FINAL PROJECT

GAZE-BASED SOUND AUGMENTATION IN AUGMENTED REALITY

20233369 JUNGHOON SEO

20239002 MINJI PARK

20236318 TAMANA PIRZAD

CONTENTS

“

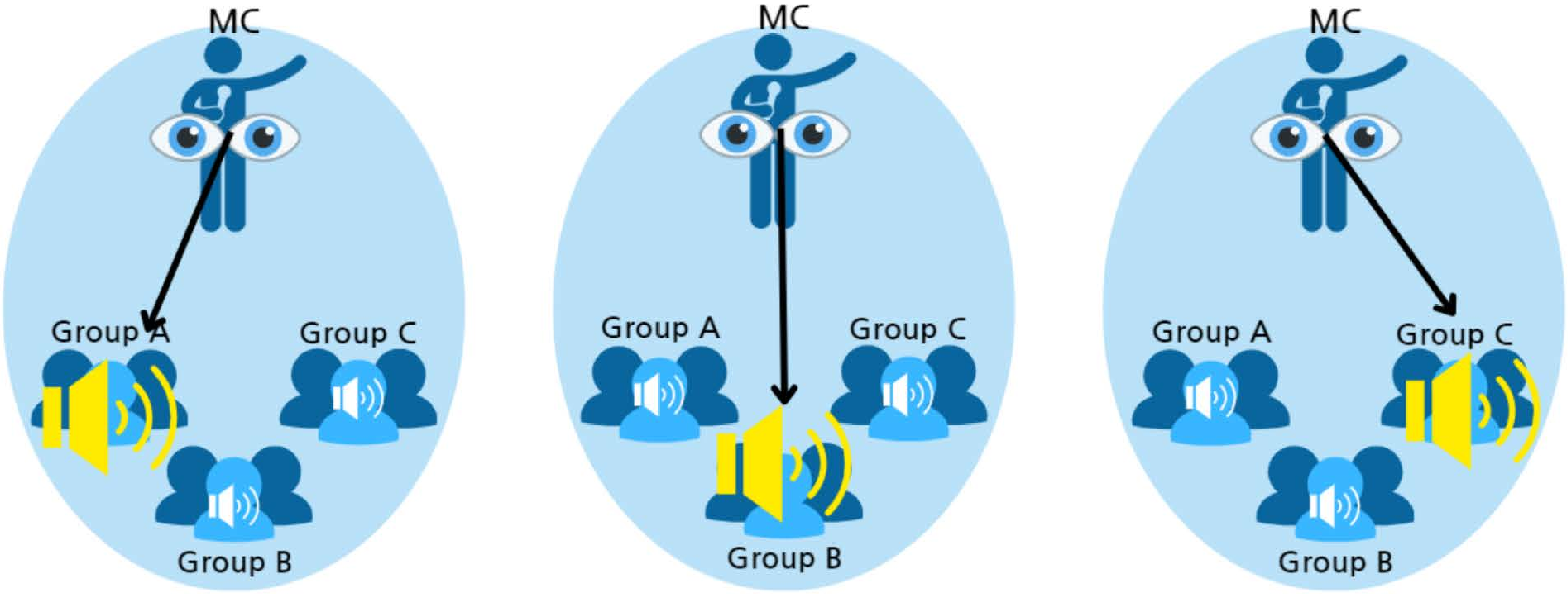
RESEARCH CONTRIBUTION

Research Objective

Literature Review

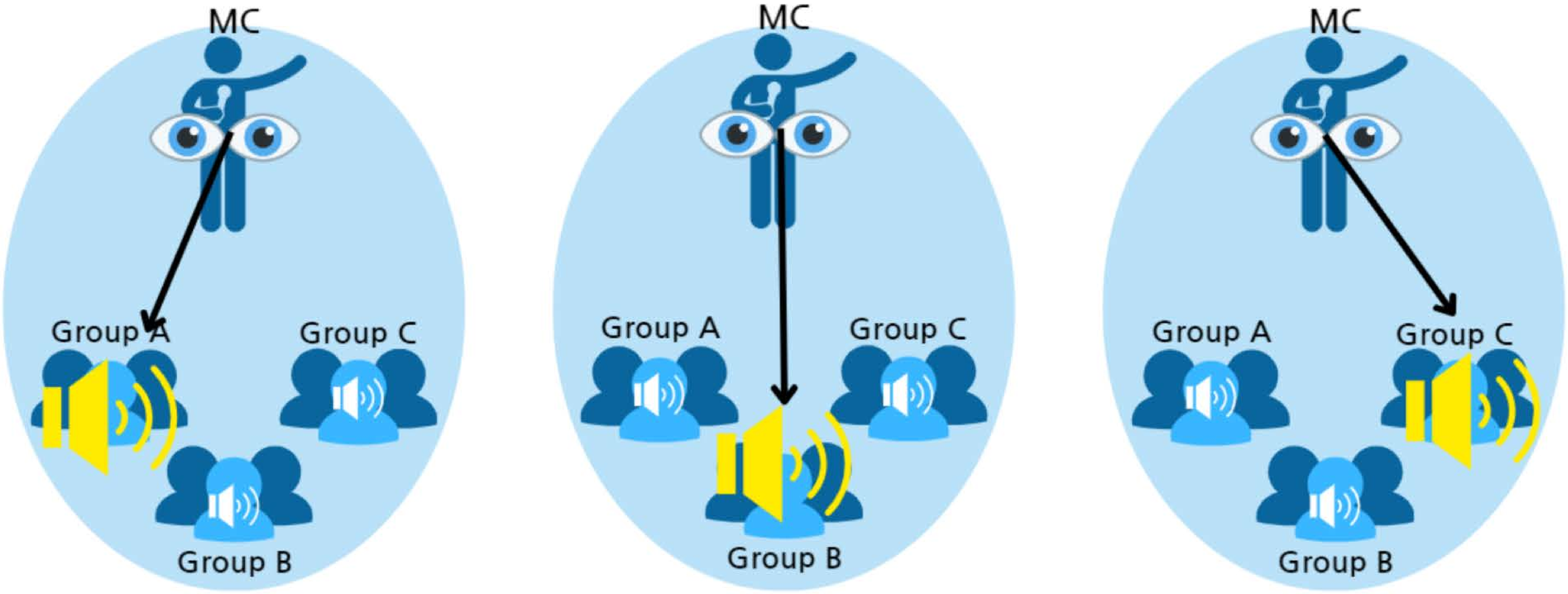
Contributions

RESEARCH OBJECTIVE



- **Goal** : To increase or decrease the sound according to the user's gaze position in real time based on gaze data

RESEARCH OBJECTIVE



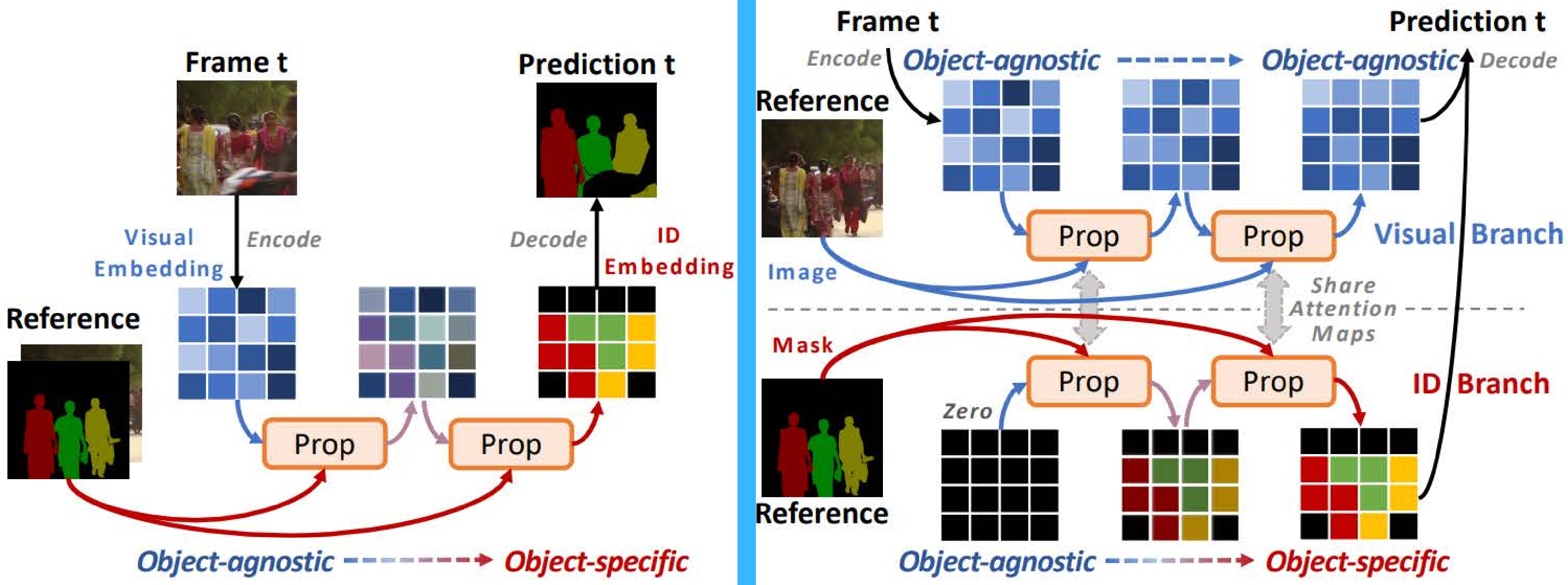
- **Grounds:** In existing reality, it is difficult to understand the conversation of a specific group due to the synthesis of various sounds

RESEARCH OBJECTIVE

- **Example**
  - Cocktail party situation
  - Group discussion
  - A situation in which a supervisor supervises the work of workers in an industrial environment
  - Rock festival performance
  - Hearing impairment, such as hearing loss

LITERATURE REVIEW - 1. DeAOT

AOT(Associative Object Tracking)



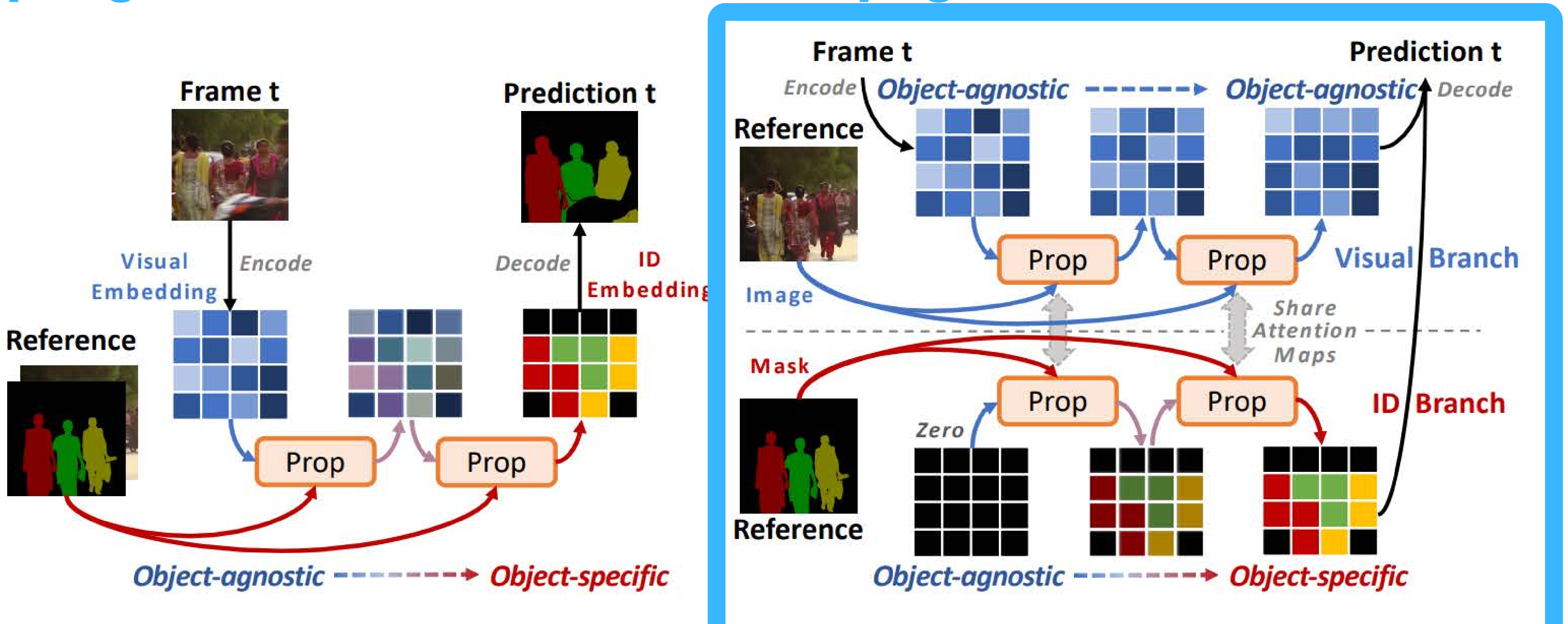
(a) AOT-like hierarchical propagation (b) Decoupling features (ours)

- AOT tracks the location and identity of objects by linking them across multiple frames of video.
- Track objects, typically using a combination of features such as color, texture, shape, or motion
- Effective at handling occlusions, scale changes, and varying lighting conditions
- Difficult when there are rapid changes in appearance or when object features are not clearly distinguishable

Decoupling Features in Hierarchical Propagation for Video Object Segmentation. Zongxin Yang, Yi Yang. 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

LITERATURE REVIEW - 1. DeAOT

DeAOT(Decoupling Feature in Hierarchical Propagation)



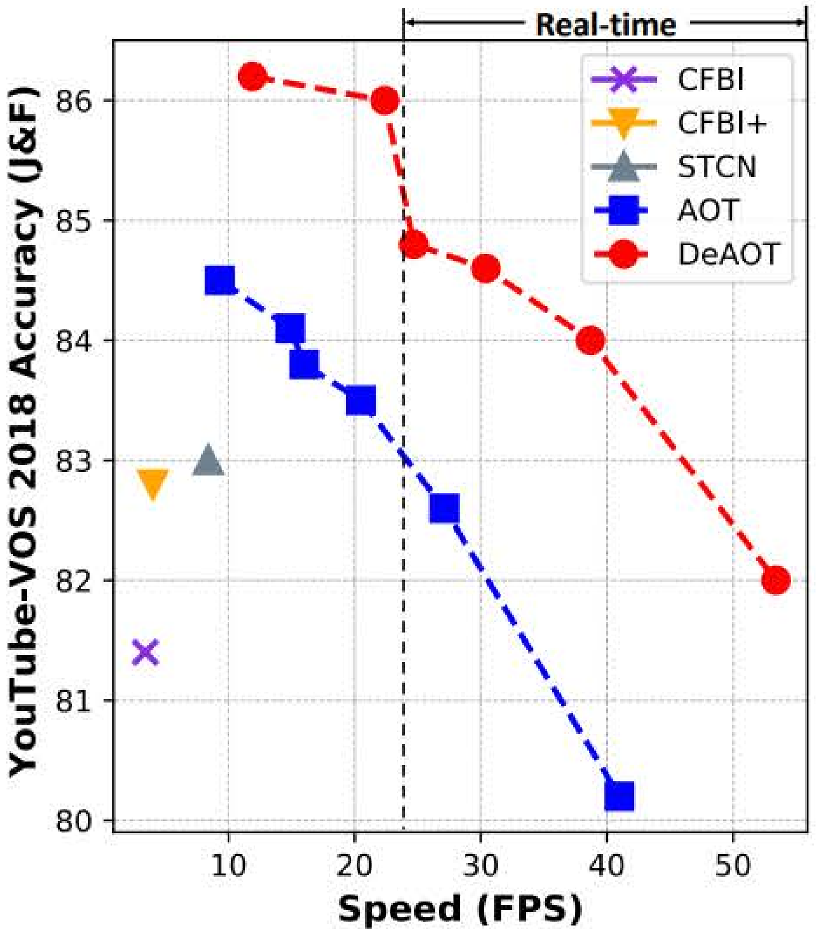
(a) AOT-like hierarchical propagation (b) Decoupling features (ours)

