

Audio-Visual Speech Source Separation and Speaker Tracking

Wenwu Wang

School of Computer Science and Electronic Engineering

University of Surrey, UK

Email: w.wang@surrey.ac.uk;

Web: <https://personalpages.surrey.ac.uk/w.wang/>

Joint work with former PhD students Qingju Liu, Yang Liu, Volkan Kilic, current PhD students Peipei Wu, Jinzheng Zhao, and collaborators

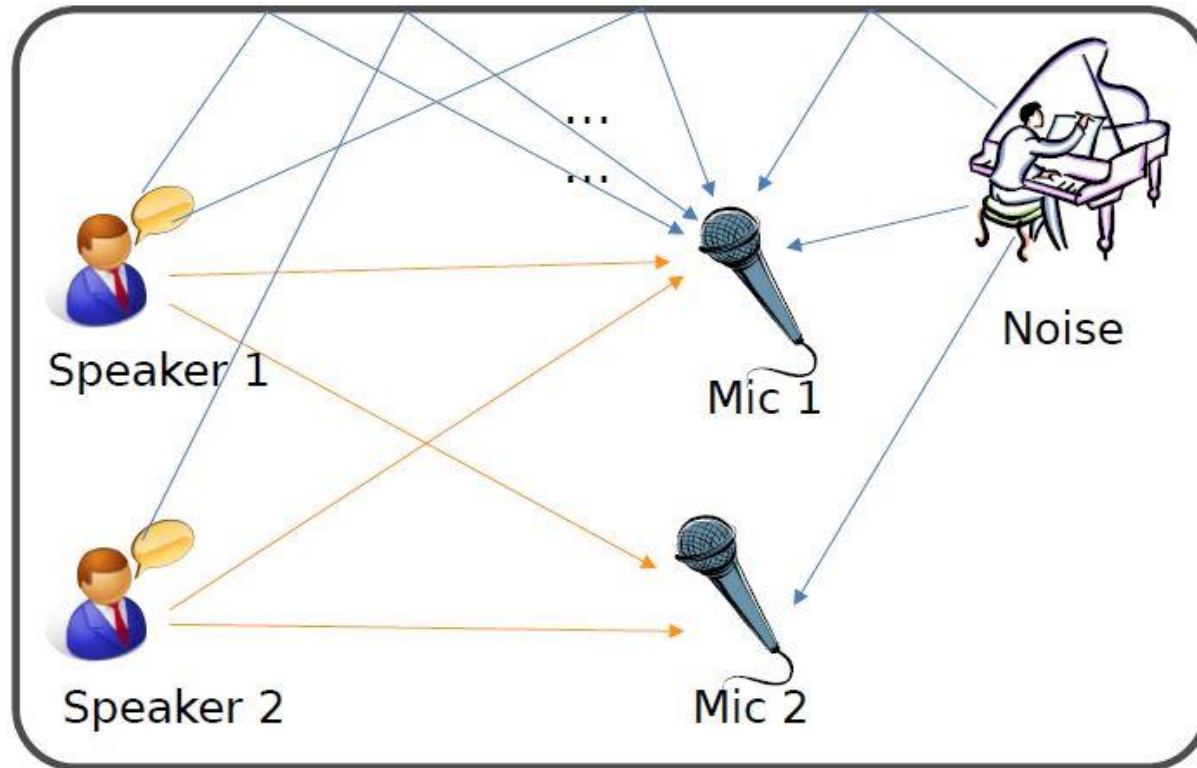
SoRAIM Winter School, Grenoble, France

21 Feb 2024



- 1 Audio-visual speech source separation
- 2 Audio-visual multi-speaker localization/tracking
- 3 Ego-centric audio-visual speaker localization/tracking
- 4 Conclusion and Future Works

Cocktail party problem



Cocktail-party problem (Cherry 1953) or *ball-room problem* (Helmholtz, 1863)

“No machine has yet been constructed to do just that [solving the cocktail party problem].” (Cherry, 1957)

Cocktail party problem may involve a few tasks:

How many speakers and where are they?

(Localization and tracking)

Who speaks and when?

(Diarization)

Multi-speaker talking simultaneously

(Speech separation)

Said What?

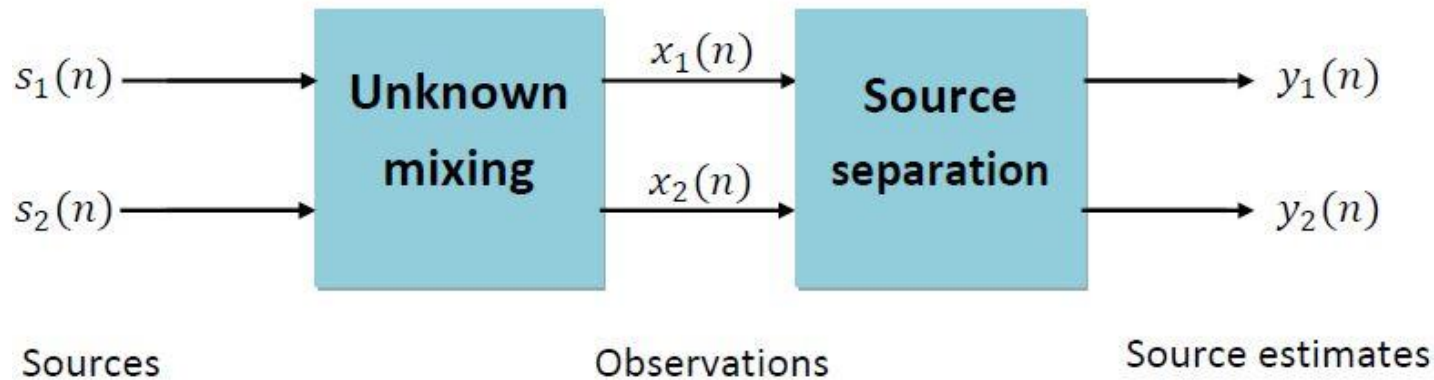
(Automatic speech recognition)

What is the environment?

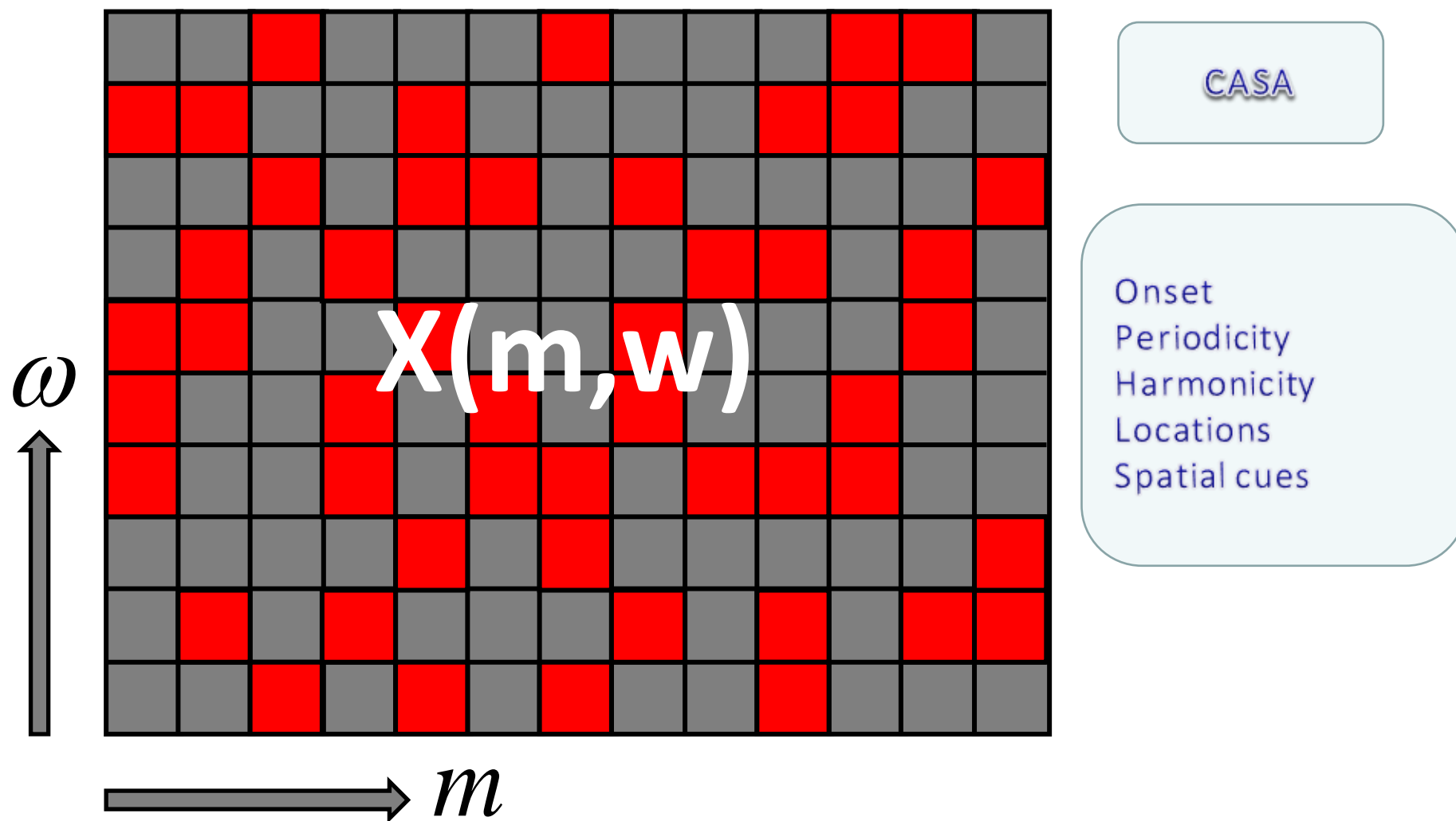
(Acoustic scene recognition, event detection, room acoustics, noise source categorization)

Speech source separation problem & potential solutions

- Potential techniques for the speech separation problem
 - Beamforming
 - Blind source separation and independent component analysis
 - Speech enhancement
 - Sparse representation and matrix factorization
 - Computational auditory scene analysis (e.g. time-frequency masking)
 - Learning based techniques
 - ...

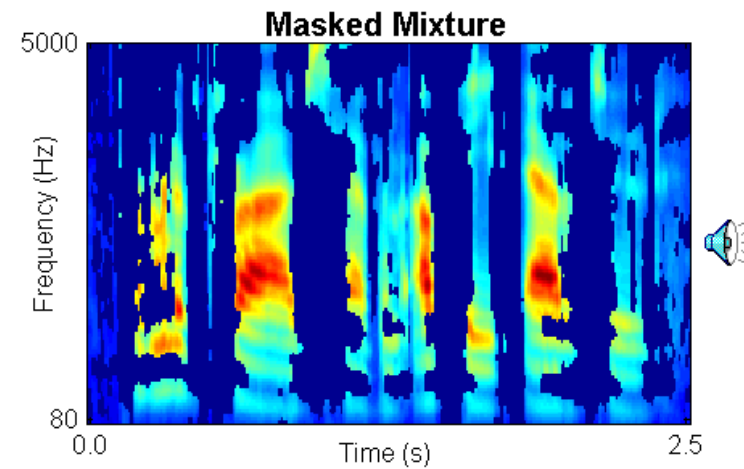
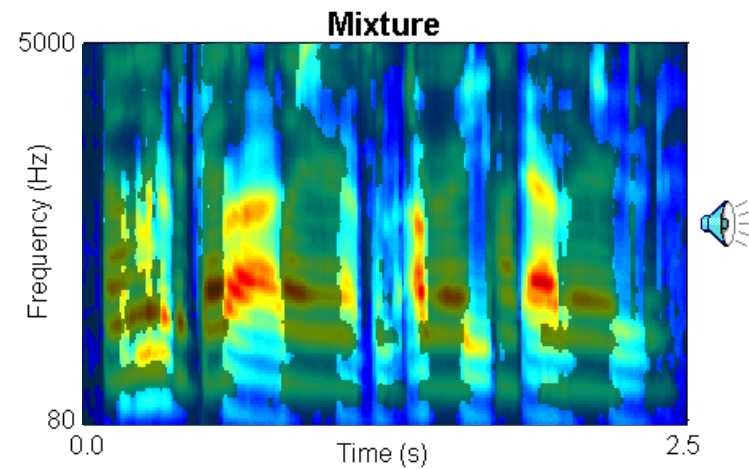
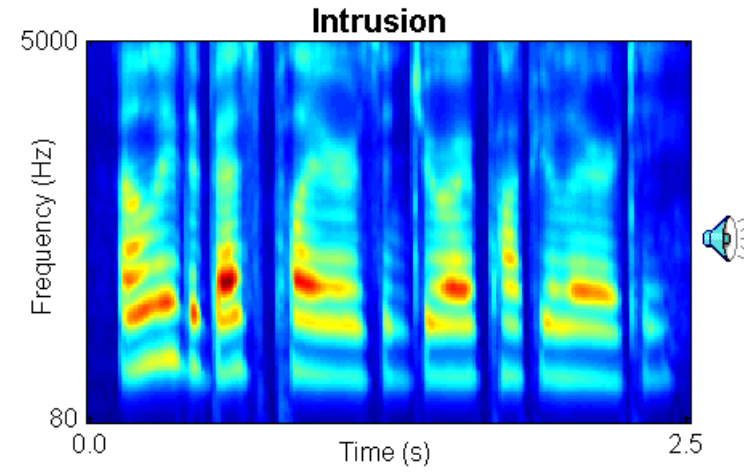
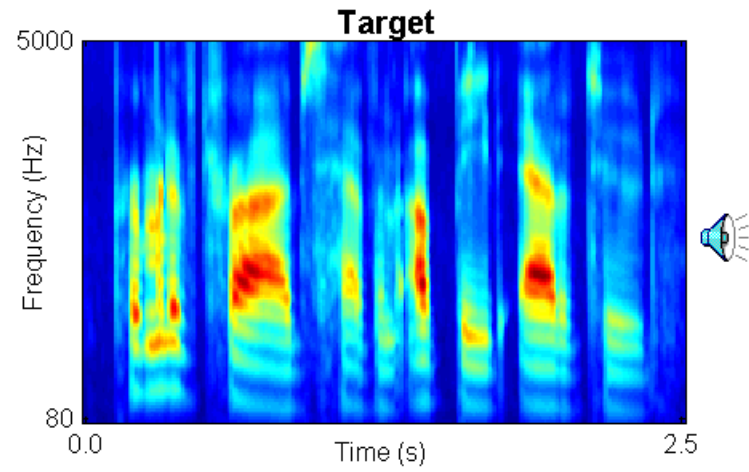


Source separation with time-frequency masking

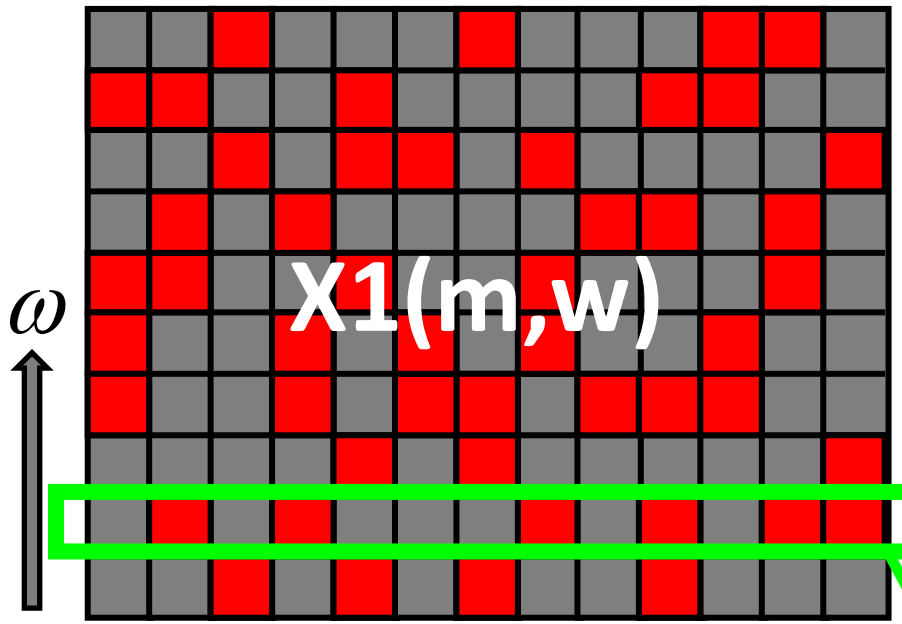


Sparsity assumption ----- each TF point is dominated by one source signal.

Speech separation with "ideal" TF masking



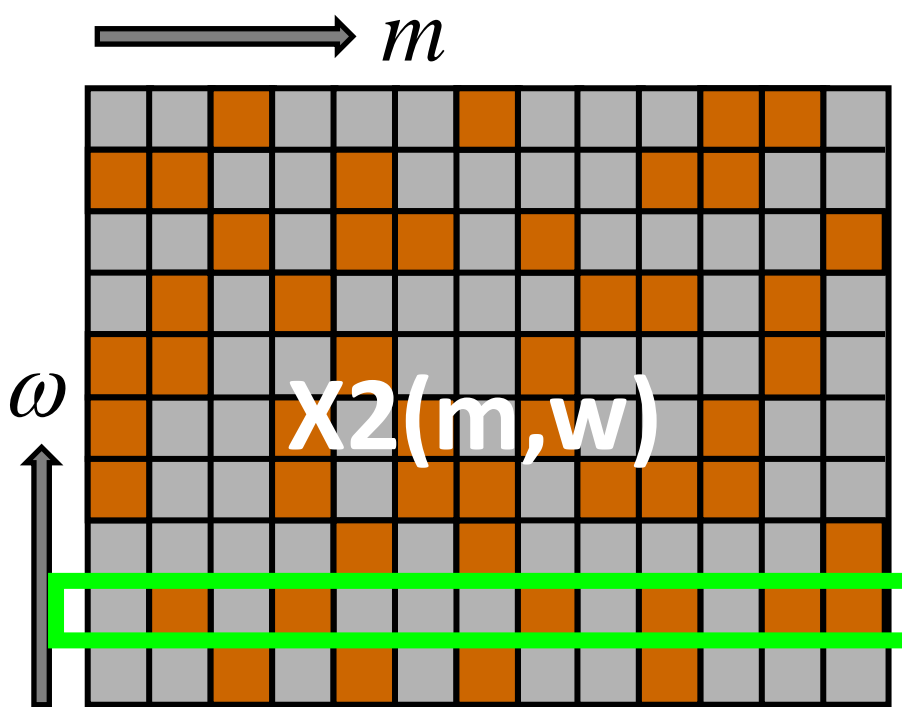
Psychophysical tests show that the ideal binary mask results in dramatic speech intelligibility improvements (Brungart et al.'06; Li & Loizou'08). Example from D.L. Wang, OSU, 2006.



$$\frac{X_1(m, \omega)}{X_2(m, \omega)} \Rightarrow \alpha(m, \omega), \beta(m, w)$$

IPD

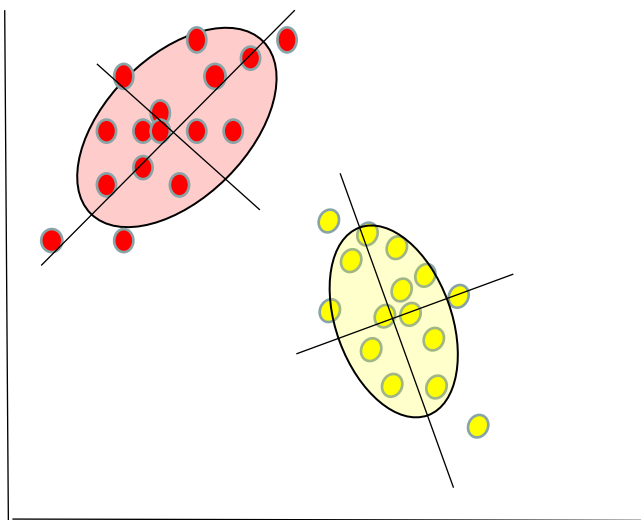
ILD



ω_2

$\beta(m, \omega_2)$

ω_2



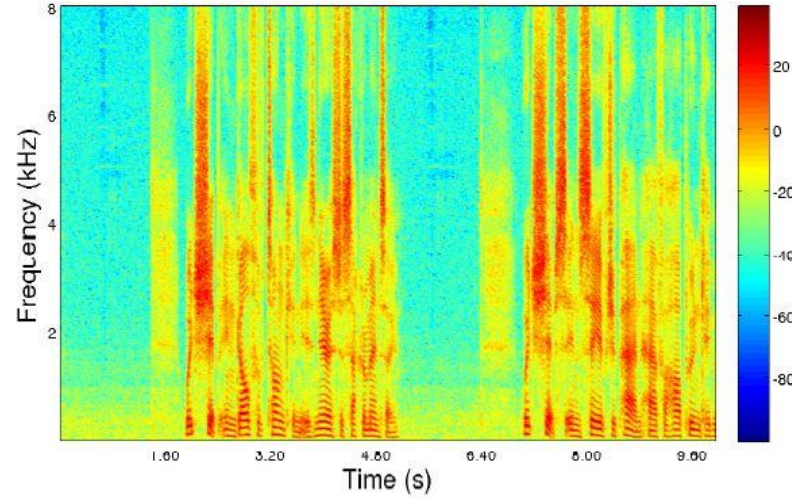
$\alpha(m, \omega_2)$



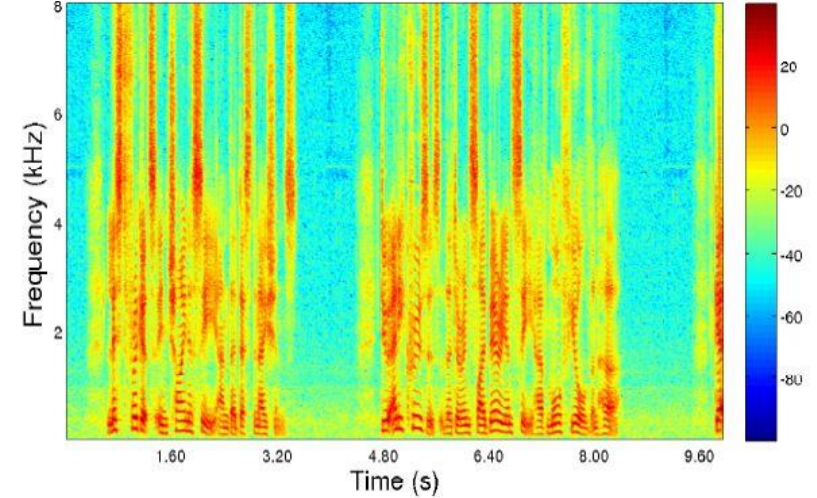
Adverse effect in speech source separation

