Separake

Source Separation with a Little Help from Echoes

Robin Scheibler Diego Di Carlo Antoine Deleforge Ivan Dokmanić

April 17, 2018

Tokyo Metropolitan University 首都大学東京

Inria

ILLINOIS

ICASSP 2018

Echoes Help Indoor Processing

- beamforming
- source localization
- self-localization

What about speech separation ?

1. Is speech separation easier with echoes than without ?
2. Full RIR vs a few early reflections ?

Echoes Help Indoor Processing

- beamforming
- source localization
- self-localization

What about speech separation ?

1. Is speech separation easier with echoes than without ?
2. Full RIR vs a few early reflections ?

Echoes Help Indoor Processing

- beamforming
- source localization
- self-localization

What about speech separation ?

1. Is speech separation easier with echoes than without ?
2. Full RIR vs a few early reflections ?

Echoes Help Indoor Processing

- beamforming
- source localization
- self-localization

What about speech separation ?

1. Is speech separation easier with echoes than without ?
2. Full RIR vs a few early reflections ?

1. Assume knowledge of a few (1-6) early echoes
2. Plug into multichannel NMF [1]
3. Three baseline scenarios
    - *Anechoic* conditions
    - *Learn* transfer functions
    - Ignore reverberation (i.e. consider 0 echoes)
4. Numerical Experiments

[1] Ozerov & Févotte, 2010

Our Methodology

1. Assume knowledge of a few (1-6) early echoes
2. Plug into multichannel NMF [1]
3. Three baseline scenarios
   - *Anechoic* conditions
   - *Learn* transfer functions
   - Ignore reverberation (i.e. consider 0 echoes)
4. Numerical Experiments

[1] Ozerov & Févotte, 2010

Our Methodology

1. Assume knowledge of a few (1-6) early echoes
2. Plug into multichannel NMF [1]
3. Three baseline scenarios
    - *Anechoic* conditions
    - *Learn* transfer functions
    - Ignore reverberation (i.e. consider 0 echoes)
4. Numerical Experiments

[1] Ozerov & Févotte, 2010

1. Assume knowledge of a few (1-6) early echoes
2. Plug into multichannel NMF [1]
3. Three baseline scenarios
   - *Anechoic* conditions
   - *Learn* transfer functions
   - Ignore reverberation (i.e. consider 0 echoes)
4. Numerical Experiments

[1] Ozerov & Févotte, 2010

1. Assume knowledge of a few (1-6) early echoes
2. Plug into multichannel NMF [1]
3. Three baseline scenarios
   - *Anechoic* conditions
   - *Learn* transfer functions
   - Ignore reverberation (i.e. consider 0 echoes)
4. Numerical Experiments

[1] Ozerov & Févotte, 2010

1. Assume knowledge of a few (1-6) early echoes
2. Plug into multichannel NMF [1]
3. Three baseline scenarios
   - *Anechoic* conditions
   - *Learn* transfer functions
   - Ignore reverberation (i.e. consider 0 echoes)
4. Numerical Experiments

[1] Ozerov & Févotte, 2010

1. Assume knowledge of a few (1-6) early echoes
2. Plug into multichannel NMF [1]
3. Three baseline scenarios
   - *Anechoic* conditions
   - *Learn* transfer functions
   - Ignore reverberation (i.e. consider 0 echoes)
4. Numerical Experiments

