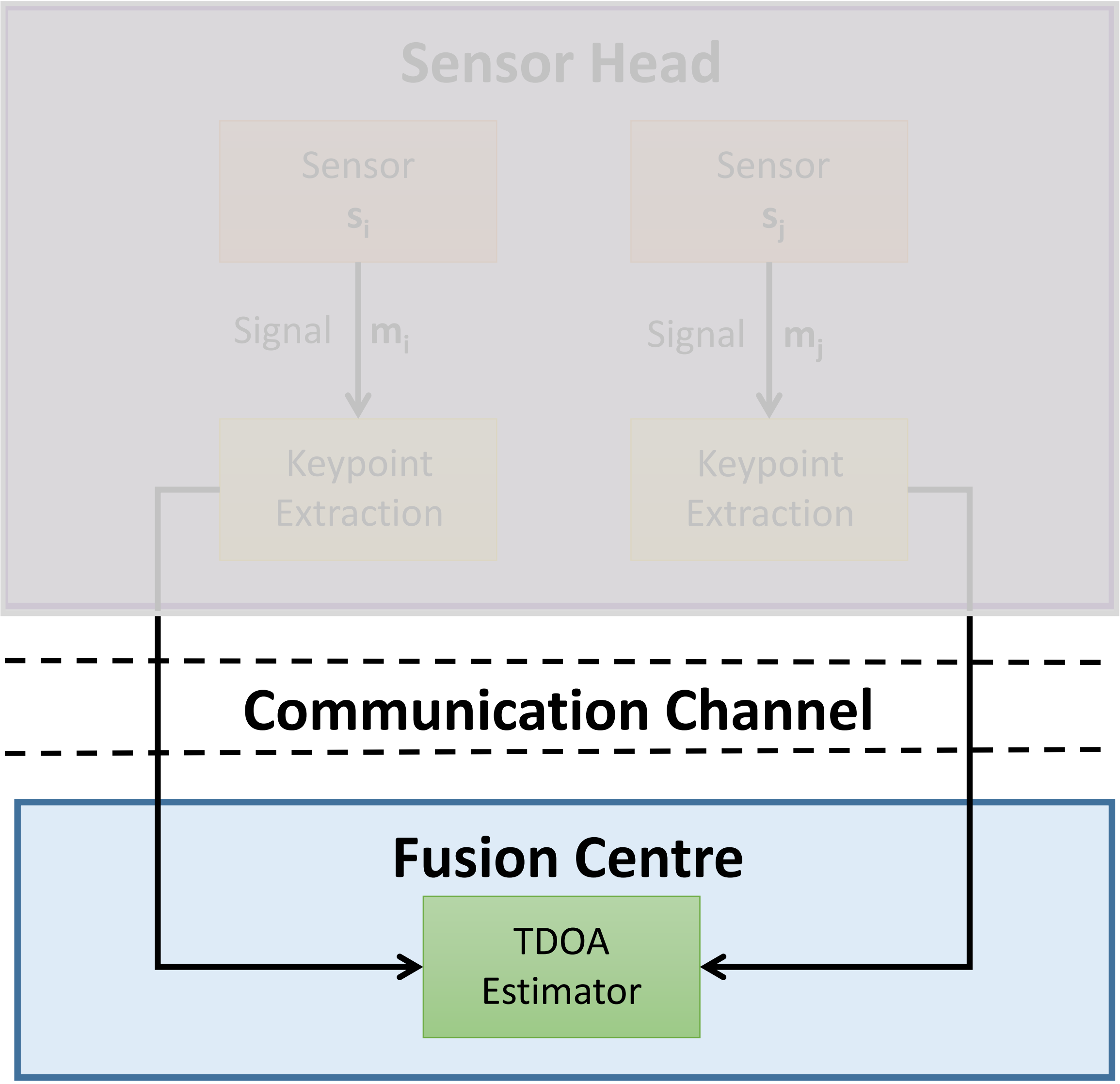


keypoints using SIFT (computer vision)

0	1	0	1	0	0	1	1	0	0

binary mask to transmit

Constraint Transmission



Using all keypoints leads to a **compression ratio** of **40:1**

Proposed Method

Binary Mask
Sensor $\mathbf{s}_i$

0	1	0	1	0	0	1	1	0	0

Binary Mask
Sensor $\mathbf{s}_j$

0	0	1	0	1	0	0	1	1	0

Generalised Cross-correlation (GCC)

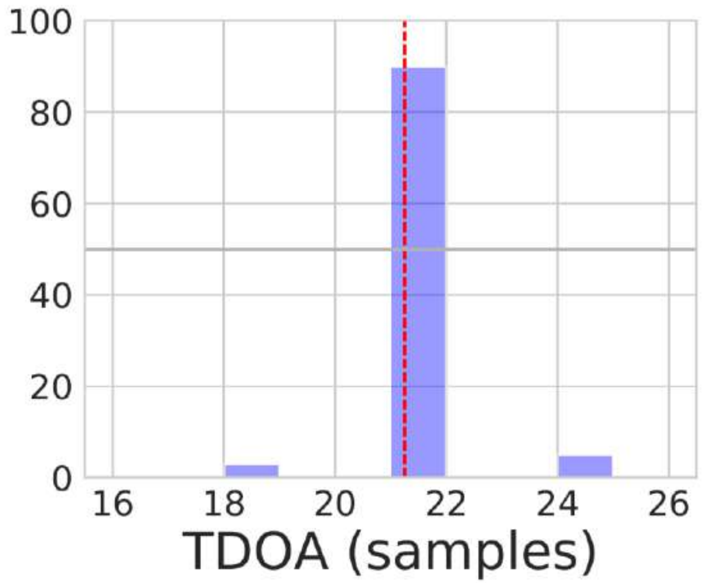
Time Difference of Arrivals (TDOA)

Initial Validation

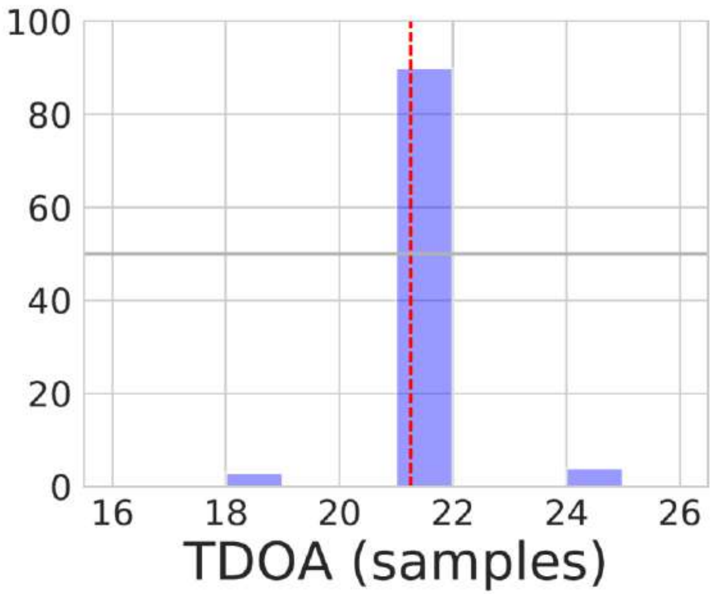
100 monte carlo simulations for a fixed microphone pair and source location

(x = 2, y = 1, z = 5)

