An Expectation-Maximization Eigenvector Clustering Approach to Direction of Arrival Estimation of Multiple Speech Sources

presented by

<u>Xiong Xiao</u>[1], *Shengkui Zhao*[2], *Thi Ngoc Tho Nguyen*[2], *Douglas L. Jones*[2], *Eng Siong Chng*[1,3], *Haizhou Li*[1,3,4]

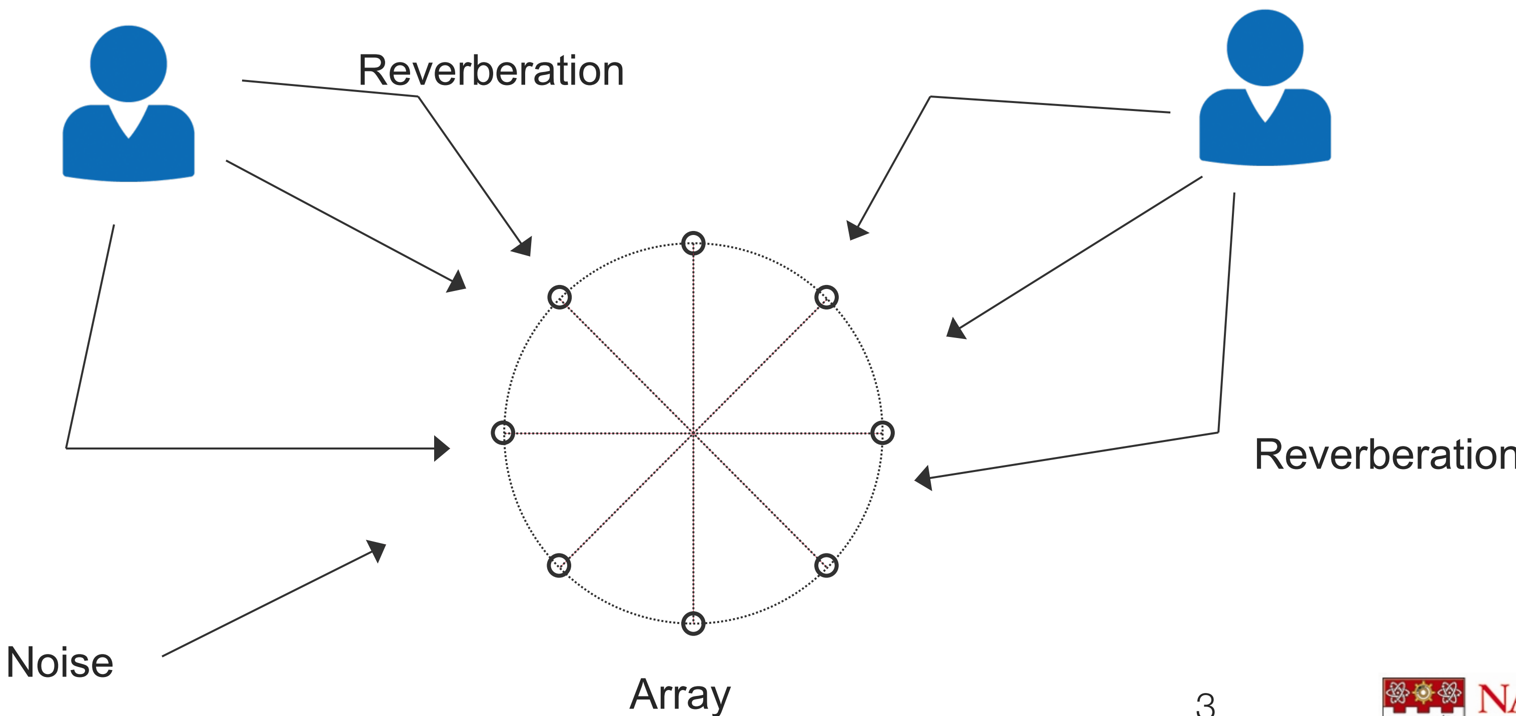
[1]Temasek Lab@NTU, Nanyang Technological University, Singapore.
[2]Advanced Digital Sciences Center, Singapore.
[3]School of Computer Engineering, Nanyang Technological University, Singapore.
[4]Department of Human Language Technology, Institute for Infocomm Research, Singapore.