[1] Ozerov & Févotte, 2010

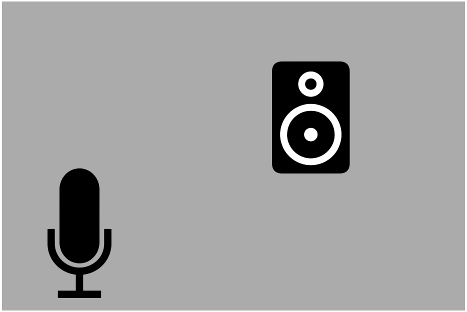
Outline

1. Approximate Propagation Model

2. NMF Algorithms

3. Results from Numerical Experiments

Approximate Propagation Model



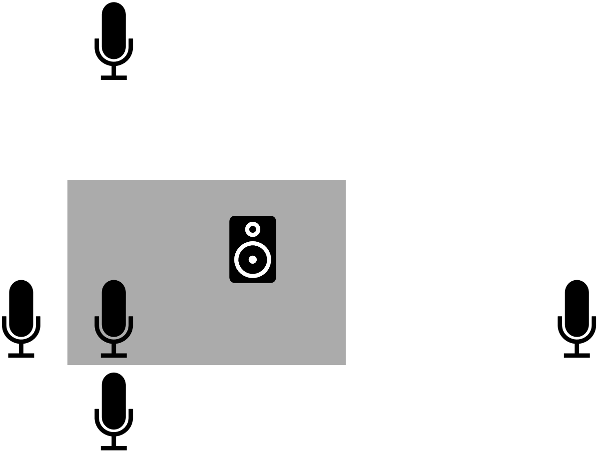
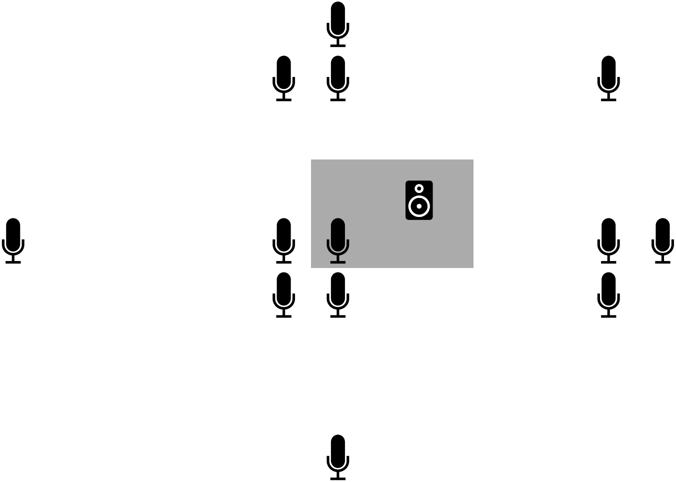
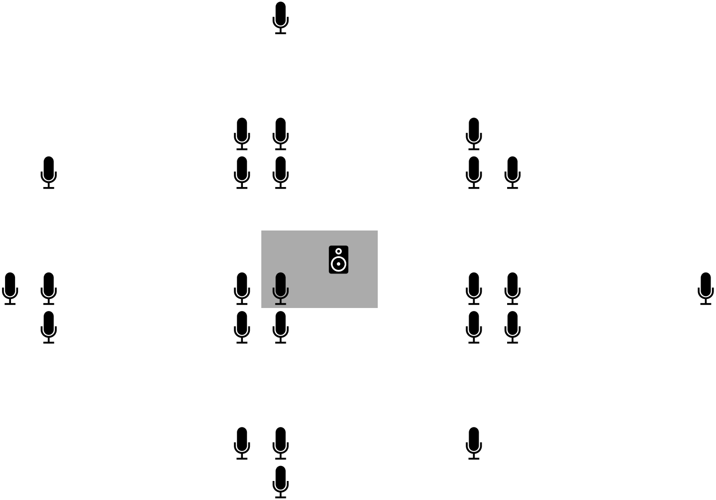
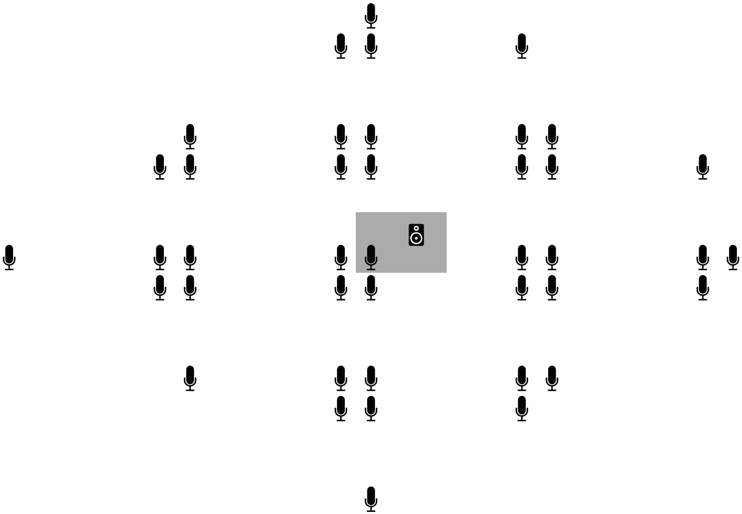