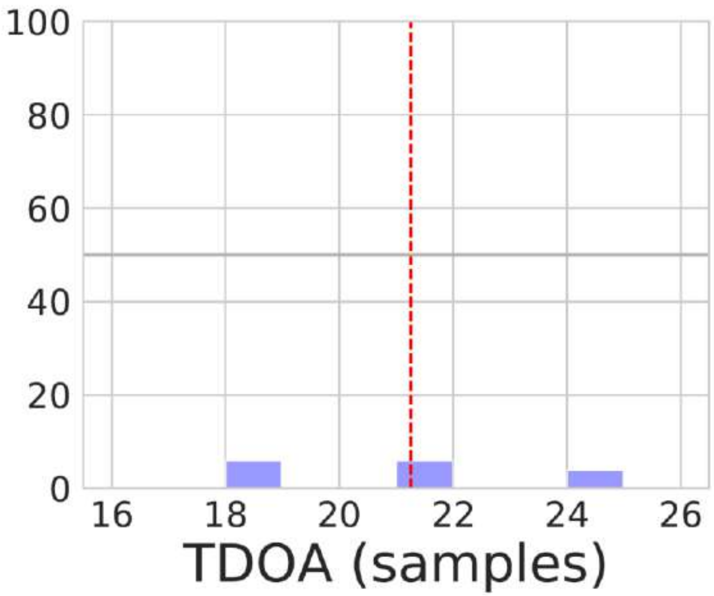
20 dB SNR



reverberation T_{60} = 0



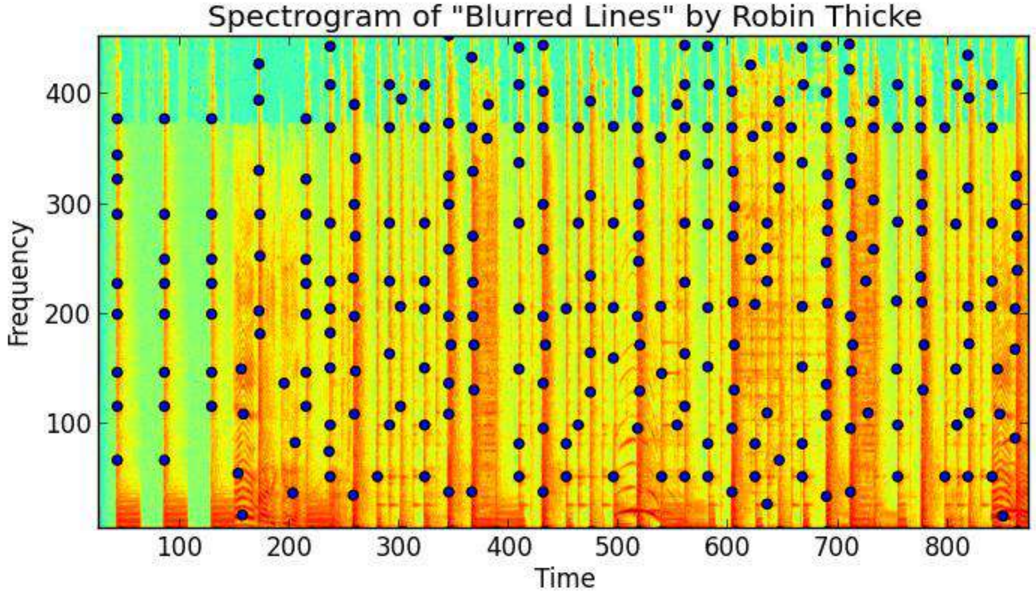
reverberation T_{60} = 0.1

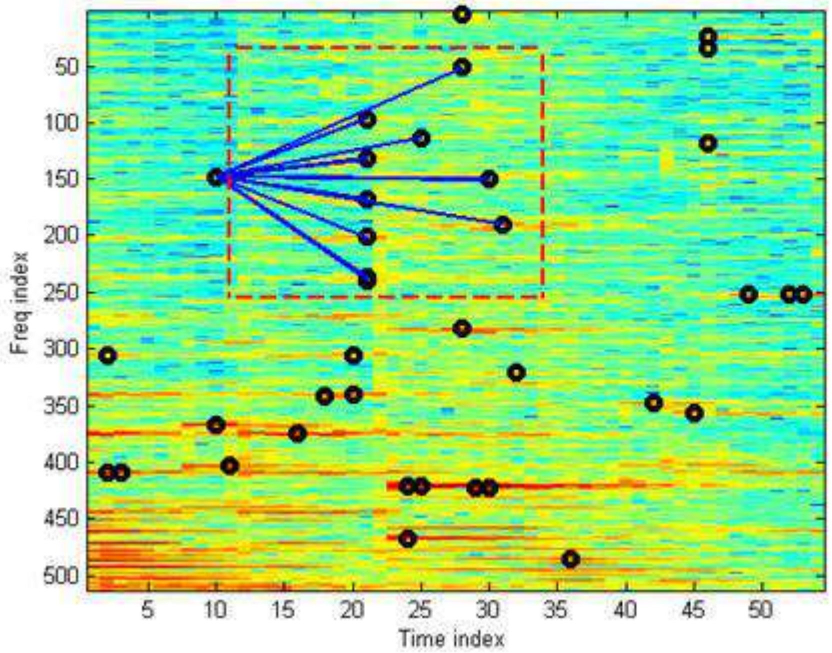


reverberation T_{60} = 0.3

Baseline: Fingerprinting

Information retrieval algorithm used mostly for song matching





It has previously been used to estimate TDOA

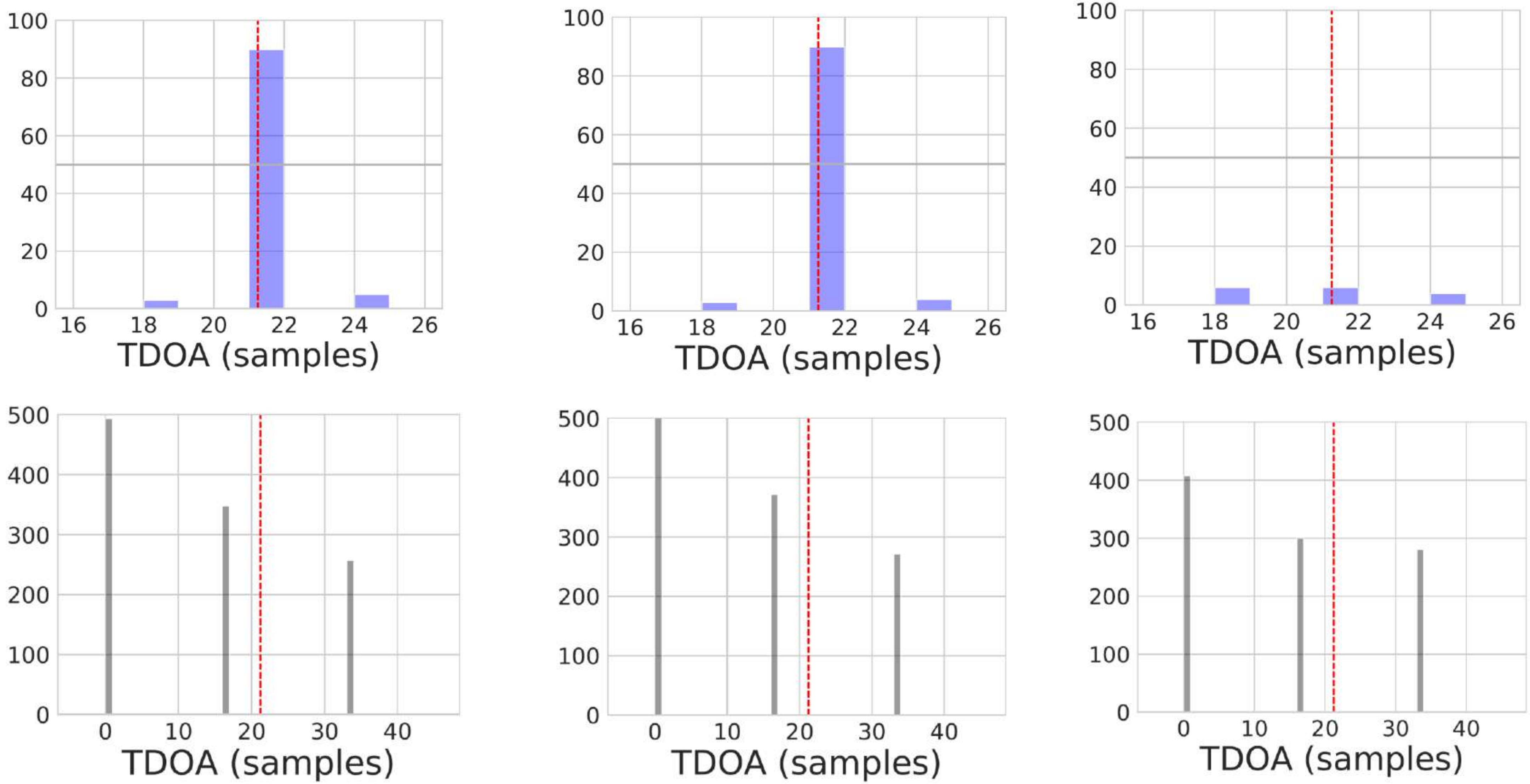
Hon et al., 2015

Fingerprinting vs Ours

reverberation T_{60} = 0 | **reverberation T_{60} = 0.1** | **reverberation T_{60} = 0.3**

Ours

Hon et al., 2015



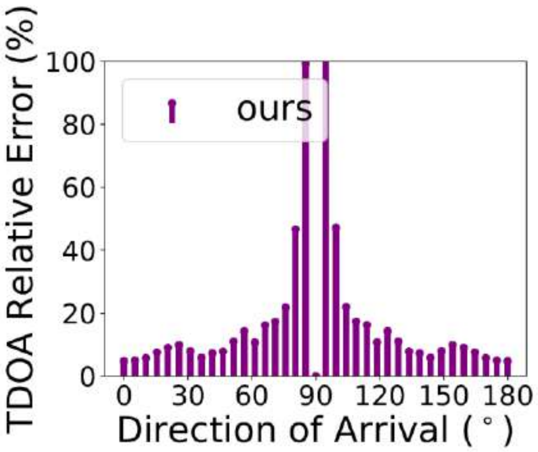
Results: 10 Different Speech Signals

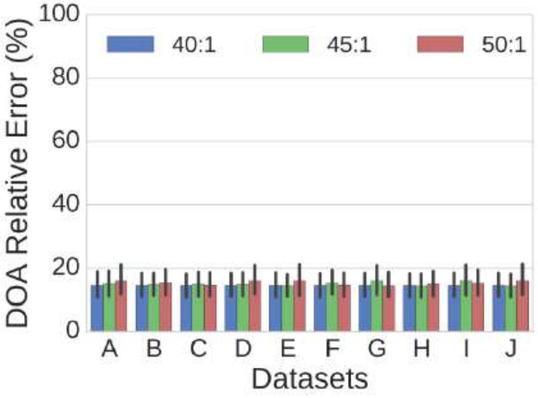
Variations of signals, reverberation, Direction of Arrival (DOA) and compression ratio