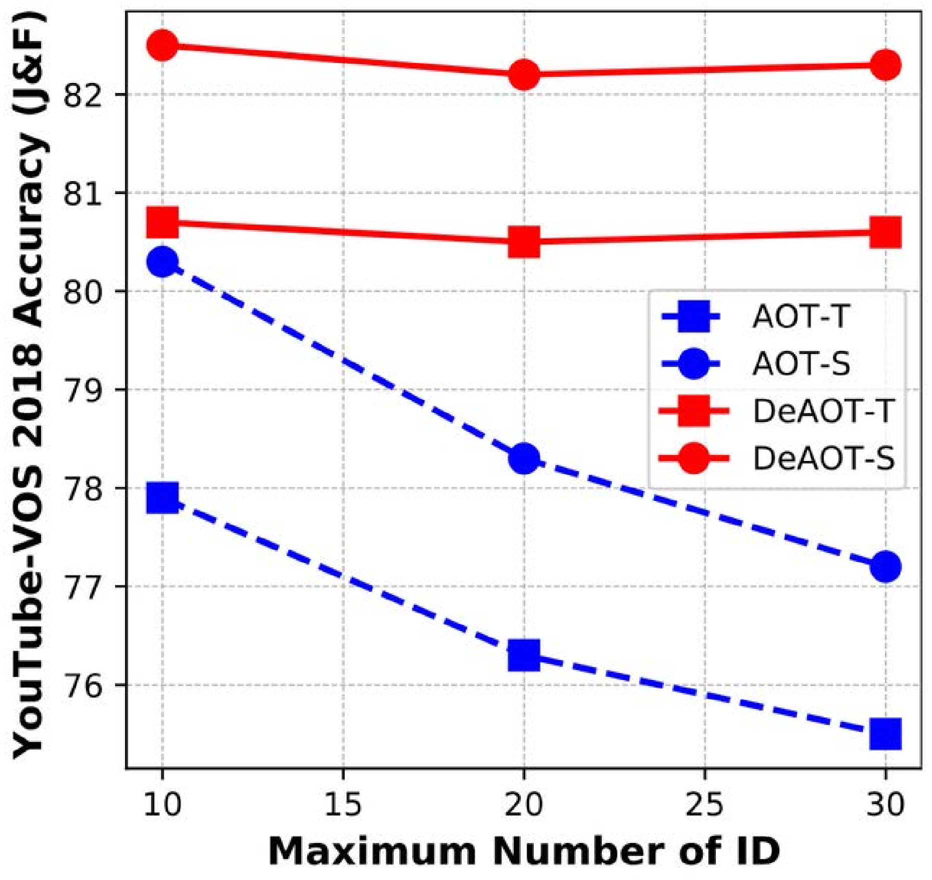
- DeAOT tracks across multiple hierarchical levels. At each level, different aspects of the object (e.g. shape, color, texture) are analyzed separately.
- AOT associates features linearly across the frame, while DeAOT separates features and processes them hierarchically, making it more flexible for handling complex changes → Excellent for fast movements

Decoupling Features in Hierarchical Propagation for Video Object Segmentation. Zongxin Yang, Yi Yang. 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

LITERATURE REVIEW - 1. DeAOT

Difference between AOT and DeAOT



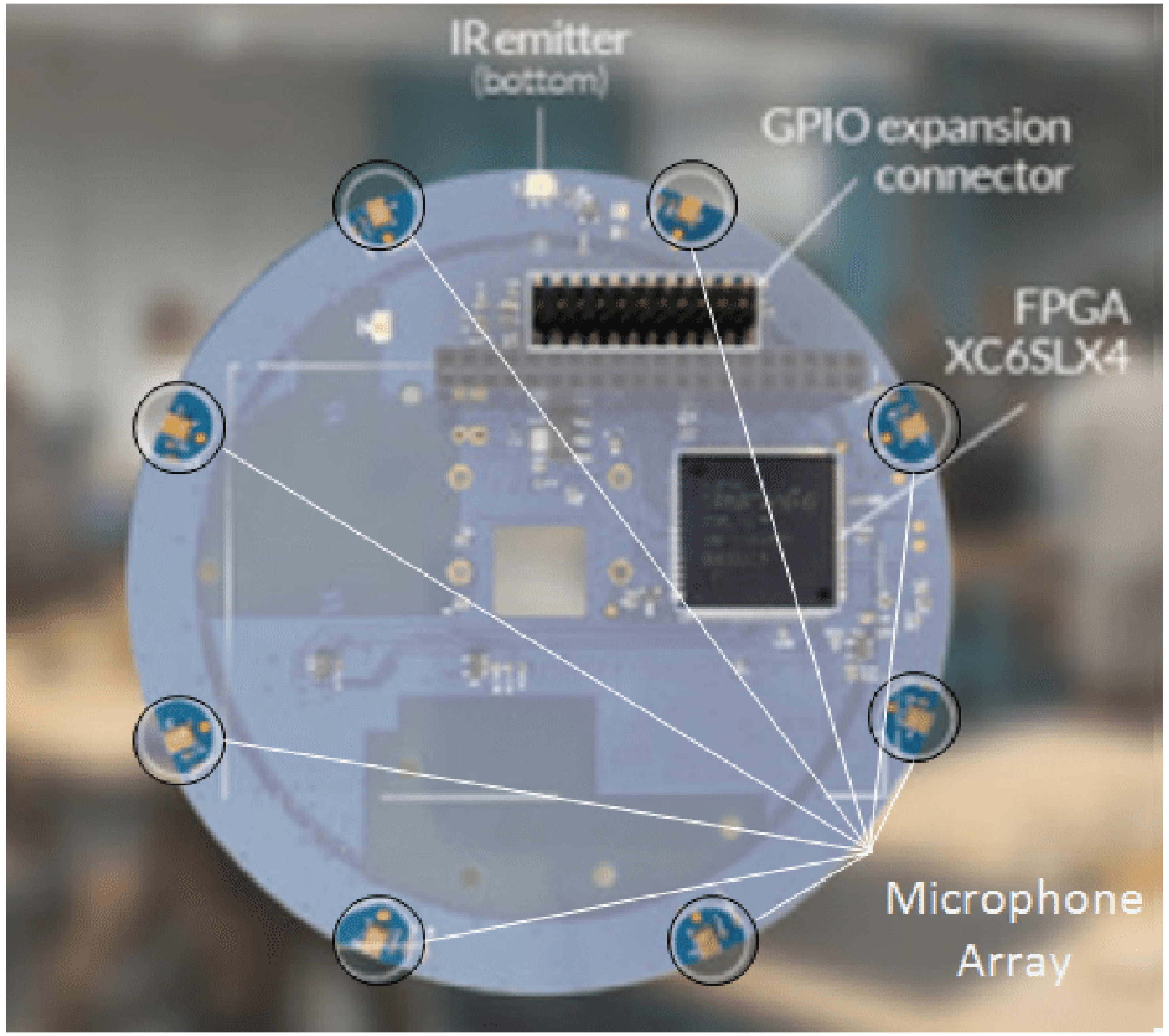


- On YouTube-VOS, DeAOT outperforms AOT in both accuracy and speed, achieving up to 86.0% at 22.4fps and 82.0% at 53.4fps
- The performance of AOT will be degraded by increasing ID's maximum number, but DeAOT doesn't show much difference

Decoupling Features in Hierarchical Propagation for Video Object Segmentation. Zongxin Yang, Yi Yang. 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

LITERATURE REVIEW - 1. DeAOT

NOVELTY - Difference between DeAOT and Our Work



- Strengthening the performance of the technology presented through DeAOT by integrating analysis through DeAOT into analysis through SAM
- DeAOT combines the SRP-PHAT-HSDA algorithm for audio signal processing (SSL, sound source localization) with visual analysis through SAM
- Presents the possibility of using DeAOT in the field of augmented hearing

Decoupling Features in Hierarchical Propagation for Video Object Segmentation. Zongxin Yang, Yi Yang. 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

LITERATURE REVIEW - 2. AV-SAM

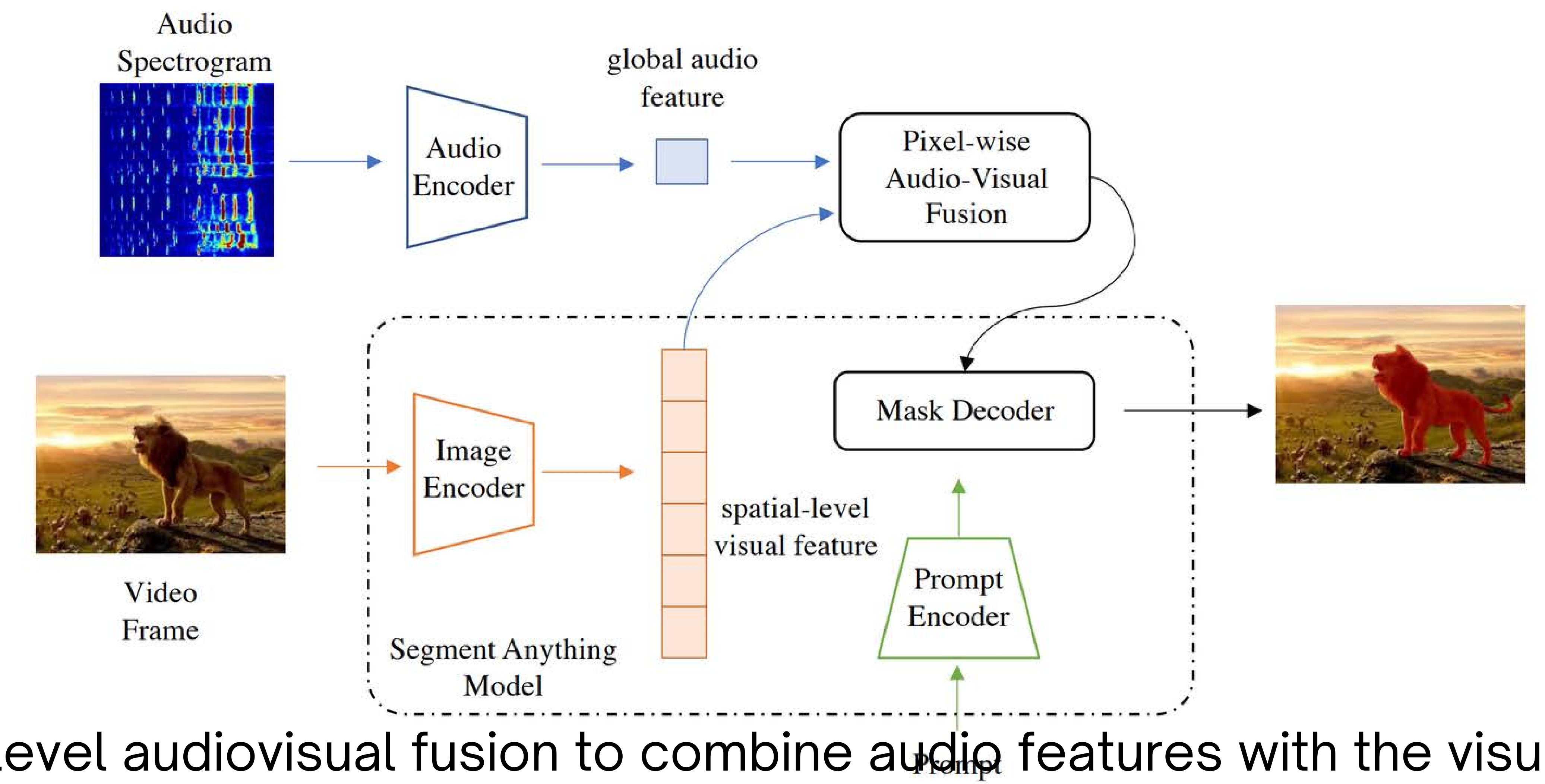
SAM(Segment Anything Model)

- It is often based on deep learning, and uses neural networks to analyze and segment various objects
- Utilizes advanced neural network architectures, such as Convolutional Neural Networks (CNNs), which are typically adept at processing image data.
- Usage: autonomous driving, medical imaging, and robotic vision

AV-SAM: Segment Anything Model Meets Audio-Visual Localization and Segmentation. Shentong Mo , Yapeng Tian. arXiv:2305.01836v1 [cs.CV] 3 May 2023

LITERATURE REVIEW - 2. AV-SAM

AV-SAM(Audio-Visual Segment Anything Model)

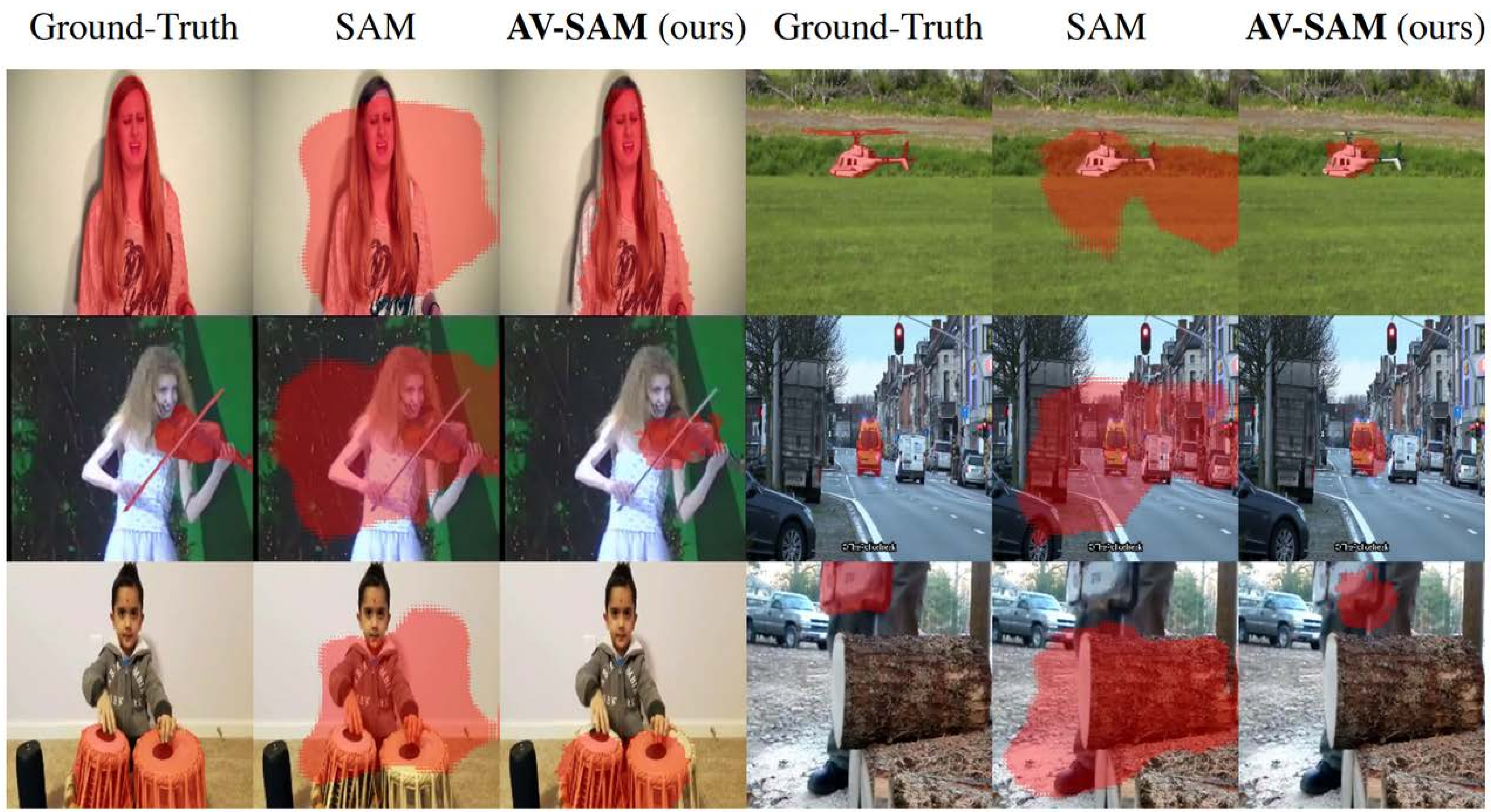


- Leverages pixel-level audiovisual fusion to combine audio features with the visual features of SAM's pre-trained image encoders
- A cross-modal representation is generated and fed into a prompt encoder and mask decoder to produce an audiovisual segmentation mask

AV-SAM: Segment Anything Model Meets Audio-Visual Localization and Segmentation. Shentong Mo , Yapeng Tian. arXiv:2305.01836v1 [cs.CV] 3 May 2023