Outline

- Task definition and literature review
- Proposed EM-based eigenvector clustering
  - Time-frequency (TF) bins selection
  - Extraction of eigenvectors
  - Generative modeling of eigenvectors
  - EM-based eigenvector clustering
- Experiments
- Conclusions

Task definition

- Detect the direction-of-arrival (DOA) of multiple speakers simultaneously in noisy and reverberant environments.
- We use a circular array: Diameter = 20cm, 8 microphones



NANYANG TECHNOLOGICAL UNIVERSITY

Review on Multiple Source Localization

- Angular spectrum based approach
  - Define an angular spectrum that is a function of DOA/TDOA.
  - Find the peaks in angular spectrum, one peak for one source. and then estimate the DOA/TDOA from the peaks.
  - Example: MUSIC [1], GCC-PHAT [2].
- Clustering based approach
  - Assume sparsity in time-frequency (TF) spectrogram representation of speech --- every TF bin is dominated by only 1 source.
  - Cluster the TF bins into several clusters. One cluster represent one source. Then estimate DOA/TDOA for each cluster.
  - Examples: normalized observation vector clustering [3], MESSL [4]

[1] R. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Transactions on Antennas and Propagation 34 (3) (1986) 276–280.
[2] C. Knapp, G. Carter, The generalized cross-correlation method for estimation of time delay, IEEE Transactions on Acoustics, Speech and Signal Processing 24 (4) (1976) 320–327.
[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *ICASSP 2006.*
[4] MichaelMandel,RonJWeiss,DanielPWEllis,etal.,"Model- based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

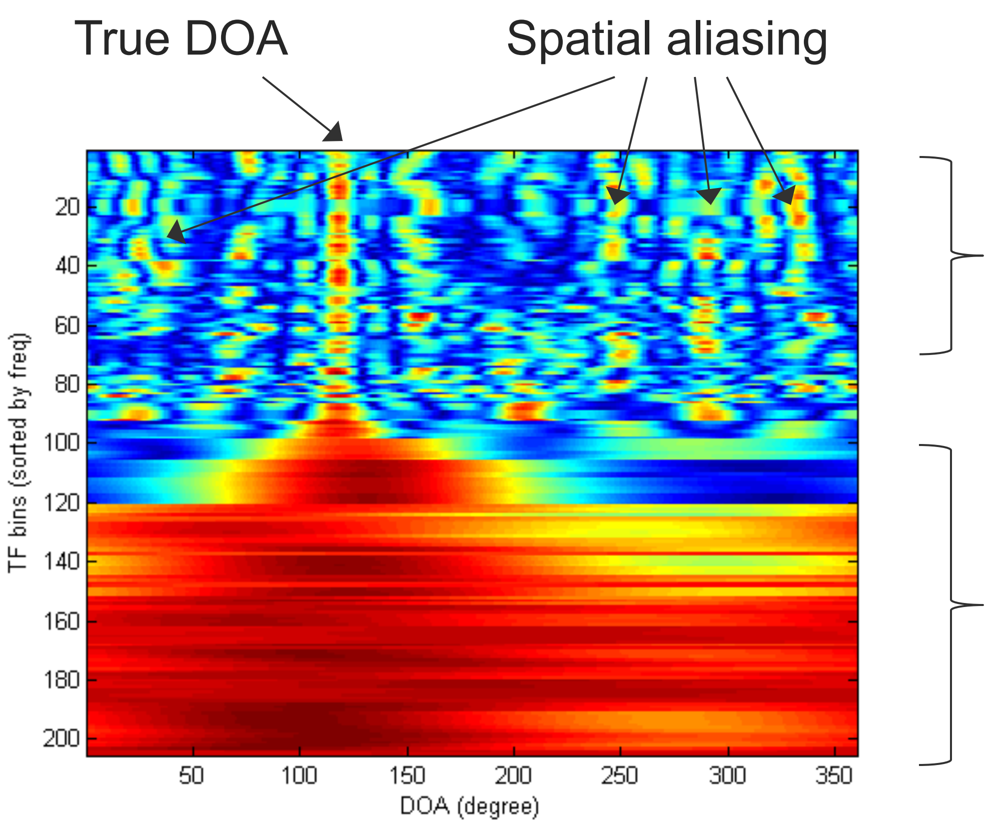
Issue 1: The correctness of the sparsity assumption