Image Microphone Model

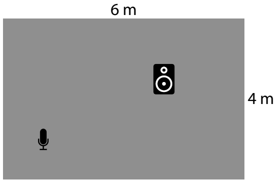




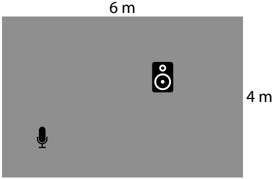


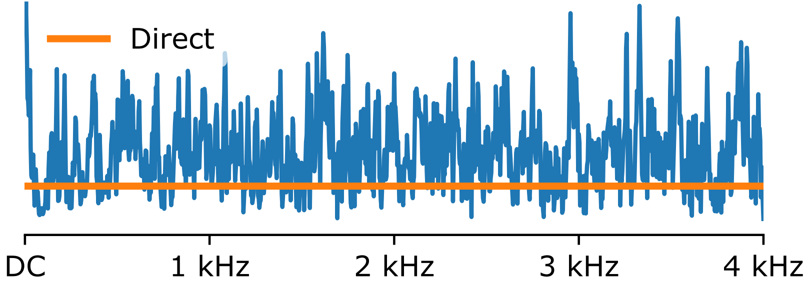
Partial Room Impulse Responses

$$h_{jm}(t) = \sum_{k=0}^{K} \alpha_{jm}^{k} \delta(t - t_{jm}^{k}) + \epsilon_{jm}(t)$$

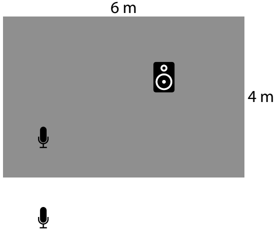


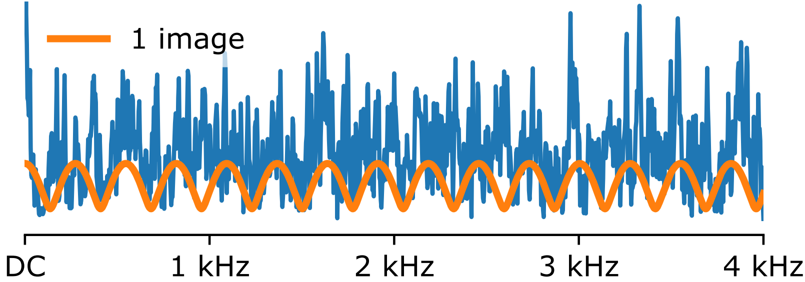
$$h_{jm}(t) = \sum_{k=0}^{K} \alpha_{jm}^{k}\delta(t - t_{jm}^{k}) + \epsilon_{jm}(t)$$





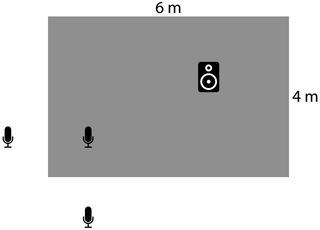
$$h_{jm}(t) = \sum_{k=0}^{K} \alpha_{jm}^{k} \delta(t - t_{jm}^{k}) + \epsilon_{jm}(t)$$