LITERATURE REVIEW - 2. AV-SAM

SAM and AV-SAM

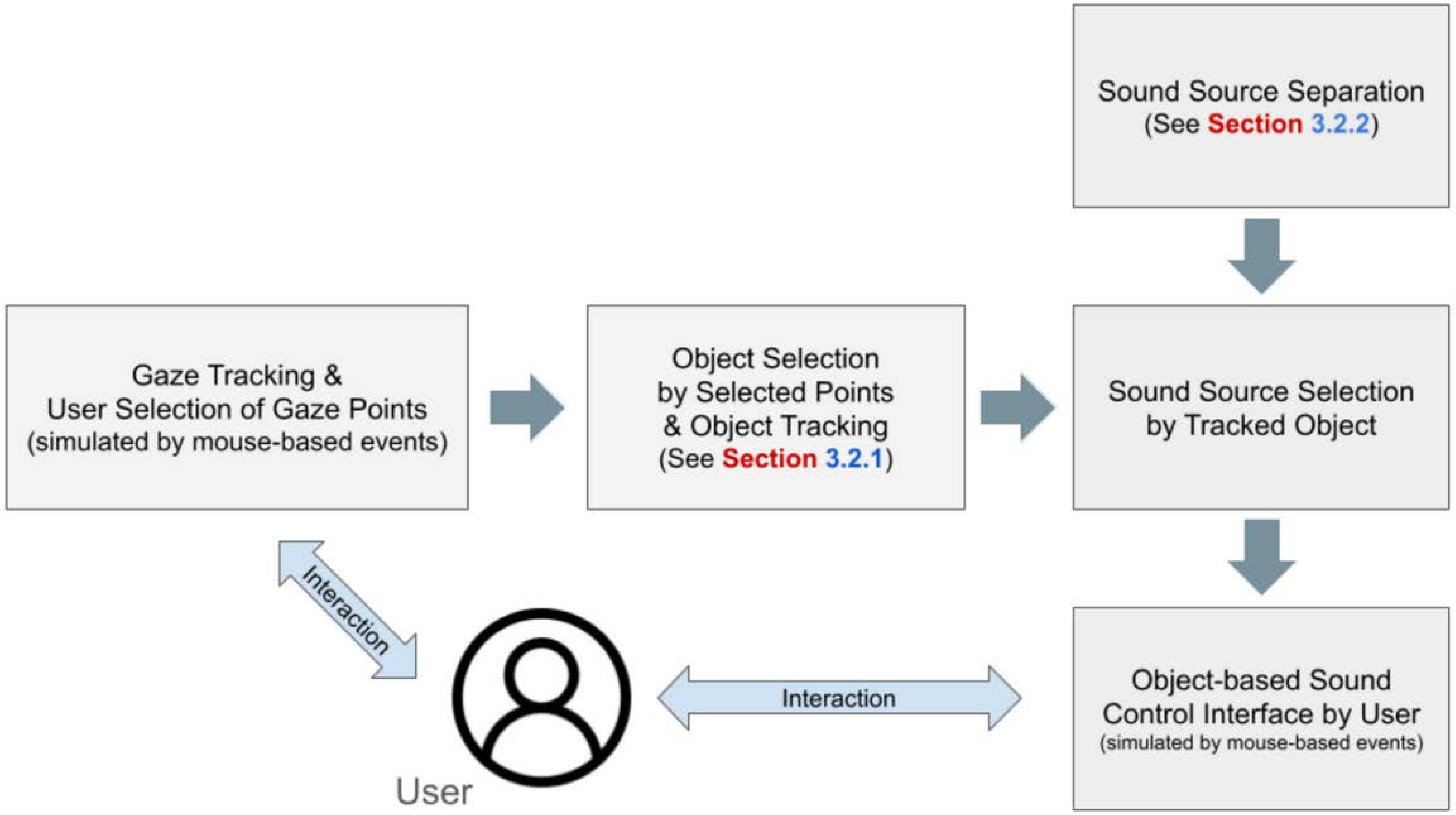


- SAM performs worse for sound objects
  - given an image of a girl playing violin, the baseline model tends to predict the mask across both the girl and the violin
- AV-SAM has been tested on Flickr-SoundNet and AVSBench datasets and has competitive performance in sound object localization and segmentation

AV-SAM: Segment Anything Model Meets Audio-Visual Localization and Segmentation. Shentong Mo , Yapeng Tian. arXiv:2305.01836v1 [cs.CV] 3 May 2023

LITERATURE REVIEW - 2. AV-SAM

NOVELTY - Difference between AV-SAM and Our work



- Strengthening the performance of the technology presented through DeAOT by integrating analysis through DeAOT into analysis through SAM
- Combining the more powerful SRP-PHAT-HSDA algorithm as a method for audio signal processing (SSL, sound source localization)

Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

LITERATURE REVIEW - 3.SRP-PHAT-HSDA(SSL)

SRP-PHAT-HSDA algorithm

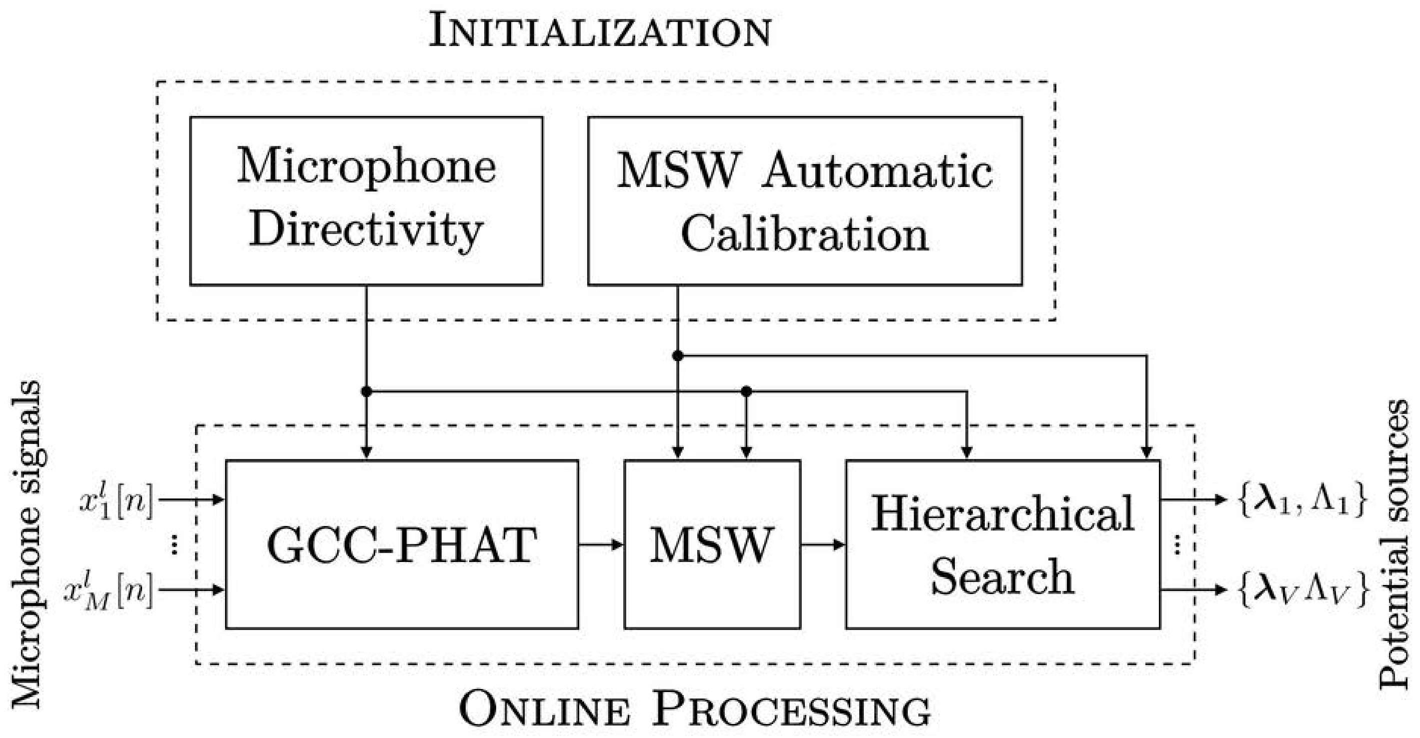


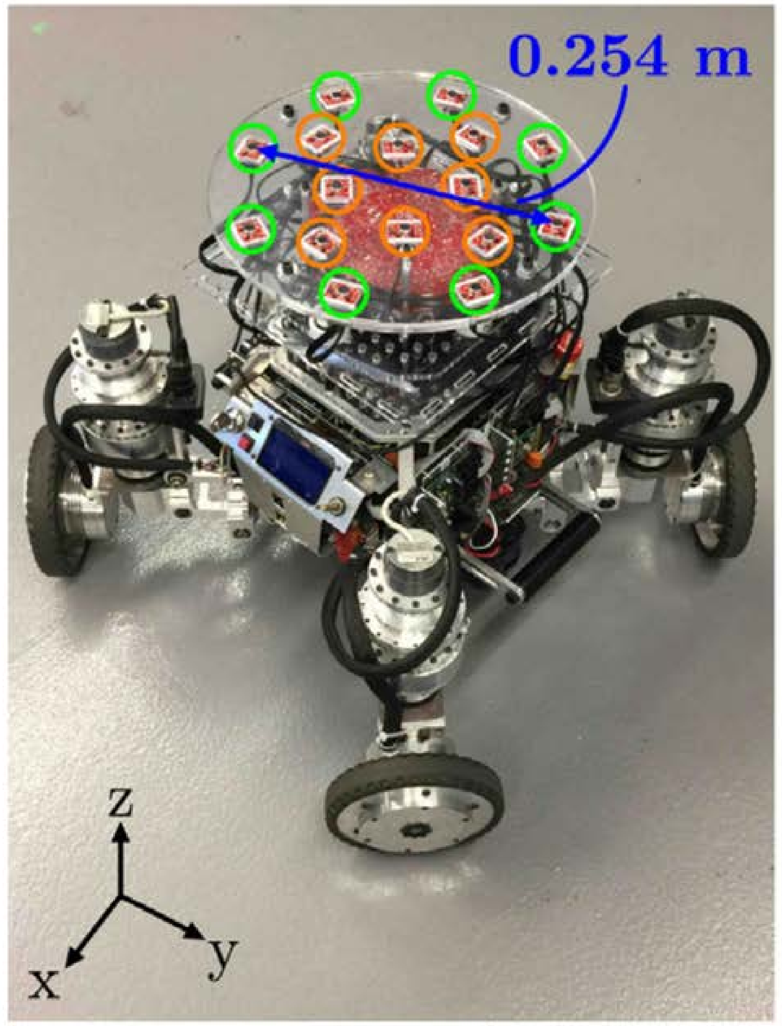
Fig. 2. Block diagram of SRP-PHAT-HSDA.

- **Steering Response Power (SRP)**
  - Calculates the power of a sound signal as if it originated from each location → Estimates the actual source location by identifying the point that produces the maximum response power
- **PHAT (Phase Transformation)**
  - Applied as a weight function to the sound source localization algorithm to improve accuracy
- **High-Resolution Spectral Density Analysis (HSDA)**
  - Improves accuracy in multi-source environments by analyzing signals at higher spectral resolution

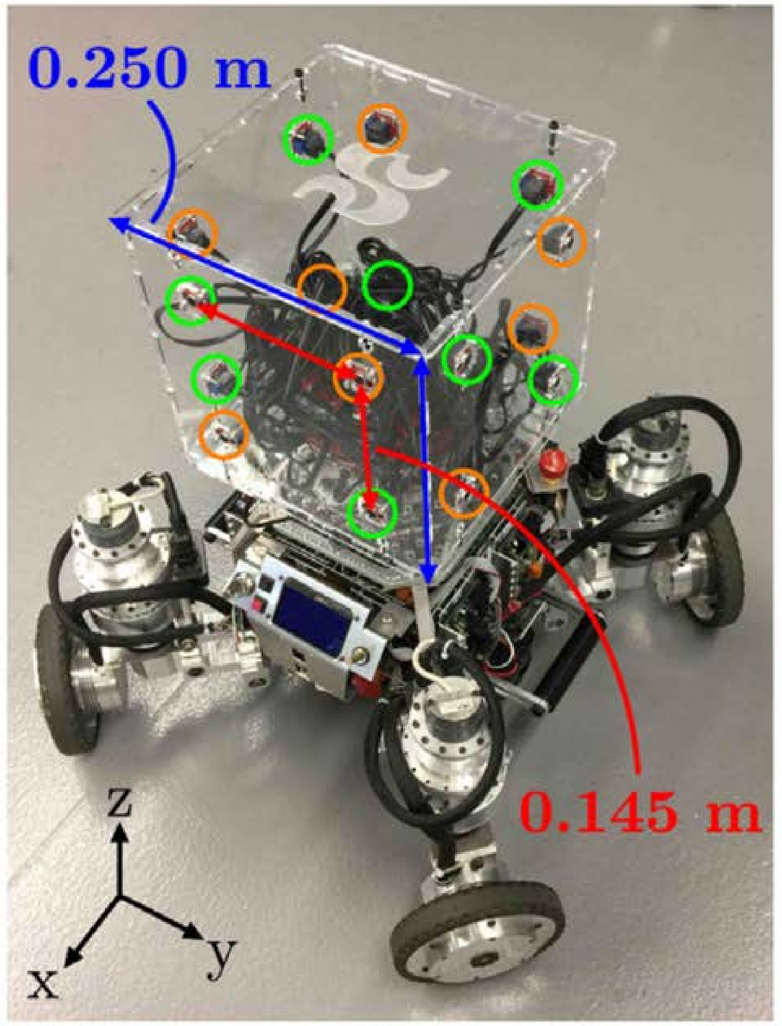
Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

LITERATURE REVIEW - 3.SRP-PHAT-HSDA(SSL)

SRP-PHAT-HSDA algorithm



(a) OMA



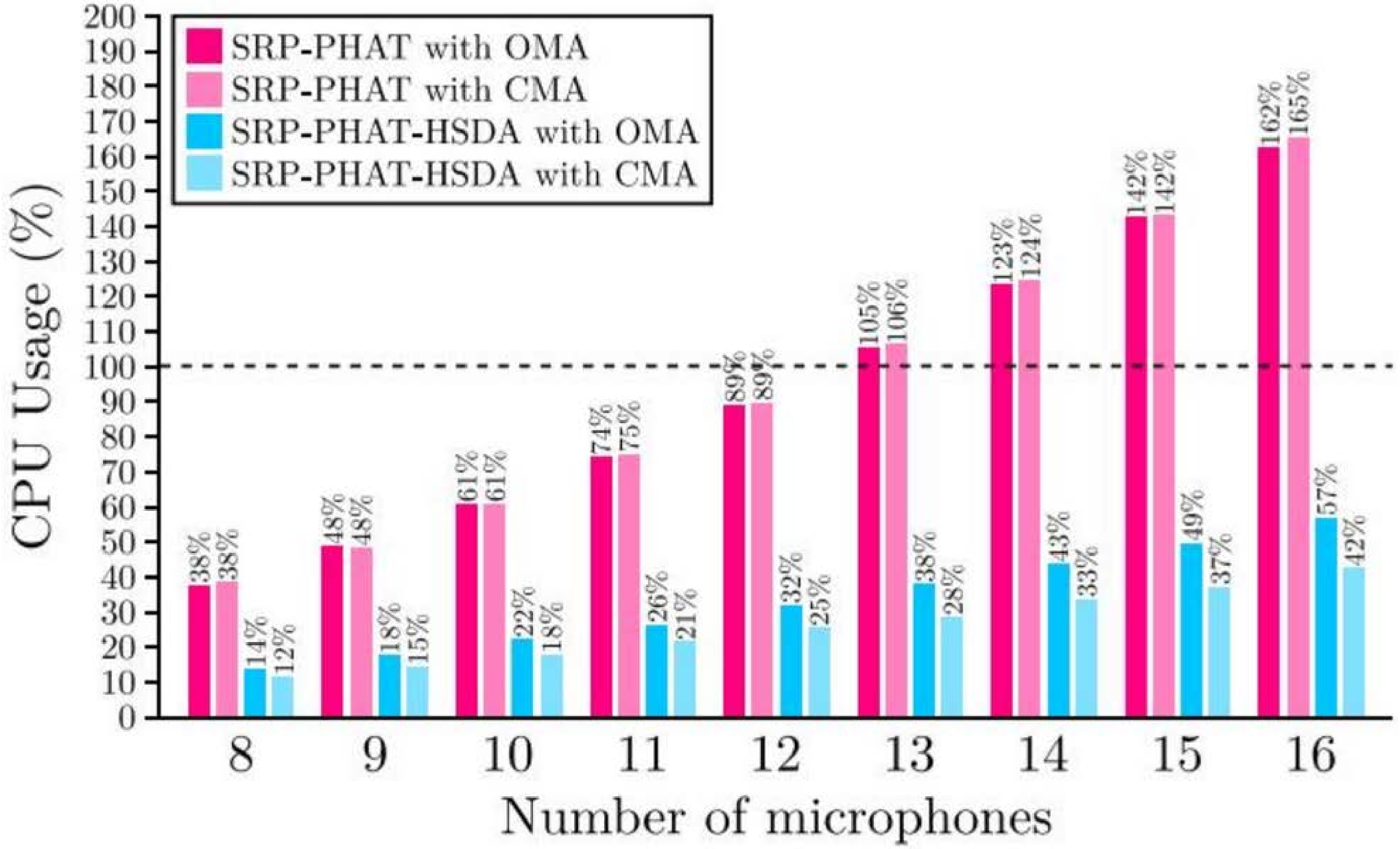
(b) CMA

- **The SRP-PHAT-HSDA algorithm combines PHAT weighting with a tuned response power approach and integrates high-resolution spectral analysis**
- Powerful and accurate technologies mainly used in the field of audio signal processing, especially sound source localization (SSL)
- Examples: audio surveillance, robotic hearing systems

Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

LITERATURE REVIEW - 3.SRP-PHAT-HSDA(SSL)