- The sparsity assumption of TF representation is not realistic for reverberant and noisy scenarios.
  - Reverberation will cause the power of the sources to spread to many future frames.
  - Ambient noise affects large proportion of TF bins.
- One solution is to model the reverberation and noise explicitly in clustering.

- We argue that it may be easier to just ignore the TF bins in poor condition (i.e. low SNR, higher reverberation).
- We propose to first select TF bins in good condition, and then apply the TF bin clustering algorithm.

Issue 2: Comparing different frequency bins

- Angular spectra at high and low frequency bins have their strength and limitation. They need to be combined to achieve accurate and robust DOA estimation.
- However, the TF bins at different frequencies cannot be compared directly. Even normalizations [3] cannot deal with spatial aliasing.
- In this work, we design an EM framework that can deal with all frequencies naturally.



Angular spectra for high frequency bins has high spatial resolution, but suffers from spatial aliasing.

Angular spectra for medium and low frequency has no spatial aliasing, but has very poor spatial resolution.

[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *ICASSP 2006.*

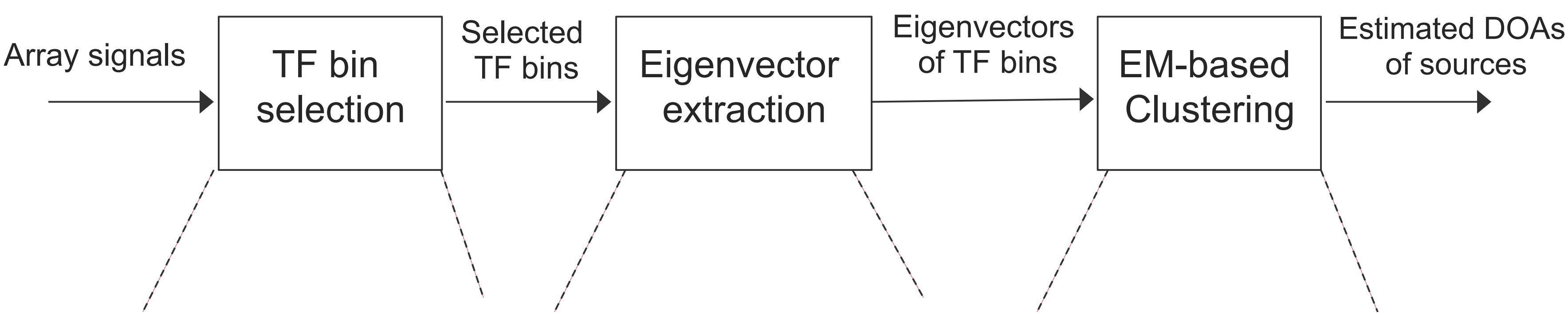
Outline

- Task definition and literature review
- Proposed EM-based eigenvector clustering
  - Time-frequency (TF) bins selection
  - Extraction of eigenvectors
  - Generative modeling of eigenvectors
  - EM-based eigenvector clustering
- Experiments
- Conclusions

System Diagram

The proposed method consists of 3 steps:

- Step 1: select TF bins with low reverberation and high SNR, and dominated by one source [5].
- Step 2: estimate the spatial covariance for the selected TF bins, get the dominant eigenvector.
- Step 3: perform EM-based clustering of the TF bins using the eigenvectors as features.