- Acoustic noise
- Reverberations

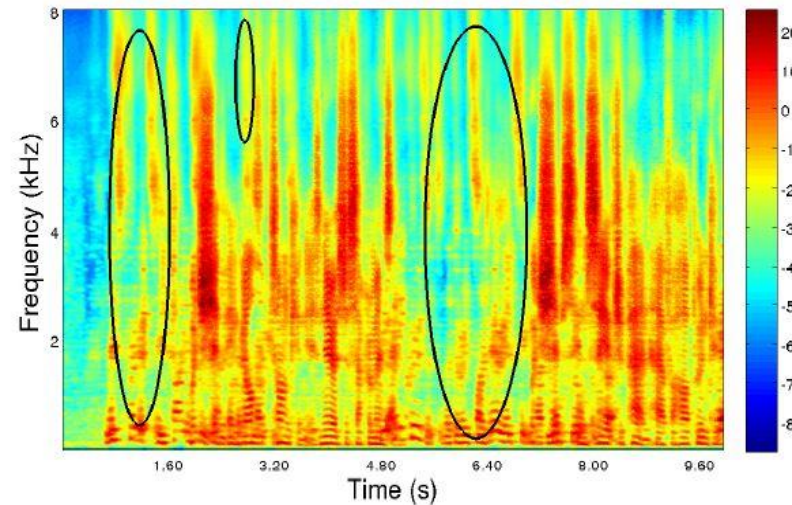
(a) Magnitude spectrum of source 1



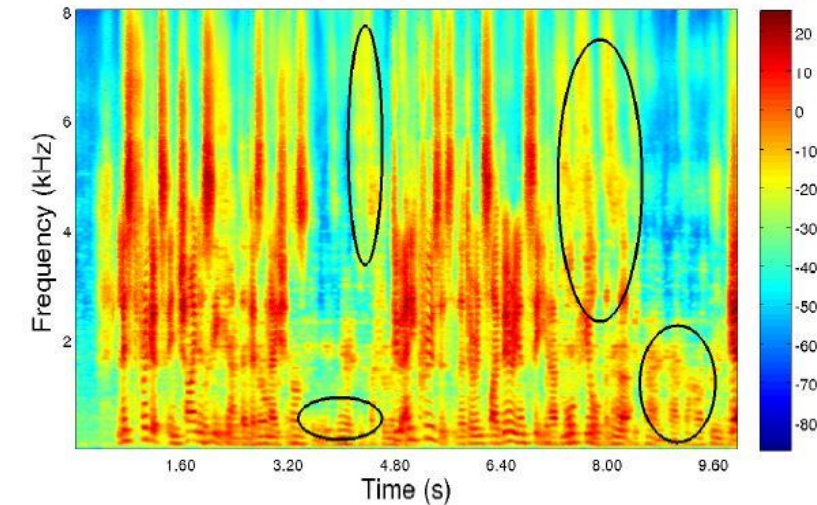
(b) Magnitude spectrum of source 2



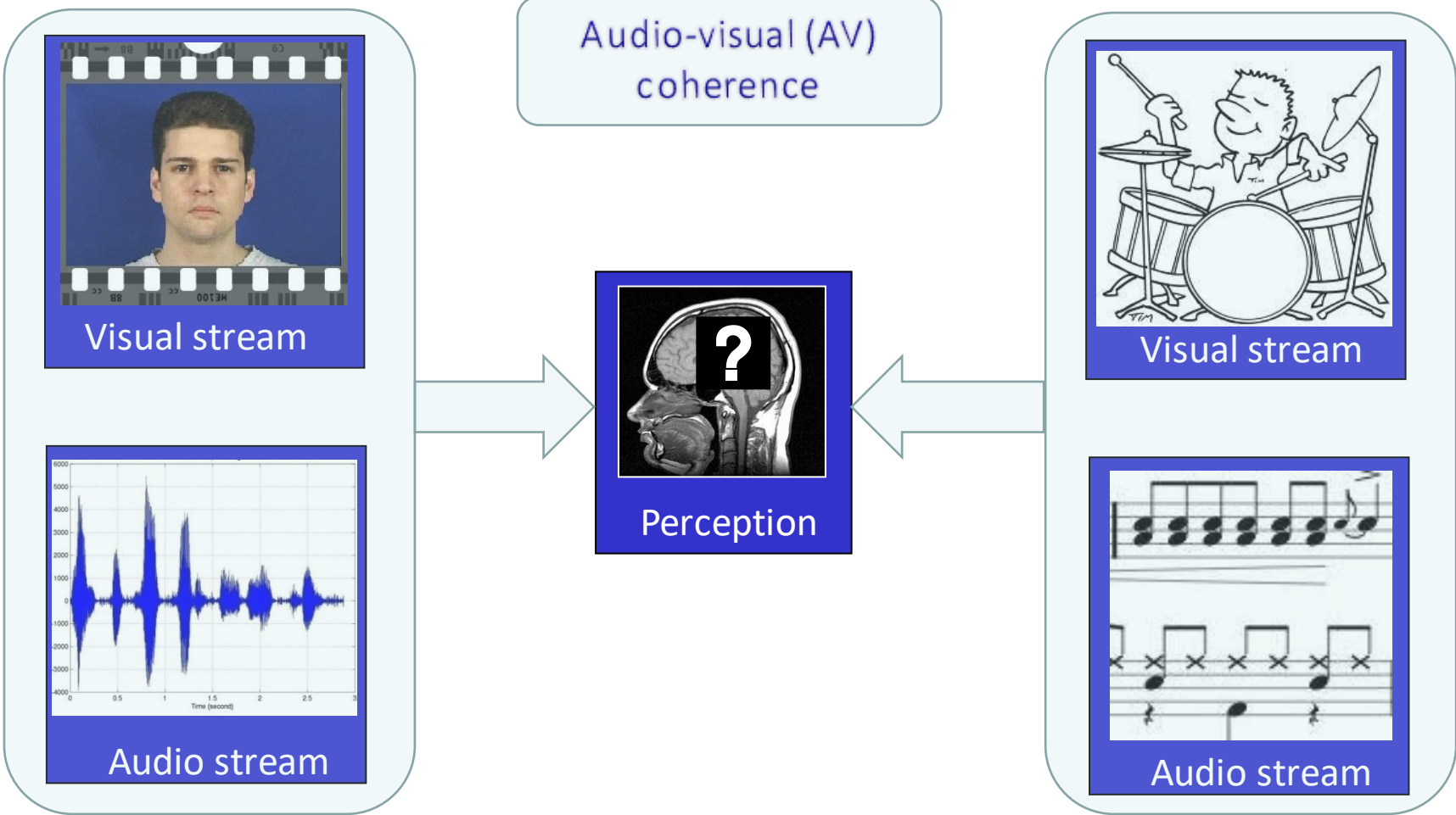
(c) Magnitude spectrum of source 1 estimate



(d) Magnitude spectrum of source 2 estimate



Audio visual speech source separation

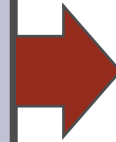


Audio-visual speech source separation

- W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video Assisted Speech Source Separation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, vol. V, pp.425-428, Philadelphia, USA, March 18-23, 2005.
- Q. Liu, W. Wang, and P. Jackson, "Use of Bimodal Coherence to Resolve Permutation Problem in Convolutional BSS," *Signal Processing*, vol. 92, no. 8, pp. 1916-1927, 2012.
- Q. Liu, W. Wang, P. Jackson, M. Barnard, J. Kittler, and J.A. Chambers, "Source separation of convolutional and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5520-5535, 2013.
- B. Rivet, W. Wang, S.M. Naqvi, and J.A. Chambers, "Audio-Visual Speech Source Separation", *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125-134, 2014.
- Q. Liu, A. Aubery, and W. Wang, "Interference Reduction in Reverberant Speech Separation with Visual Voice Activity Detection", *IEEE Transactions on Multimedia*, 2014.

Motivations and challenges

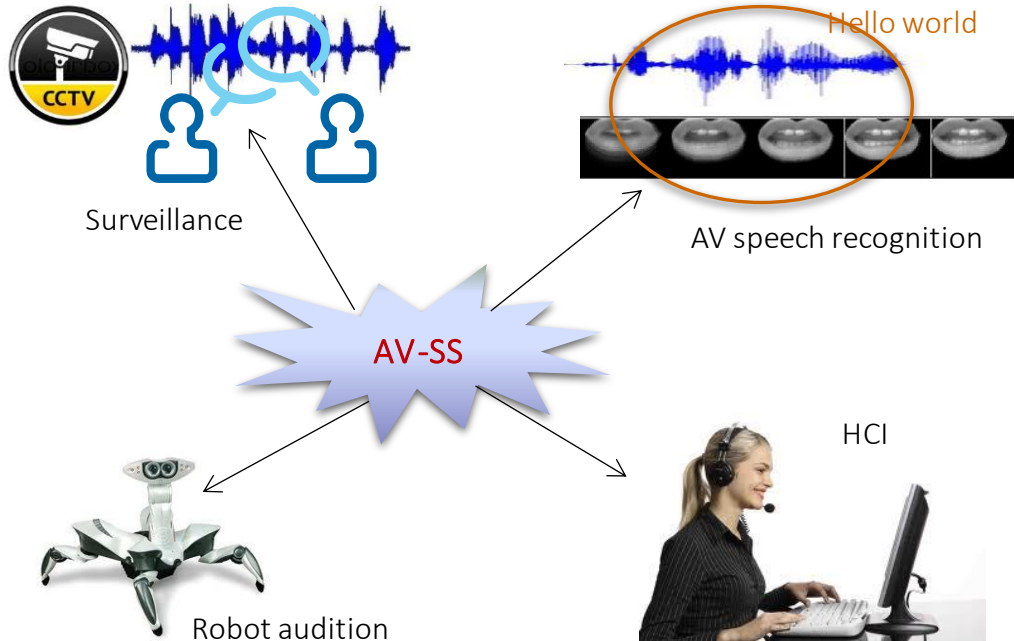
- The audio-domain SS algorithms **degrade in adverse conditions**.
- The visual stream contains **complementary information** to the coherent audio stream.



Objective

How can the visual modality be used to assist audio-domain SS algorithms in noisy and reverberant conditions?

Potential applications



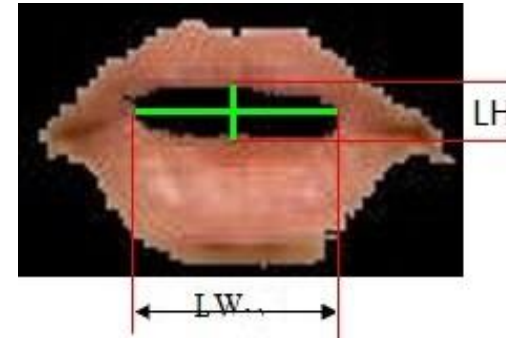
Key Challenges

- Reliable AV coherence modelling
- Bimodal differences in size, dimensionality and sampling rates
- Incorporation of AV coherence into audio-domain SS methods

Audio visual feature exaction and selection

- Visual feature extraction
 - Internal lip Width and Height
 - 2-Dimensional

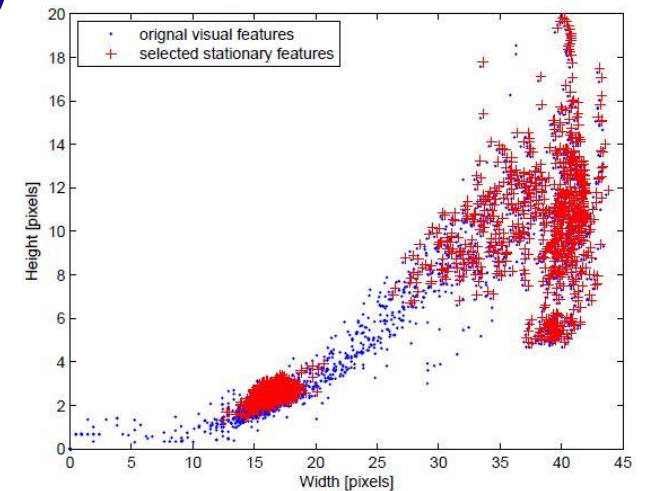
$$\mathbf{v}_T(m) = [LW(m), LH(m)]^T$$



- Audio feature extraction
 - Mel-scale Frequency Cepstrum Coefficients (MFCCs)
 - Block processing (synchronize with each video frame)
 - L-dimensional

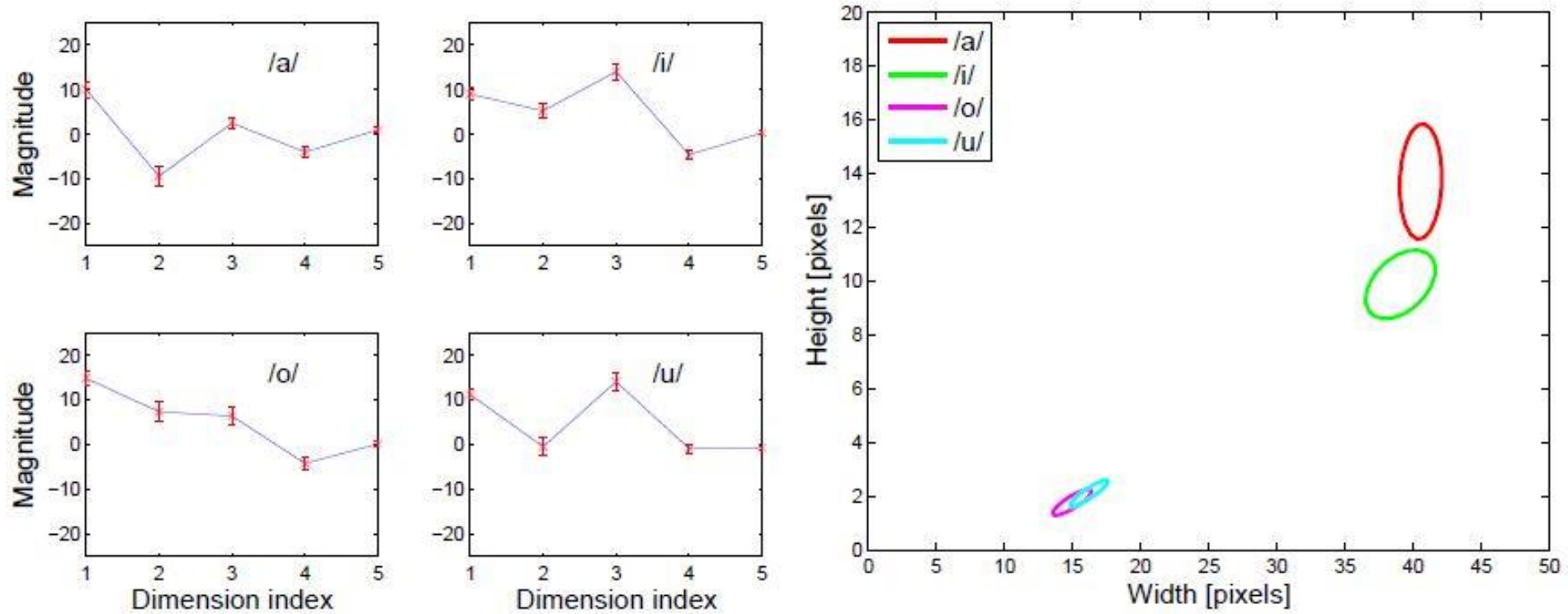
$$\mathbf{a}_T(m) = [a_{T1}(m), \dots, a_{TL}(m)]^T$$

- Audio-visual space-----Feature Selection

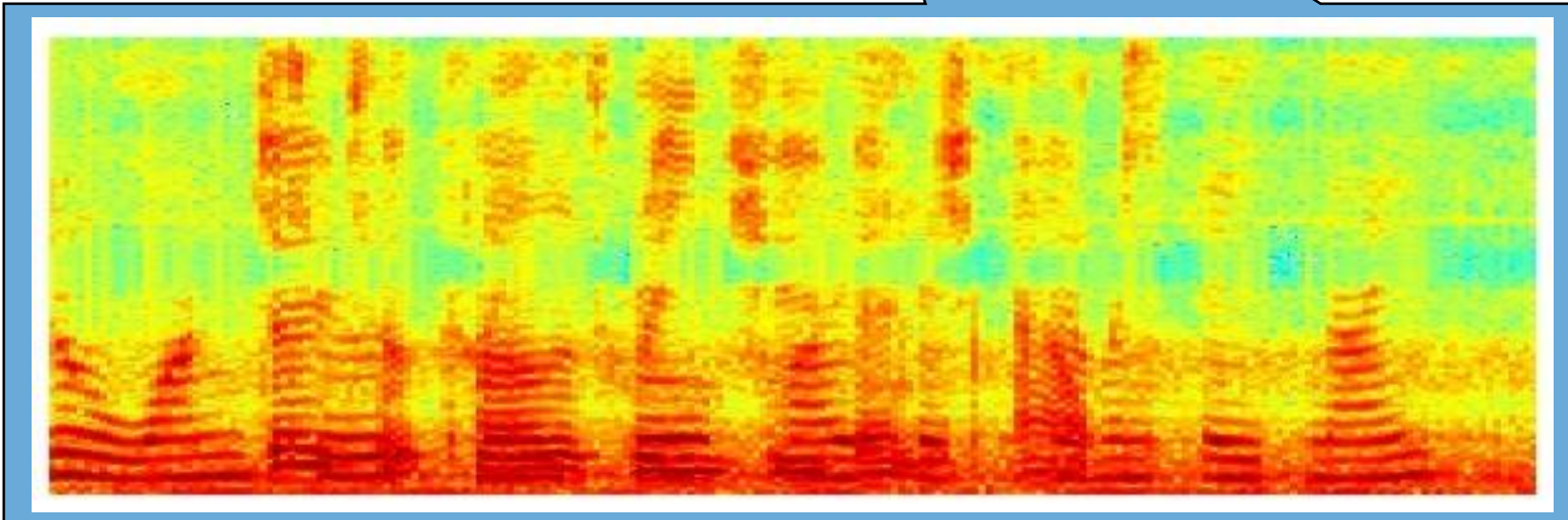
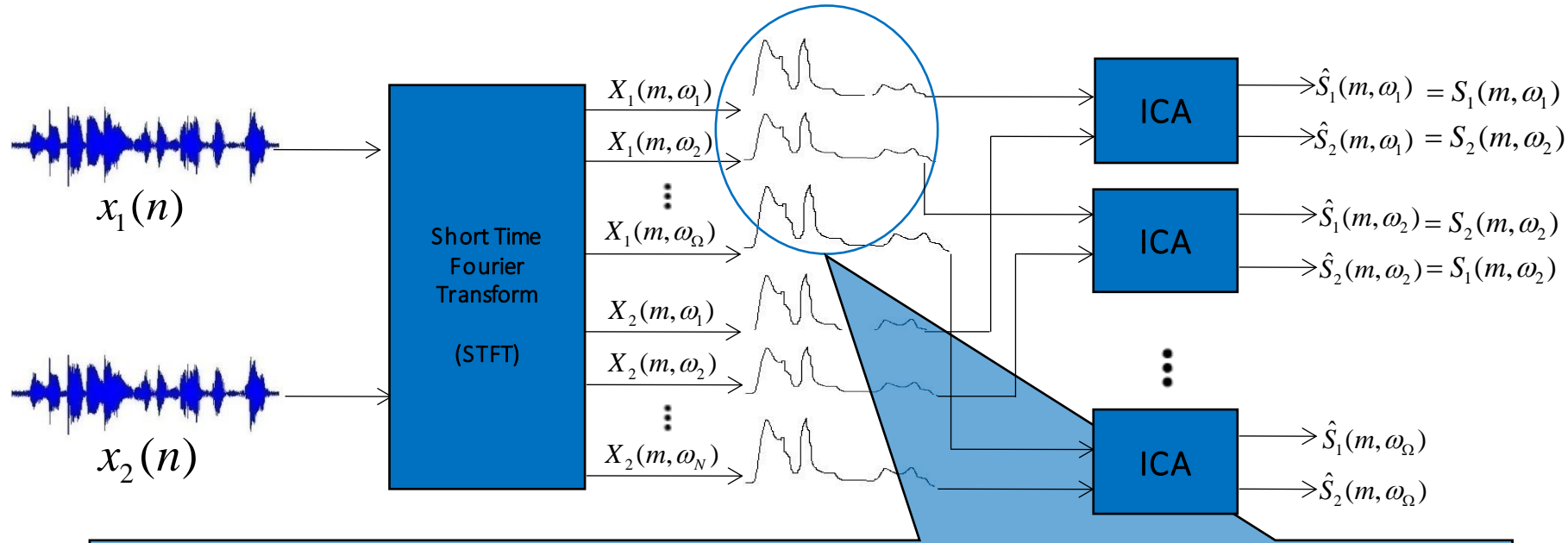