SRP-PHAT-HSDA algorithm

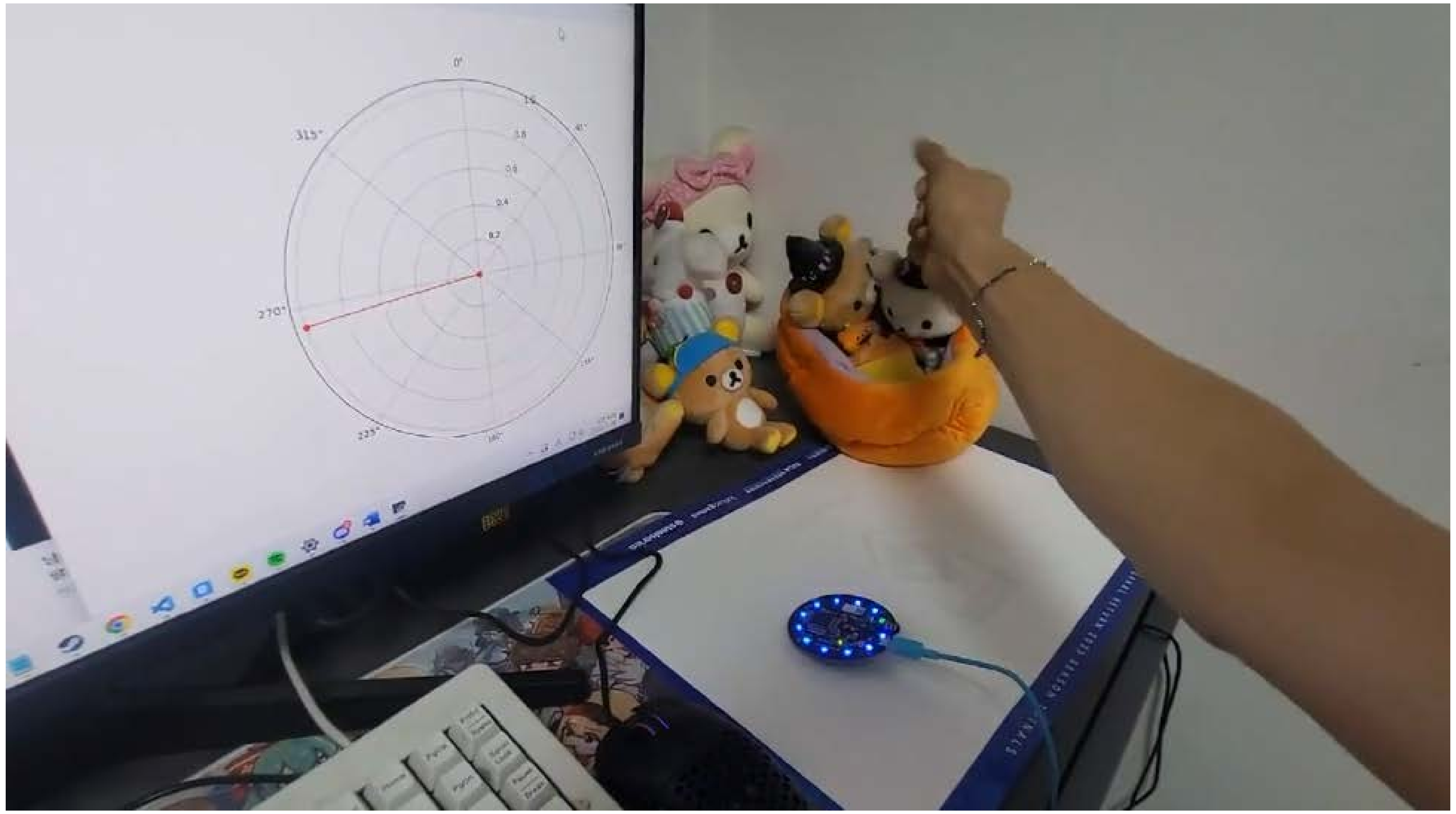


- SRP-PHAT-HSDA significantly reduces computational load due to microphone directivity model that ignores non-critical microphone pairs
- Particularly suitable for mobile robots in human-robot interaction and offers strong noise resistance and low computational cost.

Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

LITERATURE REVIEW - 3.SRP-PHAT-HSDA(SSL)

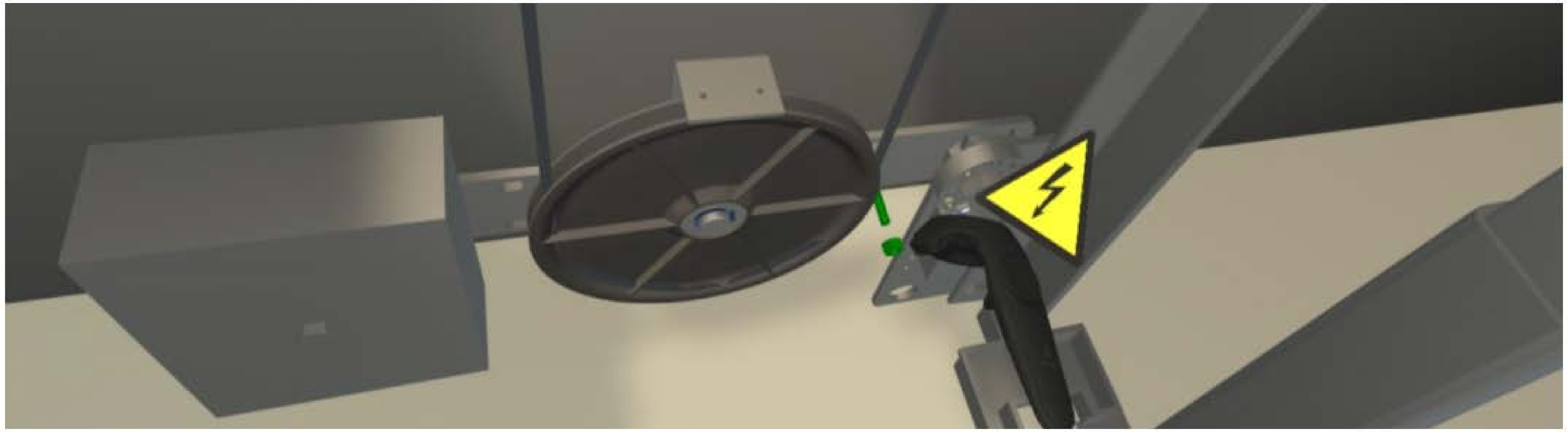
NOVELTY - Difference between SRP-PHAT-HSDA and Our Work

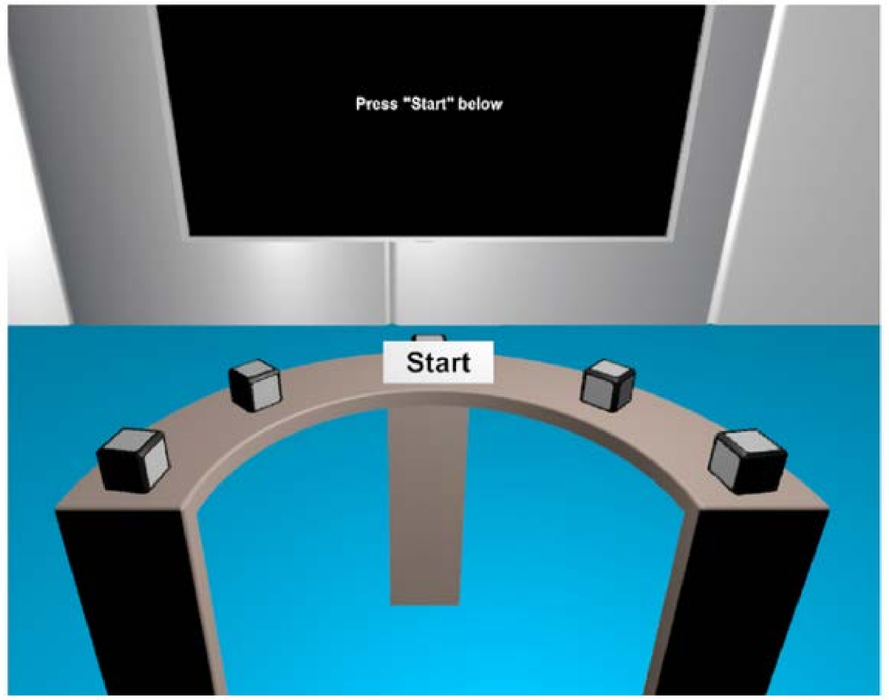


- Strengthening the performance of the presented technique by integrating analysis through SAM and DeAOT into the SRP-PHAT-HSDA algorithm
- Presents the possibility of using the SRP-PHAT-HSDA algorithm in augmented reality

Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

LITERATURE REVIEW - 4.WAYS TO USE GAZE DATA





- Research has been conducted on AR solutions for
  - maintenance using eye tracking
  - the use of gaze-based authentication
  - human movement direction classification using eye tracking

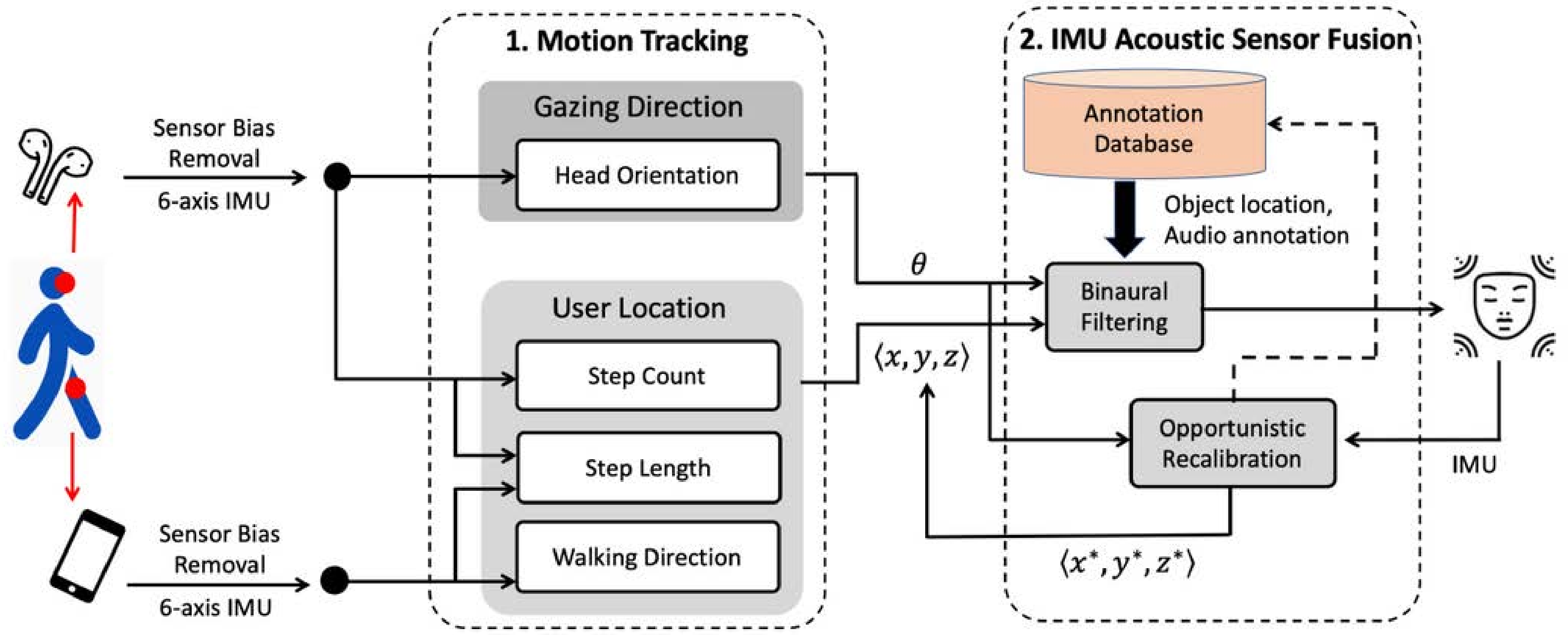
Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, Markku Turunen. 2020. Utilizing VR and Gaze Tracking to Develop AR Solutions for Industrial Maintenance. CHI 2020, April 25–30, 2020, Honolulu, HI, USA

Jonathan Liebers, Stefan Schneegass. 2020. Gaze-based Authentication in Virtual Reality. ETRA '20 Adjunct, June 2–5, 2020, Stuttgart

Julius Petterssona, Petter Falkmana. 2021. Human Movement Direction Classification using Virtual Reality and Eye Tracking. Procedia Manufacturing 51 (2020) 95–102

LITERATURE REVIEW - 5.AUGMENTED HEARING

Ear-AR and Ear-VR



- Focus on enhancing the sound information the user hears
  - Play 3D audio annotations related to the user's environment and actions
  - For the deaf and hard of hearing, it provides directional sound information through haptic feedback

Ear-AR: Indoor Acoustic Augmented Reality on Earphones. Zhijian Yang, Yu-Lin Wei, Sheng Shen, Romit Roy Choudhury. MobiCom ’20, September 21–25, 2020, London, United Kingdom

EarVR: Using Ear Haptics in Virtual Reality for Deaf and Hard-of-Hearing People. Mohammadreza Mirzaei, Peter Kan, ´ and Hannes Kaufmann. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 26, NO. 5, MAY 2020

CONTRIBUTIONS

Convergence of gaze data utilization and sound increase/decrease technology

- Presenting a way to utilize gaze data is to classify sounds and increase/decrease them by group.

Strengthening nonverbal communication in augmented reality increases similarity to real communication

- Communication in VR/AR is made more similar to reality based on non-verbal clues about gaze.

CONTRIBUTIONS

Convergence of SAM, DeAOT, SRP-PHAT-HSDA

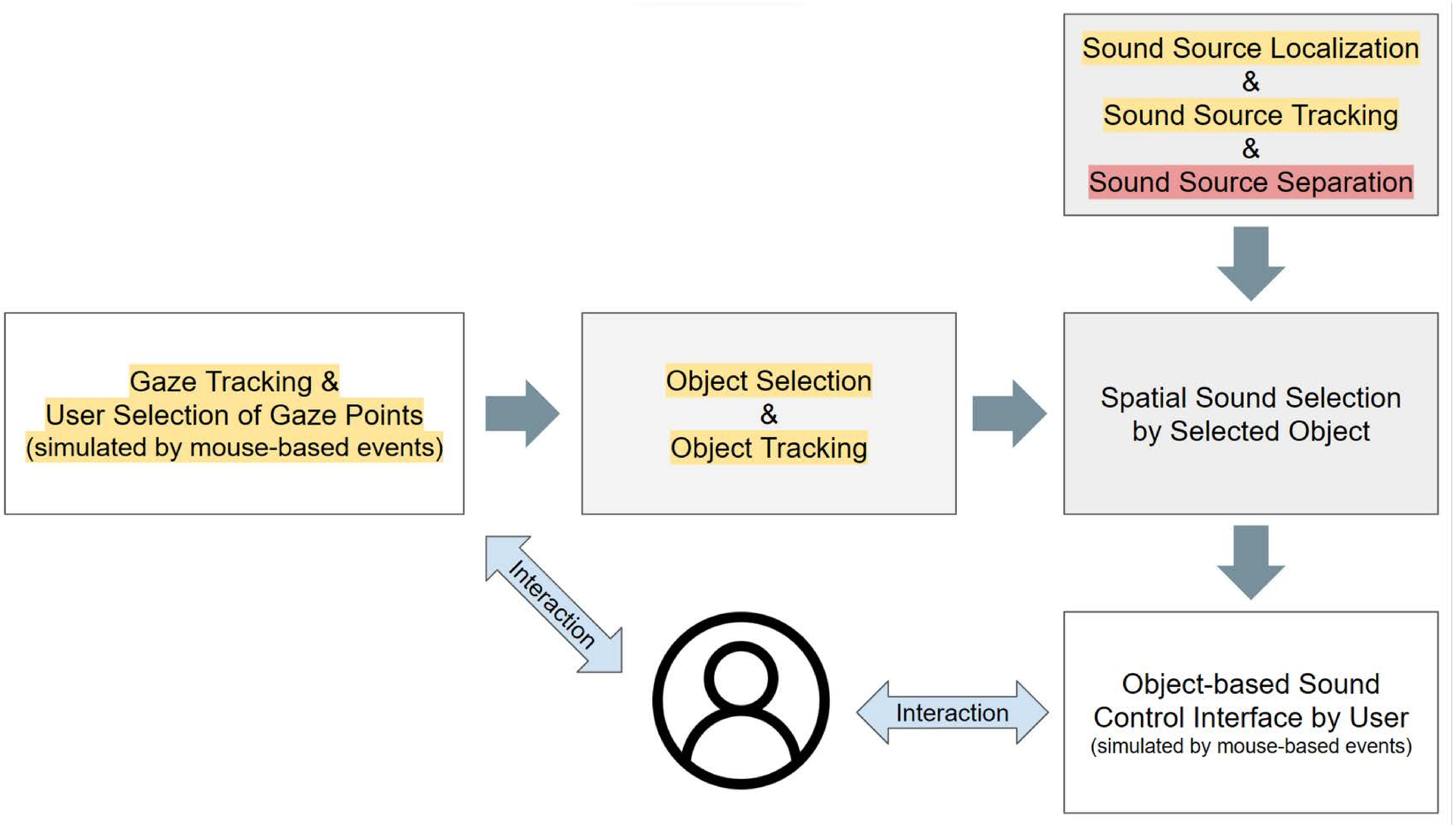
- Strengthening the performance of the technology presented through DeAOT by integrating analysis through DeAOT into analysis through SAM
- DeAOT combines the powerful SRP-PHAT-HSDA algorithm for audio signal processing (SSL, sound source localization) with visual analysis through SAM.
- Presents the possibility of using DeAOT, SAM, and SRP-PHAT-HSDA in the field of augmented hearing.