Results: 10 Different Speech Signals

Variations of signals, reverberation, Direction of Arrival (DOA) and compression ratio

reverberation T_{60} = 0.1

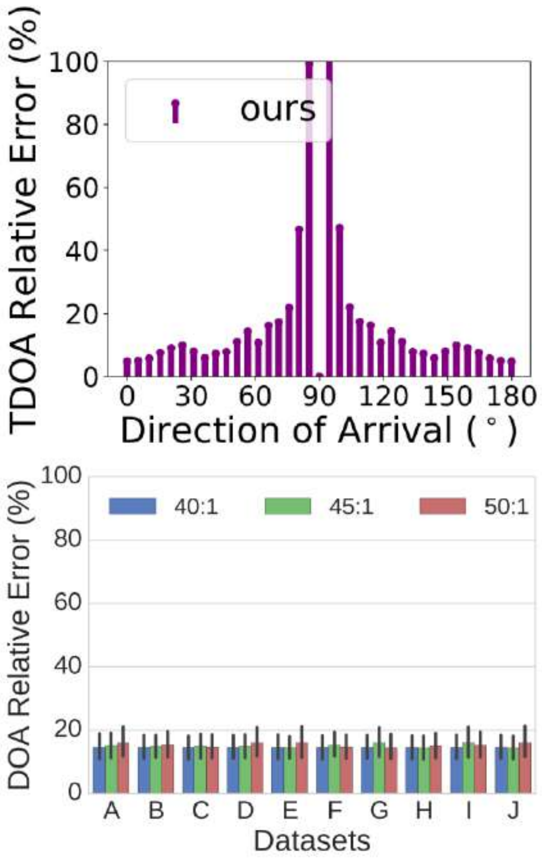




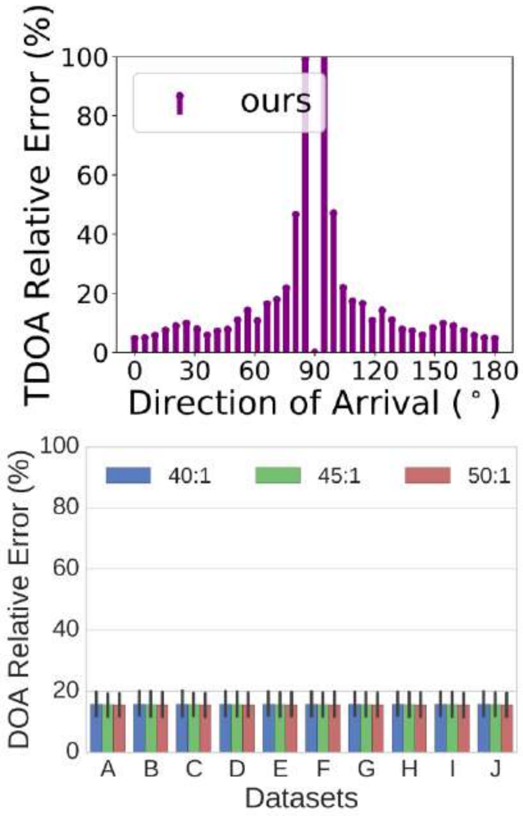
Results: 10 Different Speech Signals

Variations of signals, reverberation, Direction of Arrival (DOA) and compression ratio

reverberation T_{60} = 0.1



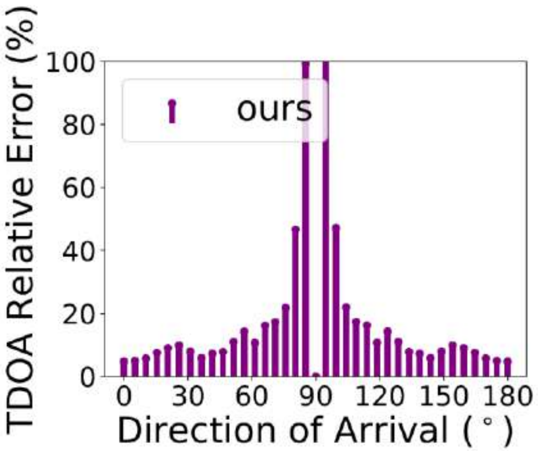
reverberation T_{60} = 0.2

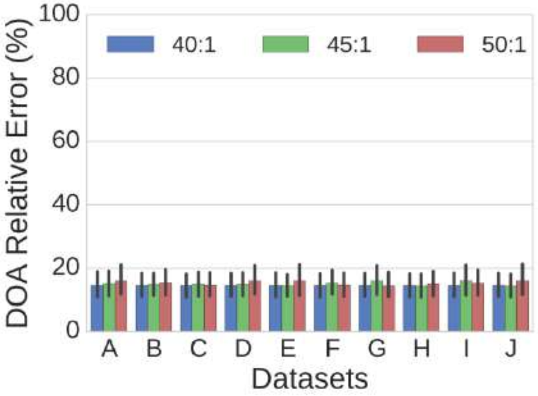


Results: 10 Different Speech Signals

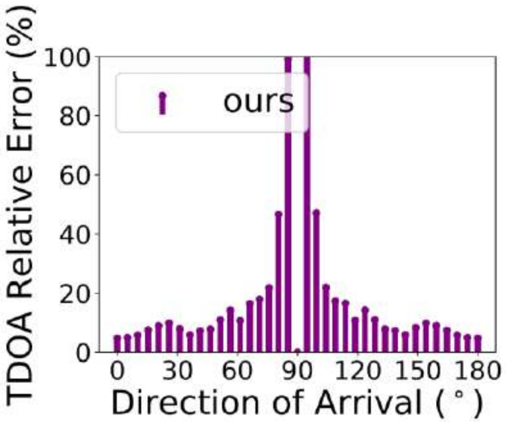
Variations of signals, reverberation, Direction of Arrival (DOA) and compression ratio

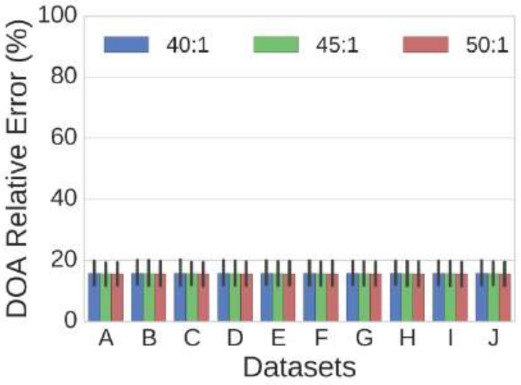
reverberation T_{60} = 0.1



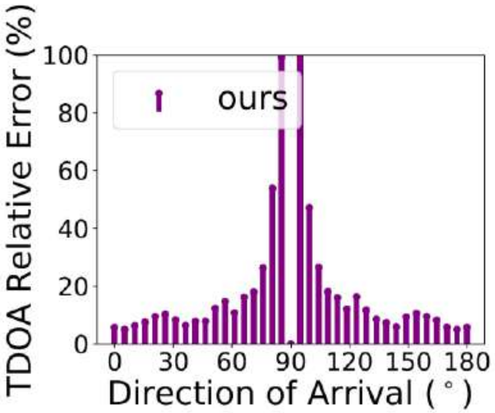


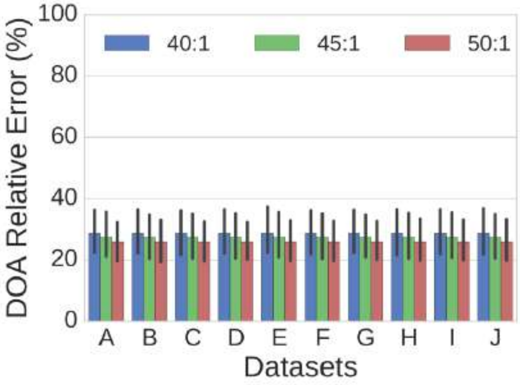
reverberation T_{60} = 0.2





reverberation T_{60} = 0.3





Summary of Contributions

1. Signal samples could be selected in order to accurately estimate TDOA/DOA, without using the entire signal.

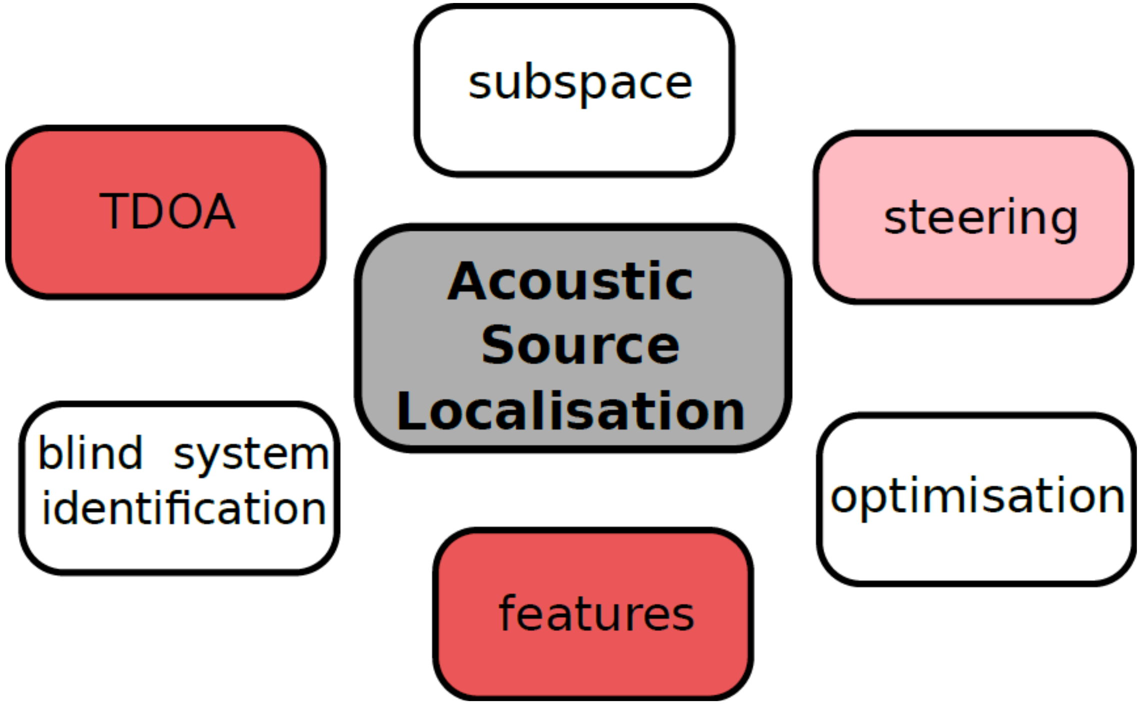
2. Computer vision techniques **(SIFT)** applied on the signal spectrogram are useful for selecting samples.

