Jointly Tracking and Separating Speech Sources Using Multiple Features and the Generalized Labeled Multi-Bernoulli Framework

SHOUFENG (FRANK) LIN

PHD CANDIDATE AND SENIOR ENGINEER

Motivations and Challenges

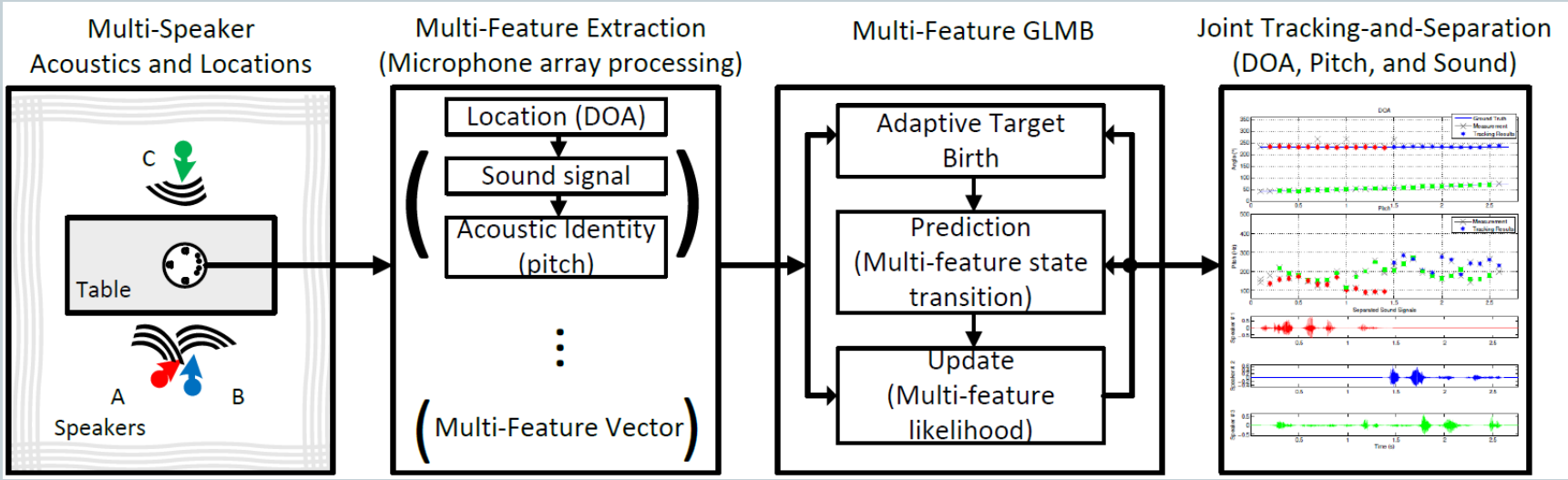
- The “cocktail party” problem
  - Concurrent speakers (how many)
  - Moving speakers (where)
  - Speech extraction (what is said)
  - Speaker identity (who said what)
  - Interference (what not)
  - Association over time (which to which)
  - etc.

Proposed Method

- Multi-feature extraction
  - Speaker localization (concurrent, moving speakers)
  - Speech separation (multiple speakers)
  - Speaker identification (pitch estimation)
- Joint online tracking
  - Bayes RFS multi-object tracking for multi-speaker and multi-feature
  - etc.

System Overview

- Proof-of-concept



Multi-feature Extraction

- Speaker localization
  - Subspace methods (MUSIC, ESPRIT, etc.)
  - Steered-response beamformers
  - TDOA based techniques
    - MCC-PHAT

$$\xi^{\mathrm{mcc-phat}}(k,\varsigma) \triangleq \prod_{(i,j)\in P} \xi_{ij}^{\mathrm{gcc-phat}}(k,\tau_{ij}(\varsigma)),$$

where

$$\xi_{ij}^{\mathrm{gcc-phat}}(k,\tau_{ij}(\varsigma)) = \int_{-\infty}^{+\infty} \Xi_{ij}^{\mathrm{gcc-phat}}(k,f)\cdot e^{\mathrm{i}2\pi f\tau_{ij}(\varsigma)} df,$$

and

$$\Xi_{ij}^{\mathrm{gcc-phat}}(k,f) = \frac{X_i(k,f)\cdot X_j^{\star}(k,f)}{|X_i(k,f)\cdot X_j^{\star}(k,f)|}.$$