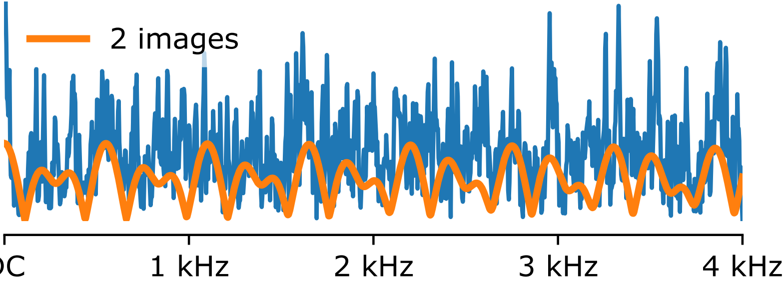




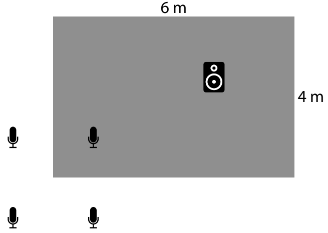
Partial Room Impulse Responses

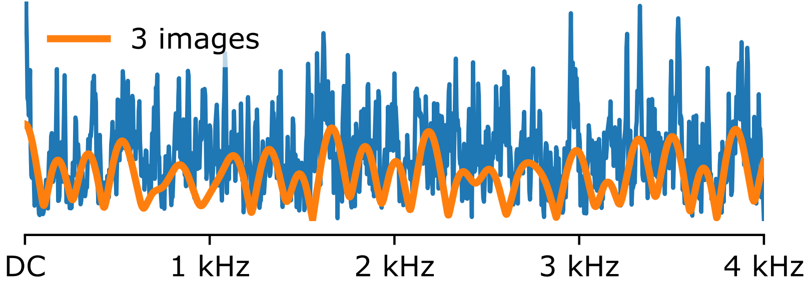
$$h_{jm}(t) = \sum_{k=0}^{K} \alpha_{jm}^{k} \delta(t - t_{jm}^{k}) + \epsilon_{jm}(t)$$

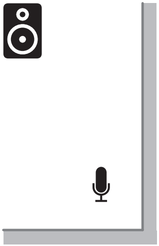


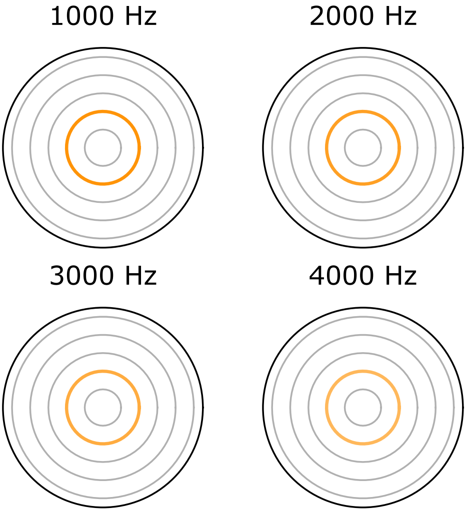


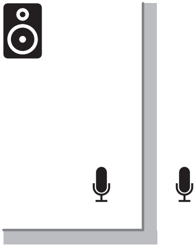
$$h_{jm}(t) = \sum_{k=0}^{K} \alpha_{jm}^{k}\delta(t - t_{jm}^{k}) + \epsilon_{jm}(t)$$