IMPLEMENTATION

Concept Video

Implementation detail

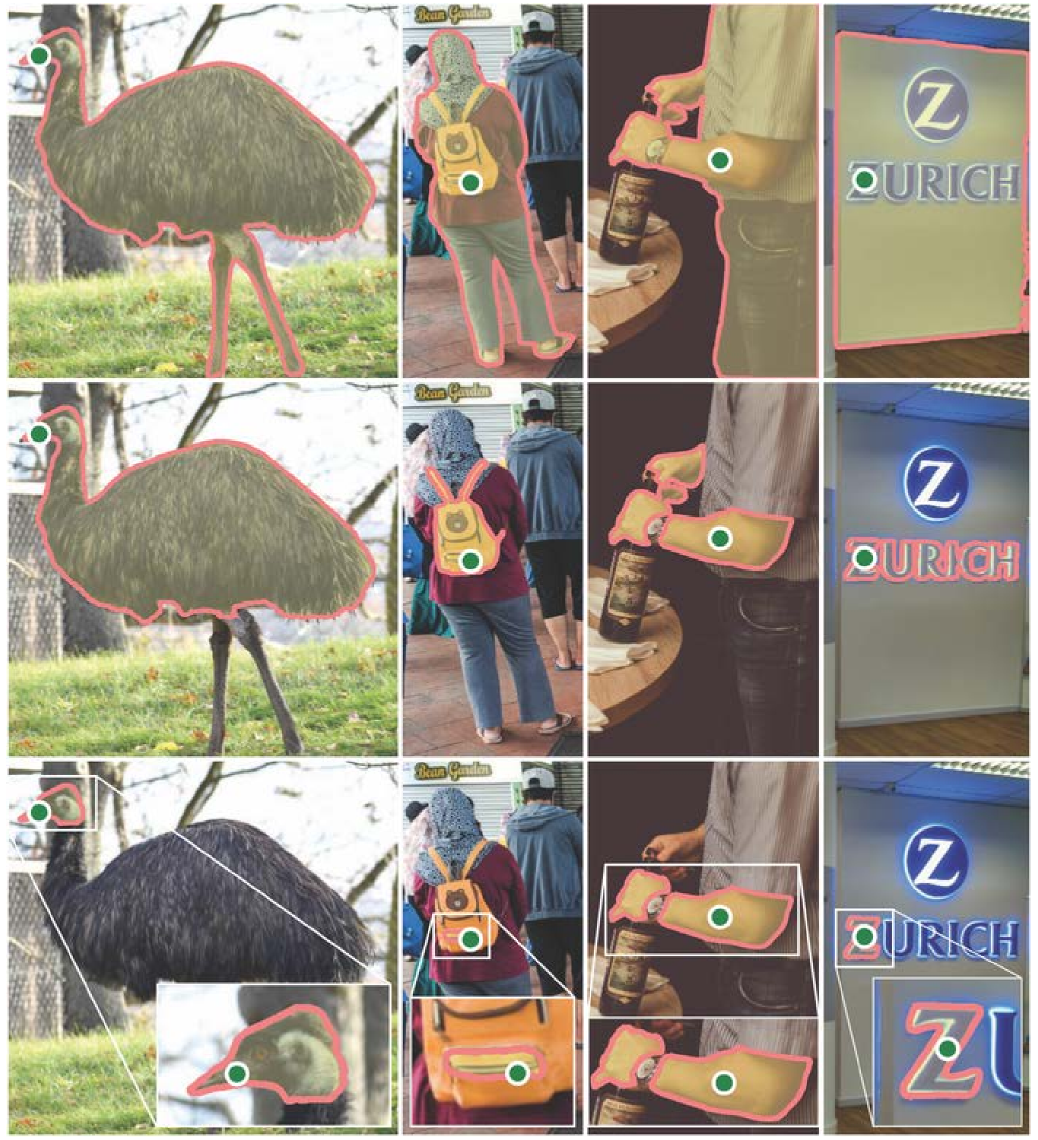
WHAT WE NEED TO DO REALLY



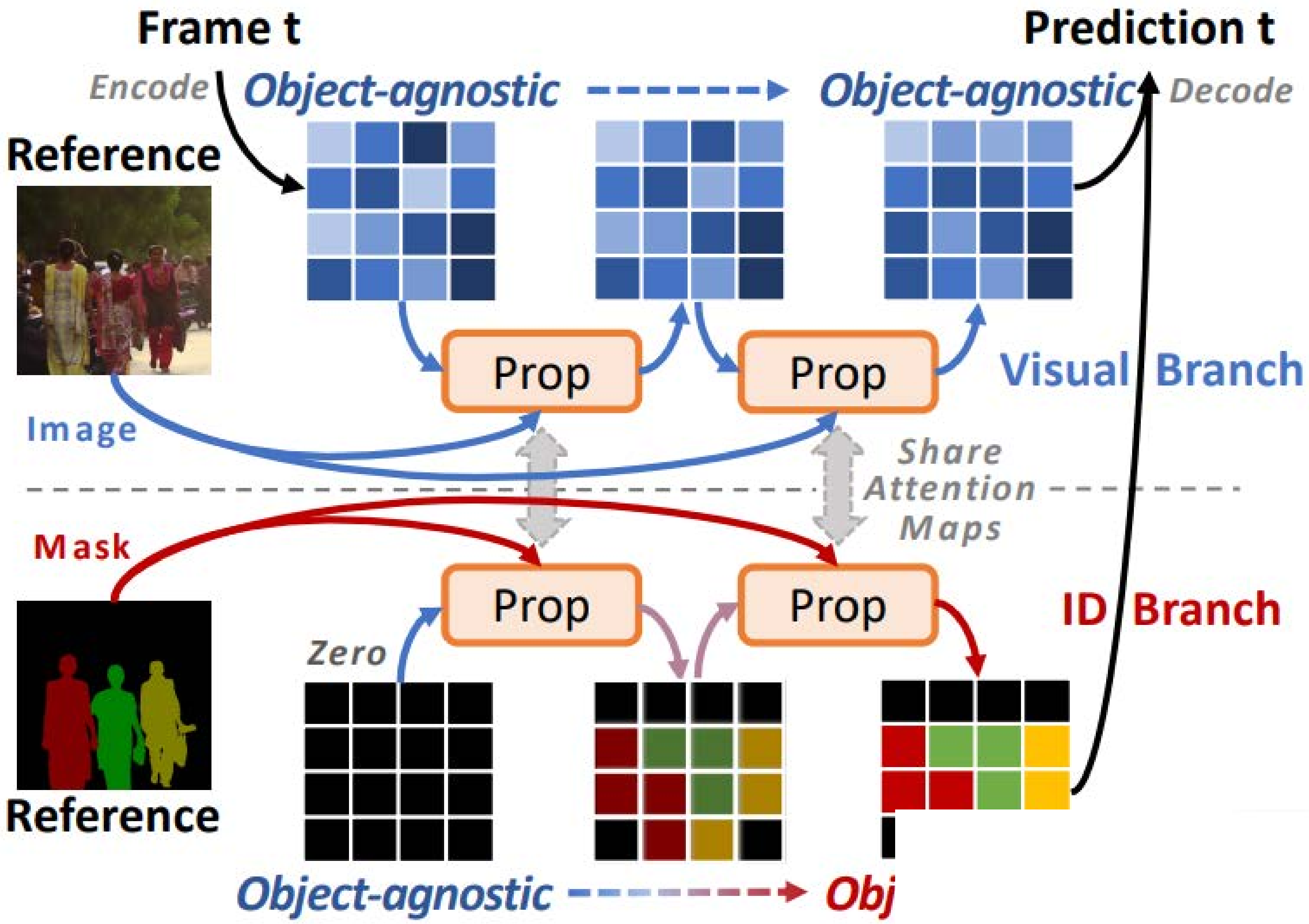
DEMONSTRATION

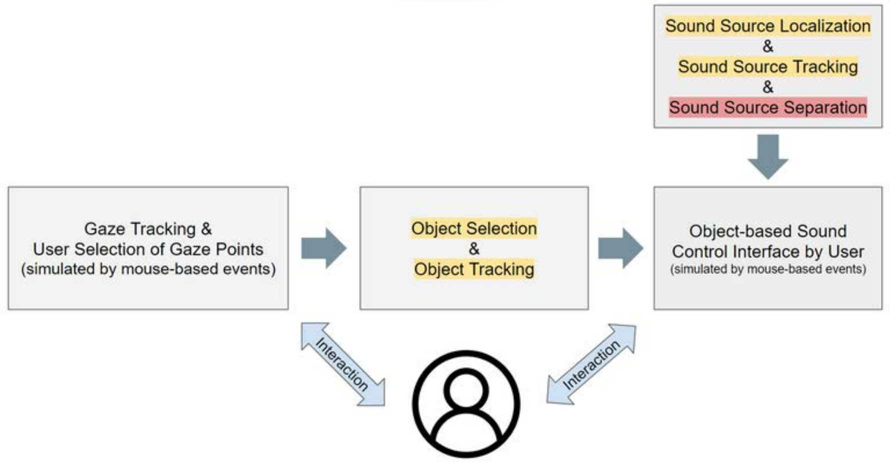
REMIND: OBJECT SELECTION & TRACKING

SAM (ICCV `23)



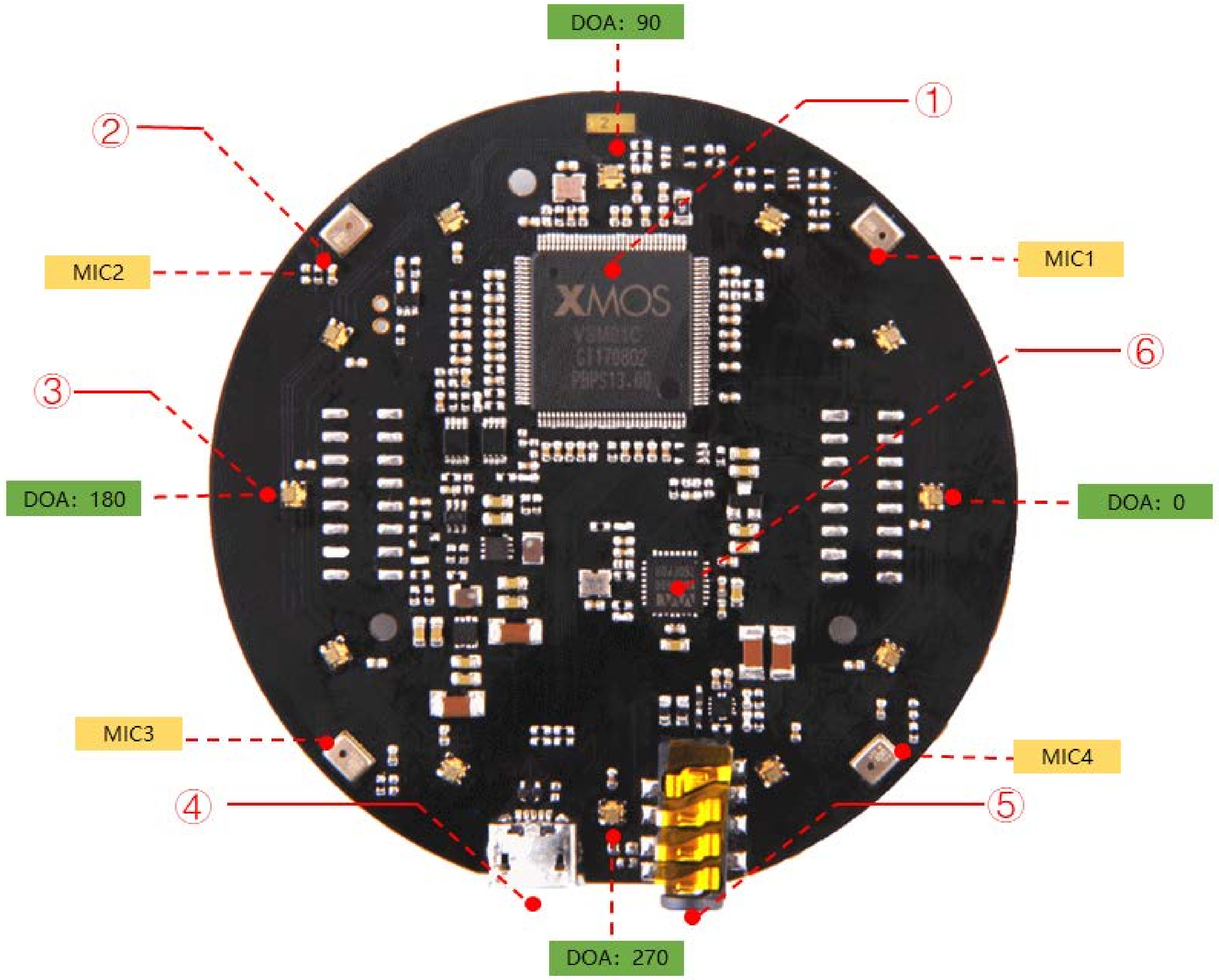
DeAOT (NeurIPS `22)





Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).
Yang, Zongxin, and Yi Yang. "Decoupling features in hierarchical propagation for video object segmentation." Advances in Neural Information Processing Systems 35 (2022): 36324-36336.

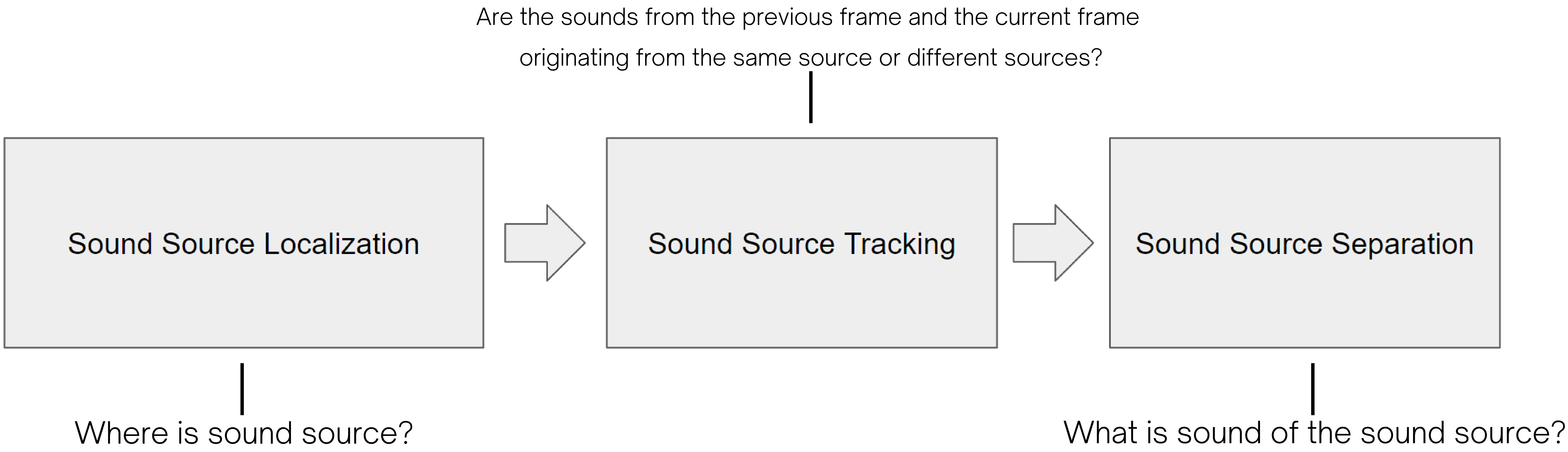
Multi-microphone Array Device



ReSpeaker Mic Array v2.0

- XVF-3000 from XMOS
- **4 digital microphones**
- Supports Far-field Voice Capture
- Speech algorithm on-chip
- 12 programmable RGB LED indicators
- Microphones: ST MP34DT01TR-M
- **Sensitivity: -26 dBFS (Omnidirectional)**
- Acoustic overload point: 120 dBSPL
- SNR: 61 dB
- Power Supply: 5V DC from Micro USB or expansion header
- Dimensions: 70mm (Diameter)
- 3.5mm Audio jack output socket
- Power consumption: 5V, 180mA with led on and 170mA with led off
- **Max Sample Rate:16Khz**