Multi-feature Extraction

- Speech Separation
  - BSS, TFM, etc.
  - Wideband Beamformer (WLS, filter-and-sum)

$$\hat{s}_{k,i}(n) = \mathbf{w}_{k,i}^{T}\,\mathbf{x}(n),$$

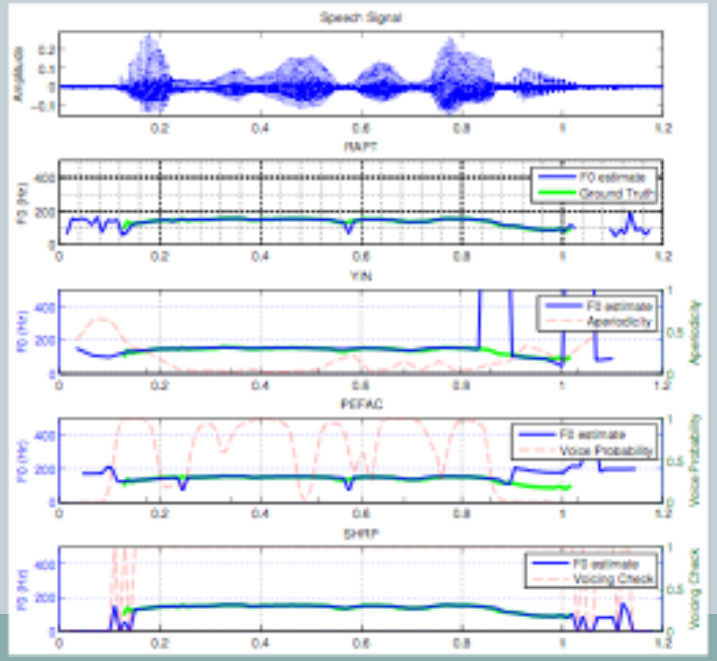
where $[\cdot]^T$ is the matrix transpose, and

$$\mathbf{x}(n) = \left[\mathbf{x}_0(n), \ldots, \mathbf{x}_{j_t}(n), \ldots, \mathbf{x}_{J_t-1}(n)\right]^T,\ j_t \in [0, J_t - 1]$$

$$\mathbf{x}_{j_t}(n) = \left[x_1(n + j_t), \ldots, x_j(n + j_t), \ldots, x_M(n + j_t)\right].$$

Multi-feature Extraction

- Speaker Identification
  - GMM, NN, etc.
  - Pitch
    - PEFAC, SHRP, YIN, RAPT, etc.
    - Example (SNR=25dB, babble noise)



Multi-Feature GLMB

- Background
  - Random Finite Set (RFS)
  - Bayes rule, Chapman-Kolmogorov equation
  - Conjugate prior
  - GLMB
  - Hypothesis and Probability density
  - Distinct label indicator
  - Label set, association map
  - Projection function

$$\pi(X|Y) = \frac{g(Y|X)\pi(X)}{\int g(Y|X)\pi(X)\delta X}$$

where

$$\int f(X)\delta X = \sum_{i=0}^{\infty} \frac{1}{i!} \int_{\mathbb{X}^i} f(\{x_1, ..., x_i\}) d(x_1, \cdots, x_i)$$

$$\pi_+(X_+) = \int f(X_+|X)\pi(X)\delta X$$

$$\pi(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \omega^{(I,\xi)} \delta_I(\mathcal{L}(\mathbf{X})) \left[p^{(\xi)}\right]^{\mathbf{X}}$$