Audio visual coherence modelling using audio mixture models

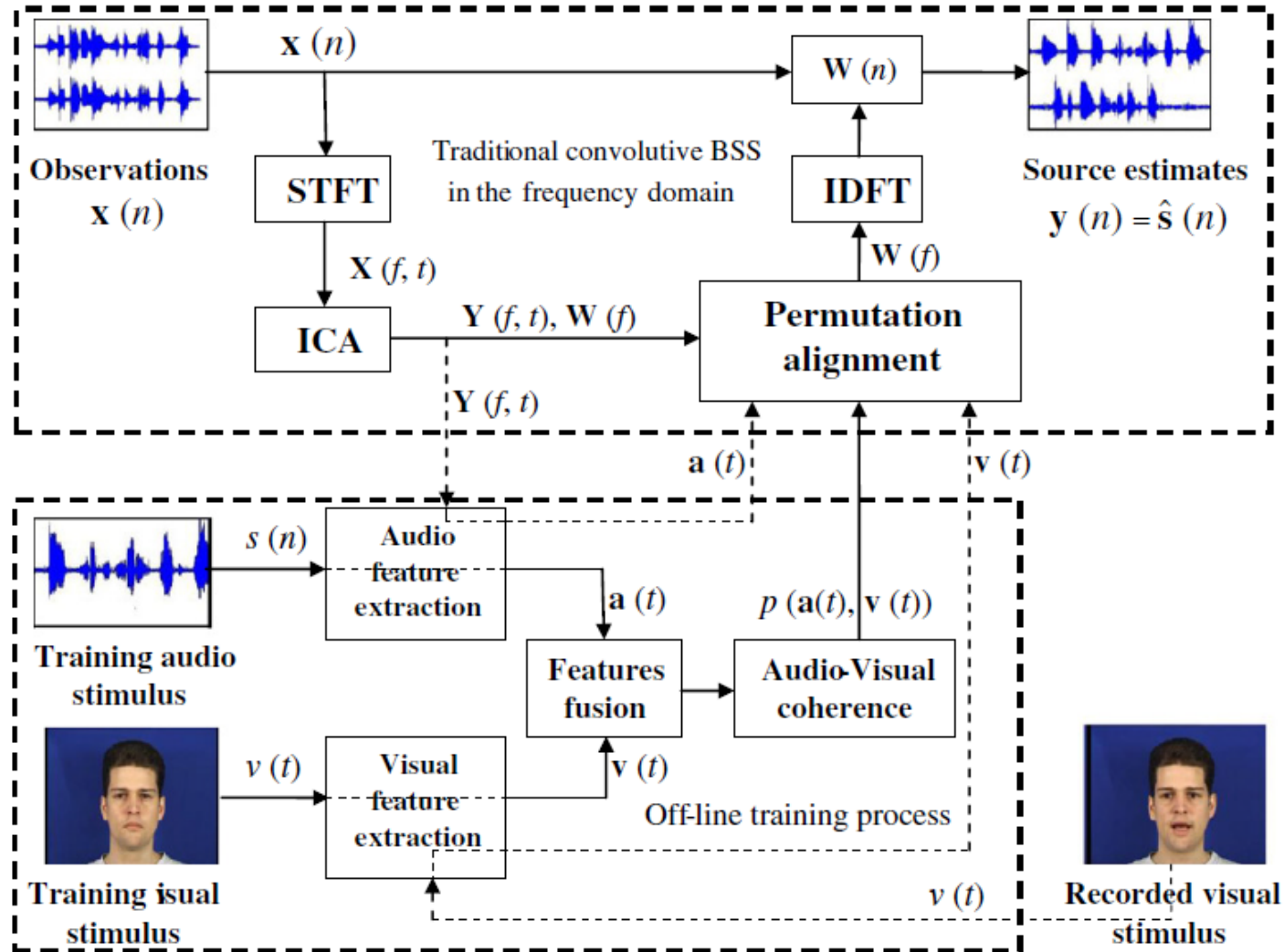
$$p(\mathbf{a}(m), \mathbf{v}(m)) = p(\mathbf{u}(m)) = \sum_{d=1}^D w_d \mathcal{N}(\mathbf{u}(m) \mid \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$$



Frequency domain speech source separation with ICA



Audio-visual independent component analysis (ICA)

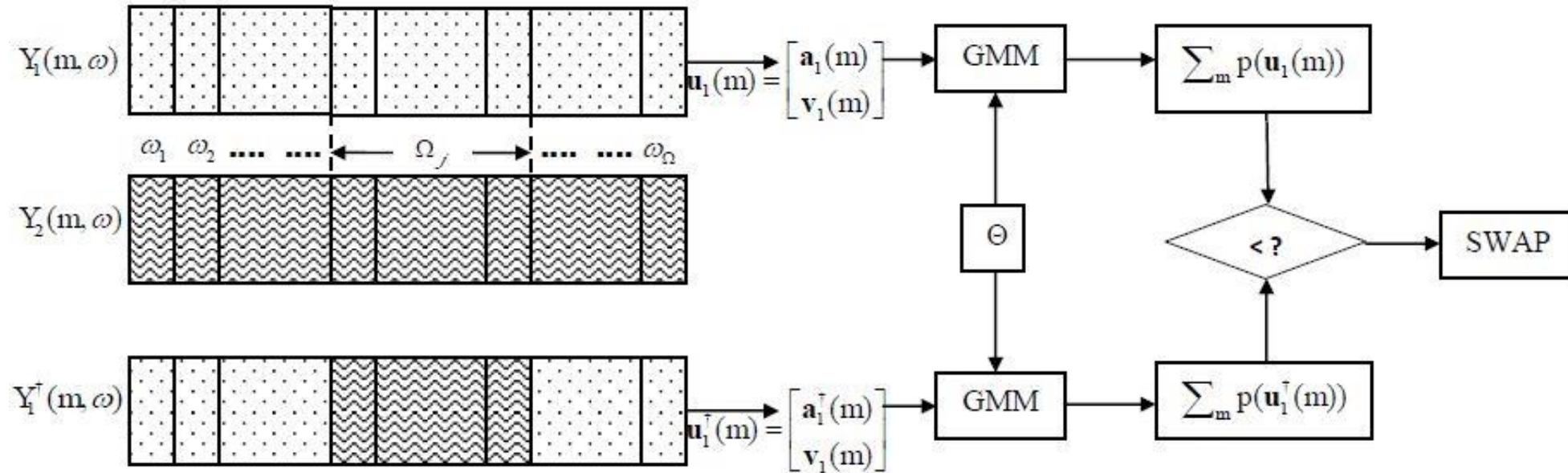


Solving the permutation problem

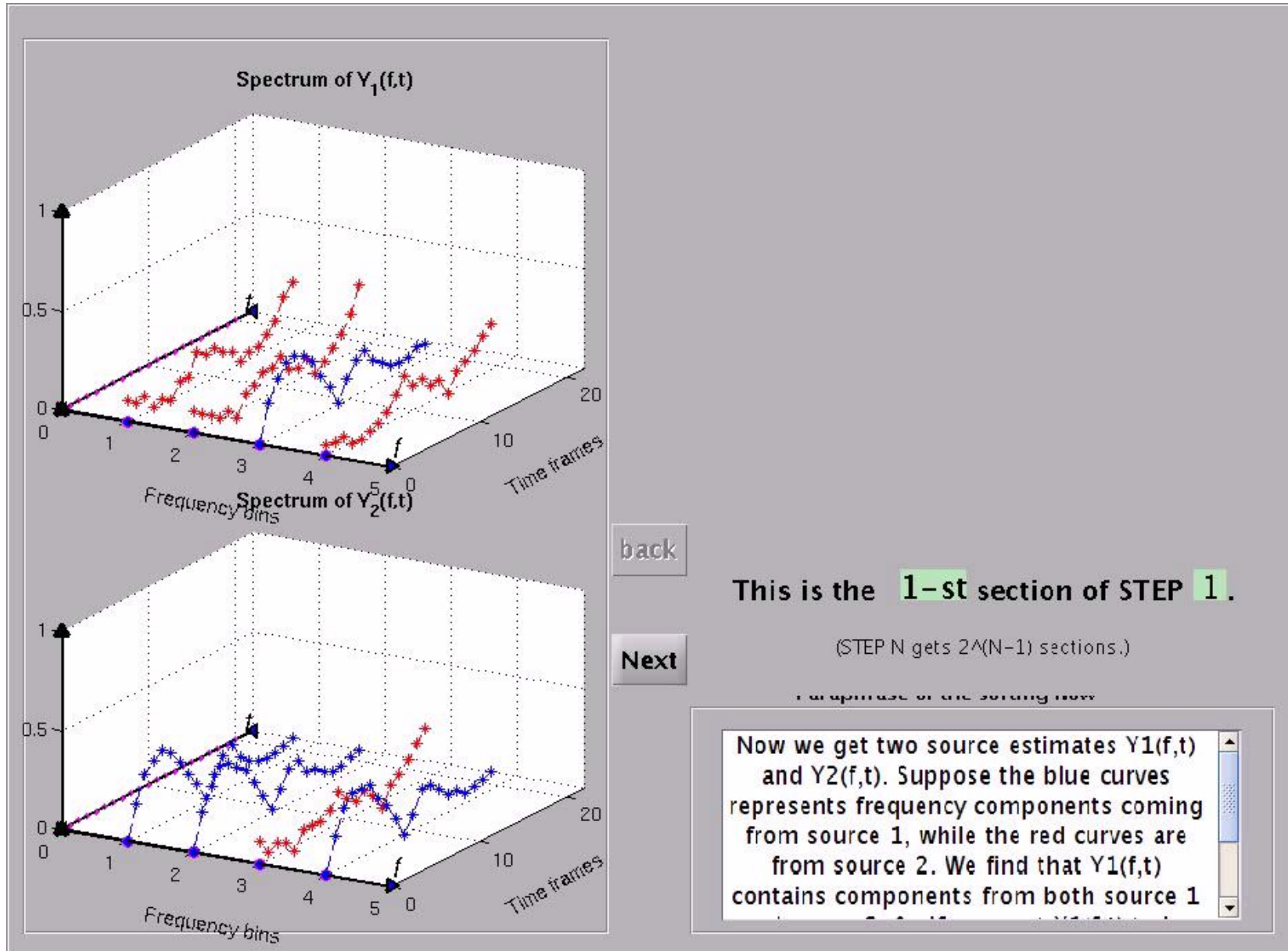
Objective

$$\hat{\mathbf{P}}(\omega) = \arg \max_{\mathbf{P}(\omega)} \sum_m \sum_{k=1}^K p(\mathbf{u}_k(m))$$

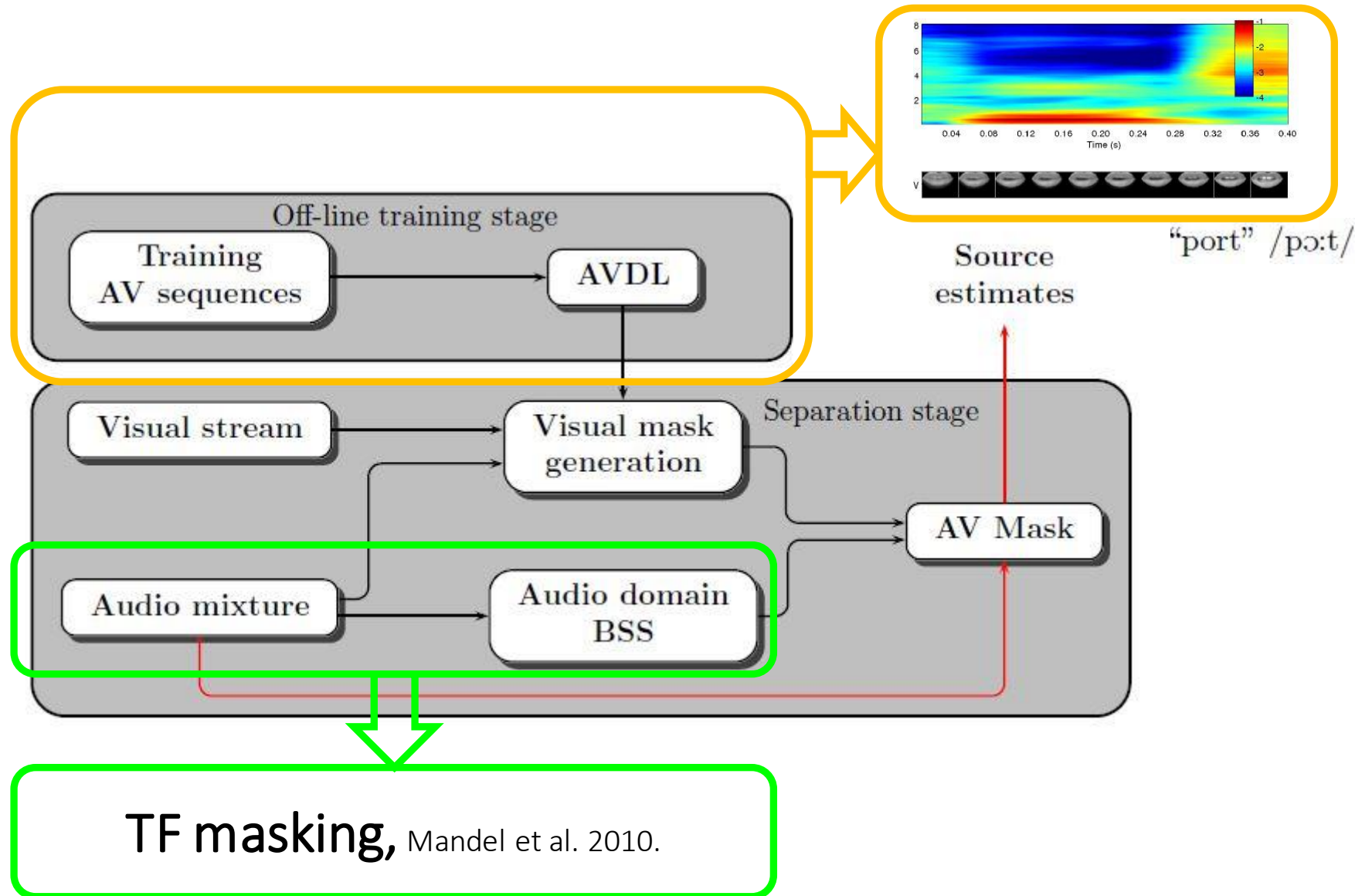
Solution: An iterative sorting scheme



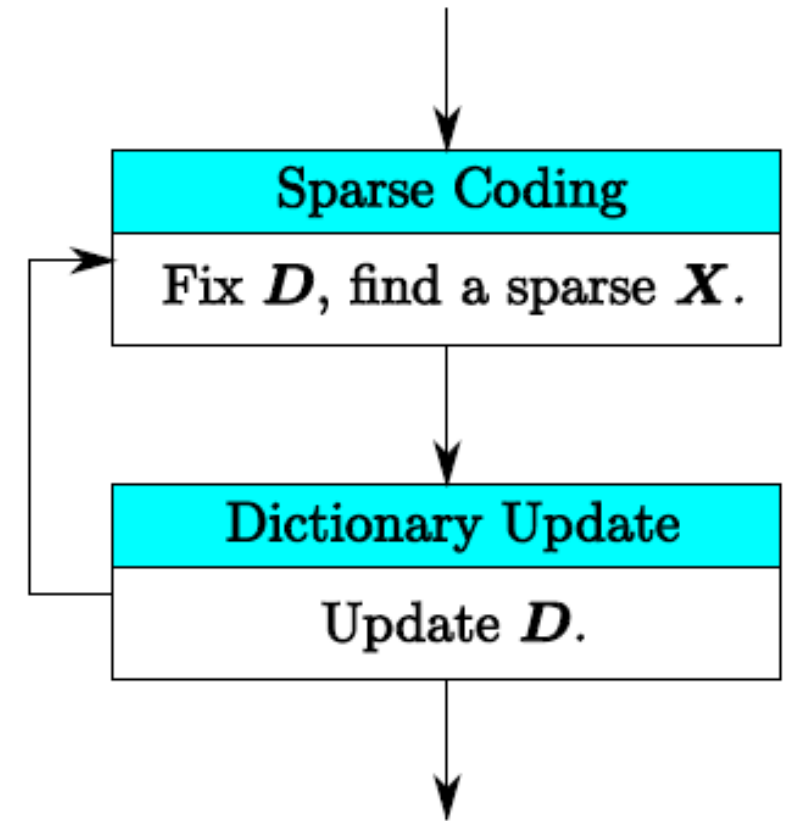
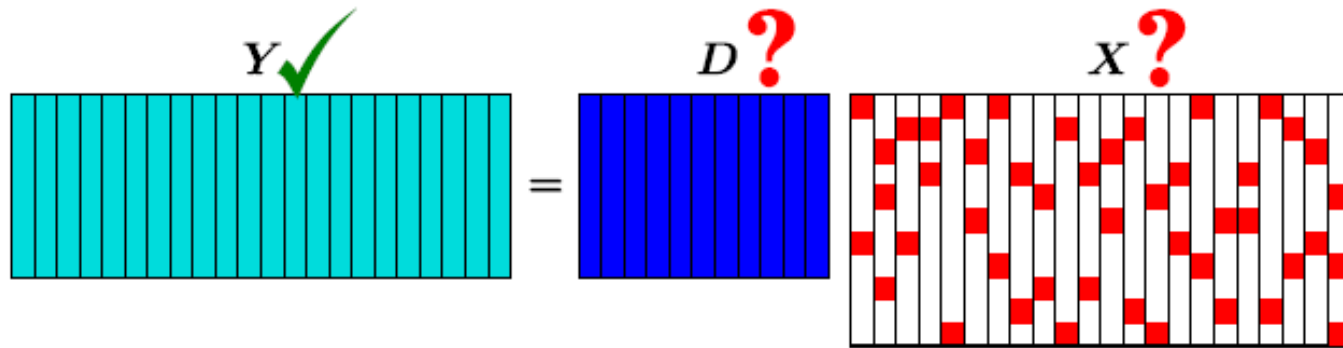
Demo for the solution to the permutation problem



Sparse representation based AV coherence modelling with dictionary learning

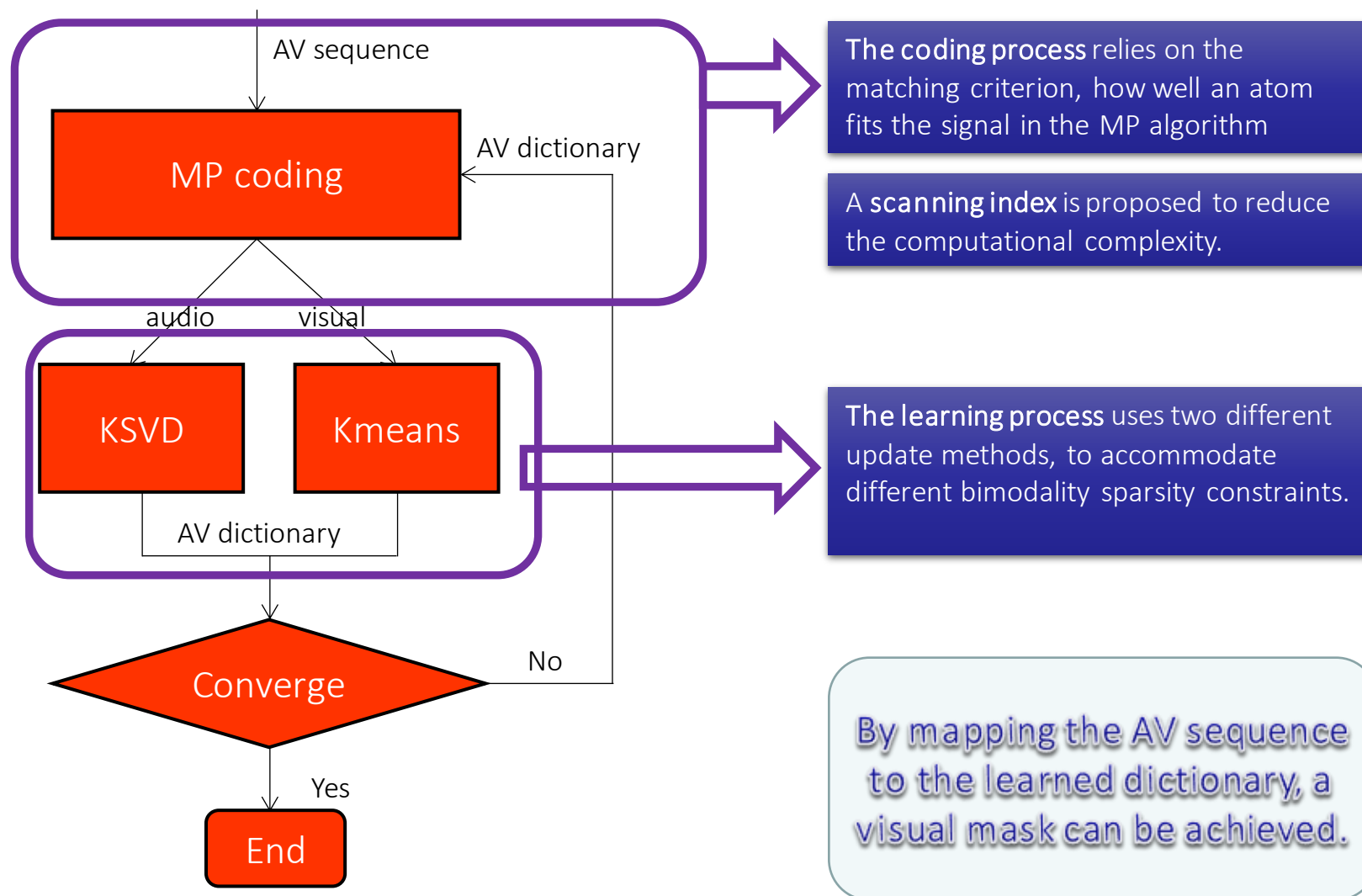


Dictionary learning



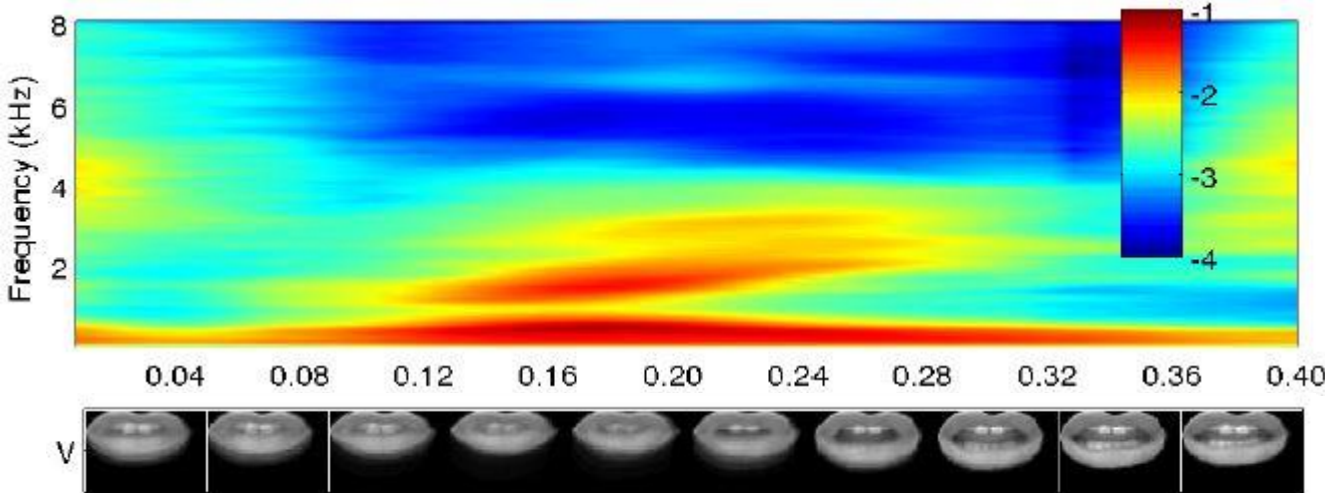
W. Dai, T. Xu, and W. Wang, "Simultaneous Codeword Optimisation (SimCO) for Dictionary Update and Learning", *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6340-6353, 2012.