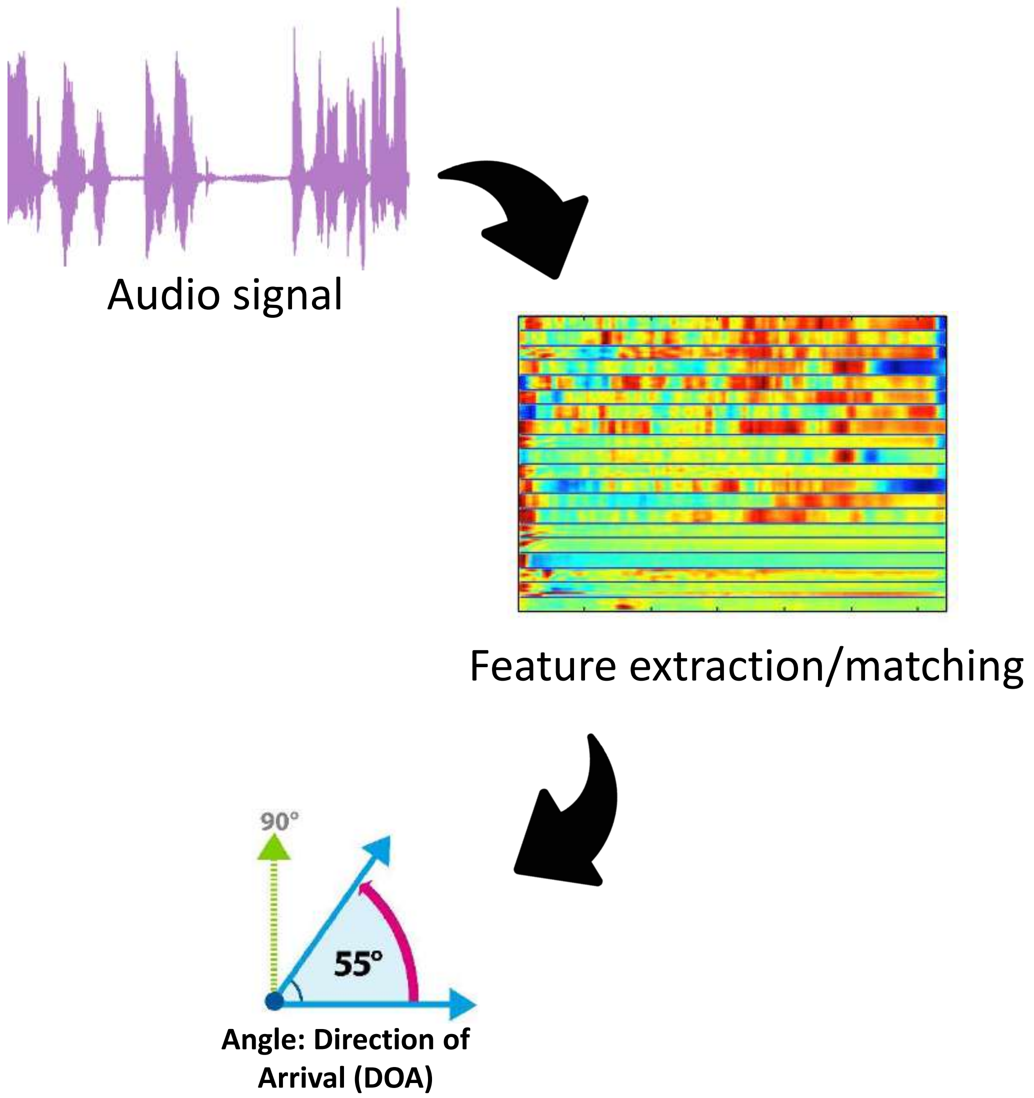
3. The proposed algorithm achieves a **compression ratio** of **40:1**

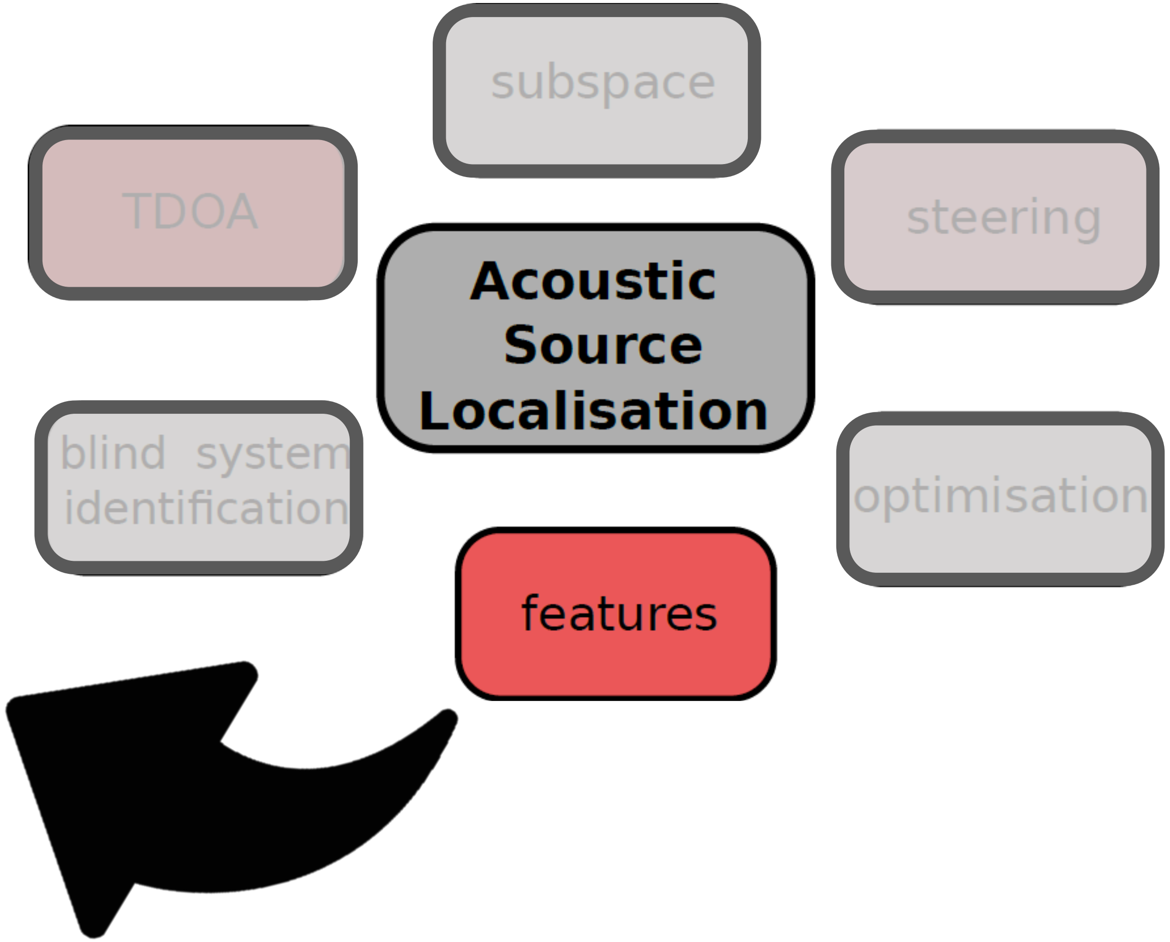
4. The algorithm outperforms the baseline, audio *fingerprinting*.