Select single source dominated TF bins [5]

- Noise tracking
- Onset detection
- Coherence test

Estimate covariance matrix:

$$\tilde{\mathbf{R}}(n,\omega) = \frac{1}{2C+1} \sum_{c=n-C}^{n+C} \mathbf{x}(c,\omega)\mathbf{x}^{H}(c,\omega).$$

Eigen-decompose covariance matrix, take the most dominant eigenvector as steering vector of the source of the TF bin.

- Cluster the eigenvectors.
- Estimate the DOA for each cluster.

(details to be shown in next few slides)

[5] TNT Nguyen, S. Zhao, and Douglas L. Jones, "Robust DOA estimation of multiple speech sources," ICASSP 2014.

NANYANG TECHNOLOGICAL UNIVERSITY

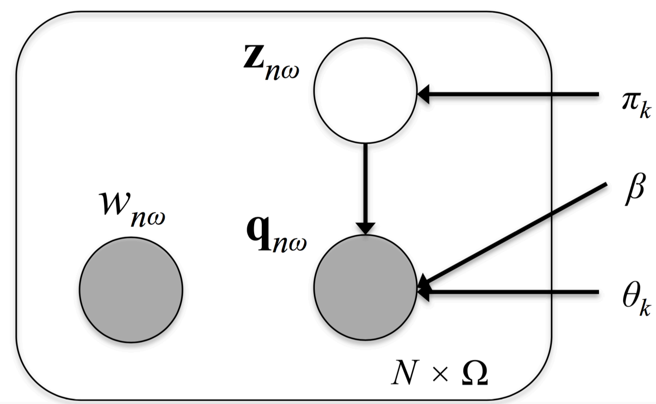
Generative modeling of eigenvectors

- Assume there are K sources
- We model the distribution of the TF bin eigenvectors by a mixture density function with K component density functions. .

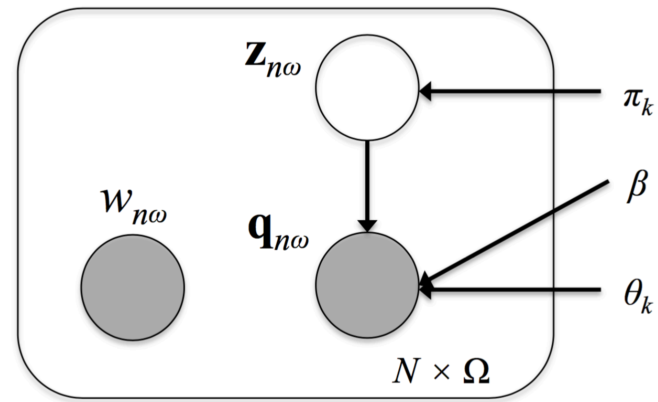
$$p(w_{n\omega}, \mathbf{q}_{n\omega}; \Theta) = \sum_{k=1}^{K} p(z_{n\omega k} = 1) p(w_{n\omega}, \mathbf{q}_{n\omega} | z_{n\omega k} = 1; \theta_k)$$

- One component of the mixture represents one source. The clustering process is equal to the maximizing of the likelihood of the observations (the eigenvector and its reliability measure).

- $\mathbf{q}_{n\omega}$ is the eigenvector (observation) at frame n and frequency ω.
- $w_{n\omega}$ is a reliability measure of $\mathbf{q}_{n\omega}$. It is also an observation.
- $\mathbf{z}_{n\omega}$ is a latent variable that denotes the cluster membership of $\mathbf{q}_{n\omega}$.
- N: total number of frames, Ω : number of frequency bins.
- K is the number of sources.