AV dictionary learning



Q. Liu, W. Wang, et al., "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5520-5535, 2013.

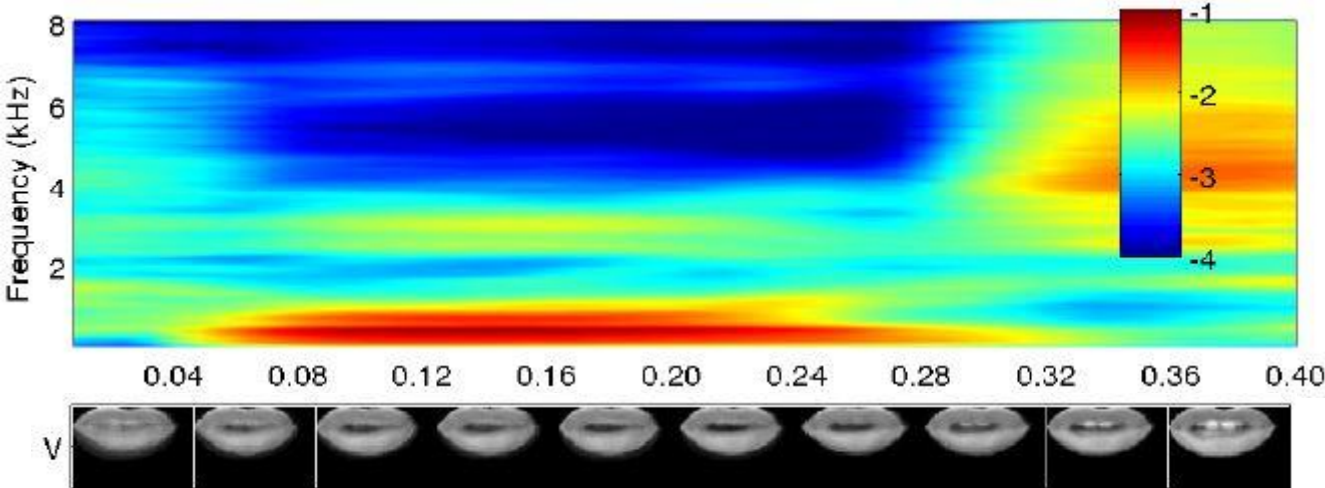
An example of the AV atoms learned from visual speech



Long Speech

Sheerman-Chase et al.
LILIR Twotalk database
2011

Lip tracking,
Ong et al. 2008



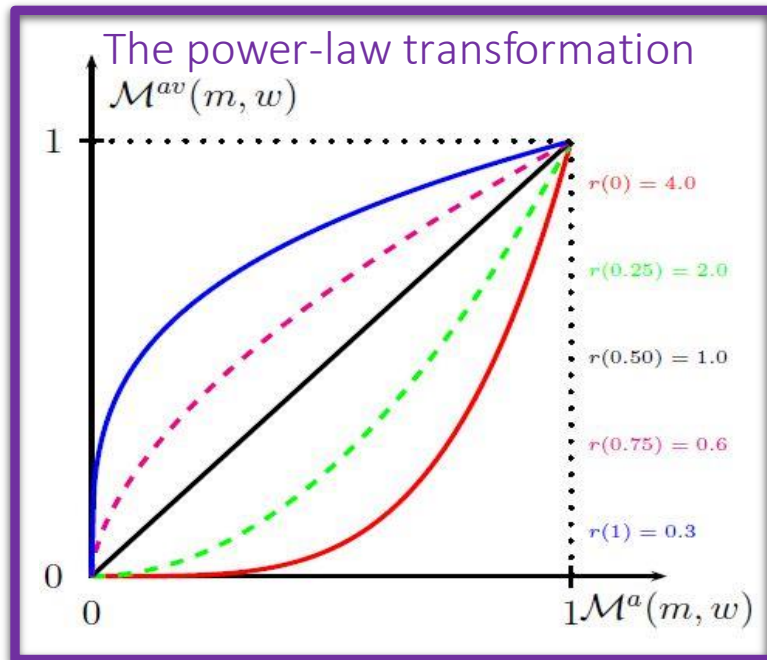
The first AV atom represents the utterance "marine" /m^əri:n/ while the second one denotes the utterance "port" /p^{ɔː}t/.

Improving audio mask with visual information

$$\mathcal{M}^{av}(m, \omega) = \mathcal{M}^a(m, \omega)^{r(\mathcal{M}^v(m, \omega))}$$

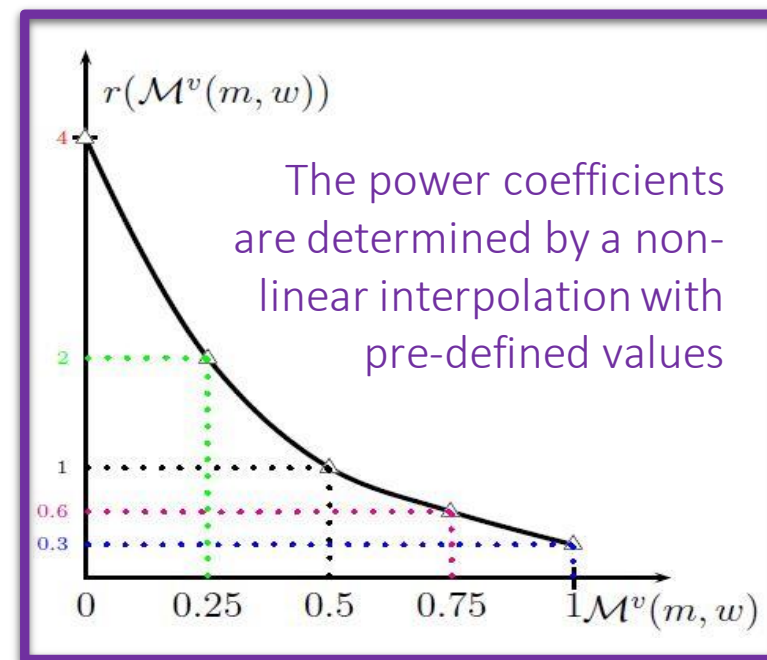
Audio mask

Statistically generated by evaluating the IPD and ILD of each TF point.



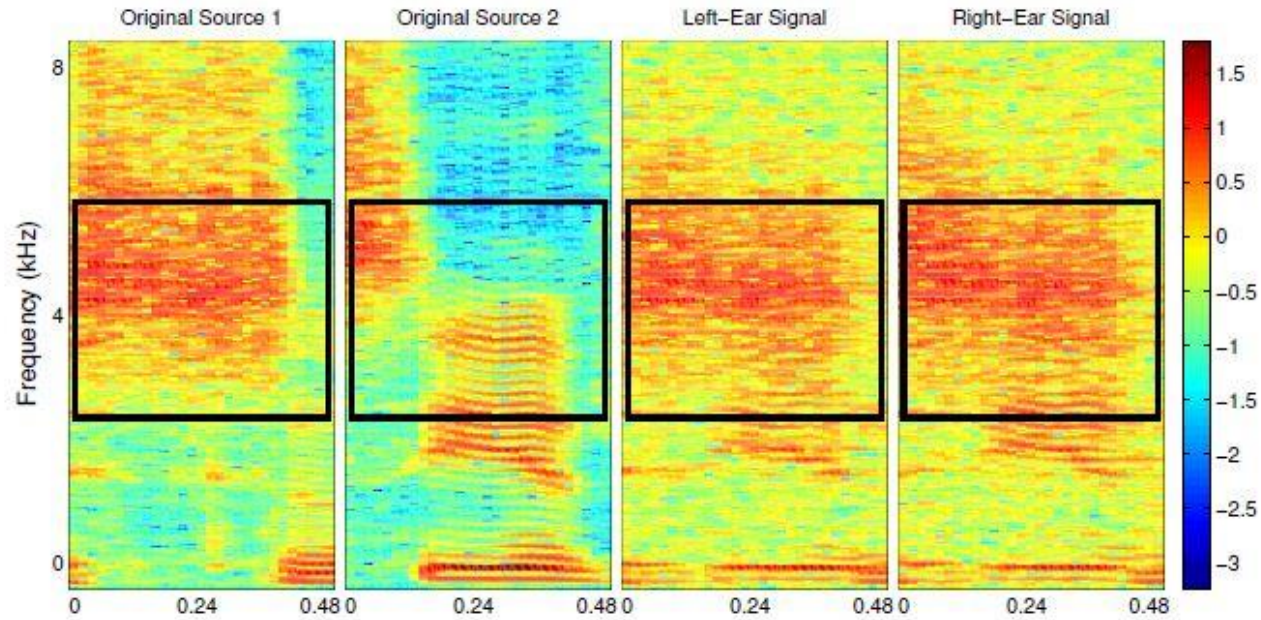
Visual mask

Mapping the observation to the learned AV dictionary via the coding stage in AVDL.



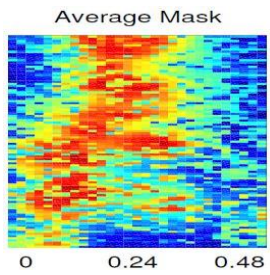
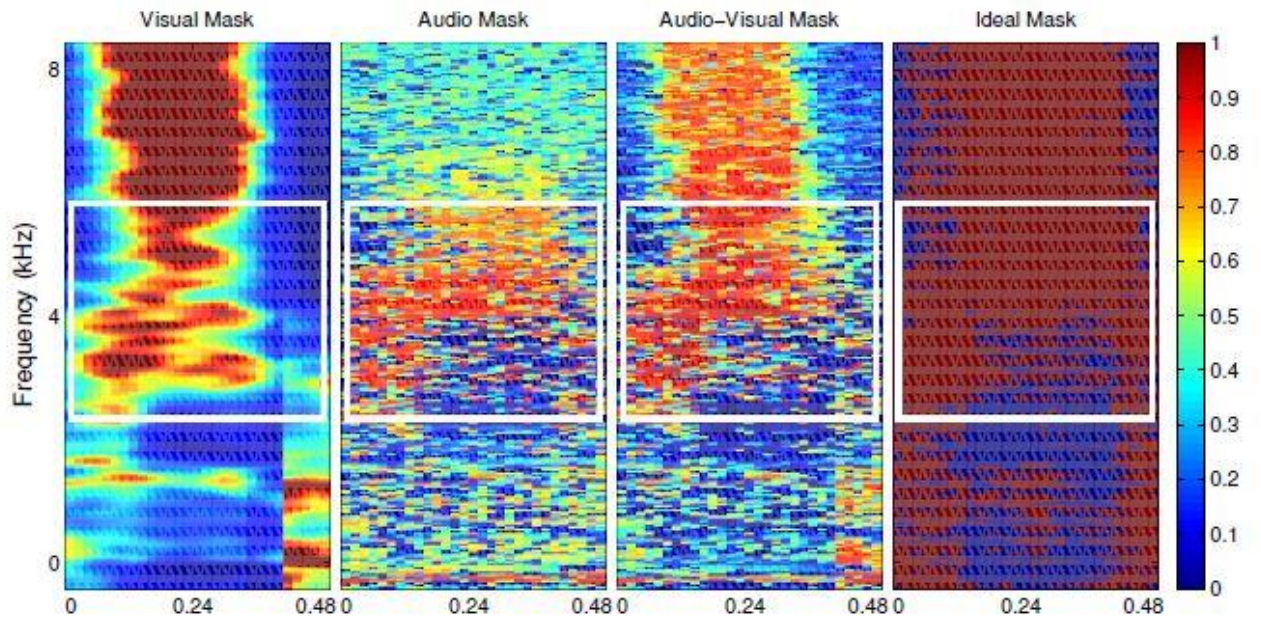
$$\mathcal{M}^v(m, \omega) = \begin{cases} 1, & \text{if } \hat{\psi}^a(m, \omega) > \psi^a(m, \omega) \\ \hat{\psi}^a(m, \omega) / \psi^a(m, \omega), & \text{otherwise.} \end{cases}$$

Audio-visual time frequency mask

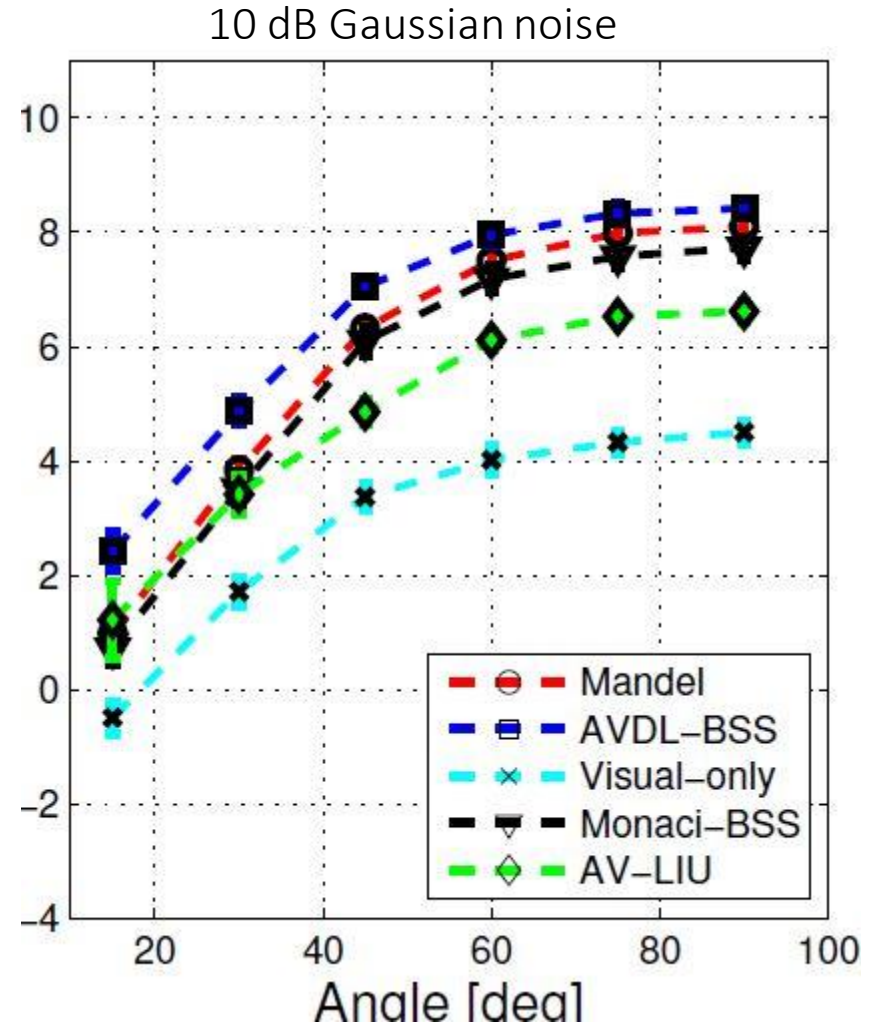
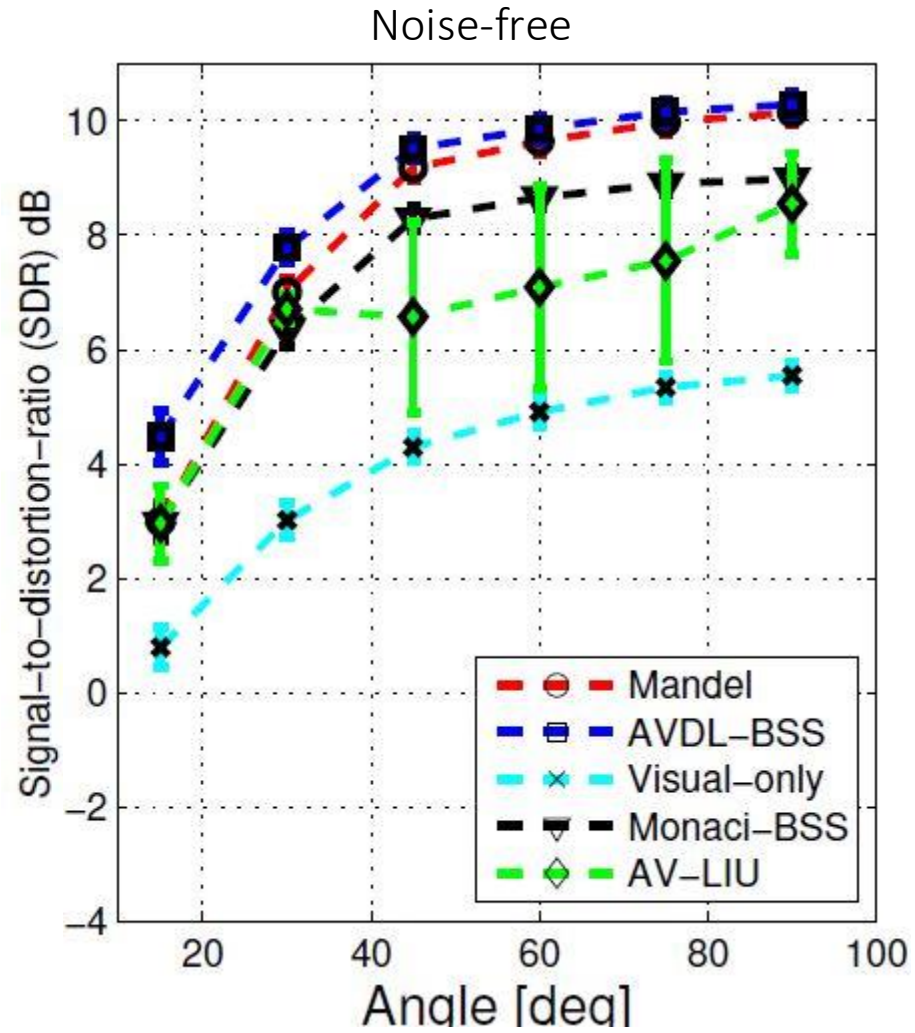


Demonstration of TF mask fusion in AVDL-BSS





























Why do we choose the power law combination, instead of, e.g., a linear combination?



Evaluations on audio-visual speech source separation



Demos

	Mixture	Ideal	Mandel	AV-LIU	AVDL-BSS	Rivet	AVMP-BSS
A							
B							
C							
D							

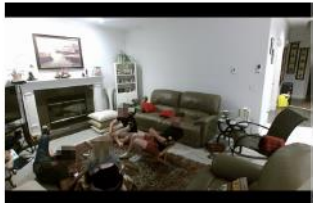
Q. Liu, W. Wang, et al., "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5520-5535, 2013.

Recent developments in speech separation and related tasks

Input	Pipeline system			End-to-End system
	Separation & Beamforming	Multi-talker Diarization	Multi-talker ASR	
Single-channel audio	Yu et al. 2017; Luo et al. 2018 Isik et al. 2016; Wang et al. 2018 Bahmaninezhad et al. 2019	Fujita et al. 2019 Medennikov et al. 2020 Kanda et al. 2022b Maiti et al. 2023;	Hershey et al. 2010 Qian et al. 2018; Gulati et al. 2020 Zhang et al. 2020; Neumann et al. 2020; Seki et al. 2018; Sklyar et al. 2021; Kanda et al. 2019, 2021b, 2022a	Delcroix et al. 2019; Mao et al. 2020 Lu et al. 2021;
Multi-channel audio	Chen et al. 2018, 2019 Gu et al. 2019	Zheng et al. 2022b Horiguchi et al. 2023 Yu et al. 2022	Subramanian et al. 2021, 2022 Erdogan et al. 2016; Ochiai et al. 2017 Watanabe et al. 2020; Shi et al. 2022; Masuyama et al. 2023	Raj et al. 2021; Yu et al. 2022
Multi-channel multi-modal	Wu et al. 2019; Gu et al. 2020, 2022 Gogate et al. 2020; Xu et al. 2021 Zhang et al. 2021; Li et al. 2021	Ding et al. 2020 Kang et al. 2020	Yu et al. 2020a,b Wu et al. 2021; Shao et al. 2022; Wang et al. 2022;	Yoshioka et al. 2019; Yu et al. 2021;

Acknowledgement to Y. Xu, Tencent US, for providing this table. “Multi-modal Multi-talker Speech Recognition, Separation and Diarization, Everything Streaming All at Once”

Data challenges in (AV) speech separation



<https://chimechallenge.github.io/chime6/>



<https://mispchallenge.github.io/mispchallenge2022>

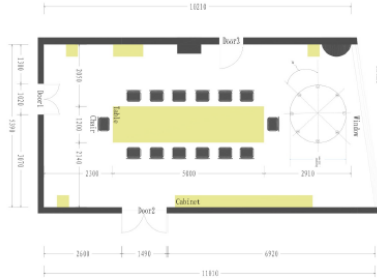


AMI meeting



<http://corpus.amiproject.org/>

M2MeT challenge -- AliMeeting



<https://www.alibabacloud.com/zh/m2met-alimeeting>

2nd COG-MHEAR Audio-Visual Speech Enhancement Challenge (AVSE)

A machine learning challenge for next-generation hearing devices

Get Started

This site provides full documentation of the challenge datasets, baseline systems and rules for participation.