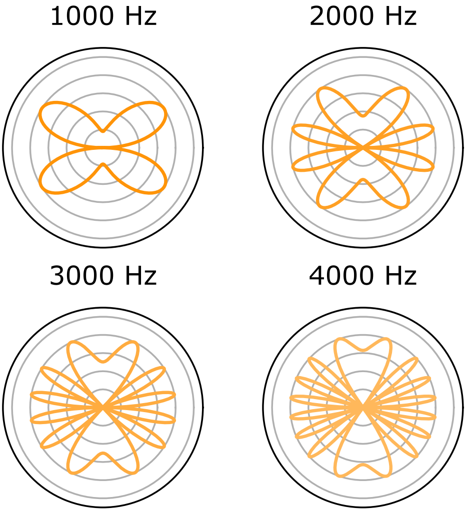


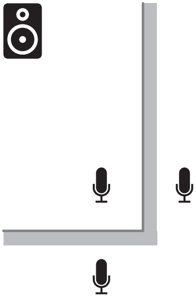


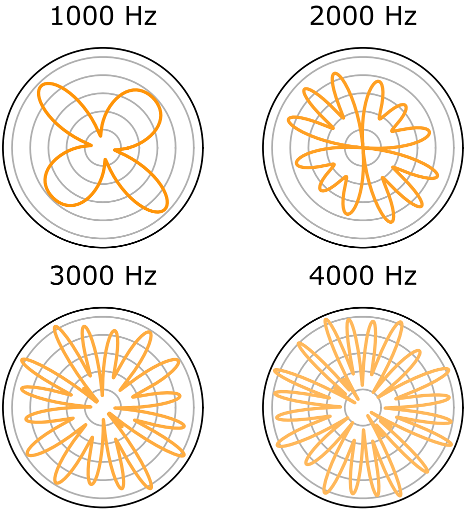



1000 Hz
2000 Hz
3000 Hz
4000 Hz



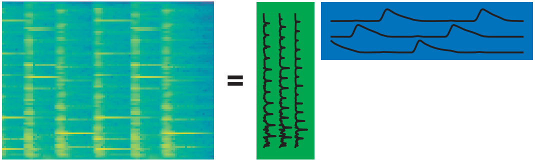

1000 Hz
2000 Hz
3000 Hz
4000 Hz




1000 Hz
2000 Hz
3000 Hz
4000 Hz

NMF Algorithms

Non-negative Spectrogram Source Model



Multiplicative Updates View (Lee & Seung 2001)

Source signal's magnitude spectrogram decomposes non-negatively

$$|\mathbf{X}_j| = \mathbf{D}_j \mathbf{Z}_j$$

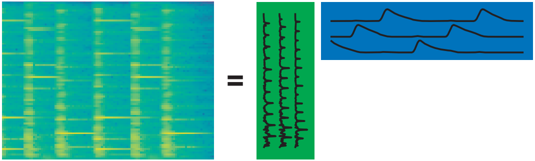
Expectation Maximization View (Ozerov & Févote 2010)

Source signal's variance spectrogram decomposes non-negatively

$$X_j[f, n] \sim \mathcal{CN}(0, (\mathbf{D}_j \mathbf{Z}_j)_{fn})$$

In this work: $\mathbf{D}_j$ is pre-trained, known dictionary

Non-negative Spectrogram Source Model



Multiplicative Updates View (Lee & Seung 2001)

Source signal's magnitude spectrogram decomposes non-negatively

$$|\mathbf{X}_j| = \mathbf{D}_j \mathbf{Z}_j$$

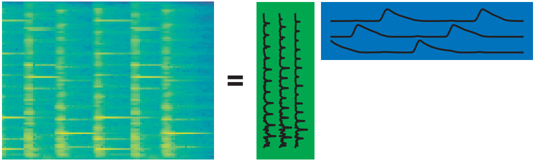
Expectation Maximization View (Ozerov & Févote 2010)

Source signal's variance spectrogram decomposes non-negatively

$$X_j[f, n] \sim \mathcal{CN}(0, (\mathbf{D}_j \mathbf{Z}_j)_{fn})$$

In this work: $\mathbf{D}_j$ is pre-trained, known dictionary

Non-negative Spectrogram Source Model



Multiplicative Updates View (Lee & Seung 2001)