Definition of component densities



- Component weights

$$p(z_{n\omega k} = 1) = \pi_k, \quad k \in [1, K],$$

$$\sum_{k=1}^{K} \pi_k = 1.$$

- Component density function

$$p(w_{n\omega}, \mathbf{q}_{n\omega} | z_{n\omega j} = 1; \theta_j) = \frac{\exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|)}{\mathcal{E}(\beta, \theta_j)}$$

  - $|\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|$ measures how much the eigenvector agrees with steering vector $\mathbf{e}_{j\omega}$.
  - The more the eigenvector agrees with the steering vector, the higher the likelihood.
  - $\mathcal{E}(\beta, \theta_j) = \int_{\mathbf{q}_{n\omega}, w_{n\omega}, \omega} \exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|) d\mathbf{q}_{n\omega} dw_{n\omega} d\omega$ is the normalization term and ensures that the distribution integrates to 1.
  - We assure that the normalization term is independent of the DOA for simplicity.

- Mixture density function

$$p(w_{n\omega}, \mathbf{q}_{n\omega}; \Theta) = \sum_{k=1}^{K} \pi_k \frac{\exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{k\omega}|)}{\mathcal{E}(\beta)}$$

EM-based clustering – iterative algorithm

- We maximize the auxiliary function of EM.
- E-step: compute the membership function of TF bins

$$\gamma_{n\omega k} = p(z_{n\omega k} = 1 | w_{n\omega}, \mathbf{q}_{n\omega}; \Theta') = \frac{\pi_k \exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{k\omega}|)}{\sum_{j=1}^{K} \pi_j \exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|)}$$

  – For efficient implementation, we can use "winner take all" membership function, i.e. each TF bin is assigned to one and only one cluster.

- M-step: estimate the DOA for each cluster based on the TF bins that are assigned to the cluster by the membership function:

$$\hat{\theta}_k = \arg \max_{\theta_k \in [0,359]} \sum_{\{n,\omega\} \in \Psi} \gamma_{n\omega k} \beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{k\omega}|$$

  – A grid search (0-359 degrees) can be performed to find the optimal DOA.

- With "winner take all" membership function, the EM algorithm becomes like k-means clustering. However, a key difference is that we use a set of template steering vectors for each cluster.

Outline

- Task definition and literature review
- Proposed EM-based eigenvector clustering
  - Time-frequency (TF) bins selection
  - Extraction of eigenvectors
  - Generative modeling of eigenvectors
  - EM-based eigenvector clustering
- Experiments
- Conclusions

NANYANG TECHNOLOGICAL UNIVERSITY

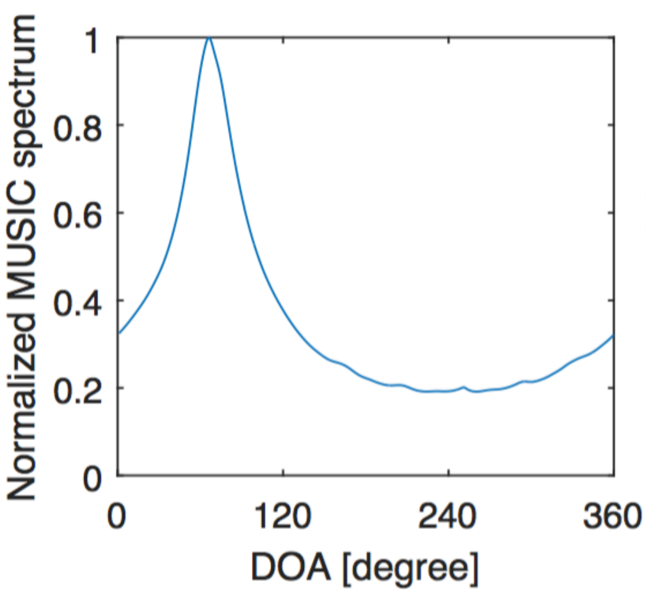
Experiments – Data Description

- Array geometry: 8-microphones circular, 20cm diameter.
- 2D DOA estimation, 0-359 degrees
- 16kHz sampling rate, 512 FFT length
- Simulated test data
  - Synthesized by convolving clean speech signal from WSJCAM0 to simulated room impulse response. Additive noise is added later.
  - T60 times from 0.3s to 0.9s. SNR from 0dB to 20dB.
- Real test data
  - Three scenarios: small meeting room (4mx3mx2.5m, T60=0.34s), pantry room (6mx5mx2.5m, T60=0.47s), and lift lobby (8mx4m,3m, T60=1.07s).
  - A male speaker's voice was recorded at 0, 45, 90 degrees.
  - A female speaker's voice was recorded at 135, 180, 225 degrees.
  - Various mixtures can be simulated by combining these recordings.
  - All test utterances are 6s in length.