<https://challenge.cogmhear.org/>

MMCSG (Multi-Modal Conversations in Smart Glasses) dataset

<https://ai.meta.com/datasets/mmcs-g-dataset/>



Other recent developments in related field

AV speech separation

- R. Gao and K. Grauman, "VISUALVOICE: Audio-Visual Speech Separation with Cross-Modal Consistency," Proc. CVPR, 2021.
A. Nagrani, et al., "Seeing Voices and Hearing Faces: Cross-modal biometric matching," in Proc. CVPR, 2018.

AV general sound separation

- C. Gan, et al. "Music Gesture for Visual Sound Separation," in Proc. CVPR, 2020.
E. Tzinis, S. Wisdom, T. Remez, and J.R. Hershey, "AudioScopeV2: Audio-Visual Attention Architectures for Calibrated Open-Domain On-Screen Sound Separation", in Proc. ECCV, 2022.

Universal sound separation

- I. Kavalerov, et al., "Universal Sound Separation," in Proc. IEEE WASPAA, 2019.
Q. Kong et al., "Universal Source Separation with Weakly Labelled Data," arXiv:2305.07447, 2023.

Text-prompted/language guided universal sound source separation

- X. Liu, et al. "Separate What You Describe: Language-Queried Audio Source Separation," in Proc. Interspeech 2022.
X. Liu, et al. "AudioSep : Separate Anything You Describe", arXiv:2308.05037, 2023.

Language guided AV source separation

- Dong, et al., "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," in ICLR 2022.
R. Tan, et al., "Language-Guided Audio-Visual Source Separation via Trimodal Consistency," in Proc. CVPR, 2023.

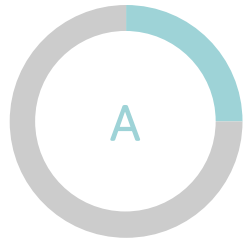
- 1 Audio-visual speech source separation
- 2 Audio-visual multi-speaker localization/tracking
- 3 Ego-centric audio-visual speaker localization/tracking
- 4 Conclusion and Future Works

Audio-visual multi-speaker localization/tracking

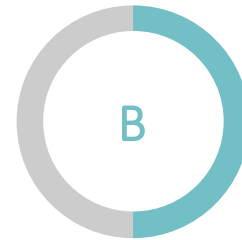
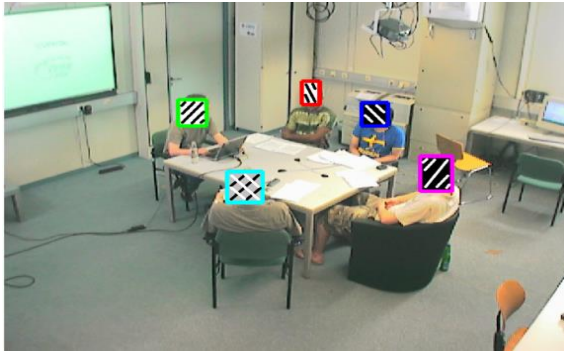
Aim: estimate the number of speakers and their positions from audio-visual data



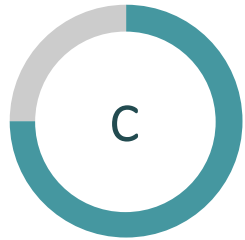
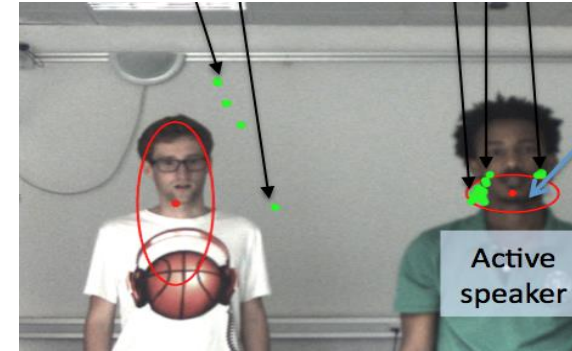
Application examples



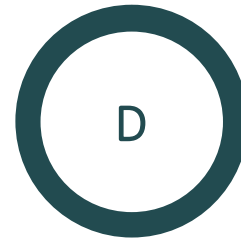
Video Conferencing



Speaker Identification



Human-computer/robot Interaction



Monitoring



Challenges of Multi-Speakers Tracking

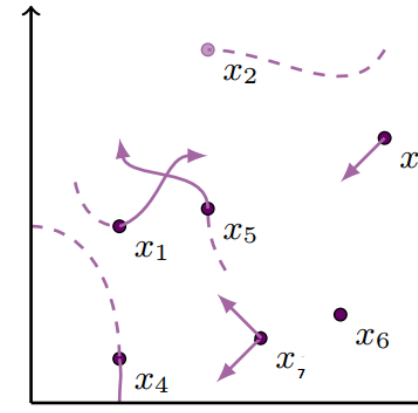
1. Unreliable Measurements

- a. miss-detection
- b. occlusion
- c. noise
- d. clutters

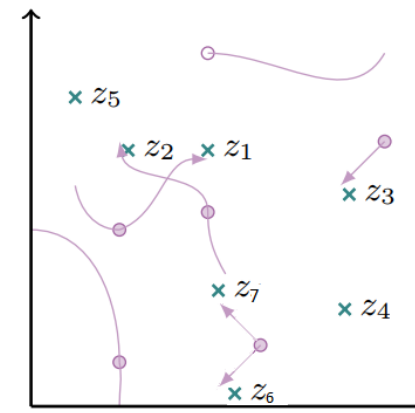
2. Unknown and Varying Number of Speakers

- a. born speaker
- b. spawned speaker
- c. dead speaker

3. Disordered Speakers



Object state space



Measurement space

Tracking Modalities

i.e., color, contour, texture and motion, scale-invariant feature transform, neural network learned features.

Benefit: Low measurement noise.

Drawback: Limited tracking area.

Visual Cues

Audio Cues

i.e., beamforming, super-resolution spectral estimation and time delay estimation.

Benefit: Easy to distinguish occluded speakers and unlimited broad tracking area.

Drawback: High measurement noise.

Modality Fusion

With a single modality, the targets may not be detected. Multi-modality fusion provides an effective solution to improve the tracking performance.

Methods for Multi-Speaker Tracking

Three main methods for multi-target tracking:

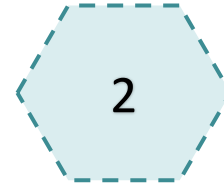


Joint probabilistic data association filter (JPDAF)

i.e. Kalman filter, Extended Kalman Filter [T. E. Fortmann et al. 1983] and particle filter [Khan et al. 2004].

Benefit: Easily implement.

Drawback: Speaker state is only calculated by nearby measurement.

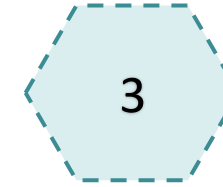


Multiple hypothesis tracking (MHT)

i.e. Kalman-Consensus Filter [Olfati-Saber et al. 2009] and particle filter [Kim et al. 2006].

Benefit: Tracking unknown number of speakers.

Drawback: Assume that speakers are detected.



Random finite set (RFS)

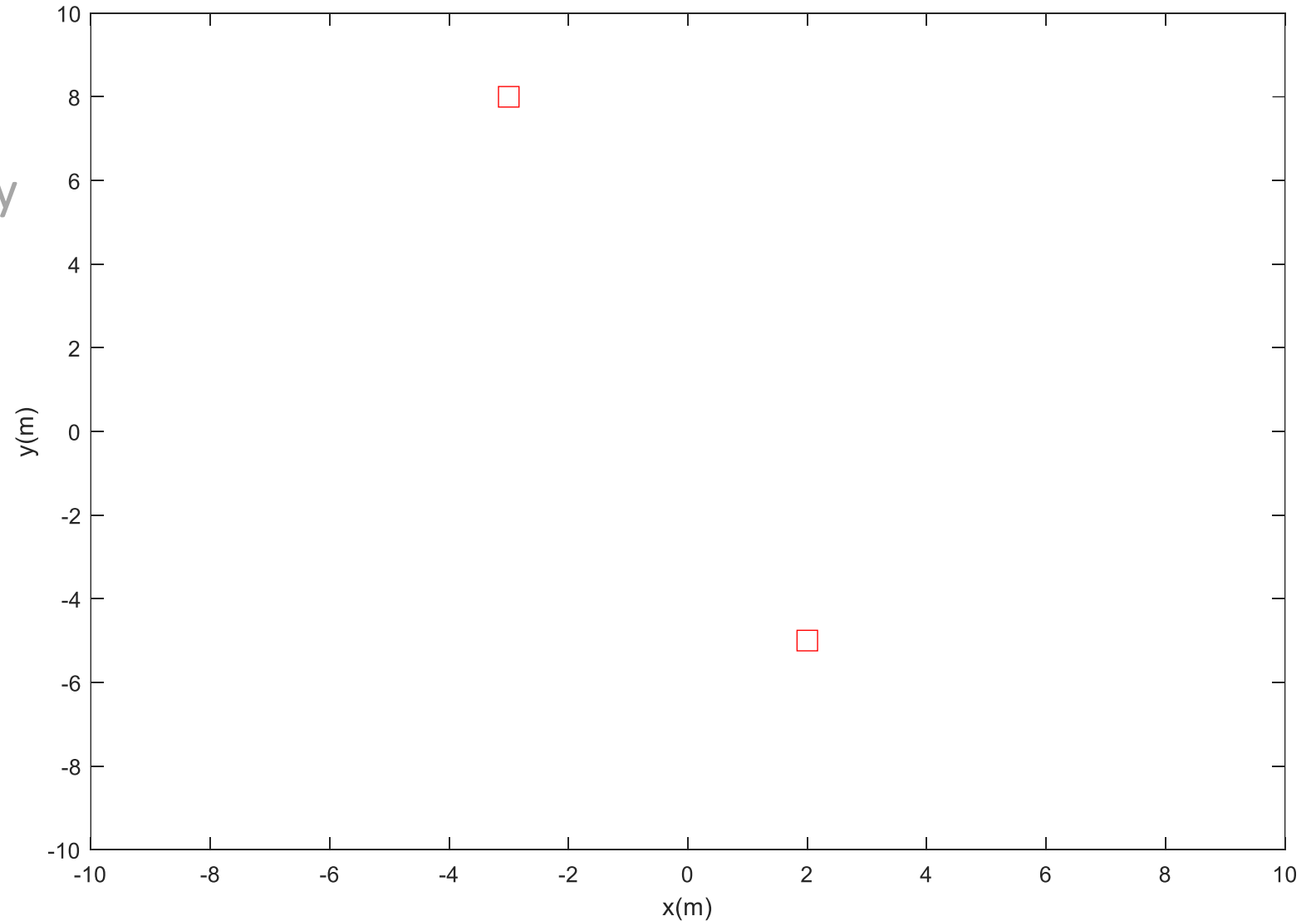
i.e. PHD filter [R. Mahler et al. 2003] and Bernoulli filter [R. Mahler et al. 2007].

Benefit: Tracking unknown number of speakers and speakers can disappear.

Drawback: High computational cost

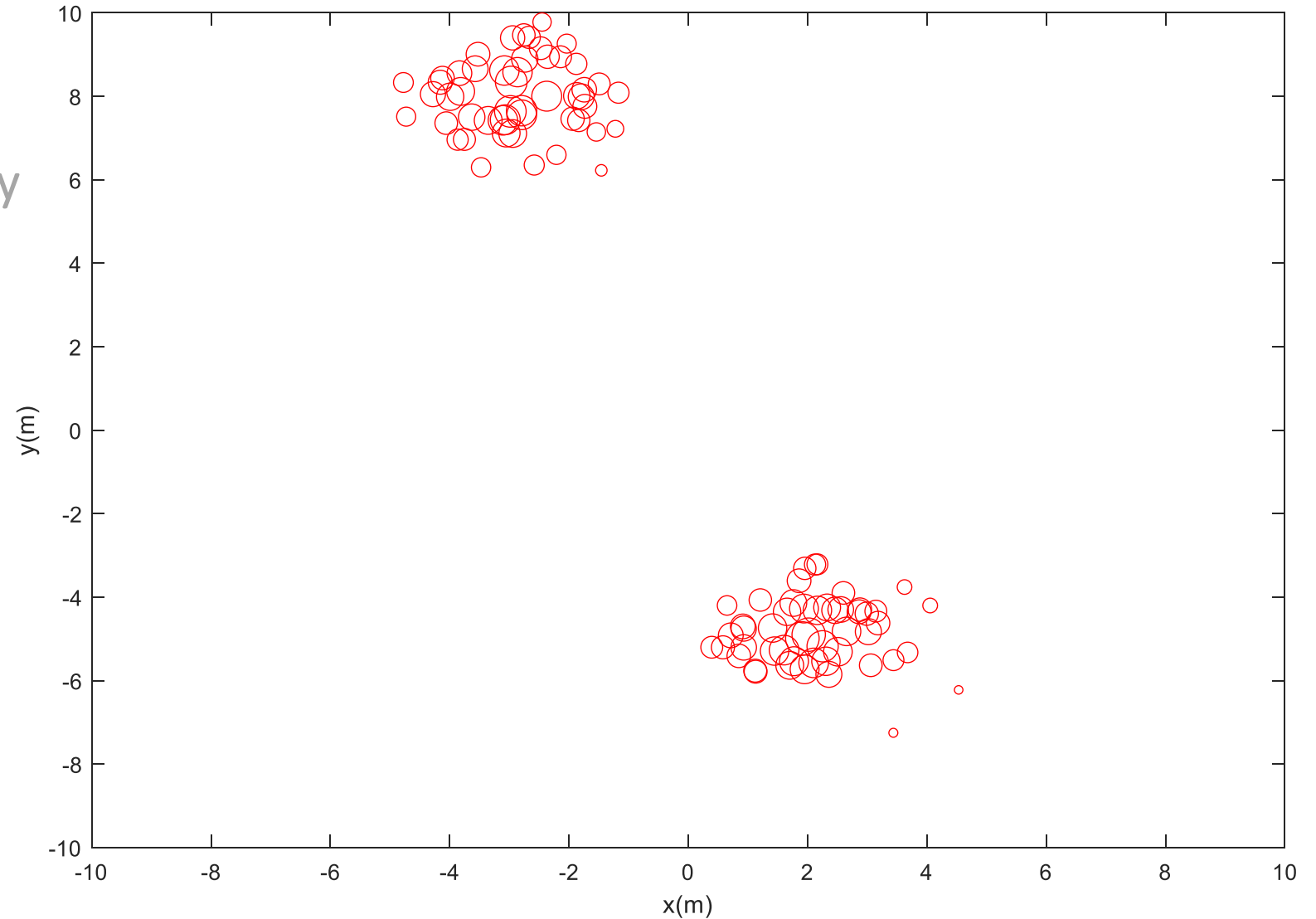
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



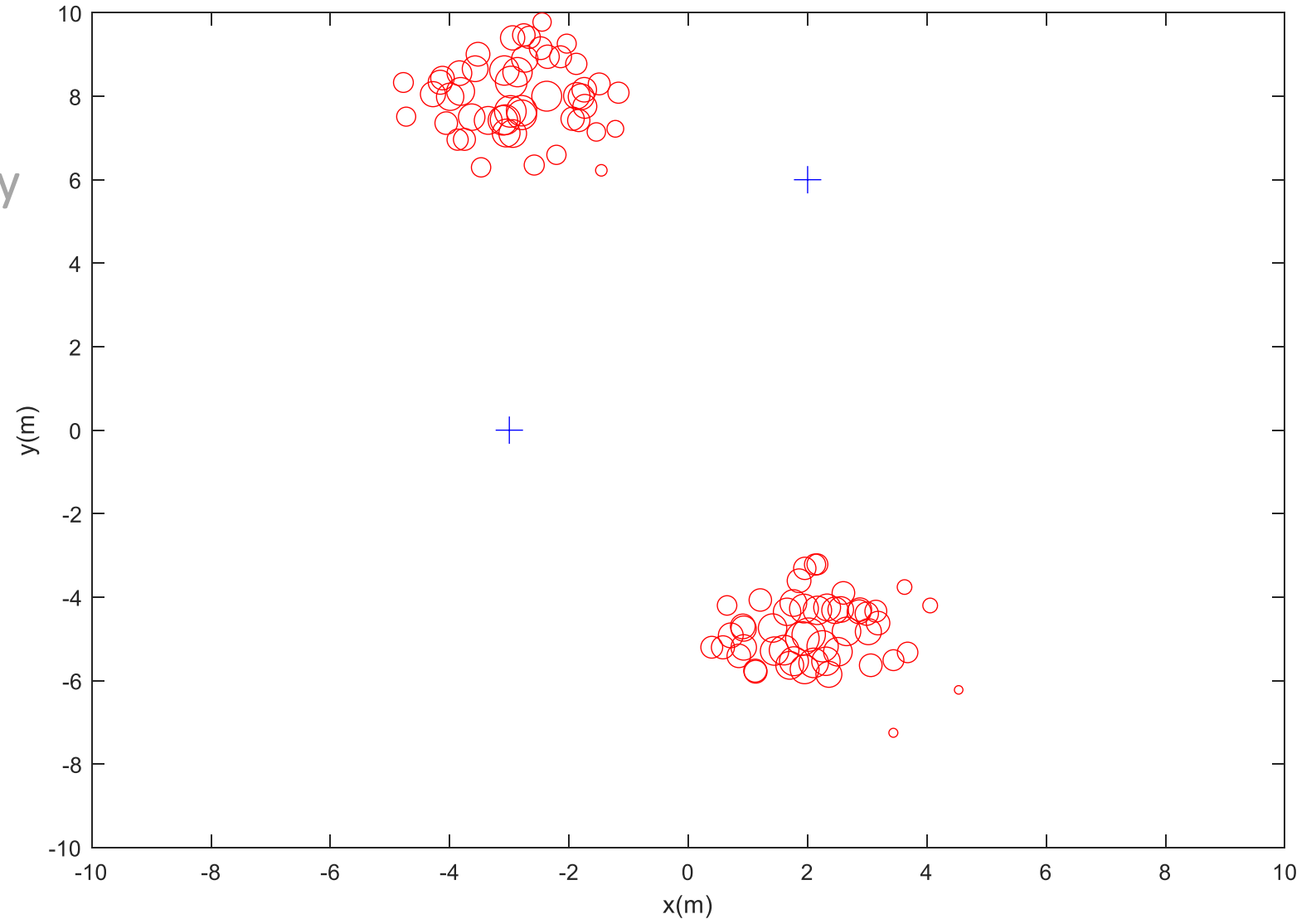
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



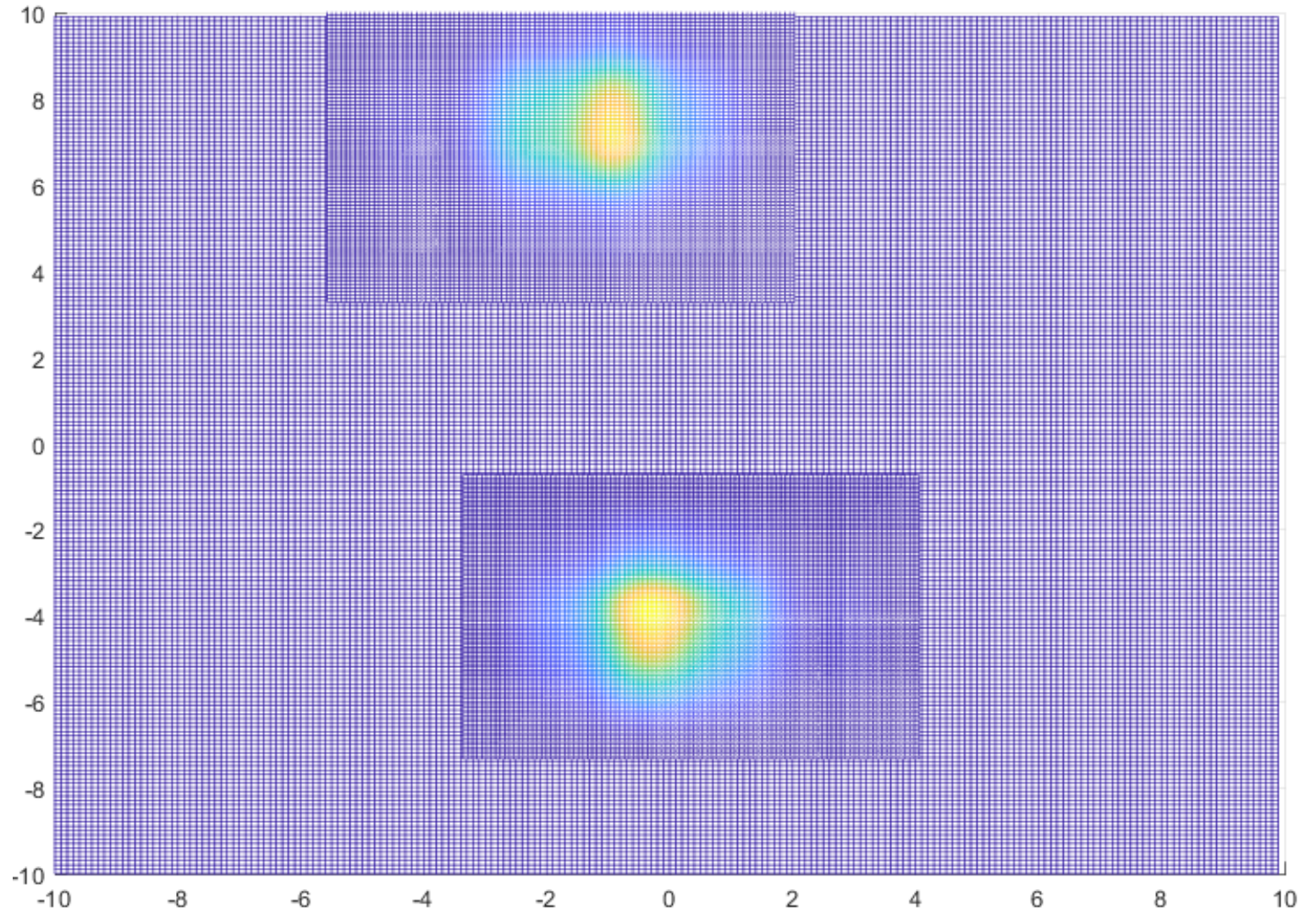
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