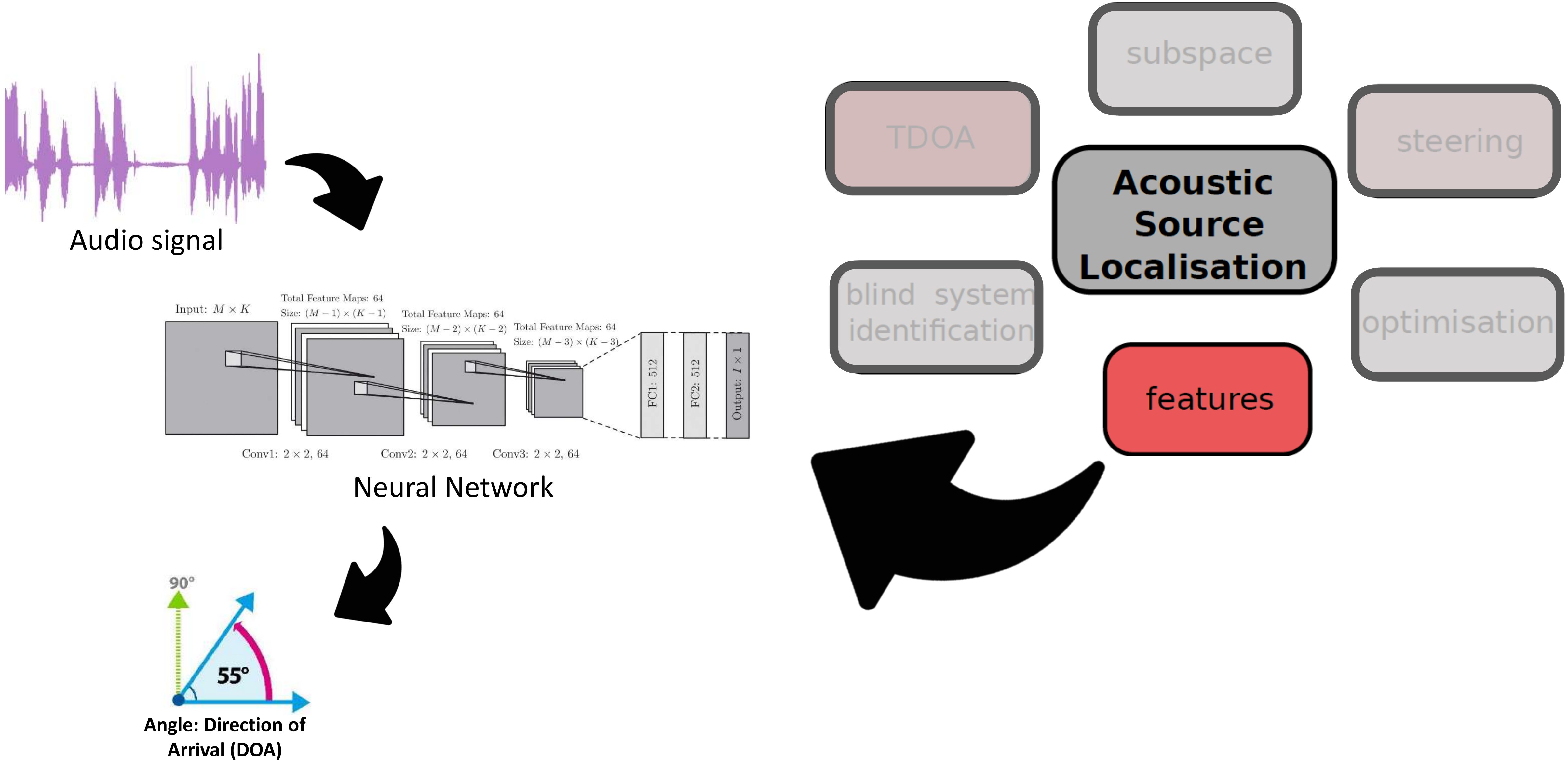
5. The proposed method is suitable for speech signals.

Contribution III
Data Available for Training (Chapter 6)


subspace
TDOA
steering
Acoustic Source Localisation
blind system identification
optimisation
features

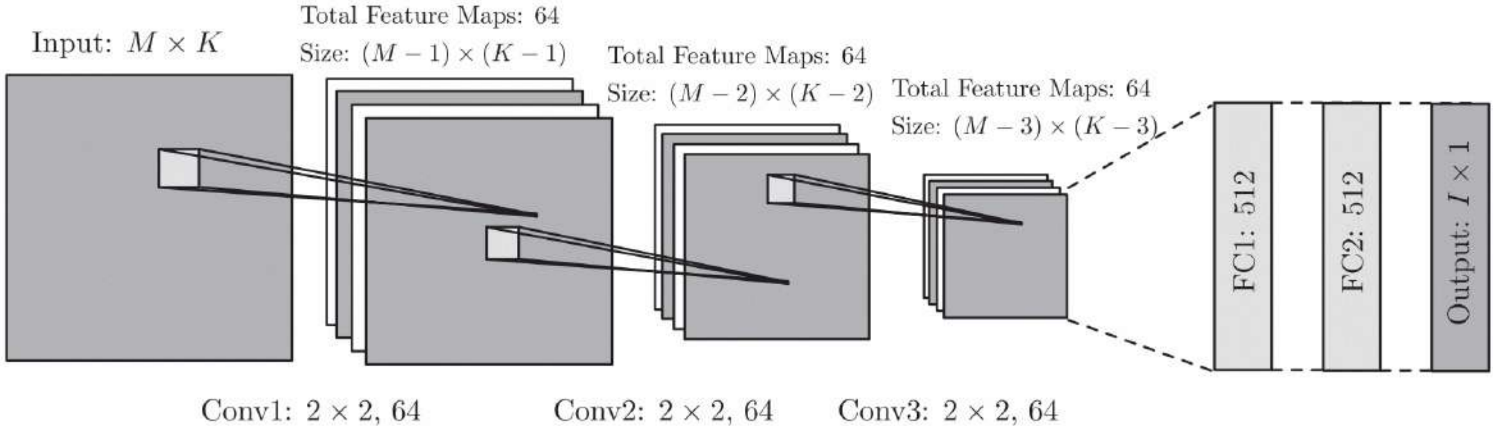

Audio signal
Feature extraction/matching
90°
55°
Angle: Direction of Arrival (DOA)


subspace
TDOA
steering
Acoustic Source Localisation
blind system identification
optimisation
features



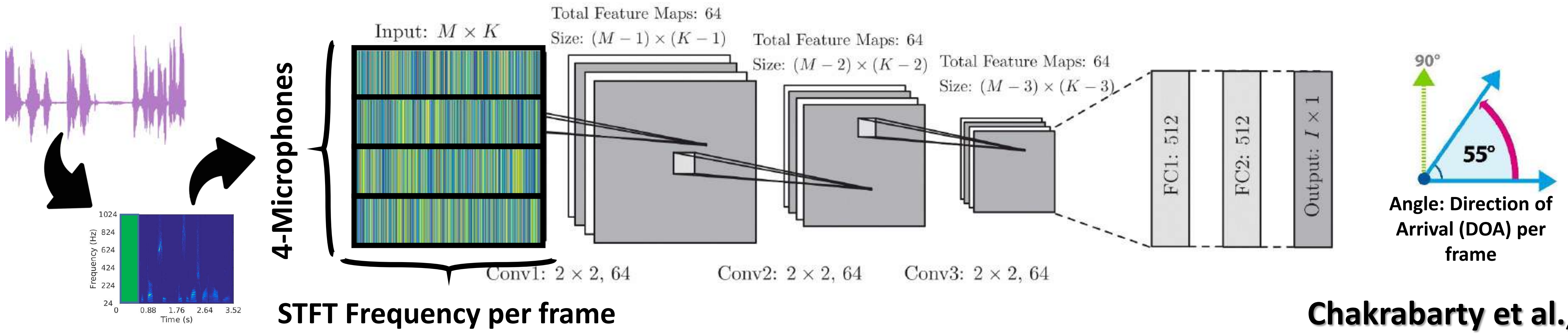
Audio signal
Input: M × K
Total Feature Maps: 64
Size: (M − 1) × (K − 1)
Total Feature Maps: 64
Size: (M − 2) × (K − 2)
Total Feature Maps: 64
Size: (M − 3) × (K − 3)
FC1: 512
FC2: 512
Output: I × 1
Conv1: 2 × 2, 64
Conv2: 2 × 2, 64
Conv3: 2 × 2, 64
Neural Network
90°
55°
Angle: Direction of Arrival (DOA)
subspace
TDOA
steering
Acoustic Source Localisation
blind system identification
optimisation
features