Multi-Feature GLMB

- Multi-feature GLMB Recursion: Update

$$\pi(\mathbf{X}|Z) = \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \sum_{\theta\in\Theta(I)} \omega^{(I,\xi,\theta)}(Z)\delta_I(\mathcal{L}(\mathbf{X}))\left[p^{(\xi,\theta)}(\cdot|Z)\right]^{\mathbf{X}}$$

- Multi-object Multi-feature likelihood function

$$\mathrm{g}(z_{\theta(\ell)}|\mathrm{x},\ell) \triangleq \mathrm{g}(\hat{\zeta}_{\theta(\ell)}|\zeta,\ell)\cdot\mathrm{g}(\hat{F}_{0\theta(\ell)}|F_0,\ell)$$

Multi-Feature GLMB

- Multi-feature GLMB Recursion: Prediction

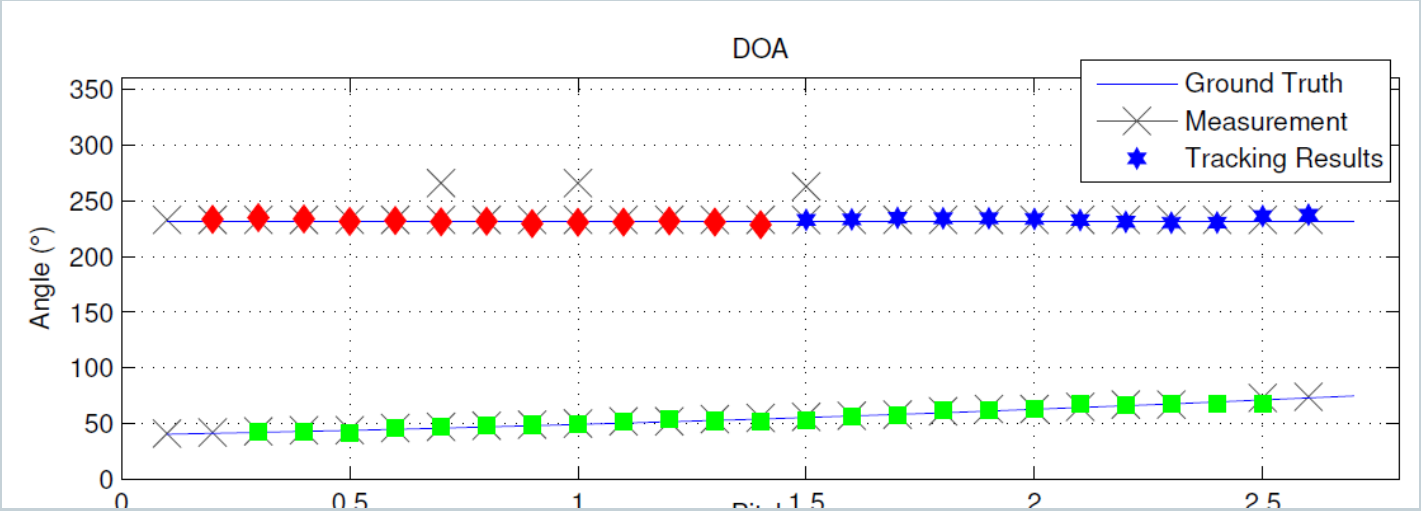
$$\pi_+(\mathbf{X}_+) = \Delta(\mathbf{X}_+) \sum_{(I_+,\xi)\in\mathcal{F}(\mathbb{L}_+)\times\Xi} \omega_+^{(I_+,\xi)} \delta_{I_+}(\mathcal{L}(\mathbf{X}_+)) \left[p_+^{(\xi)}\right]^{\mathbf{X}_+}$$

- Multi-object Multi-feature state transition function

$$\mathrm{f}(\mathrm{x}|\cdot,\ell) = 1_{\mathrm{x}}(\zeta)\cdot \mathrm{f}(\zeta|\cdot,\ell)\cdot 1_{\mathrm{x}}(F_0)\cdot \mathrm{f}(F_0|\cdot,\ell)$$