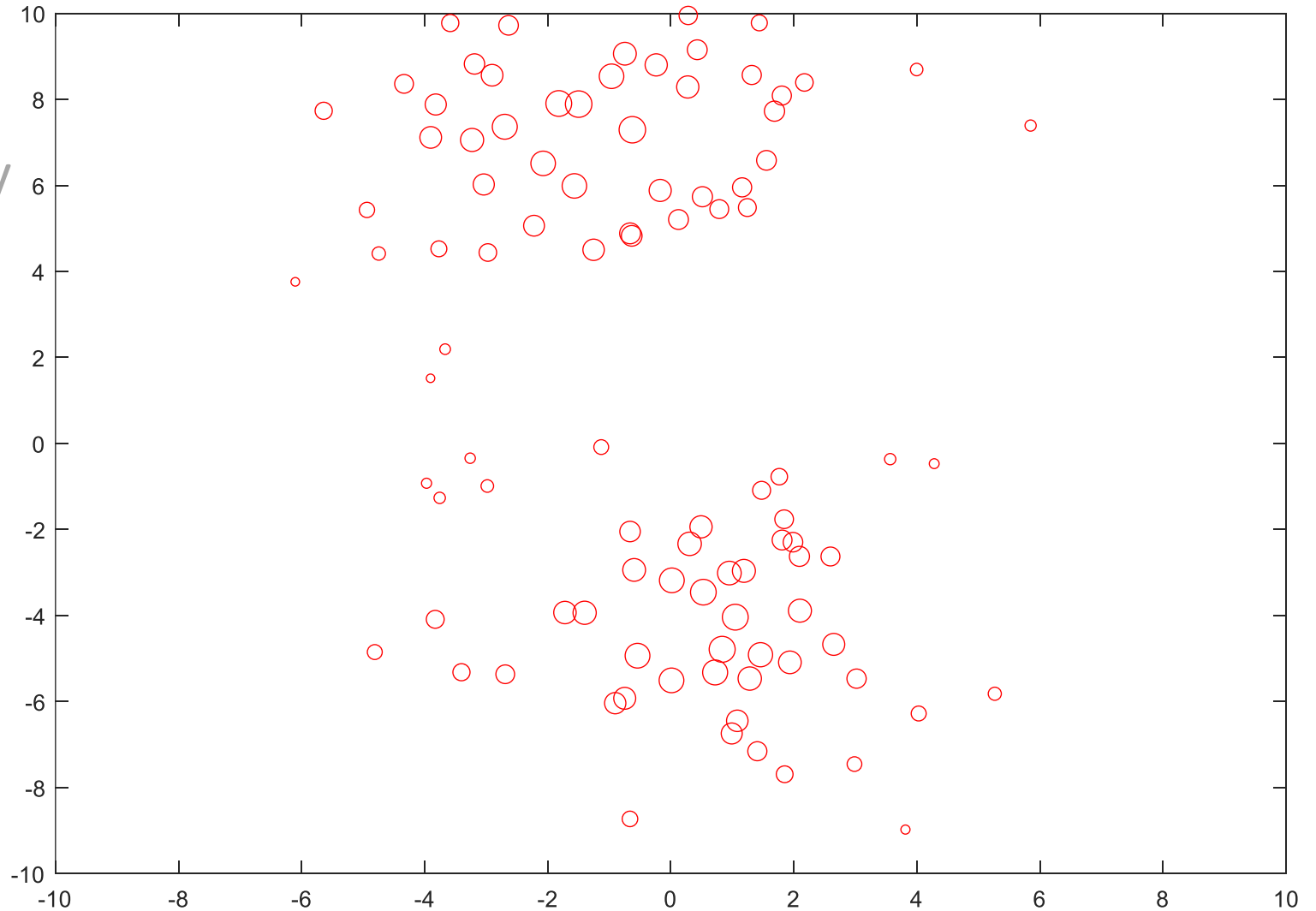
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



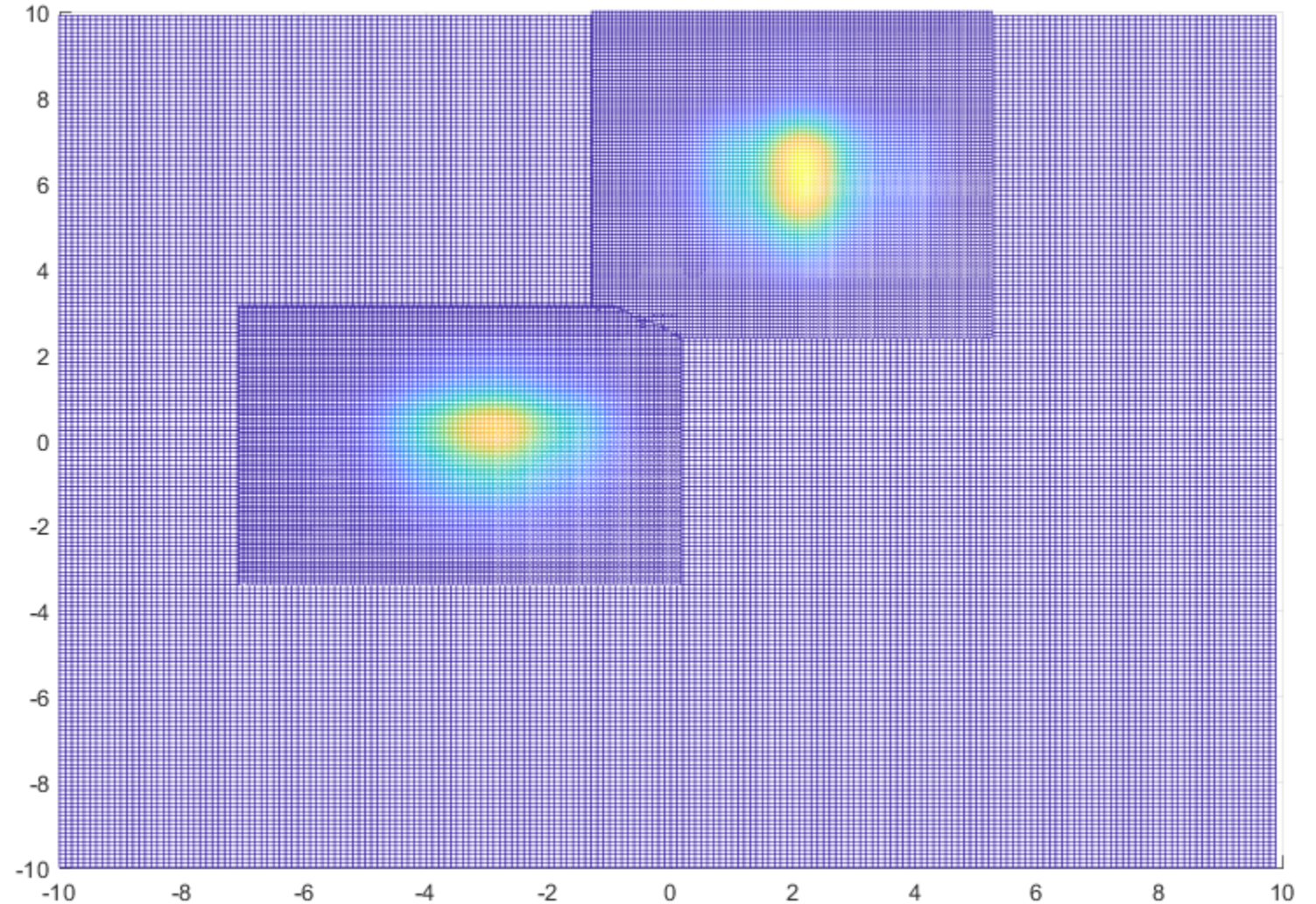
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



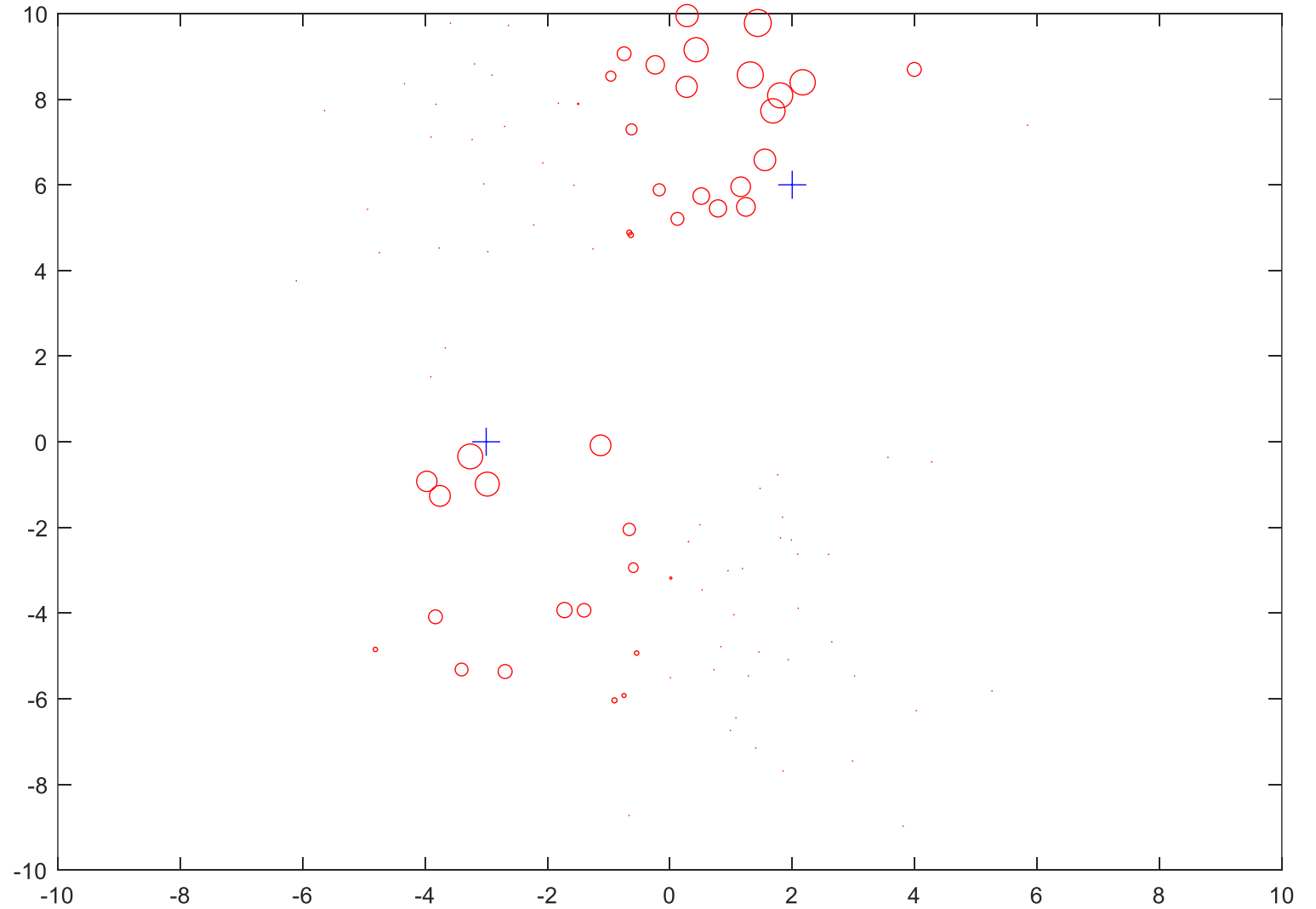
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



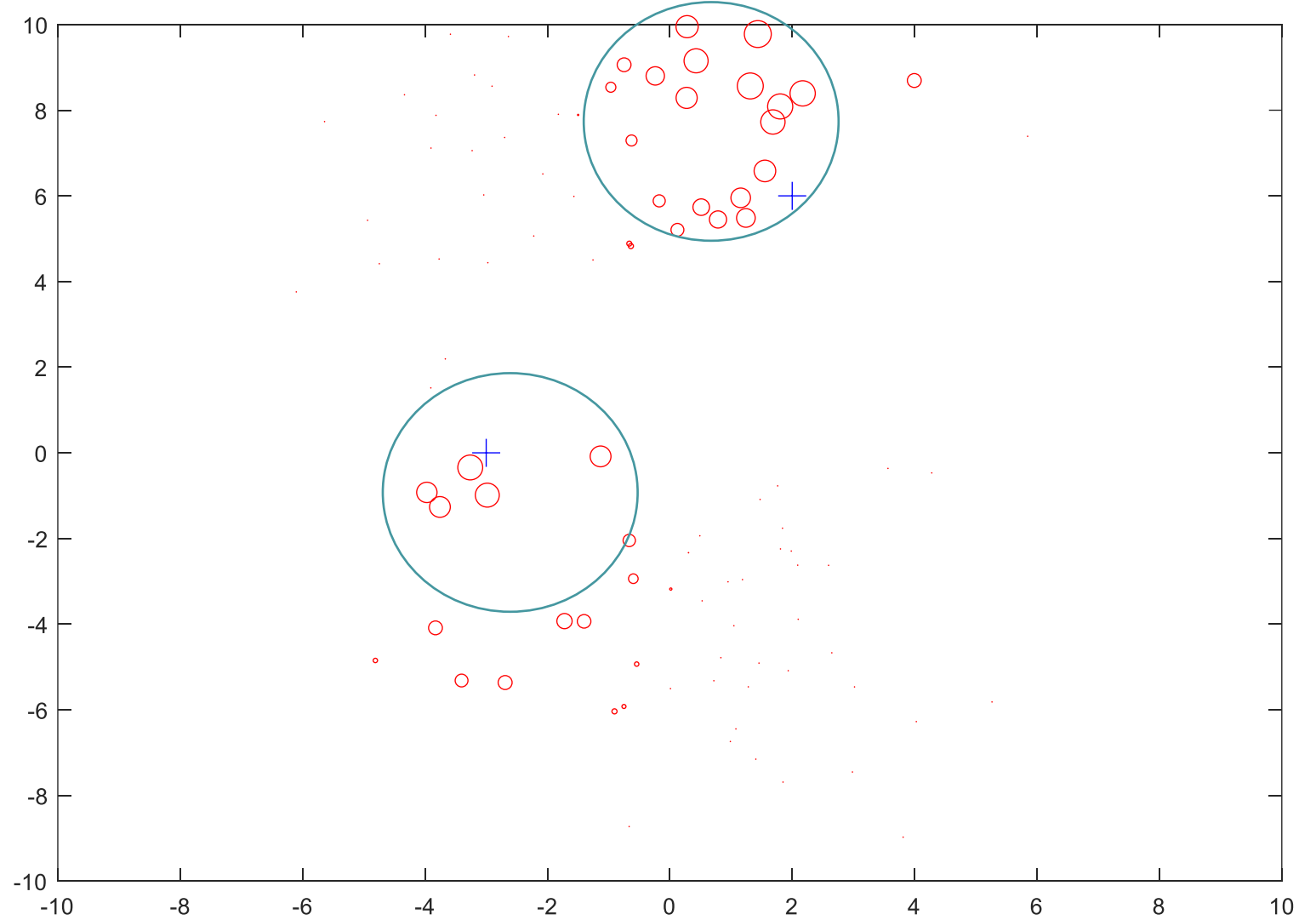
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



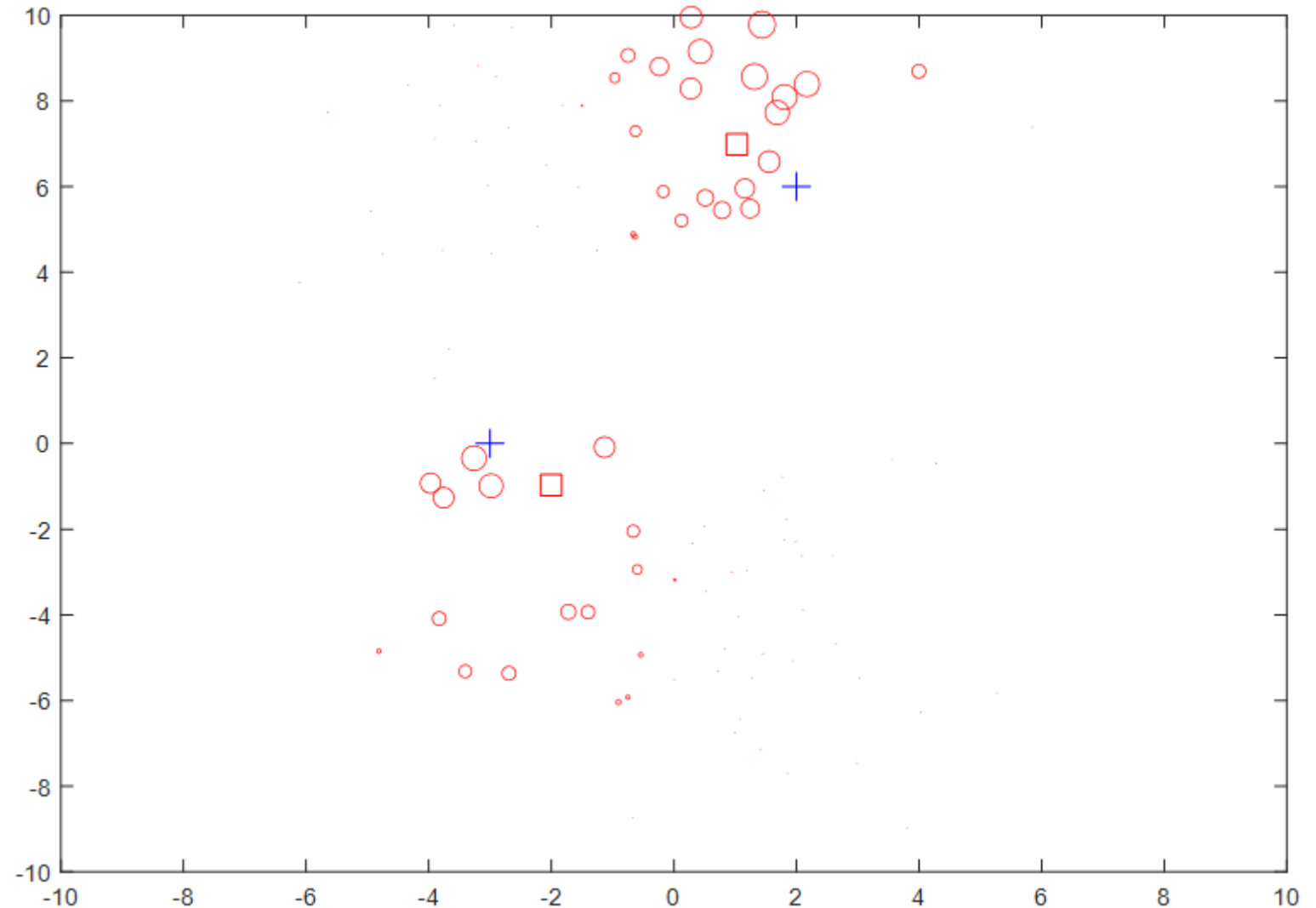
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



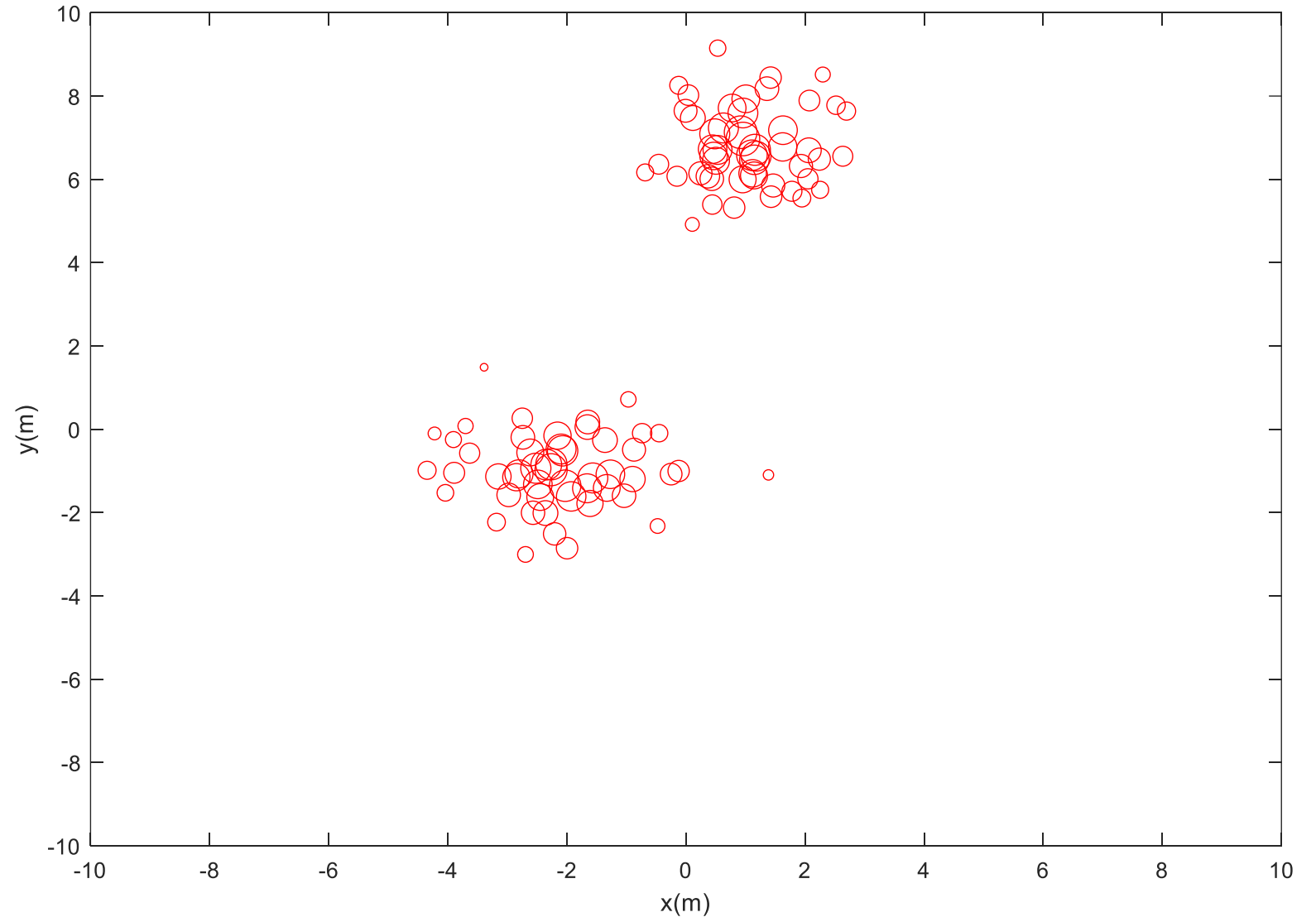
Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Steps of PHD filter

1. Predict particles
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Gaussian mixture probability hypothesis density filter

Advantage

1. A closed form solution.
2. Without a clustering step.

Drawback

1. Assume that the system is non-linear and non-Gaussian

Sequential Monte Carlo probability hypothesis density filter

Advantage

1. High accuracy for the non-linear and non-Gaussian problem, such as speaker tracking.

Drawback

1. Weight degeneracy problem.
2. Need a clustering step.

Steps of AV-SMC-PHD filter

1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



V. Kilic, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-Shift and Sparse Sampling Based SMC-PHD Filtering for Audio Informed Visual Speaker Tracking", *IEEE Transactions on Multimedia*, vol. 18, no. 10, October 2016.