Baseline: CNN for DOA Estimation



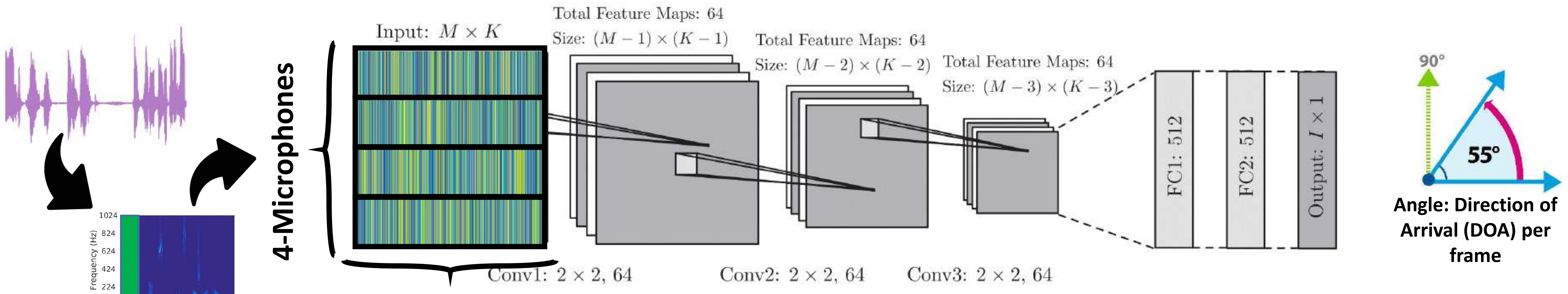
Chakrabarty et al., 2017

Baseline: CNN for DOA Estimation

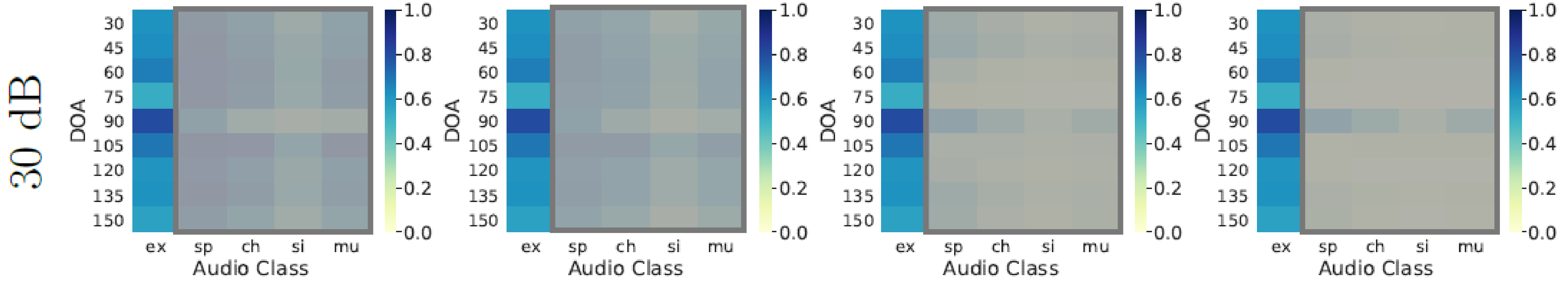


Chakrabarty et al., 2017

Baseline: Train with noise, test with speech

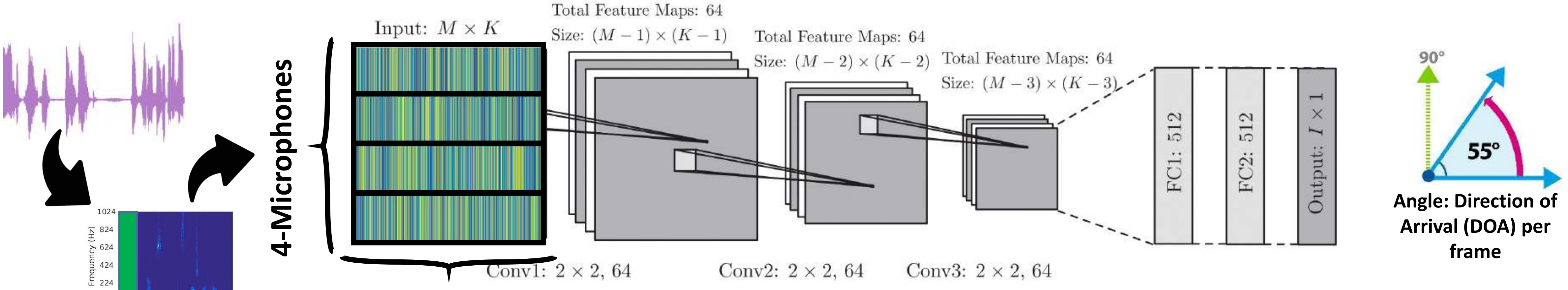


Chakrabarty et al., 2017



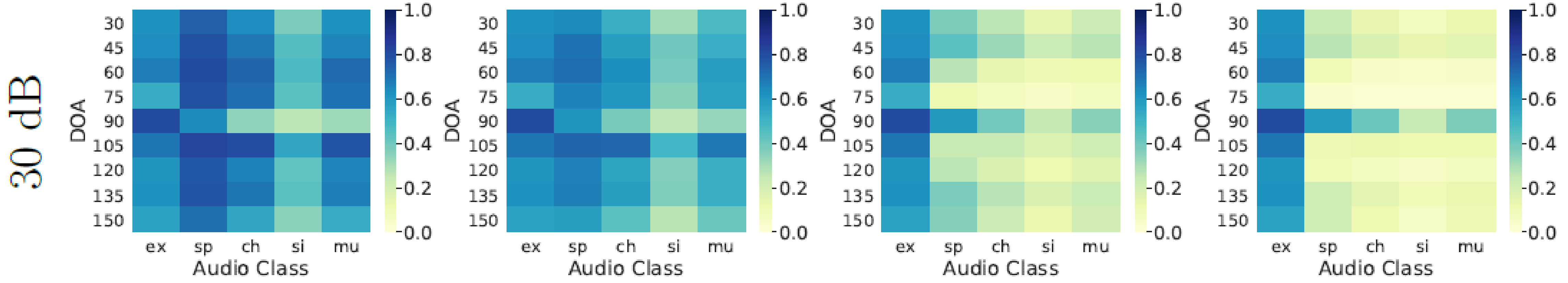
Reverb 0 s   Reverb 0.1 s   Reverb 0.2 s   Reverb 0.3 s

Baseline: Train with noise, test with speech



STFT Frequency per frame

Chakrabarty et al., 2017



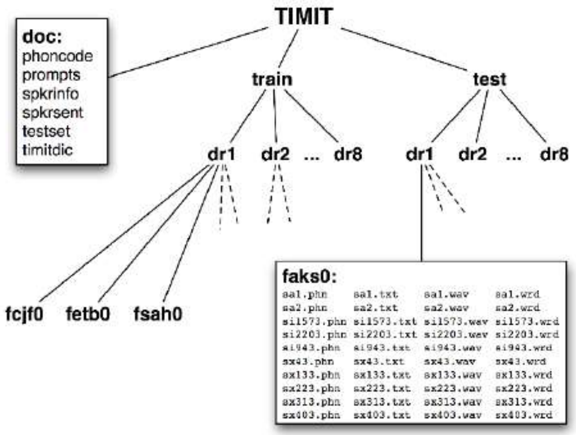
Reverb 0 s | Reverb 0.1 s | Reverb 0.2 s | Reverb 0.3 s

PROBLEM: We want to be able to obtain accurate DOA for all audio classes

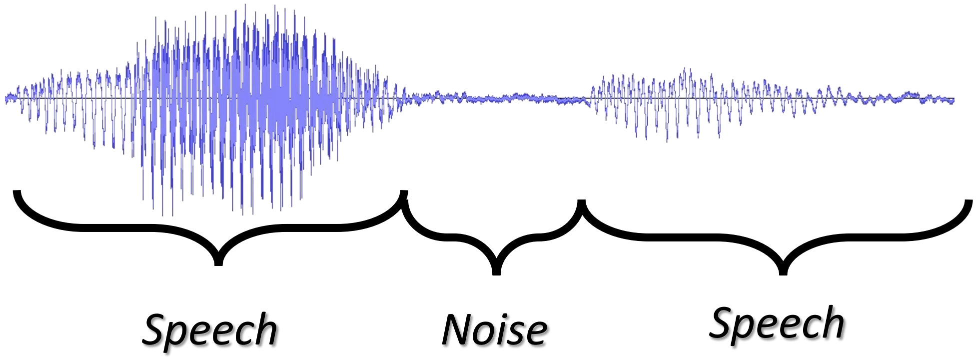
Hypothesis III

Using **speech** and **music** data for training will provide **more accurate DOA estimation** than using noise, as in the state of the art.

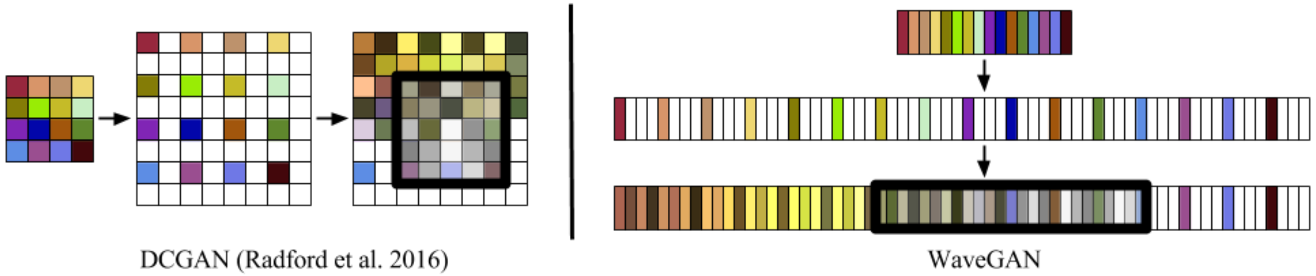
Variations of Speech and Music



Dataset



Voice Activity Detector (VAD)



Donahue et al., 2019

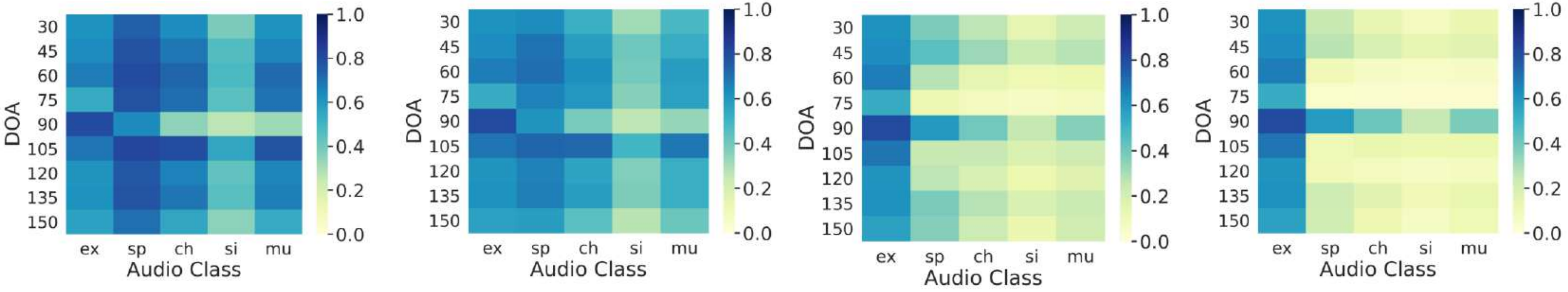
Generative Adversarial Training (GAN)