Source signal's magnitude spectrogram decomposes non-negatively

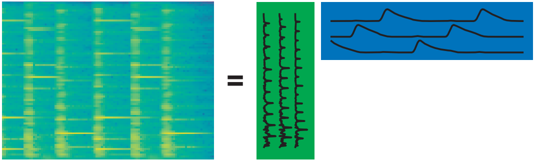
$$|\mathbf{X}_j| = \mathbf{D}_j \mathbf{Z}_j$$

Expectation Maximization View (Ozerov & Févote 2010)

Source signal's variance spectrogram decomposes non-negatively

$$X_j[f, n] \sim \mathcal{CN}(0, (\mathbf{D}_j \mathbf{Z}_j)_{fn})$$

In this work: $\mathbf{D}_j$ is pre-trained, known dictionary



Multiplicative Updates View (Lee & Seung 2001)

Source signal's magnitude spectrogram decomposes non-negatively

$$|\mathbf{X}_j| = \mathbf{D}_j \mathbf{Z}_j$$

Expectation Maximization View (Ozerov & Févote 2010)

Source signal's variance spectrogram decomposes non-negatively

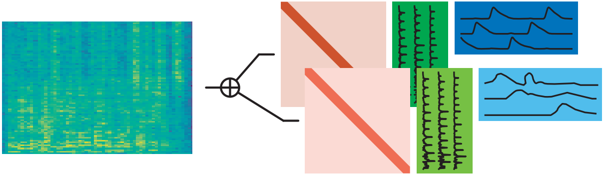
$$X_j[f, n] \sim \mathcal{CN}(0, (\mathbf{D}_j \mathbf{Z}_j)_{fn})$$

In this work: $\mathbf{D}_j$ is pre-trained, known dictionary

Multiplicative Updates - NMF

Microphone magnitude spectrogram model

$$\widehat{\mathbf{V}}_m = \sum_j \mathrm{diag}(|\widehat{\mathbf{H}}_{mj}|)\mathbf{D}_j\mathbf{Z}_j$$



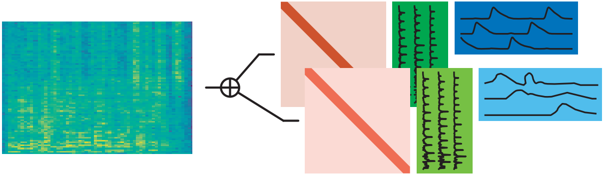
Minimize *Itakura-Saito* divergence

$$C_{\mathrm{MU}}(\mathbf{Z}_j) = \sum_{mfn} d_{\mathrm{IS}}(V_m[f,n] \mid \widehat{V}_m[f,n]) + \gamma \sum_j \|\mathbf{Z}_j\|_1$$

- Efficient multiplicative update rules (Ozerov & Févotte 2010)
- Regularization needed for large number of latent variables

Microphone magnitude spectrogram model

$$\widehat{\mathbf{V}}_m = \sum_j \operatorname{diag}(|\widehat{\mathbf{H}}_{mj}|)\mathbf{D}_j\mathbf{Z}_j$$



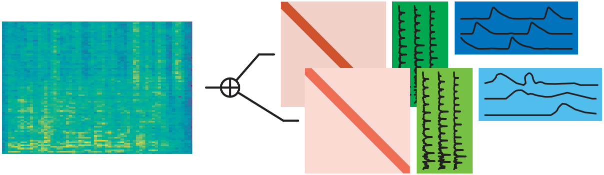
Minimize *Itakura-Saito* divergence

$$C_{\text{MU}}(\mathbf{Z}_j) = \sum_{mfn} d_{\text{IS}}(V_m[f,n] \mid \widehat{V}_m[f,n]) + \gamma \sum_j \|\mathbf{Z}_j\|_1$$

- Efficient multiplicative update rules (Ozerov & Févotte 2010)
- Regularization needed for large number of latent variables