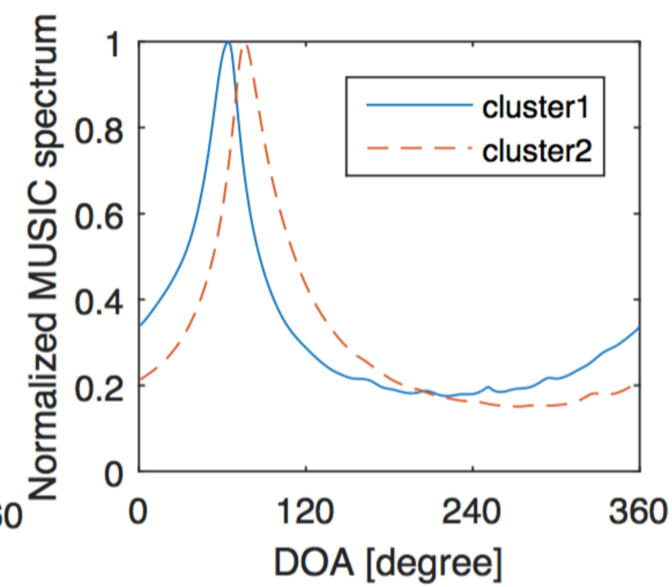
Ability to resolve closely spaced speakers

- MUSIC spatial spectrum is not able to resolve two closely spaced speakers (65 and 72 degrees). Only one peak in the spectrum.
- The proposed EM based clustering produces two clusters, each corresponds to one source.