https://chrisdonahue.com/wavegan_examples/

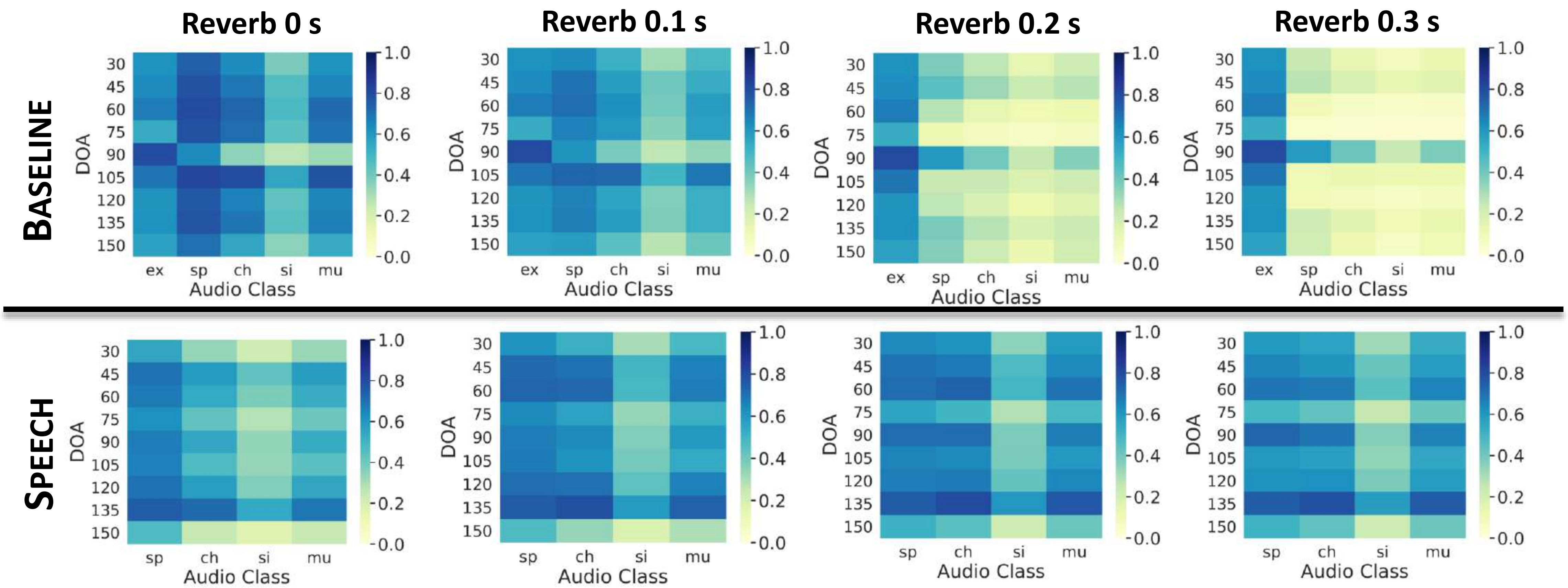
Speech vs Music vs Baseline

Reverb 0 s | Reverb 0.1 s | Reverb 0.2 s | Reverb 0.3 s

BASELINE



Speech vs Music vs Baseline



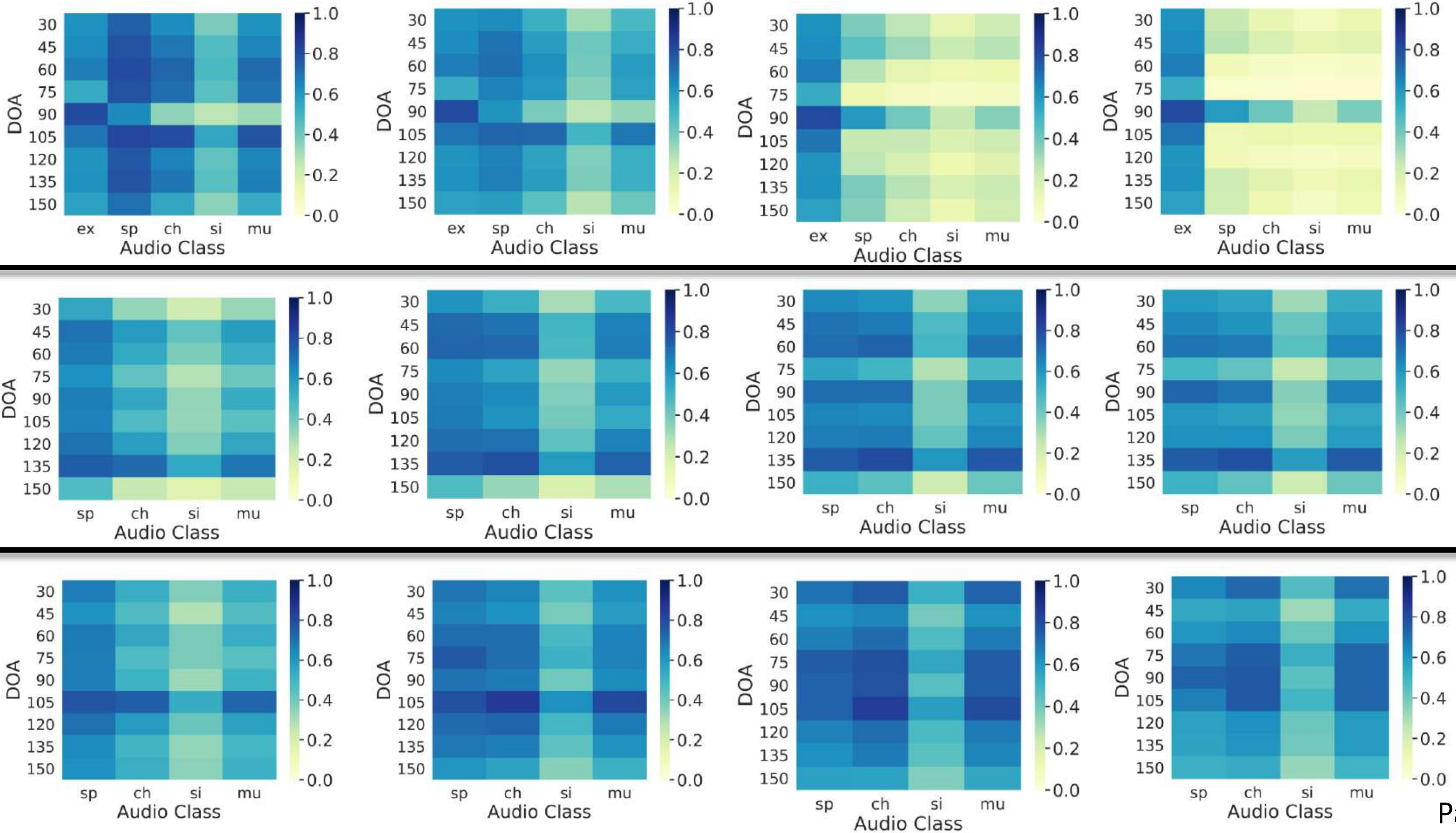
Speech vs Music vs Baseline

Reverb 0 s | **Reverb 0.1 s** | **Reverb 0.2 s** | **Reverb 0.3 s**

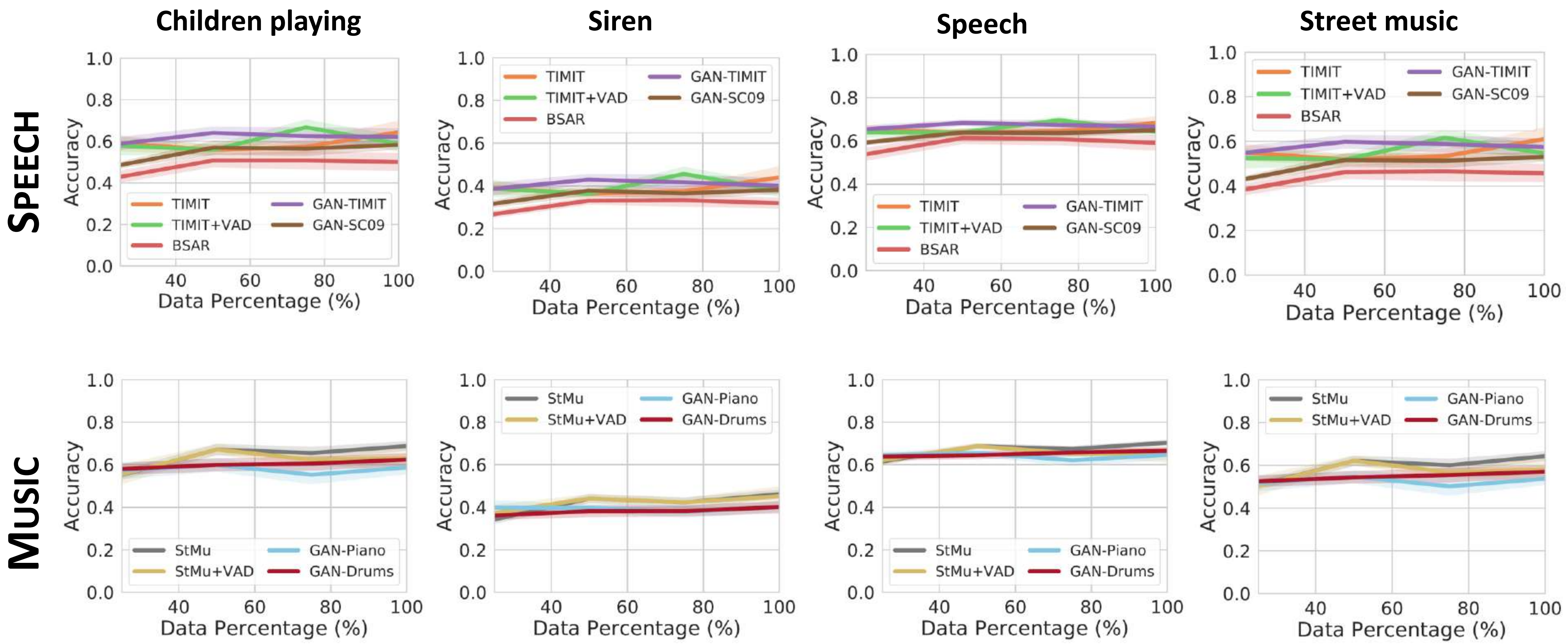
BASELINE

SPEECH

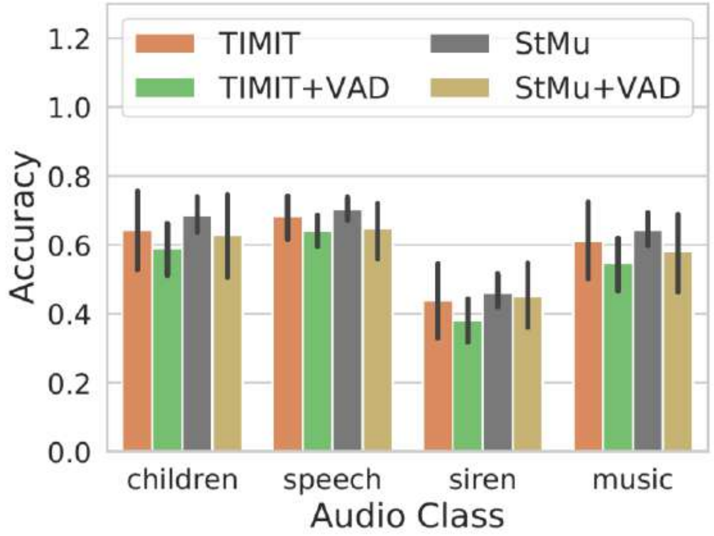
MUSIC



Reducing amount of training data



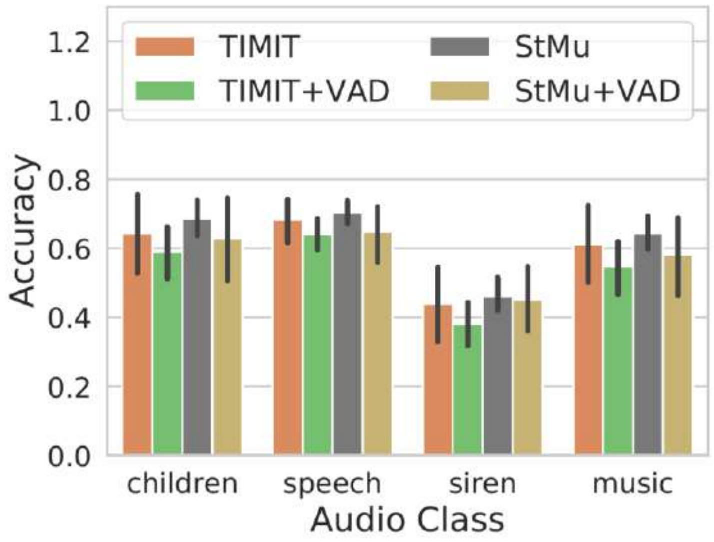
Speech vs Music



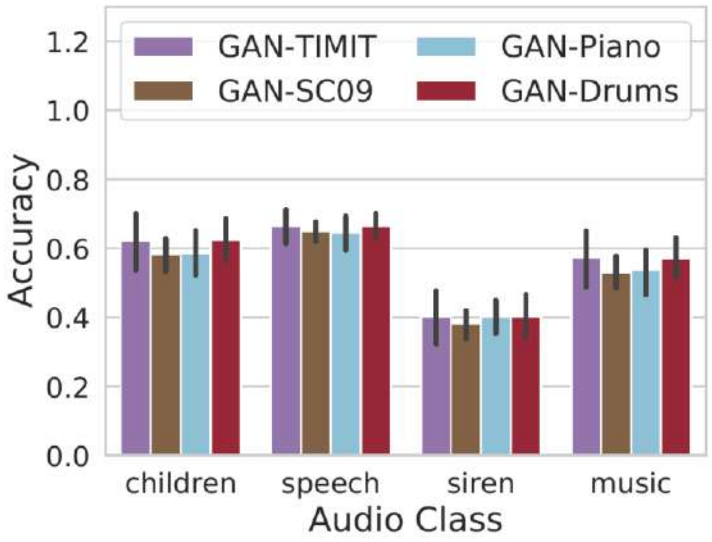
Trained with Datasets

When using data from datasets,
music is better than speech

Speech vs Music



Trained with Datasets

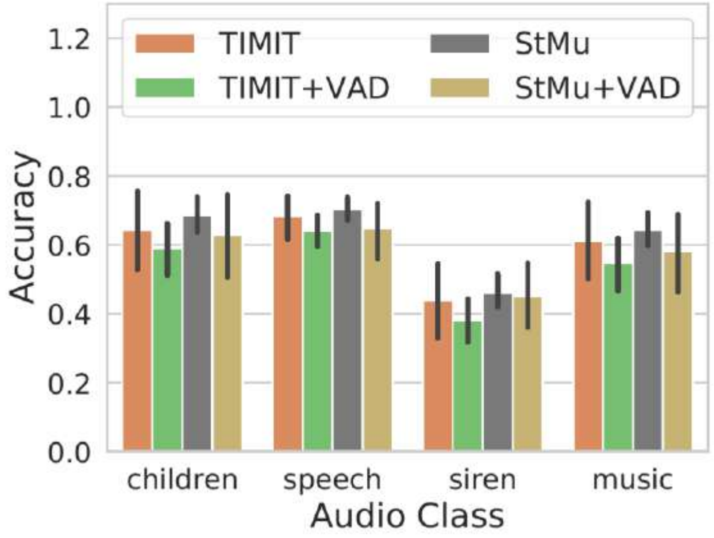


Trained with GANs

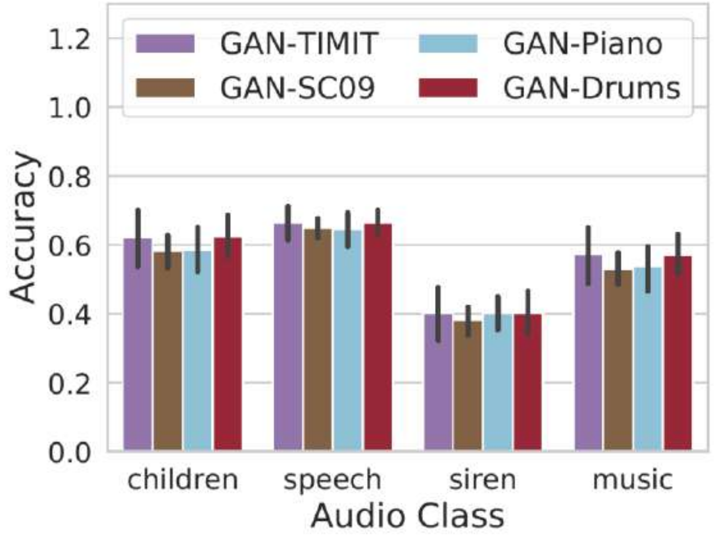
When using data from datasets, **music is better than speech**

When using data from GANs, **speech is better than music**

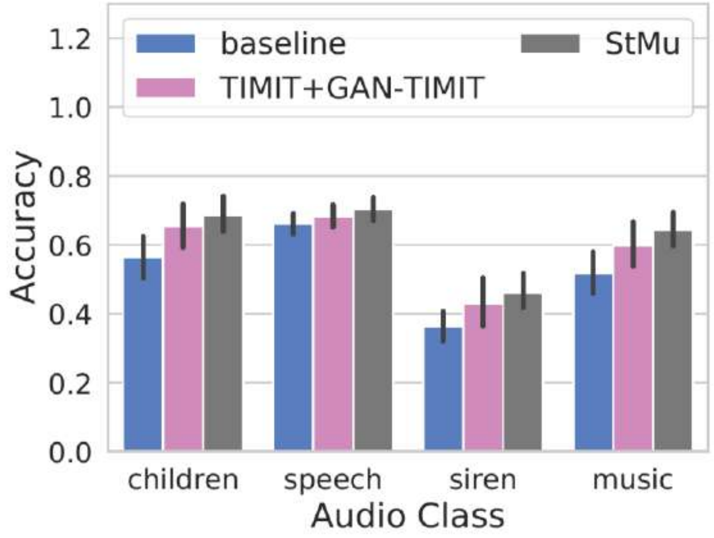
Speech vs Music



Trained with Datasets



Trained with GANs



Best vs Baseline

When using data from datasets, **music is better than speech**

When using data from GANs, **speech is better than music**

Both are better than the baseline

Summary of Contributions

1. Training a CNN with either speech or music data is an improvement over the state of the art, which uses noise for training.

2. Training with music produces an average improvement of **19%** with respect to the state of the art, while speech produces an improvement of **17%**.

3. Synthetic data generated using a **GAN** is as effective in training as using datasets.

4. Music data performs better than speech data for training when obtained using real sound recordings: however, when they are synthetically generated using a GAN, speech data produces better results than music data.

5. Using **25%** of the training data is as effective as using 100% of it.

Conclusion

This thesis presented work on **Acoustic Source Localisation (ASL) in constrained environments**. The three constraints studied were the number and configuration of sensors; the signal samples; and training data, with the main findings summarised as follows:

1. In regard to the number and configuration of sensors, accuracy can be maintained at state-of-the-art levels (SRP) while **reducing computation sixfold**.
2. In regard to signal sampling, the algorithm presented in this work outperforms an audio fingerprinting baseline while maintaining a **compression ratio of 40:1**.
3. In regard to training data, music training data is used to record an **improvement of 19%** against a noise data baseline **using only 25% of the training data.**