Microphone magnitude spectrogram model

$$\widehat{\mathbf{V}}_m = \sum_j \mathrm{diag}(|\widehat{\mathbf{H}}_{mj}|)\mathbf{D}_j\mathbf{Z}_j$$



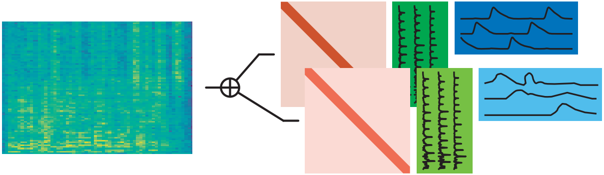
Minimize *Itakura-Saito* divergence

$$C_{\mathrm{MU}}(\mathbf{Z}_j) = \sum_{mfn} d_{\mathrm{IS}}(V_m[f,n] \mid \widehat{V}_m[f,n]) + \gamma \sum_j \|\mathbf{Z}_j\|_1$$

- Efficient multiplicative update rules (Ozerov & Févotte 2010)
- Regularization needed for large number of latent variables

Microphone magnitude spectrogram model

$$\widehat{\mathbf{V}}_m = \sum_j \mathrm{diag}(|\widehat{\mathbf{H}}_{mj}|)\mathbf{D}_j\mathbf{Z}_j$$



Minimize *Itakura-Saito* divergence

$$C_{\mathrm{MU}}(\mathbf{Z}_j) = \sum_{mfn} d_{\mathrm{IS}}(V_m[f,n] \mid \widehat{V}_m[f,n]) + \gamma \sum_j \|\mathbf{Z}_j\|_1$$

- Efficient multiplicative update rules (Ozerov & Févotte 2010)
- Regularization needed for large number of latent variables

Probabilistic Model

Source are complex Gaussian with low-rank spectrogram

$$X_j[f,n] \sim \mathcal{CN}(0, (\mathbf{D}_j \mathbf{Z}_j)_{fn})$$

Microphone signals have variance

$$\mathbf{\Sigma_y}[f,n] = \widehat{\mathbf{H}}[f]\, \mathbf{\Sigma_x}[f,n]\, \widehat{\mathbf{H}}^H[f] + \mathbf{\Sigma_b}[f,n],$$

Minimize Negative Log-likelihood

$$C_{\text{EM}}(\mathbf{Z}_j) = \sum_{fn} \operatorname{trace}\left(\mathbf{y}[f,n]\mathbf{y}[f,n]^H \mathbf{\Sigma}_\mathbf{y}^{-1}[f,n]\right) + \log\det \mathbf{\Sigma_y}[f,n]$$

Efficiently minimized by Expectation-Maximization algorithm (Ozerov & Févotte 2010)

Probabilistic Model

Source are complex Gaussian with low-rank spectrogram

$$X_j[f,n] \sim \mathcal{CN}(0, (\mathbf{D}_j \mathbf{Z}_j)_{fn})$$

Microphone signals have variance

$$\mathbf{\Sigma_y}[f,n] = \widehat{\mathbf{H}}[f]\,\mathbf{\Sigma_x}[f,n]\,\widehat{\mathbf{H}}^H[f] + \mathbf{\Sigma_b}[f,n],$$

Minimize Negative Log-likelihood

$$C_{\mathrm{EM}}(\mathbf{Z}_j) = \sum_{fn} \operatorname{trace}\left(\mathbf{y}[f,n]\mathbf{y}[f,n]^H \mathbf{\Sigma_y}^{-1}[f,n]\right) + \log\det \mathbf{\Sigma_y}[f,n]$$

Efficiently minimized by Expectation-Maximization algorithm (Ozerov & Févotte 2010)