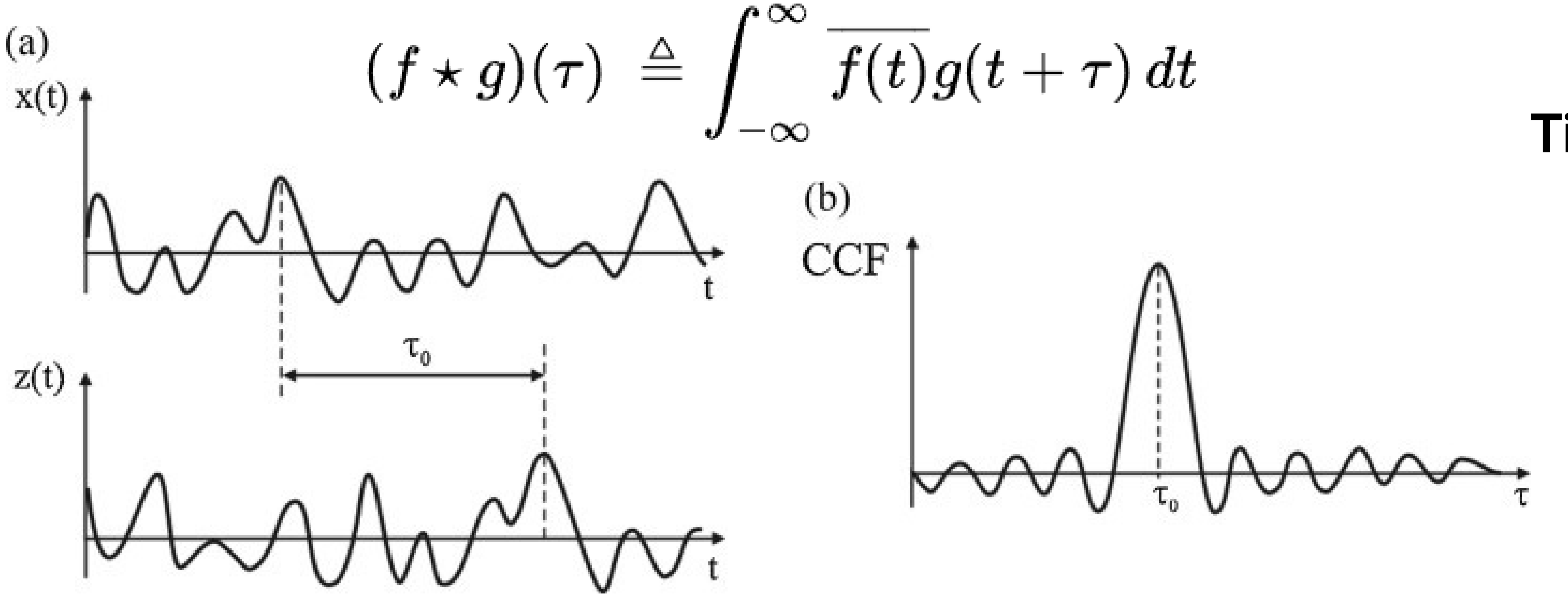
Steps


Are the sounds from the previous frame and the current frame originating from the same source or different sources?
Sound Source Localization
Sound Source Tracking
Sound Source Separation
Where is sound source?
What is sound of the sound source?

1. Sound Source Localization

Cross-correlation

$$(f \star g)(\tau) \triangleq \int_{-\infty}^{\infty} \overline{f(t)}\, g(t+\tau)\, dt$$



Time delay between two signals

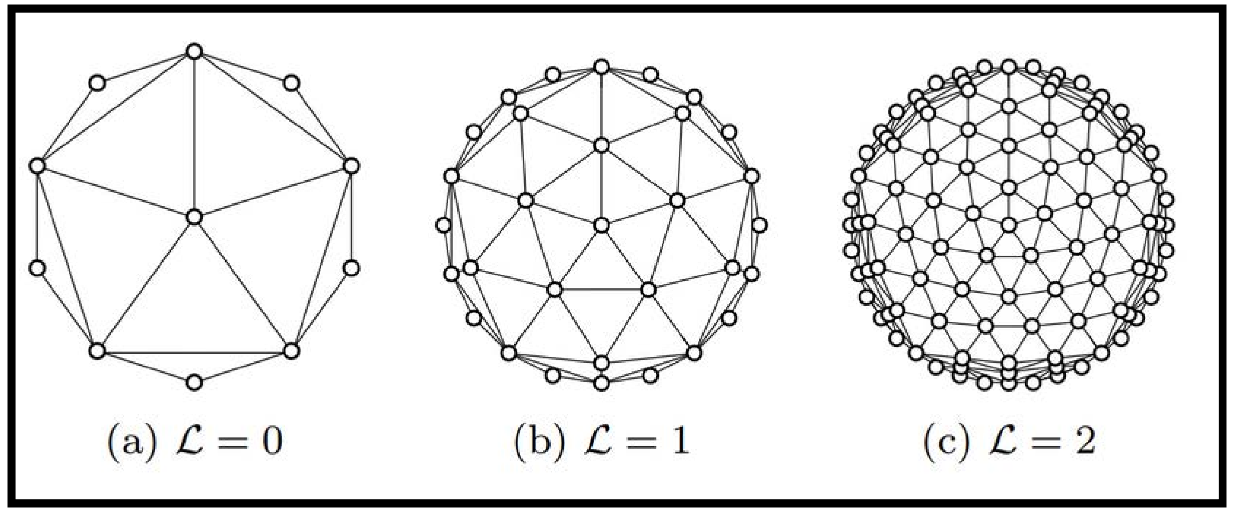
$$\boxed{\tau_{\text{delay}} = \underset{t \in \mathbb{R}}{\arg\max}((f \star g)(t))}$$

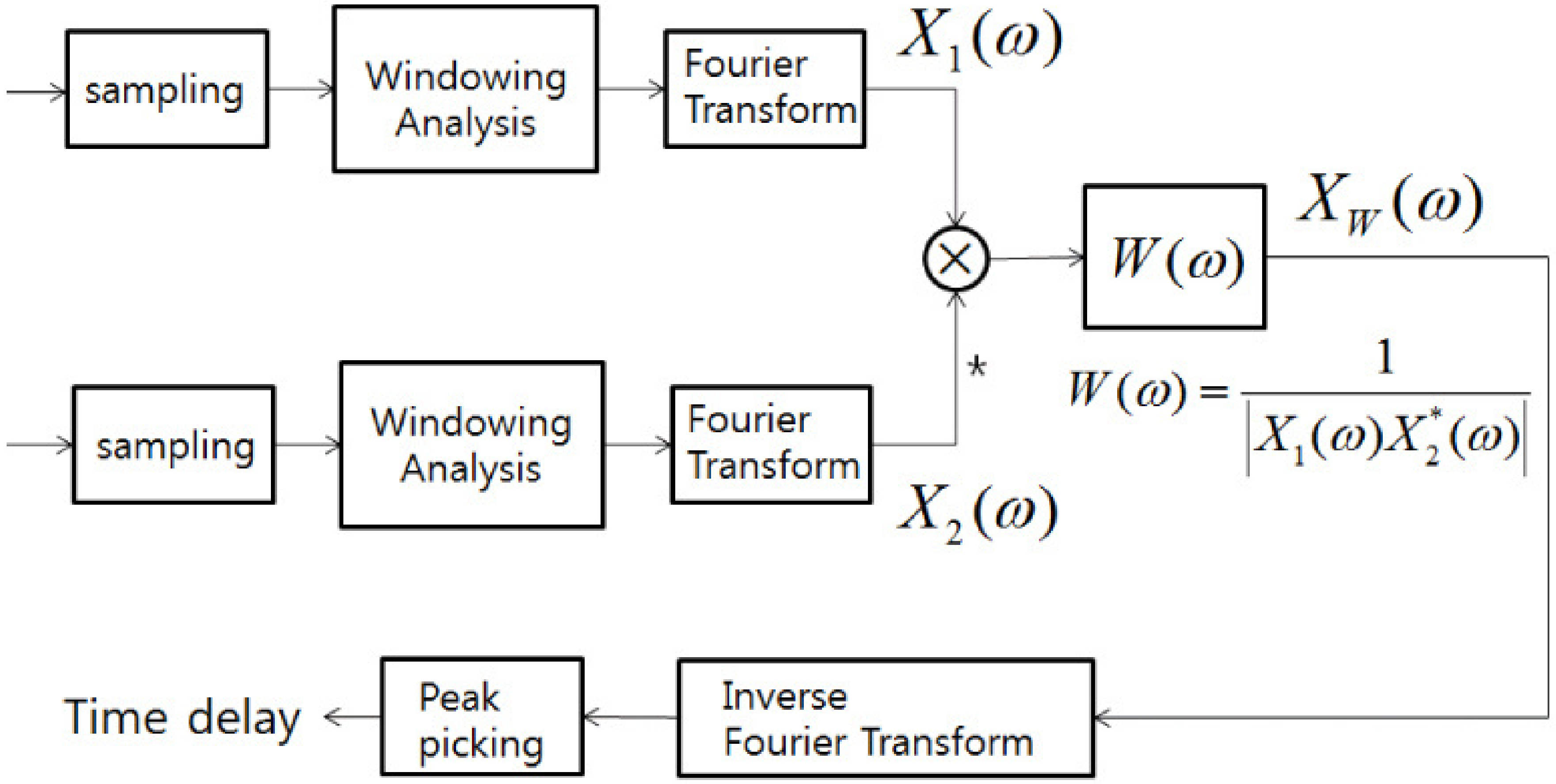
How about N signals?

How about M unknown common part?

Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

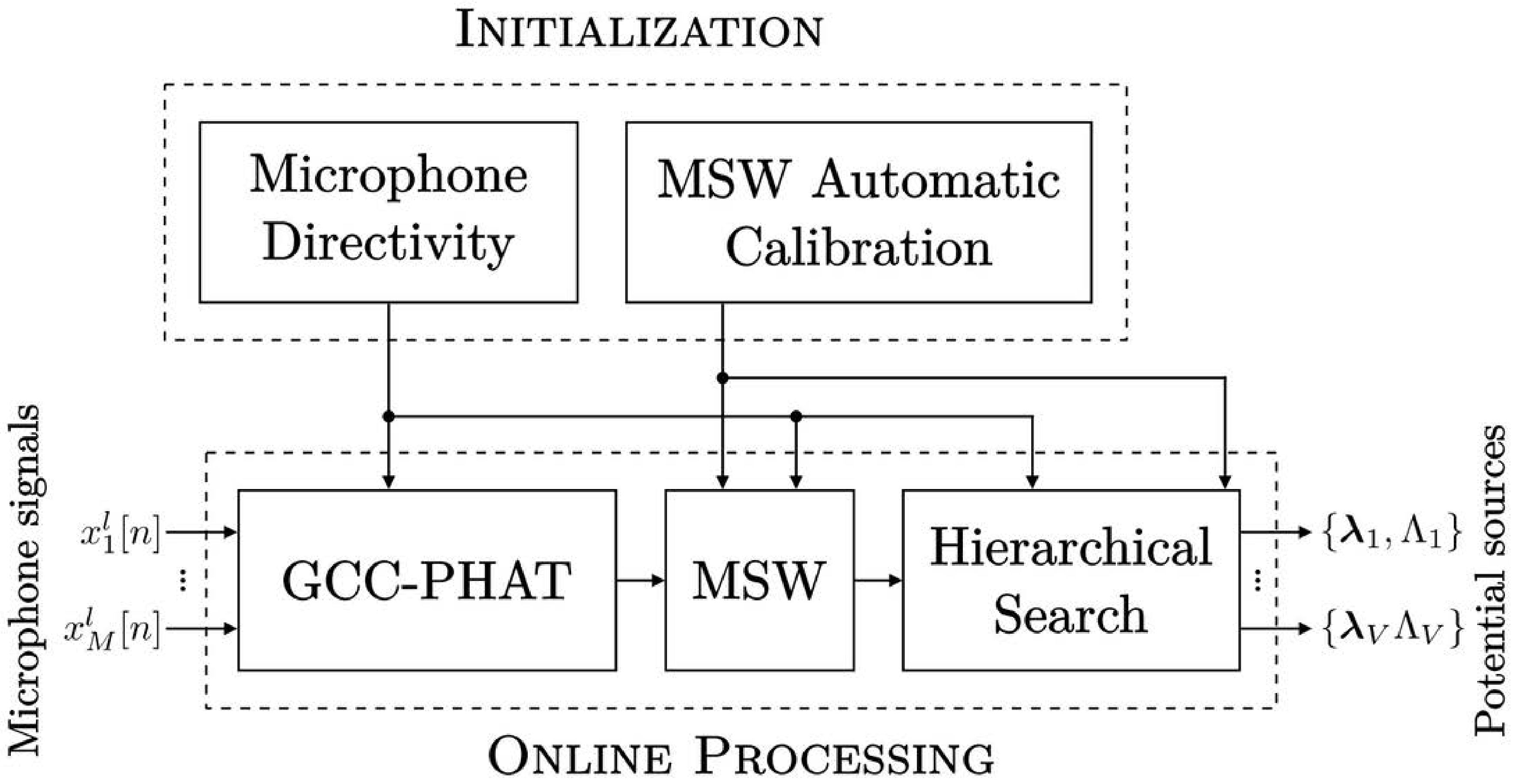
1. Sound Source Localization





Generalized cross-correlation

(Estimation of Time Difference of Arrival)



SRP-PHAT-HSDA

(Hierarchical search of Direction of Arriva)

Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

2. Sound Source Tracking

$$\boxed{\mathbf{x}_i^l = \mathbf{F}\mathbf{x}_i^{l-1} + \mathbf{B}\mathbf{u}_i^l + \mathbf{w}_i}$$

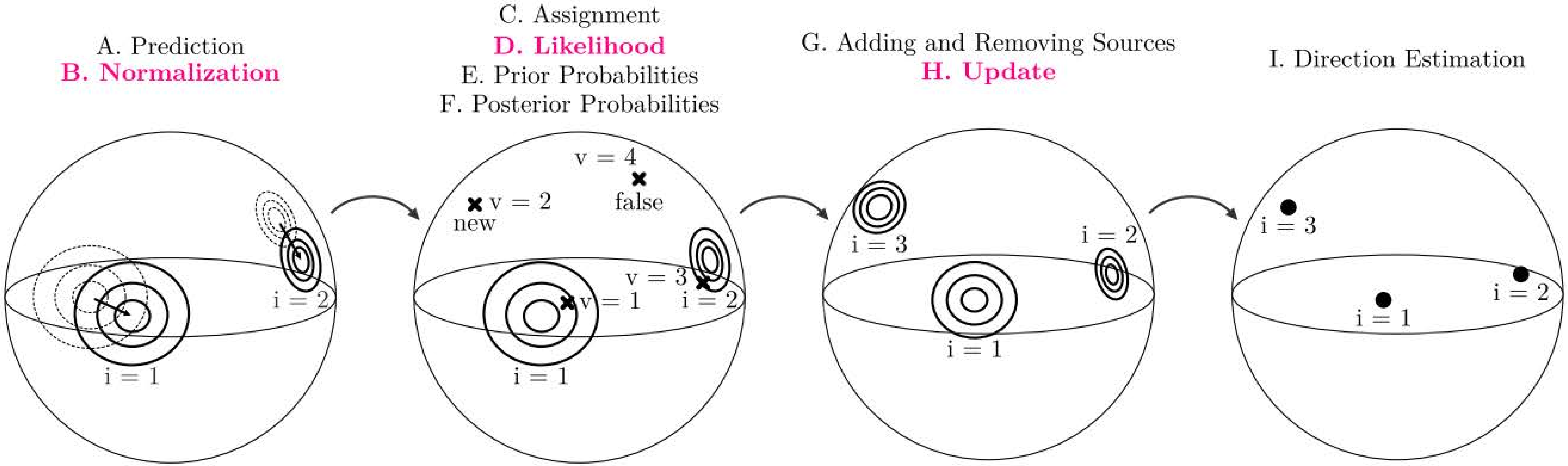
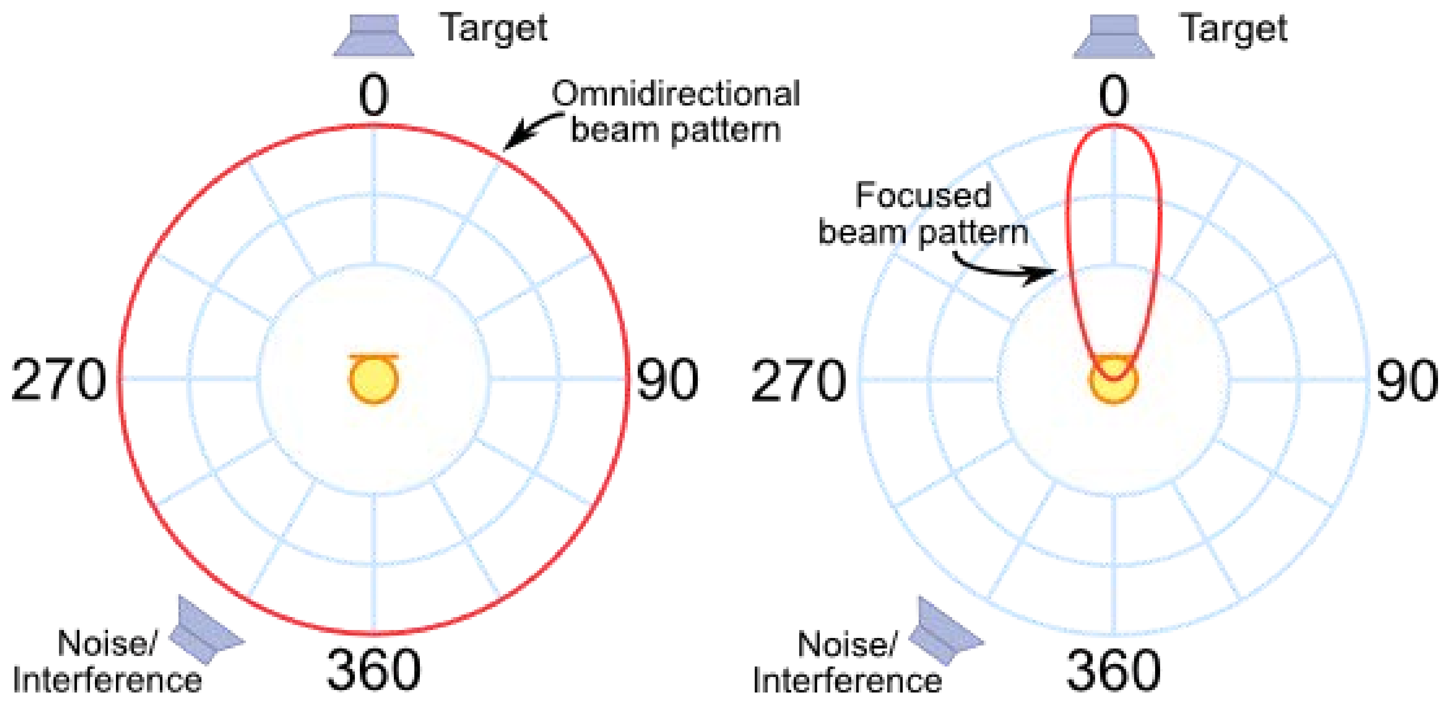