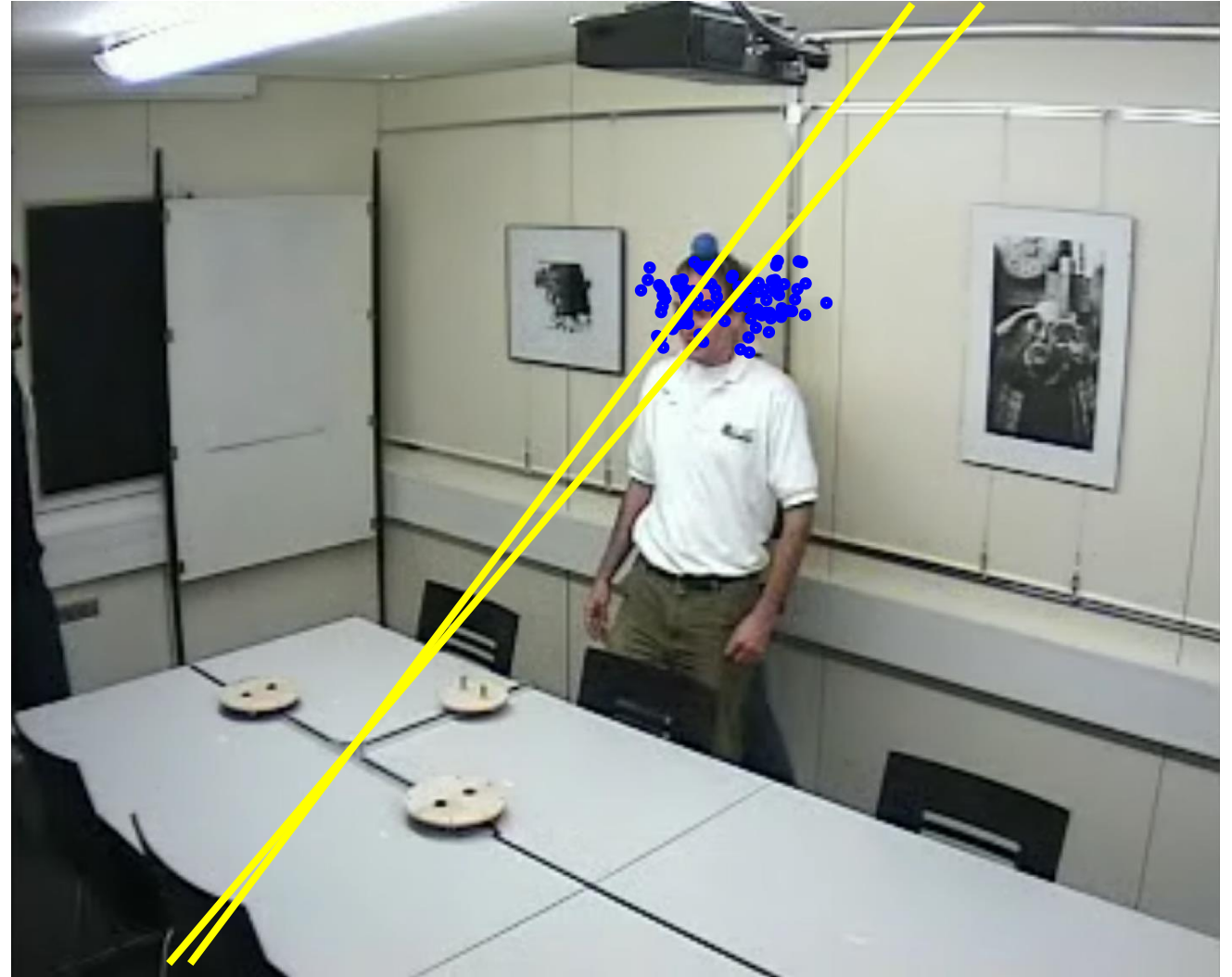
Steps of AV-SMC-PHD filter

1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Steps of AV-SMC-PHD filter

1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Steps of AV-SMC-PHD filter

1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Steps of AV-SMC-PHD filter

1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Steps of AV-SMC-PHD filter

1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Steps of AV-SMC-PHD filter

1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



Steps of AV-SMC-PHD filter

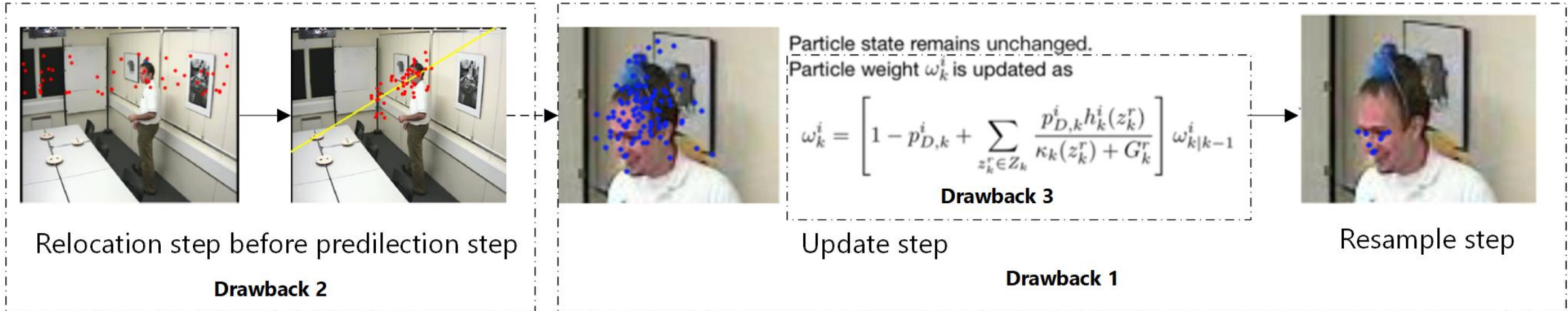
1. Relocation step
2. Update particles
 - 2.1 Calculate likelihood density
 - 2.2 Update particle weights
3. Cluster particles
4. Resample particles



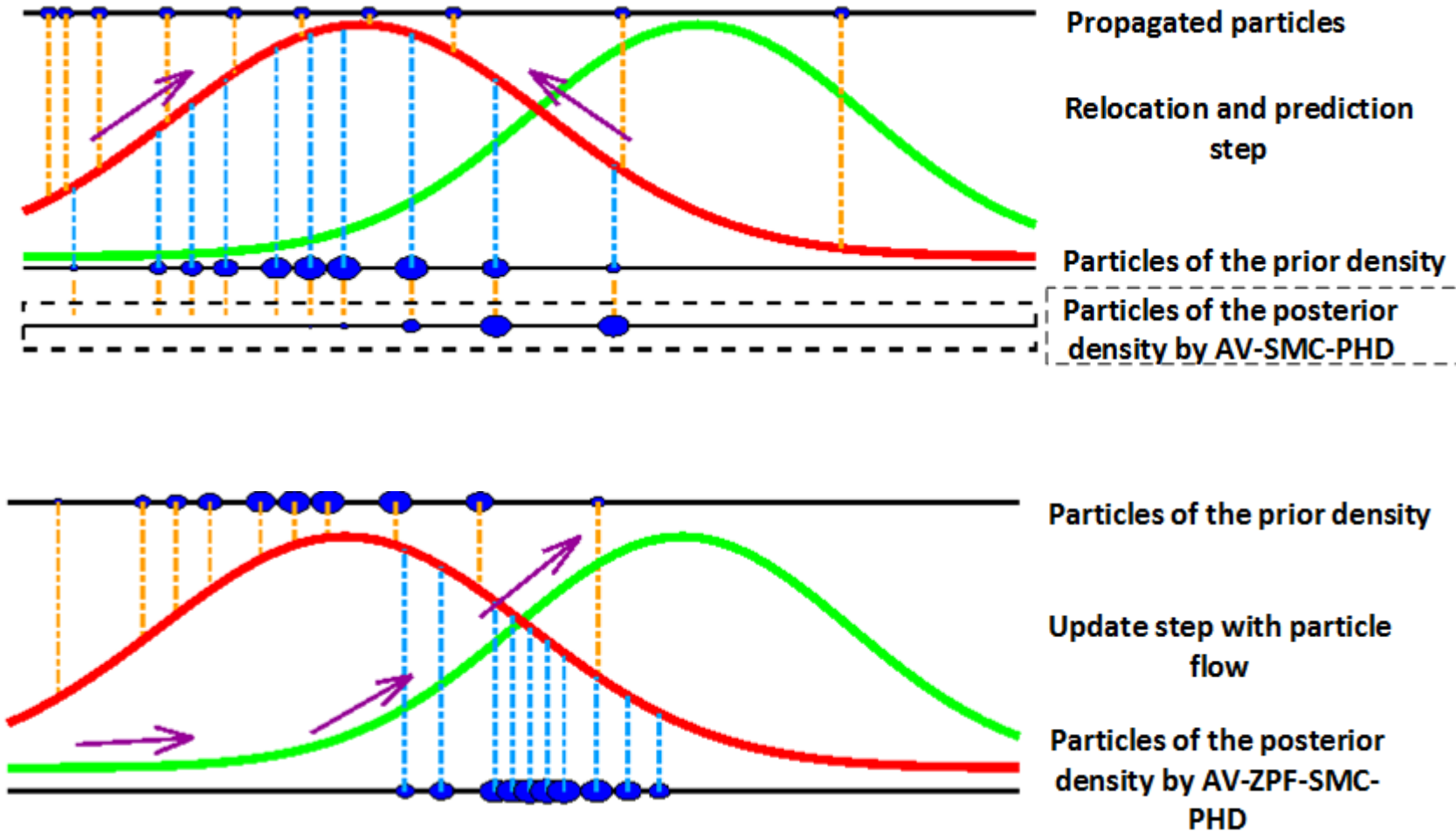
V. Kilic, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-Shift and Sparse Sampling Based SMC-PHD Filtering for Audio Informed Visual Speaker Tracking", *IEEE Transactions on Multimedia*, vol. 18, no. 10, October 2016.

Drawback of the SMC-PHD filter

1. Weight degeneracy (major issue)
2. Re-location step only with audio measurements (minor issue)
3. Update step only with visual measurements (minor)



Weight Degeneracy and Particle Flow



Y. Liu, V. Kilic, J. Guan, and W. Wang, "Audio-visual particle flow SMC-PHD filtering for multi-speaker tracking", *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 934-948, 2020

Steps of Particle flow

1. Select particles
2. Calculate the variance and mean of particles
3. Calculate particle flow
4. Update particles
5. Update pseudo time
6. Repeat steps 2-5



Steps of Particle flow

1. Select particles
2. Calculate the variance and mean of particles
3. Calculate particle flow
4. Update particles
5. Update pseudo time
6. Repeat steps 2-5



Steps of Particle flow

1. Select particles
2. Calculate the variance and mean of particles
3. Calculate particle flow

$$f_{k,\lambda}^i = \frac{dm_k^i}{d\lambda} = A_k^i m_k^i + b_k^i$$

where

$$A_k^i = -\frac{1}{2} P_k^i (H_k^i)^T (\lambda H_k^i P_k^i (H_k^i)^T + R)^{-1} H_k^i,$$

$$b_k^i = (I + 2\lambda A_k^i) [(I + \lambda A_k^i) P_k^i (H_k^i)^T R^{-1} z_k^r + A_k^i \bar{m}_k^i]$$

4. Update particles
5. Update pseudo time
6. Repeat steps 2-5



Steps of Particle flow

1. Select particles
2. Calculate the variance and mean of particles
3. Calculate particle flow
4. Update particles

$$\Delta m_{k|k-1}^i = f_{k,\lambda}^i \Delta \lambda + v_k^i w_k^i$$

5. Update pseudo time
6. Repeat steps 2-5



Steps of Particle flow

1. Select particles
2. Calculate the variance and mean of particles
3. Calculate particle flow
4. Update particles
5. Update pseudo time

$$\lambda \in [0, \Delta\lambda, 2\Delta\lambda, \dots, N_\lambda\Delta\lambda]$$

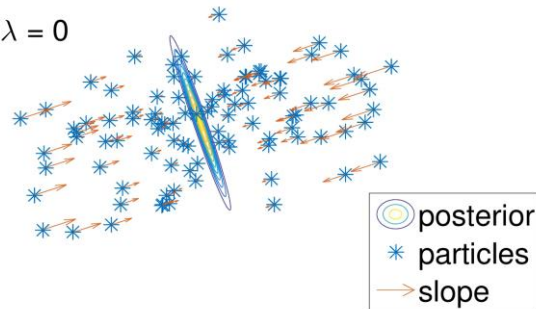
6. Repeat steps 2-5



Steps of Particle flow

1. Select particles
 2. Calculate the variance and mean of particles
 3. Calculate particle flow
 4. Update particles
 5. Update pseudo time
- $\lambda \in [0, \Delta\lambda, 2\Delta\lambda, \dots, N_\lambda\Delta\lambda]$
6. Repeat steps 2-5

particle flow: $\lambda = 0$



Particle flow

Zero-diffusion particle flow:

Benefit

Easy for implementation.

Drawback

Prior and likelihood density are Gaussian.

Non-zero diffusion particle flow:

Benefit

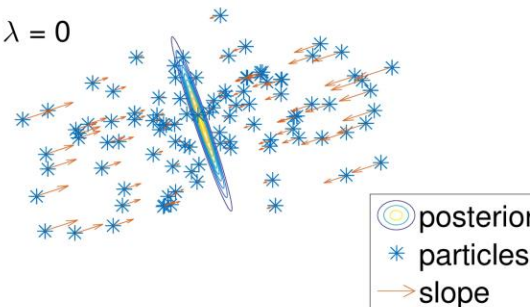
Likelihood density can be non-Gaussian.

Drawback

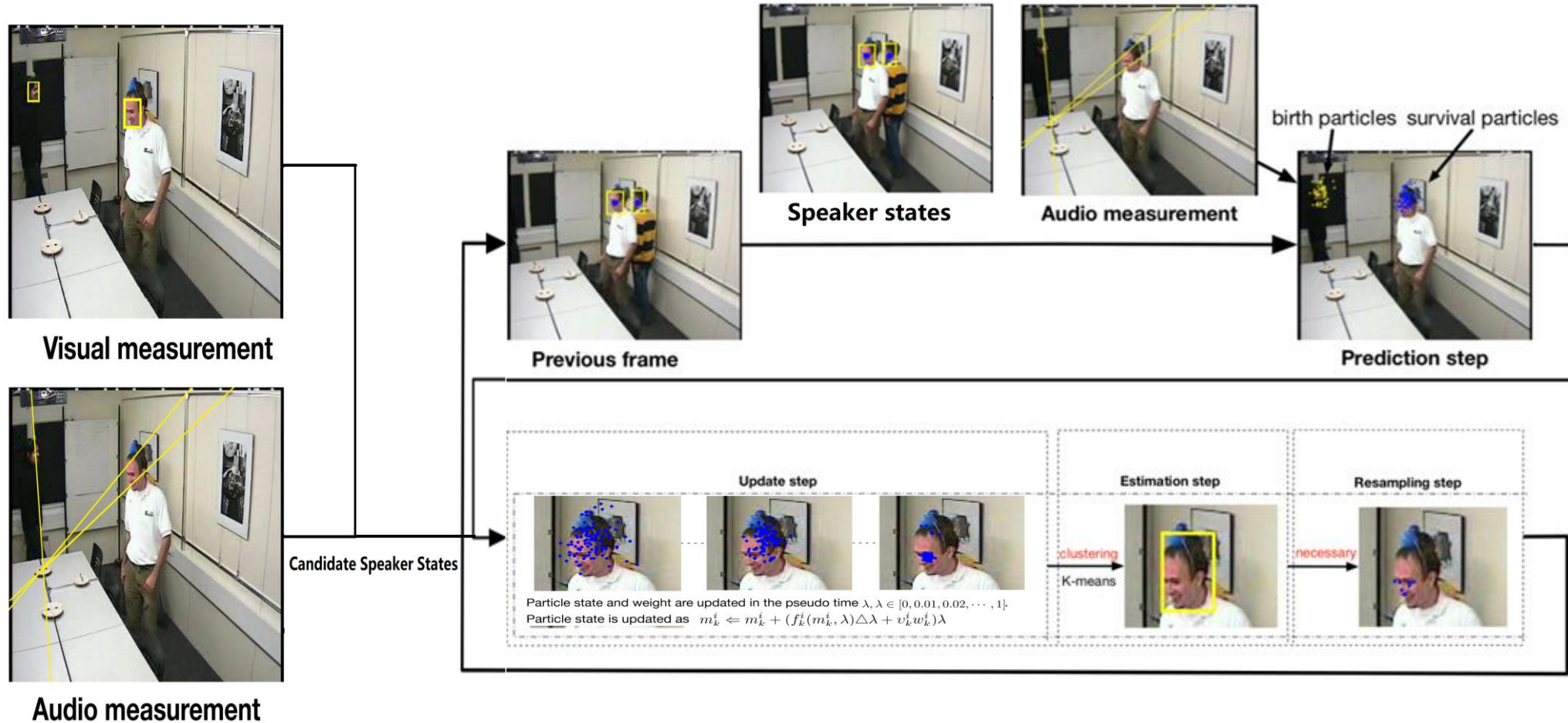
Sensitive to parameters

Particle flow is used to adjust the particle states and weights before the update step.