MUSIC



EM-clustering



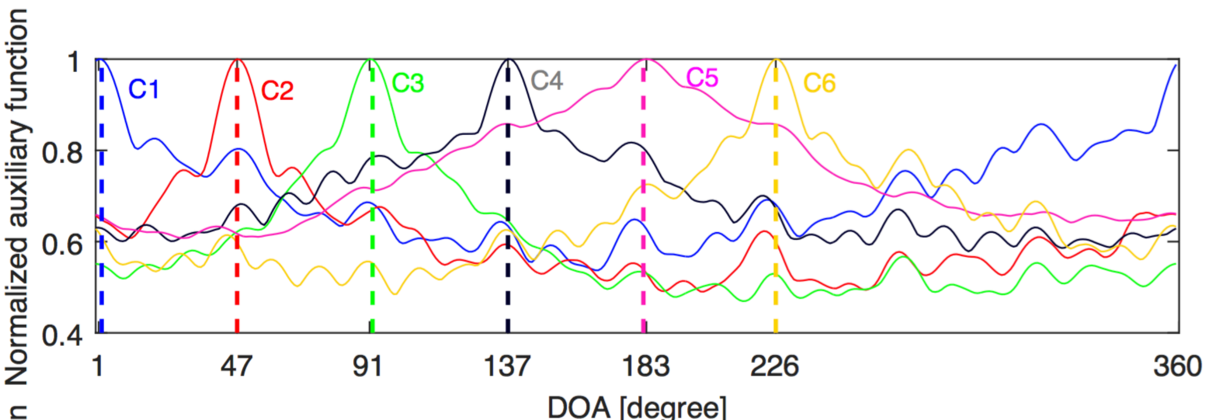
True DOA = 65 and 72 degrees

NANYANG TECHNOLOGICAL UNIVERSITY

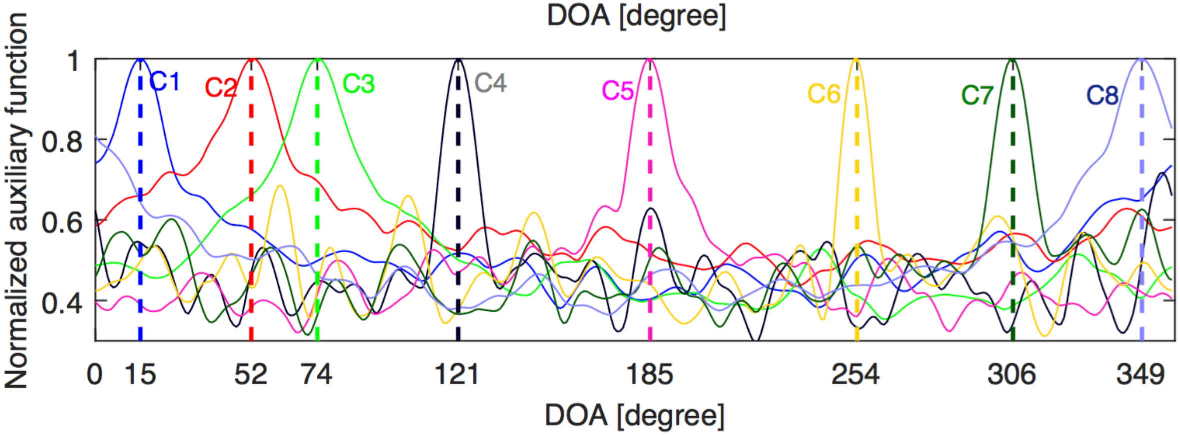
Ability to detect multiple sources

- The proposed EM-based clustering is able to detect up to 6 real sources and 8 simulated sources accurately.

Plot of normalized auxiliary function of clusters



- 6 real sources,
- T60=0.47s.
- DOAs = 0/45/90/135/180/225



- 8 simulated sources,
- T60=1.0s, SNR = 20dB
- DOAs = 13/50/74/118/186/254/306/349

Results on Simulated Data (2 sources)

- MUSIC: angular spectrum based method
- Algorithm [14]: k-means clustering of TF bins proposed in [15]. (Typo in the paper)
  - Select TF bins (same as this work).
  - Find DOA of each TF bin, then cluster the DOAs.
  - Due to spatial aliasing, only uses TF bins below 1700Hz.
- Proposed: the EM-based eigenvector clustering proposed in this paper.

Room	Method	SNR=20dB	SNR=10dB	SNR=0dB
Small	MUSIC	15.51	18.30	23.76
	Algorithm[14]	14.26	18.41	20.50
	Proposed	**2.78**	**2.84**	**6.75**
Medium	MUSIC	13.67	9.41	12.37
	Algorithm[14]	18.19	10.97	14.12
	Proposed	**3.76**	**5.15**	**3.07**
Large	MUSIC	7.87	4.99	14.94
	Algorithm[14]	11.70	10.10	11.49
	Proposed	**1.17**	**2.04**	**9.47**

The numbers are root mean square of DOA estimation errors are compared.

[15] TNT Nguyen, S. Zhao, and Douglas L. Jones, “Robust DOA estimation of multiple speech sources,” in *ICASSP 2014.*

Results on Real Data (2 sources)

- DOA ground truth is obtained by using MUSIC for each single-source recordings individually.
- Significant improvement over MUSIC and the k-means algorithm.

		Testing Environment		
		small	pantry	lift
Method	MUSIC	33.43	28.7	62.66
	Algorithm[14]	40.74	42.92	70.02
	Proposed	**1.7**	**0.92**	**13.02**

T60 of small room = 0.34s
T60 of pantry room = 0.47s
T60 of lift = 1.07s

Conclusions

- We propose an EM-based eigenvector clustering methods for multi-source DOA estimation.
  - Selection of TF bins that are dominated by one source.
  - Multiple steering vectors for each cluster to use information in all frequency bins.
- Significant improvement can be obtained compared to MUSIC and a k-means clustering method.
- Future work
  - Automatic determination of number of sources
  - Better modeling of eigenvectors
  - Extend to source separation (cluster all TF bins)

Thank you!