Test Scenario

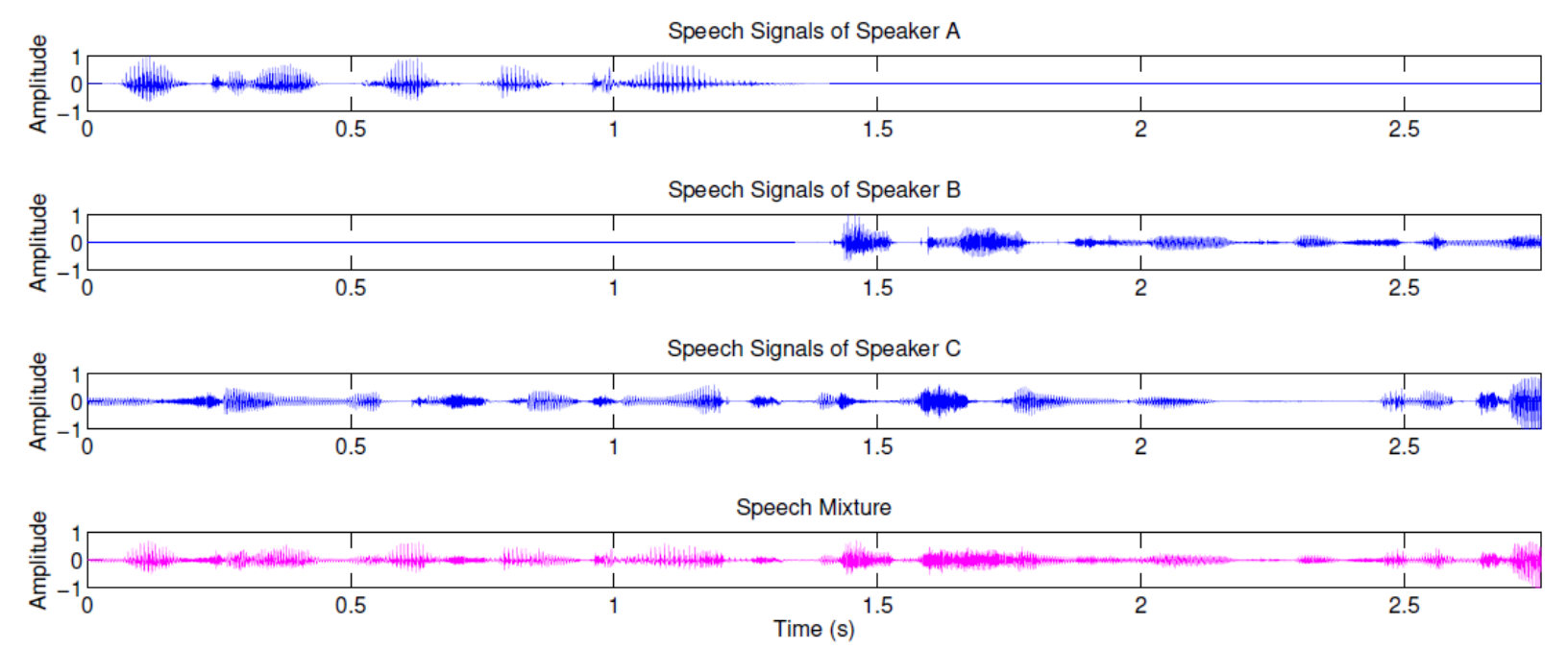
- Set-up
  - UCA with 8 microphones, diameter 0.1m
  - 3 speakers (static and moving)