particle flow: $\lambda = 0$

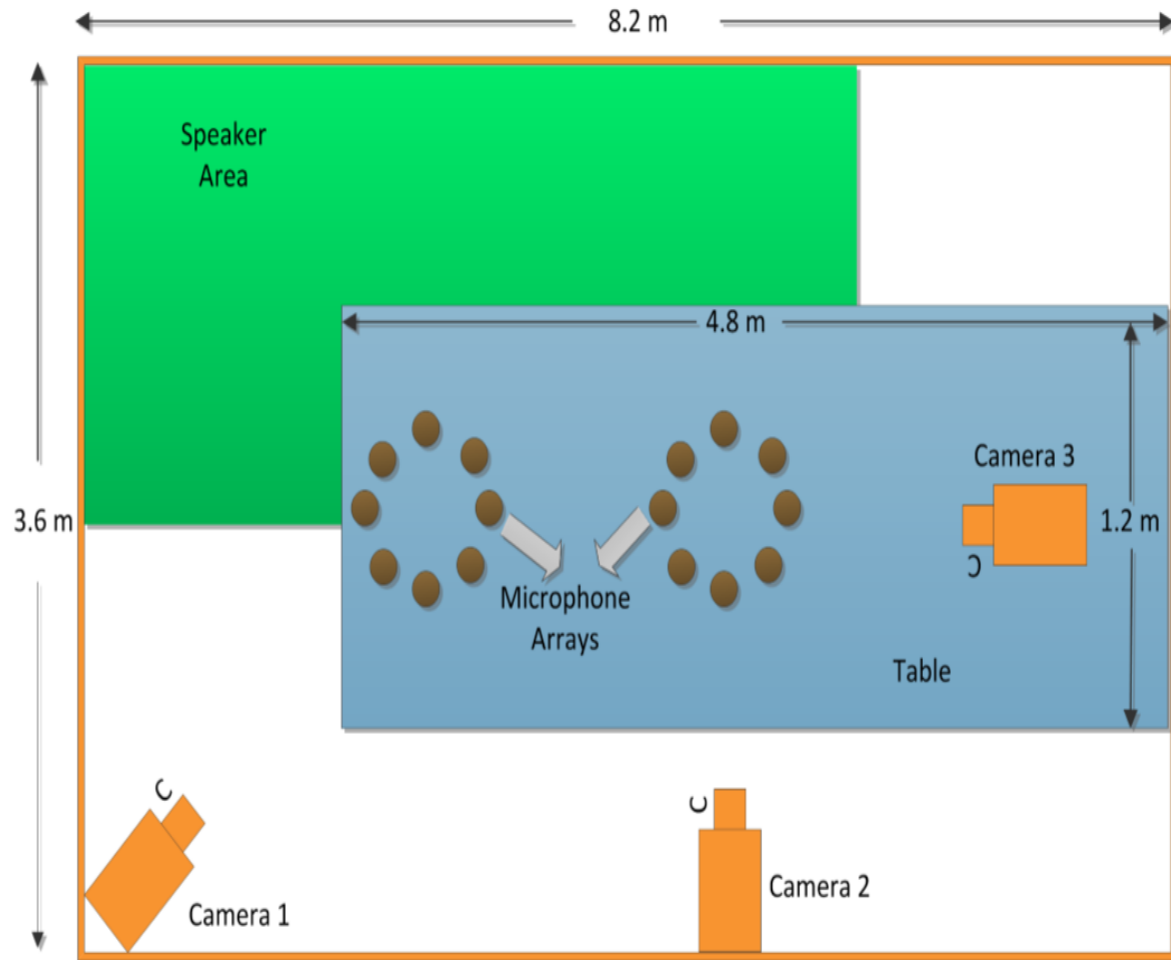


AV-PF-SMC-PHD filter

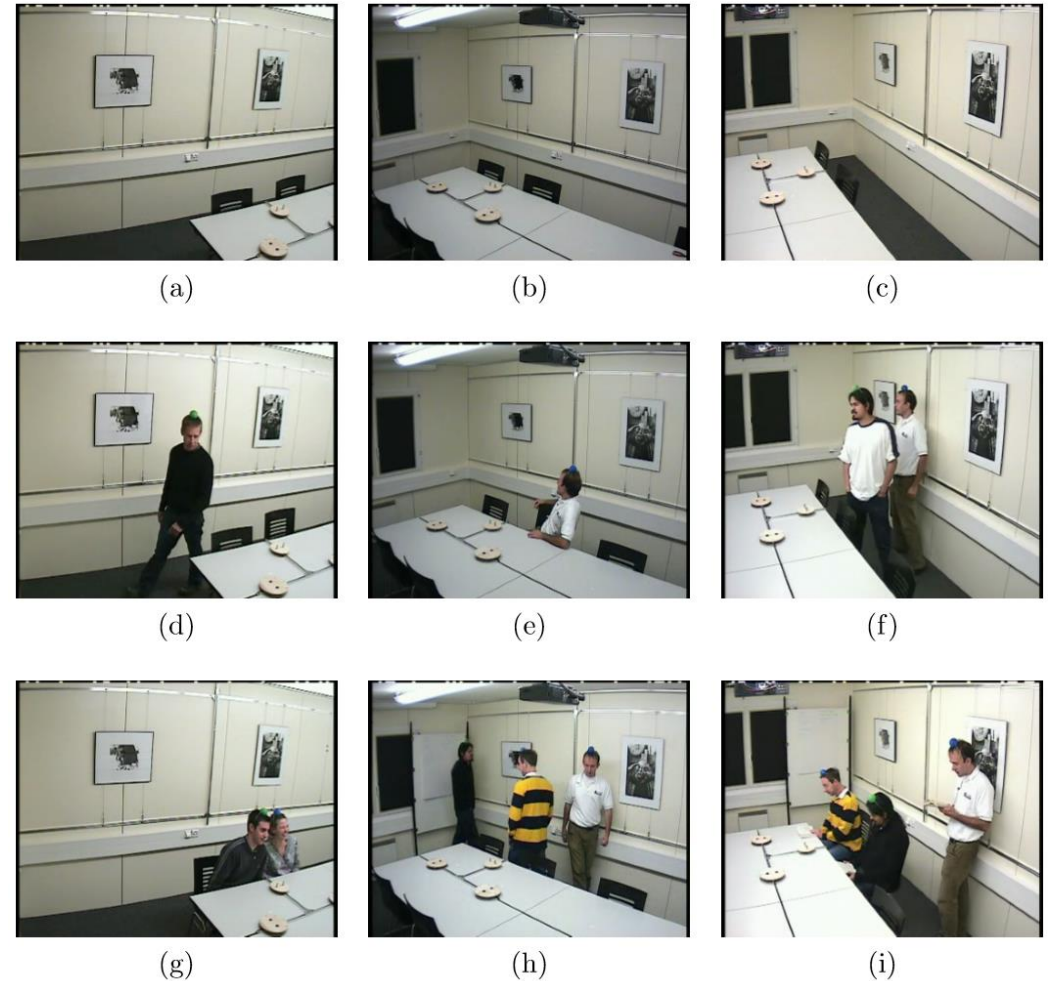


Y. Liu, V. Kilic, J. Guan, and W. Wang, "Audio-visual particle flow SMC-PHD filtering for multi-speaker tracking", *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 934-948, 2020

AV16.3



Physical setup of the AV16.3 corpus.



Some frames from the AV16.3 dataset.

Performance Metrics

Optimal Sub-pattern Assignment (OSPA) is used to evaluate the performance of the proposed and baseline algorithms

$$\text{OSPA}(\{\tilde{\mathbf{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \{\tilde{\mathbf{m}}_k^{\tilde{j}}\}_{\tilde{j}=1}^{\tilde{\mathcal{N}}_k}) = \sqrt[a]{\frac{\min_{\pi \in \Pi_{\tilde{\mathcal{N}}_k, \tilde{N}_k}} \sum_{j=1}^{\tilde{N}_k} \bar{d}^{(c)}(\tilde{\mathbf{m}}_k^j, \tilde{\mathbf{m}}_k^{\pi(j)})^a + c^a(\tilde{\mathcal{N}}_k - \tilde{N}_k)}{\tilde{\mathcal{N}}_k}}$$

ESS is widely applied to evaluate the severity of weight degeneracy problem

$$\text{ESS} = \frac{(\sum_{i=1}^{N_k} \omega_k^i)^2}{\sum_{i=1}^{N_k} (\omega_k^i)^2}$$

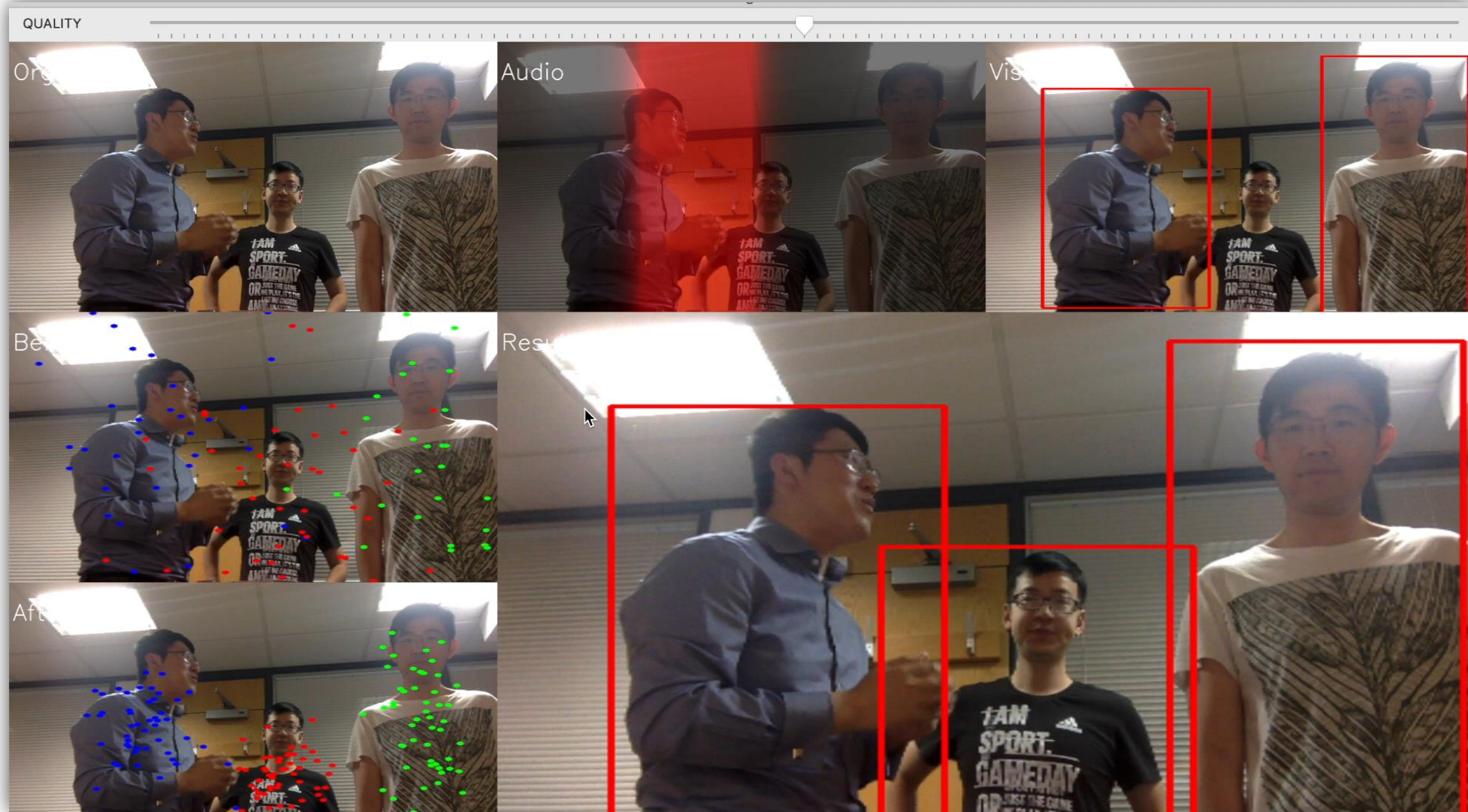
The experimental results for ZPF, SMS, PPF, GPF, SMC, ASMC and NPF in terms of the OSPA error

Seq	ZPF	SMS	PPF	GPF	SMC	ASMC	NPF
45(1)	17.60	23.40	24.50	23.12	29.46	26.07	17.65
45(2)	18.55	23.16	22.26	22.71	29.47	25.97	18.60
45(3)	19.54	23.80	24.34	23.76	28.43	26.41	19.50

Computational cost (s/Sequence) comparison for ZPF, SMS, PPF, GPF, ASMC and NPF

Seq	ZPF	SMS	PPF	GPF	SMC	ASMC	NPF
45	263.4	162.2	237.4	490.7	93.1	121.5	197.5
	$N_k N_\lambda$	$U_k N_k$	$N_k N_\lambda$	$U_k N_k N_\lambda$	$U_k N_k$	$U_k N_k$	$N_k N_\lambda$

Tracking on the live demo



- 1 Audio-visual speech source separation
- 2 Audio-visual multi-speaker localization/tracking
- 3 Ego-centric audio-visual multi-speaker localization/tracking
- 4 Conclusion and Future Works

Ego-centric AV speaker tracking



The listener (e.g. robot), who is walking, wears a RGB camera, a depth camera and a microphone array.

The speaker, who is speaking, is also moving.

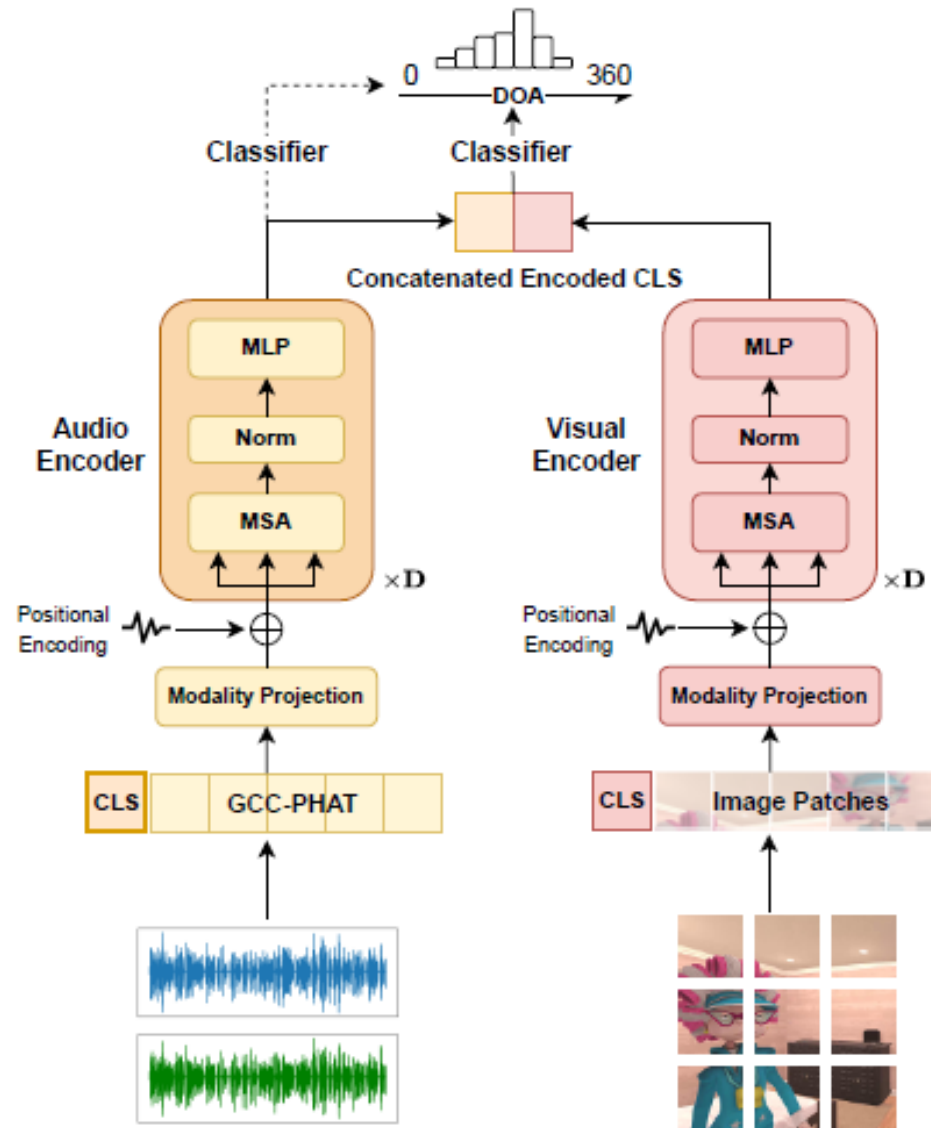
The aim of this task is to predict the position or direction, e.g. Direction of Arrival (DOA) of the moving speaker relative to the listener.

Ego-centric AV speaker tracking

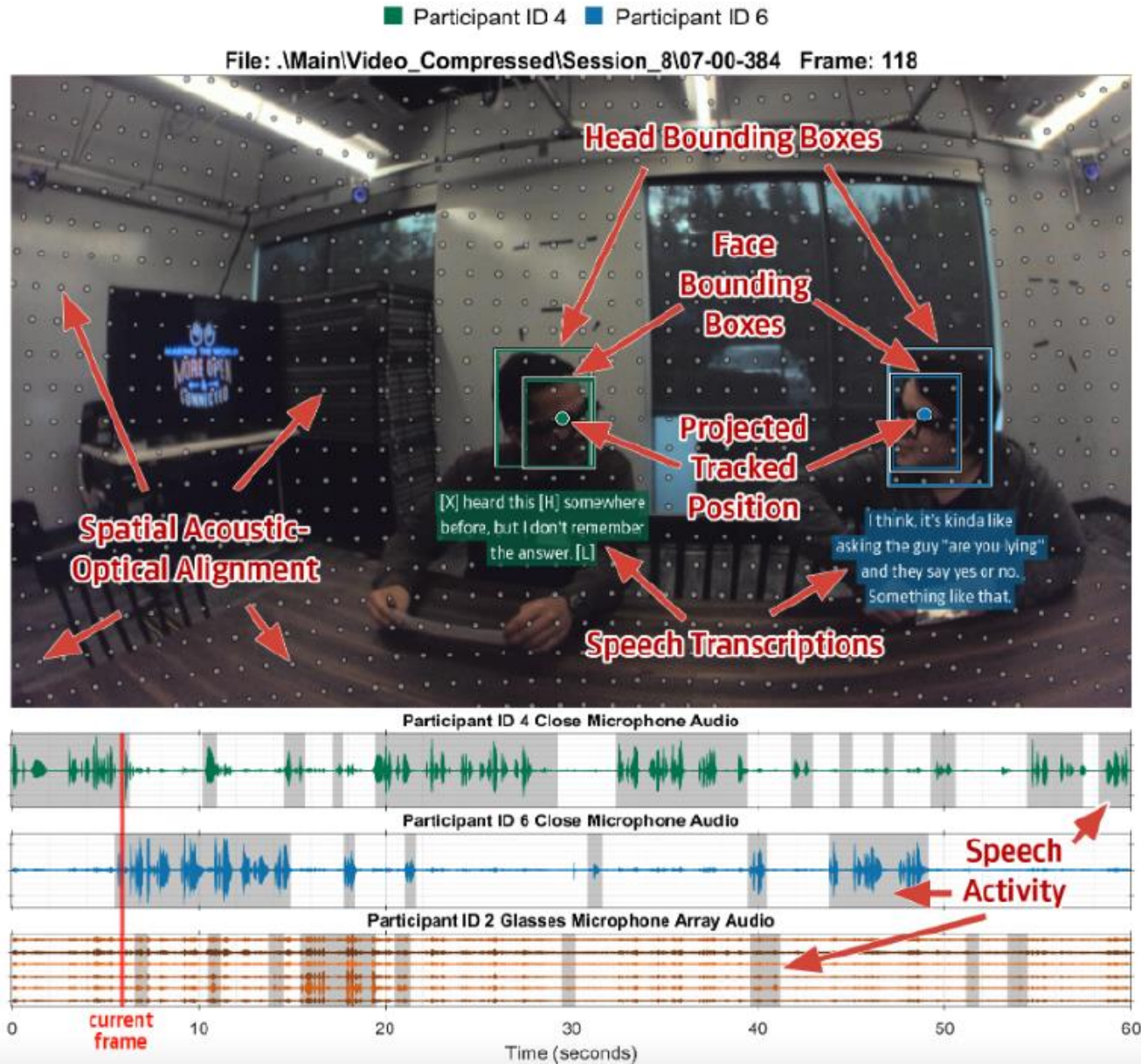


We show the sequences of the AV16.3 dataset and our simulated Ego-AVSL dataset. It can be seen that the main difference between the egocentric scenario and the conventional scenario is that, the speaker is not always in the camera view of the listener due to the movement of the listener. The speaker absence poses a significant challenge of data fusion.

Ego-centric AV speaker tracking



Ego-centric AV speaker tracking – EasyCom dataset



Ego-centric AV speaker tracking

Methods	Mean E1	Std1	Mean E2	Std2
AV(cor)	16.77	12.63	6.56	8.77
AV(spec)	8.81	9.63	6.21	6.89
DOA	129.82	18.26	46.45	21.50
DOA+image	66.81	7.89	36.48	8.97
AV-rawaudio	40.14	10.55	140.75	19.58
Ours [†]	9.33	12.78	4.72	7.15
Ours [‡]	8.00	10.31	4.49	7.53

J. Zhao, Y. Xu, X. Qian, W. Wang, "Audio Visual Speaker Localization from EgoCentric Views", submitted.

<https://arxiv.org/pdf/2309.16308.pdf>

- 1 Audio-visual speech source separation
- 2 Audio-visual multi-speaker localization/tracking
- 3 Ego-centric audio-visual speaker localization/tracking
- 4 Conclusion and Future Works

1

Audio-Visual Speech Source Separation

- We have presented method for audio-visual coherence modelling and incorporate such information to improve speech source separation.
- For AV coherence modelling, we could use statistical models or dictionary learning models.
- Such information could be incorporated into conventional source separation methods such frequency domain ICA or time-frequency masking.

2

Audio-Visual Multi-Speaker Tracking

- Audio-visual particle flow is used to migrate the particles smoothly from the prior to the posterior density.
- A novel relocation step is proposed.
- A novel particle flow assisted by the label information. A novel AV likelihood function. Clustering is replaced by taking the mean of the labelled particles.

3

Ego-centric AV Speaker Tracking

- We have developed a transformer-based system and also created a simulated dataset for ego-centric scenario.
- The ego-centric tracking scenario is very complicated. There are many problems which may happen in real applications including motion blur, speaker disappearance, occlusions, surrounding noise, poor illumination conditions. For now, we mainly focus on speaker disappearance, occlusions and audio noise.

Potential future works

1

Visual-text prompted speech/audio event separation

Visual or text guidance as to which speech source to be separated.

2

Ego-centric audio-visual speaker tracking

Dealing with high-percentage missing measurements to improve the tracking robustness and accuracy.

3

Prompt based speech source localization and tracking

Instructions could be given by a robot as to which speech source needs to be localized/tracked.

THANK YOU