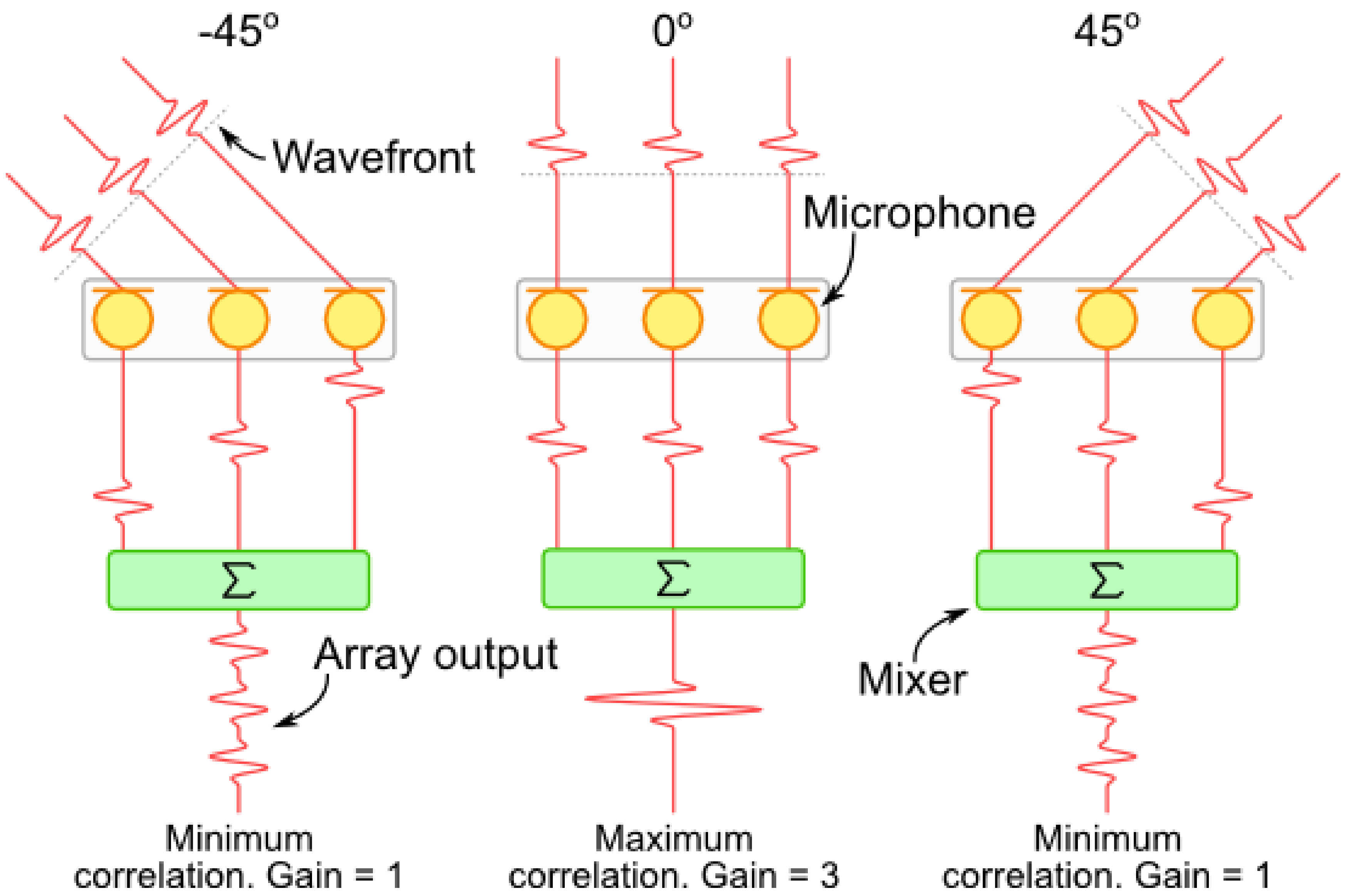
Figure 8: Tracking simultaneous sound sources using M3K. Tracked sources are labeled $i = 1, 2, 3$ and potential sources are labeled $v = 1, 2, 3, 4$.

Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. François Grondin, François Michaud D. Robotics and Autonomous Systems 113 (2019) 63–80

3. Sound Source Separation



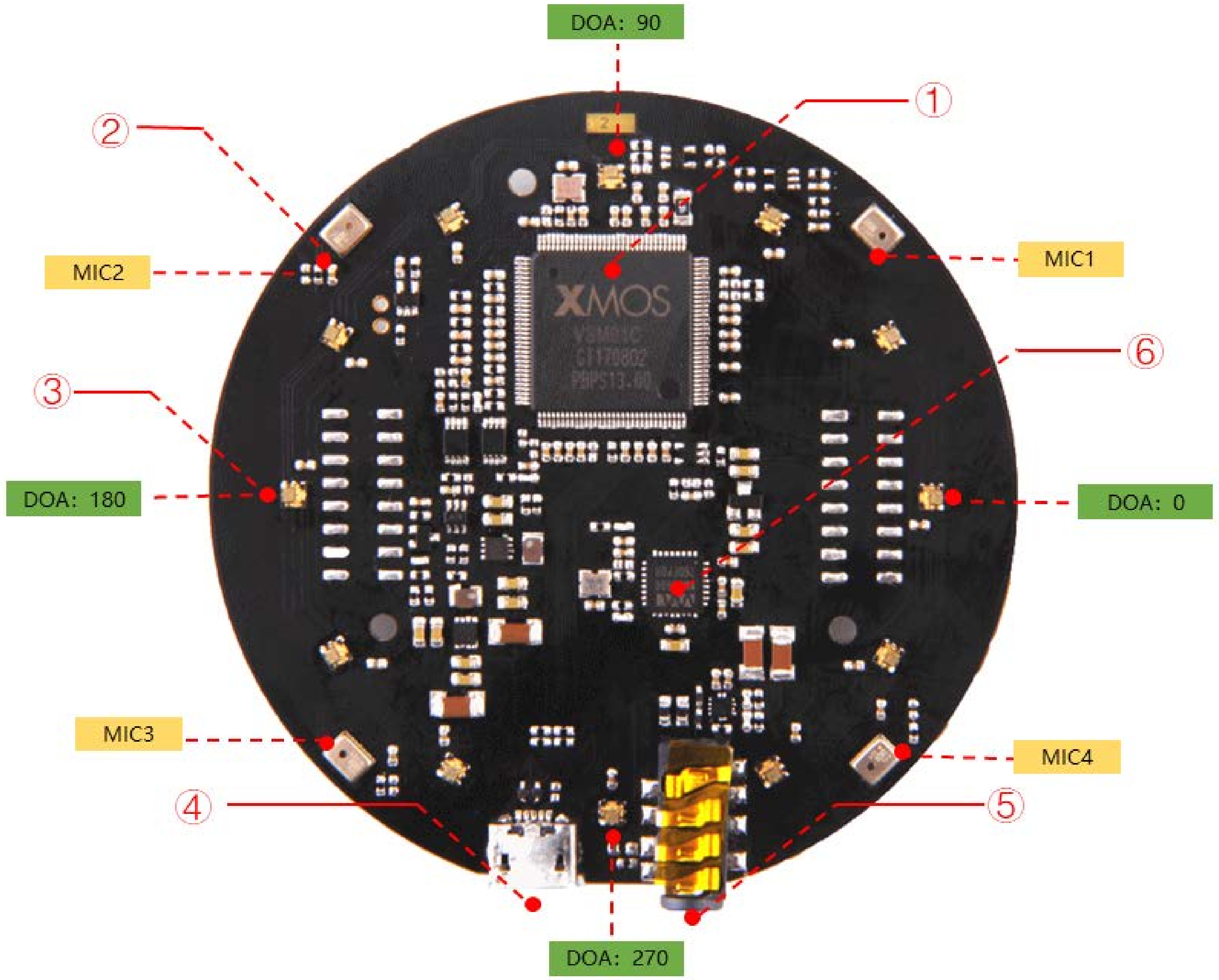
Concept of Beamforming



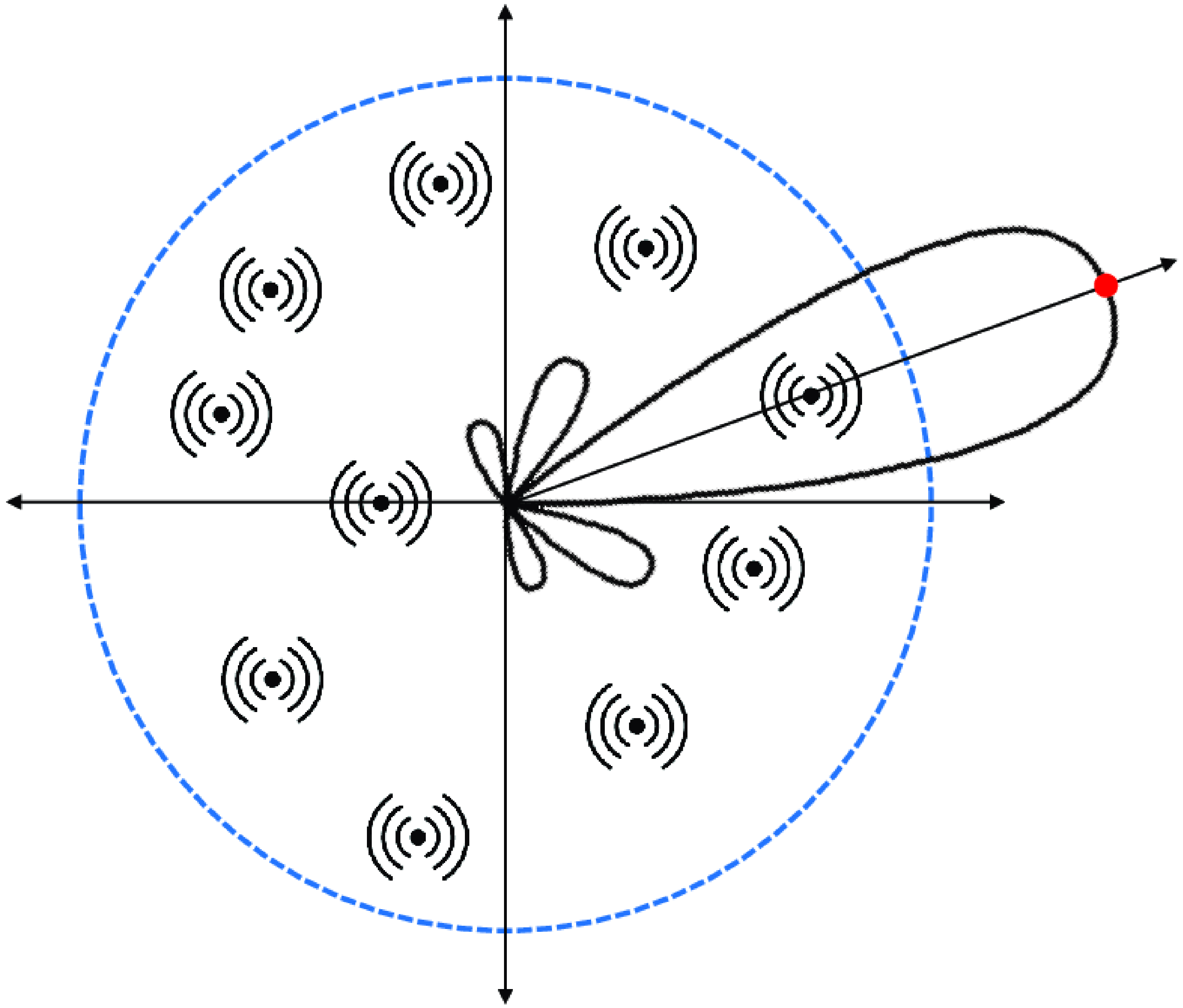
Classifcal Beamforming (Delay-and-Sum Beamforming)

http://www.labbookpages.co.uk/audio/beamforming/delaySum.html

3. Sound Source Separation

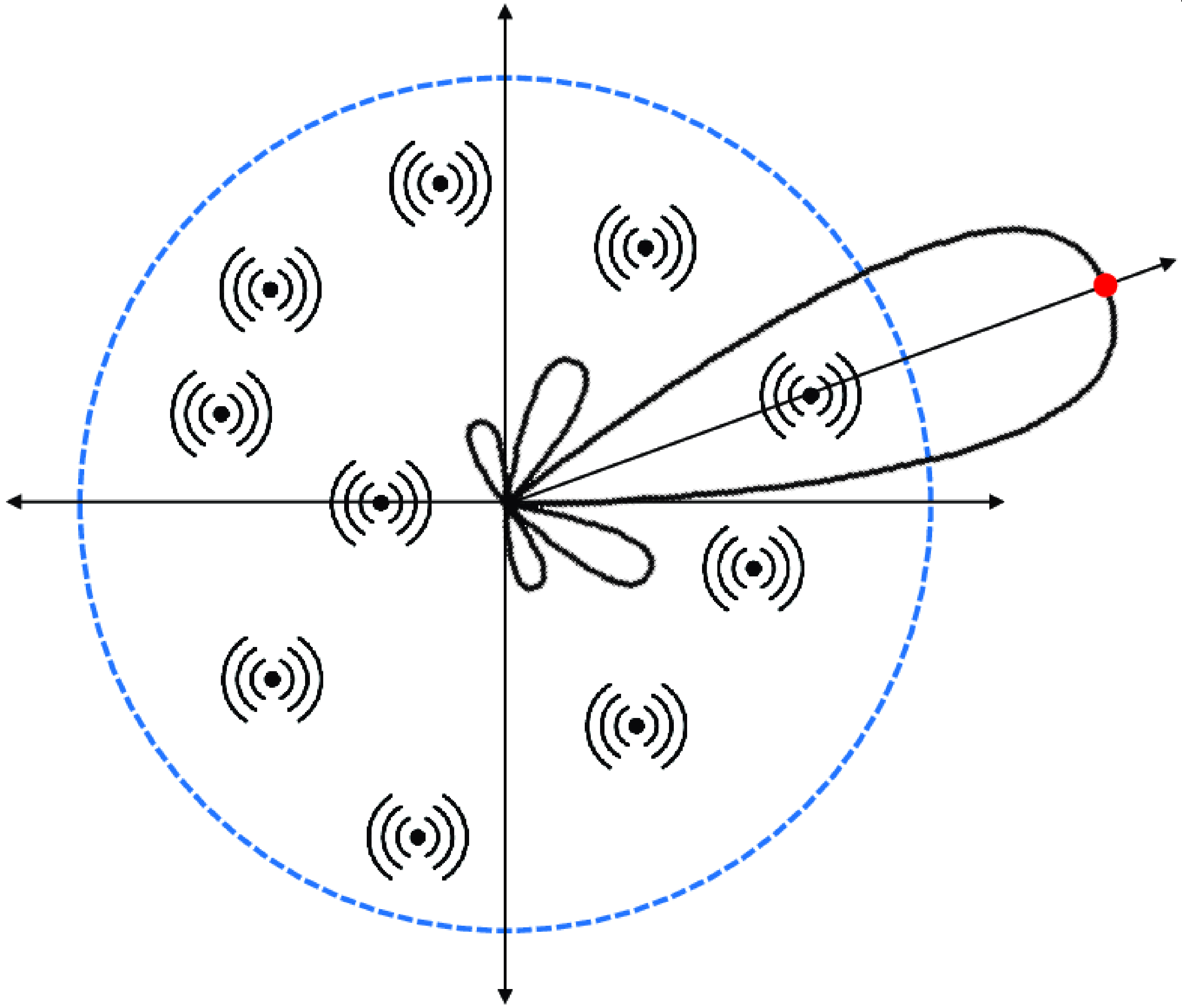


Our device



Beamforming with Arbitrary Sensor Configuration

3. Sound Source Separation



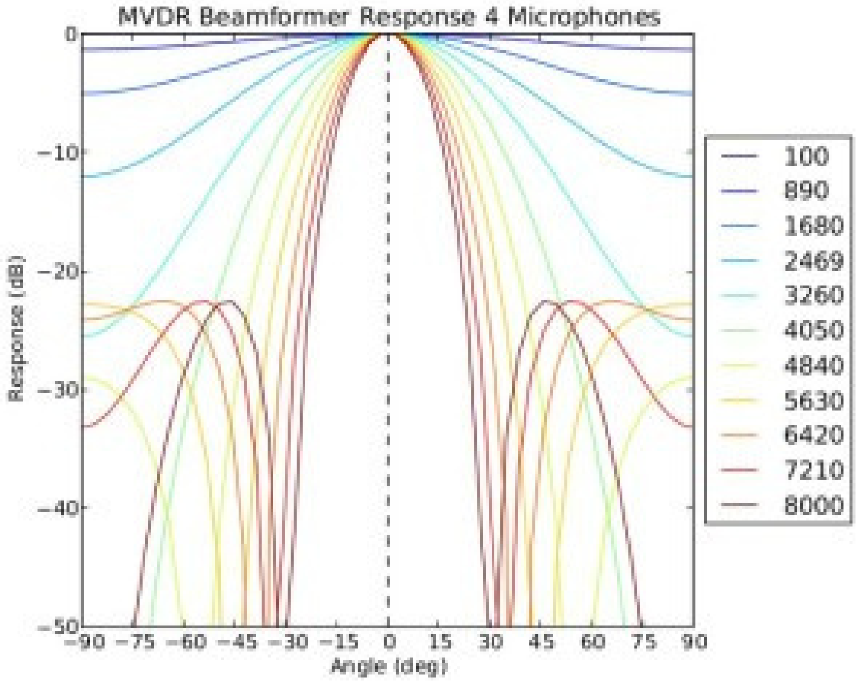
Beamforming with Arbitrary Sensor Configuration