- Metrics: OSPA, PEASS

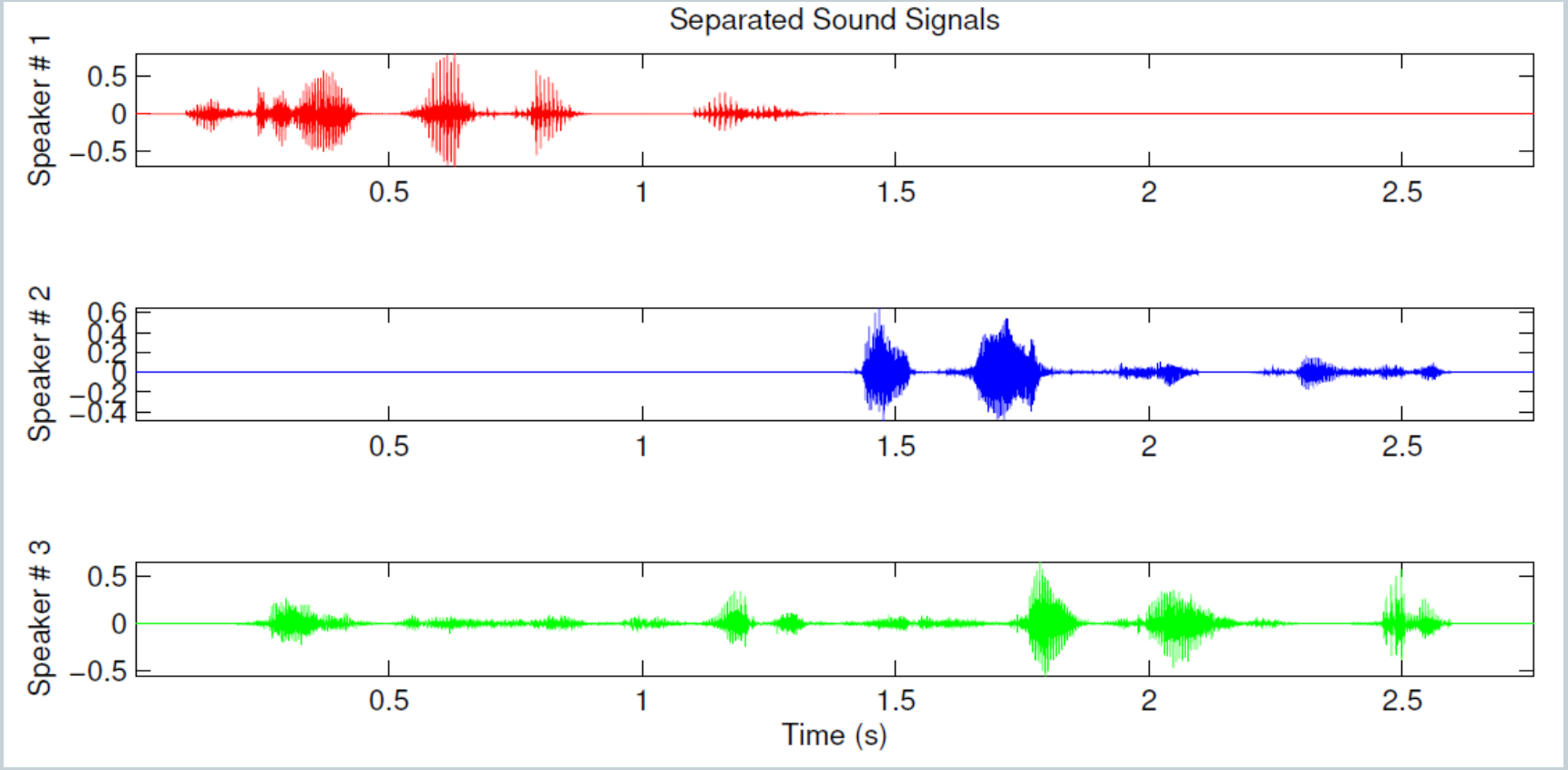
Test Scenario

- Mixture of speech signals



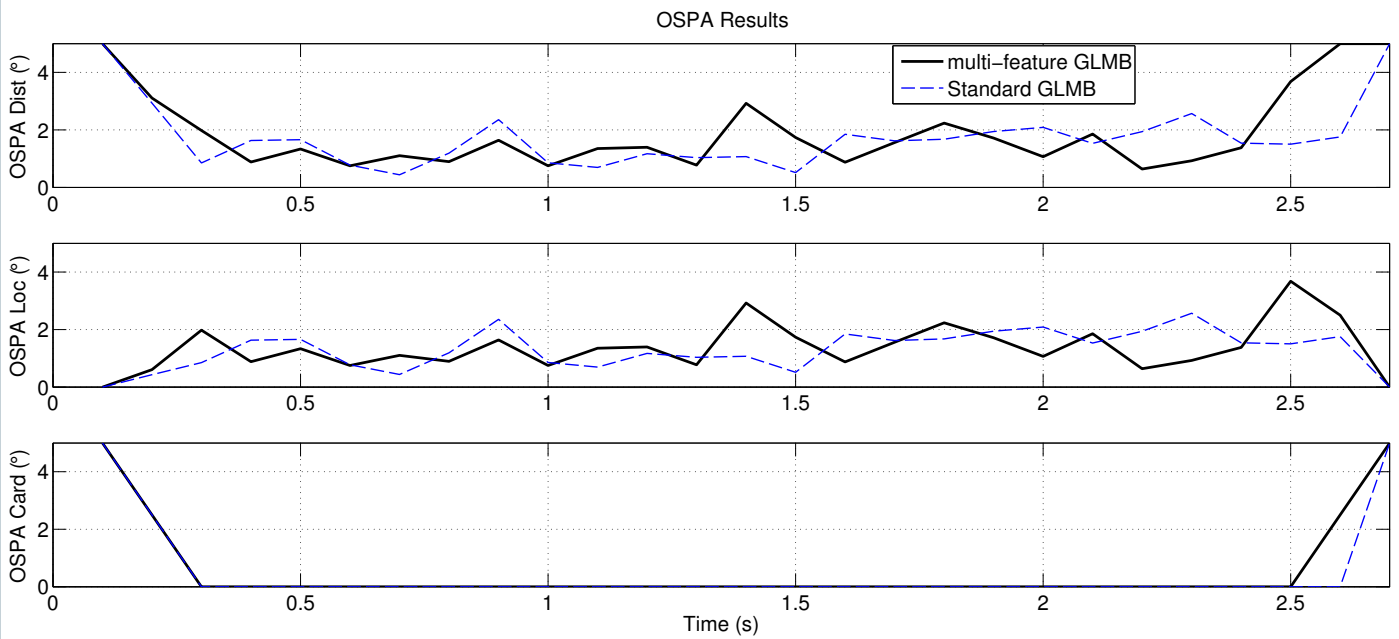
Test Scenario

- Separated speech signals



Test Scenario

- Test results - OSPA for DOAs



Test Scenario

- Test results - PEASS for speech signals

Method	Speaker	OPS	TPS	IPS	APS
Proposed	1	48.75	57.03	71.19	49.11
	2	32.69	29.35	72.06	35.61
	3	36.02	35.73	65.65	37.71
UCBSS	$< 1, 2 >$	18.66	45.84	43.21	24.33
	3	25.00	6.10	83.97	3.50
DUET	$< 1, 2 >$	18.73	38.82	16.38	50.43
	3	24.97	51.16	32.40	44.32

Future Works

- Reverberation and noise
- Robust feature extraction methods
  - E.g. pitch, location, sound extraction
- Other tracking techniques
  - E.g. track-before-detect, compare MHT, JPDA, etc.

Acknowledgement

- Australian Postgraduate Award
- Australian Government Research Training Program Scholarship
- Supervisions at Curtin University

About the Author

- PhD Candidate (final year)
- Senior Engineer (industrial experience)
  - Lead Electrical Engineer with General Electric
  - Staff Analog Engineer with National Instruments
  - RF Engineer with Huawei Technologies
  - etc.

- https://www.linkedin.com/in/franklins/
- ee.linsf@gmail.com

Thanks!
Questions and Answers

- Thanks to the reviewers' helpful comments.
- Thank you all for attending.
- What's the main idea of this work?
  - Using the multi-feature GLMB framework, to jointly separate and track multiple features of speakers.
- Is the localization (MCC-PHAT) reverberation robust?
  - The short answer is yes.
- Other questions?