Typical Technique

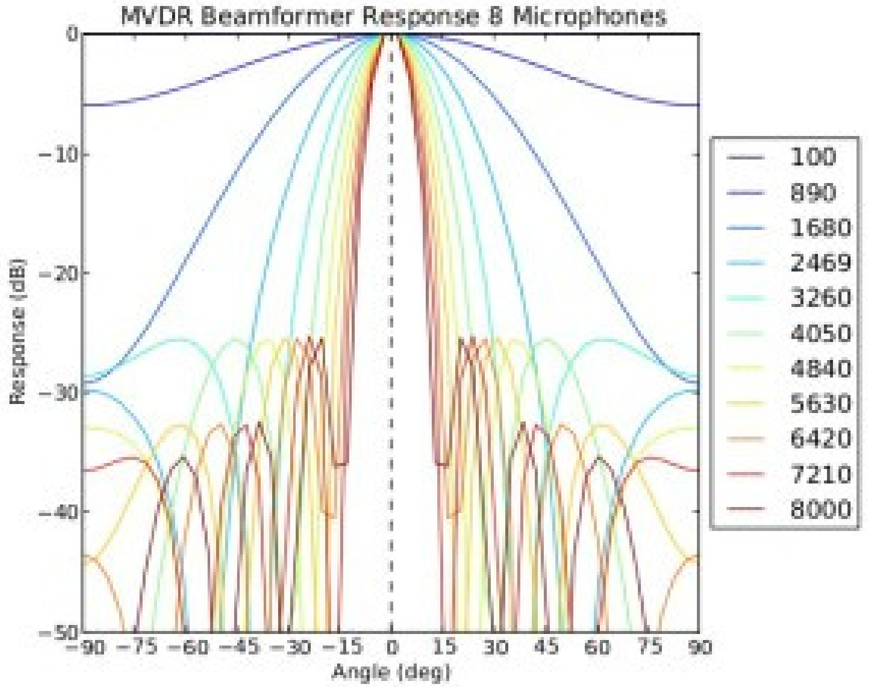
-> Minimum Variance Distortionless Response (MVDR)

$$\mathbf{w}_{\mathrm{MVDR}}(f) = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathsf{H}}(f)\boldsymbol{\Phi}_{\mathbf{YY}}(f)\mathbf{w}(f)$$
$$\text{s.t.}\,\mathbf{w}(f)^{\mathsf{H}}\mathbf{d}(f) = 1$$

4 Microphones

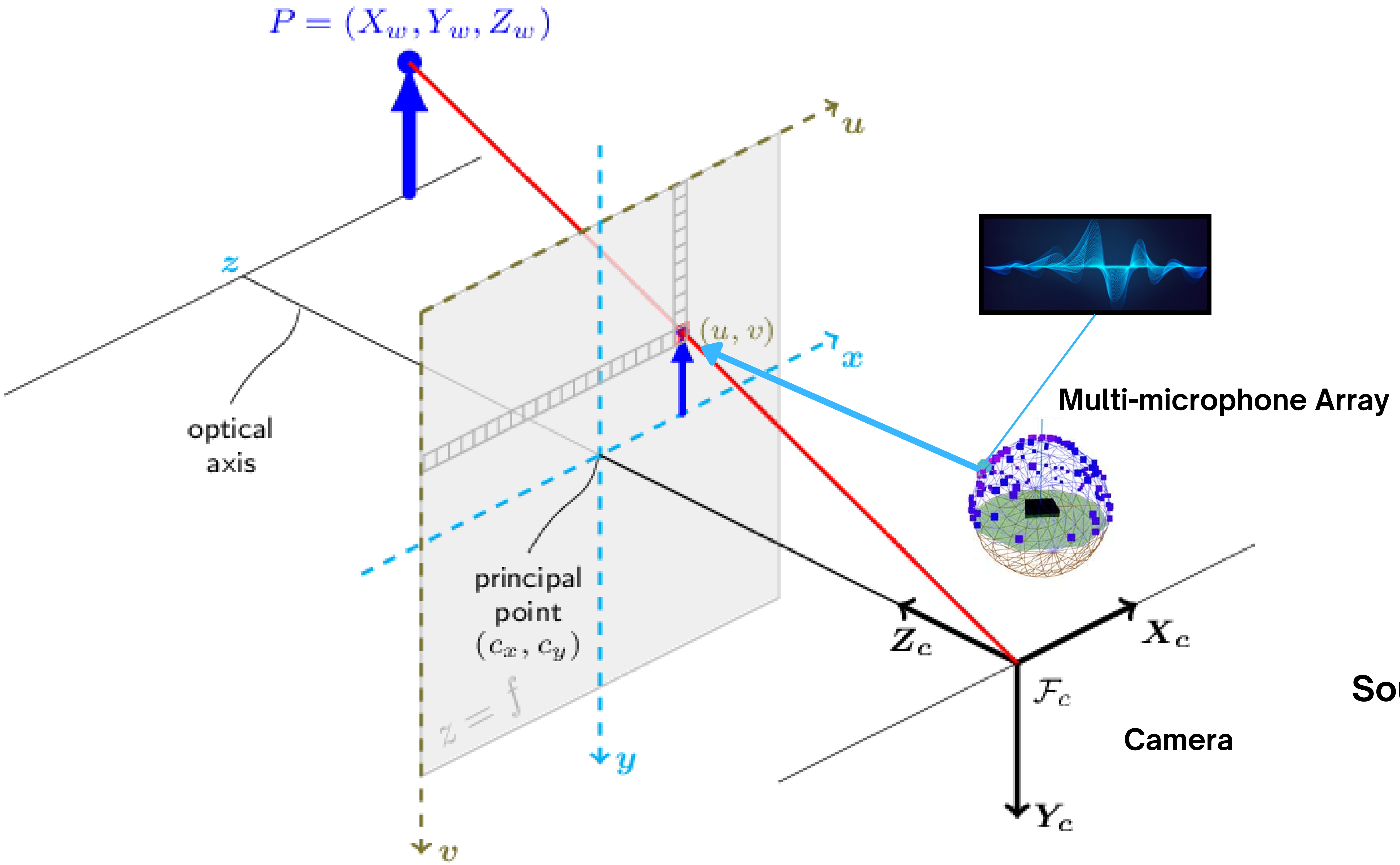


8 Microphones



- Souden, Mehrez, Jacob Benesty, and Sofiene Affes. “On optimal frequency-domain multichannel linear filtering for noise reduction.” IEEE Transactions on audio, speech, and language processing 18, no. 2 (2009): 260-276.
- Erdogan, Hakan, John R. Hershey, Shinji Watanabe, Michael I. Mandel, and Jonathan Le Roux. “Improved mvdr beamforming using single-channel mask prediction networks.” In Interspeech, pp. 1981-1985. 2016.
- https://nateanl.github.io/2021/07/21/mvdr-tutorial/

4. Spatial Sound Mapping



- **Camera**
  - Intrinsic Parameters
  - Extrinsic Parameters
- **Microphone Array**
  - Relative Position to Camera
- **Sound**
  - Sound Direction

Sound Source Position on Pixel Space

EVALUATION AND FUTURE WORK

Technical evaluation

User Study

Example Applications

Yang, Zongxin, and Yi Yang. "Decoupling features in hierarchical propagation for video object segmentation." Advances in Neural Information Processing Systems 35 (2022): 36324-36336.

Technical evaluation

Object selection

Since collecting gaze points is replaced with mouse-based clicks...

What: Appropriate selection of target based on click

Performance metric: Correctly Selected Objects / Intended Selection

How: Data gathered through **user study**

Visual object tracking

What: System's ability to continuously track selected object's dynamic movement

Performance metric: IoU score per frame -> average IoU score over time of selection

How: Compare system's predicted bounding box to ground truth (manually annotated per frame)

https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

User Study Plan

Determine the **effectiveness of features** and gauge **user satisfaction** with system's **performance, usability and intuitiveness**

Participants

- Around 10 participants
- Varying backgrounds and experience with AR applications

Analysis

Currently: No results; technical evaluation and user studies were not conducted due to time and technical constraints

User Study Plan

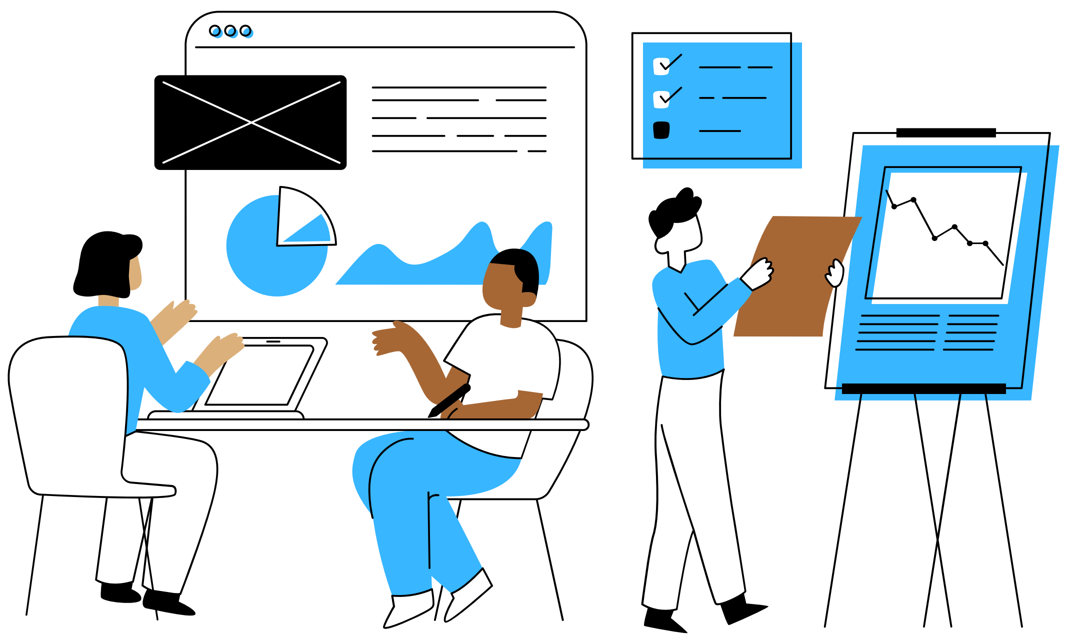
Procedure

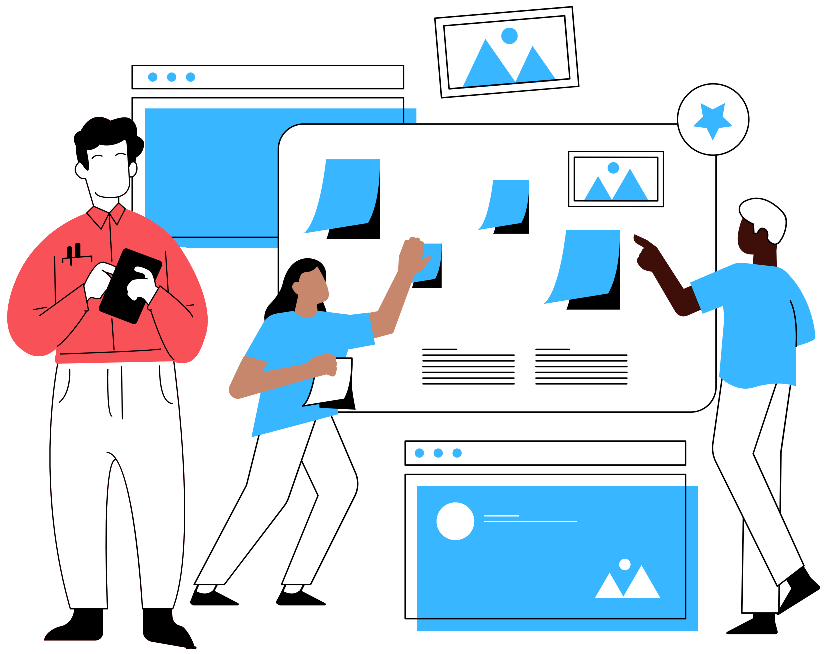
- Participants seated behind laptop displaying video of simulated meeting
- All individuals in meeting are visible to participant
- Participants are guided to click on conversationalist they want to listen to
- System should select corresponding person -> track them -> amplify their voices
- Participant repeats this several times until meeting ends

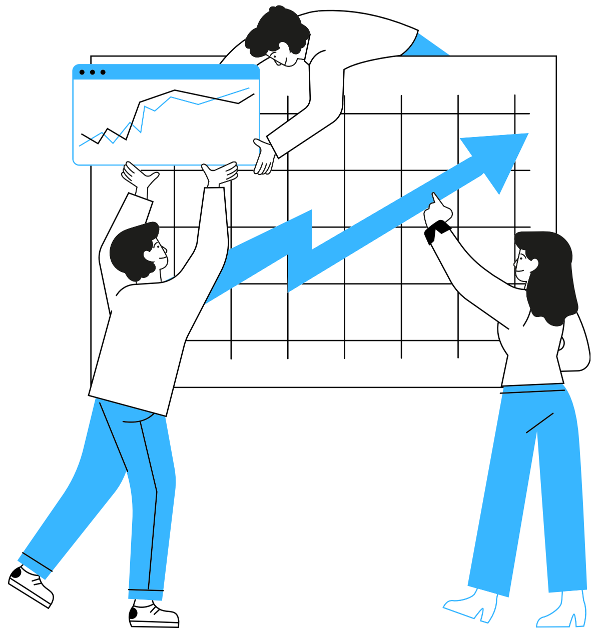
- Participant scores on a scale of 1 to 5 for the following:
  - **Intuitiveness**
  - **System's selection accuracy**
  - **Audibility of selected person**
  - **System's responsiveness**
- Opinions on perceived effectiveness and suggestions for improvement are collected throughout

Application Example

Use Case Scenario







Conference room setting

Group meeting where multiple parties must interact and brainstorm towards their shared goals

Use of proposed solution

Head of the meeting must be able to keep up with different parties Uses proposed solution AR technology to distinguish conver-sations

Ideal results

Use of proposed solution improves interpersonal communication and the meeting produces positive results

Main takeaways

Conclusions

- Integration of gaze-based sound augmentation in AR for enhancing real-time auditory focus
- Effectiveness of selectively amplifying sound based on user gaze, addressing the 'cocktail party' problem

Future work: exploring the implementation in real-world scenarios using AR, and conducting the proposed user study

Contributions

- Combining of sound data and gaze neutralization technology
  - Novel way of using gaze data to classify sounds
- Combining of sound source-based intelligent systems using CNN with AR
- Enhancing of non-verbal communication in AR and strengthening it

Q&A

THANK YOU FOR LISTENING