Probabilistic Model

Source are complex Gaussian with low-rank spectrogram

$$X_j[f,n] \sim \mathcal{CN}(0, (\mathbf{D}_j \mathbf{Z}_j)_{fn})$$

Microphone signals have variance

$$\boldsymbol{\Sigma}_{\mathbf{y}}[f,n] = \widehat{\mathbf{H}}[f]\,\boldsymbol{\Sigma}_{\mathbf{x}}[f,n]\,\widehat{\mathbf{H}}^H[f] + \boldsymbol{\Sigma}_{\mathbf{b}}[f,n],$$

Minimize Negative Log-likelihood

$$C_{\mathrm{EM}}(\mathbf{Z}_j) = \sum_{fn} \operatorname{trace}\left(\mathbf{y}[f,n]\mathbf{y}[f,n]^H \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}[f,n]\right) + \log\det \boldsymbol{\Sigma}_{\mathbf{y}}[f,n]$$

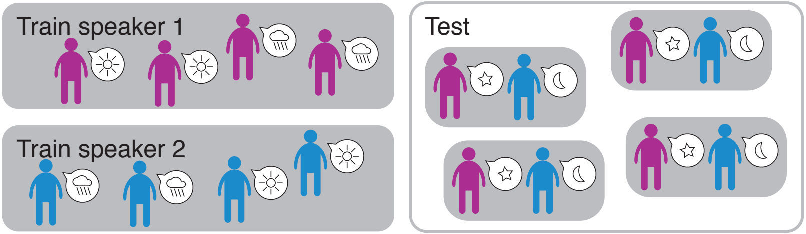
Efficiently minimized by Expectation-Maximization algorithm (Ozerov & Févotte 2010)

Pre-trained Dictionaries

Speaker Dependent

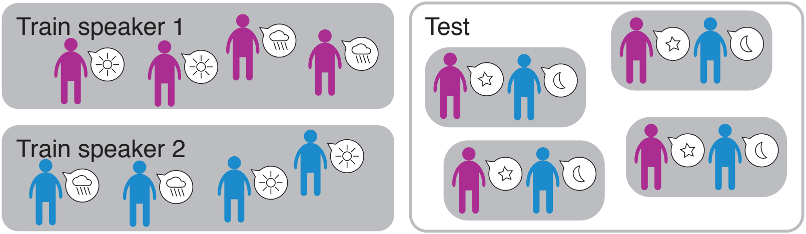
Universal

Speaker Dependent

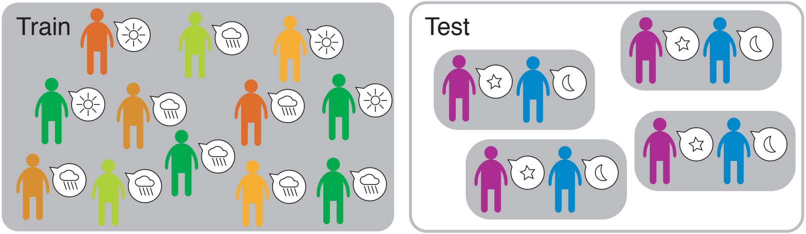


Universal

Speaker Dependent



Universal



Remark 1: Anechoic separation *cannot* work!

$$\widehat{\mathbf{V}}_m = \sum_j \mathbf{D}_j \mathbf{Z}_j \quad \rightarrow \quad \widehat{\mathbf{V}}_m = \sum_j \mathbf{D}\mathbf{Z}_j = \mathbf{D} \sum_j \mathbf{Z}_j$$

Remark 2: TF makes universal dict. speaker specific

$$\widehat{\mathbf{V}}_m = \sum_j (\mathbf{H}_{mj}\mathbf{D})\mathbf{Z}_j$$

Remark 3: EM-NMF with Universal Dictionary

- Unclear how to enforce sparsity in EM (to us)
- Left for future work

Remark 1: Anechoic separation *cannot* work!

$$\widehat{\mathbf{V}}_m = \sum_j \mathbf{D}_j \mathbf{Z}_j \quad \rightarrow \quad \widehat{\mathbf{V}}_m = \sum_j \mathbf{D} \mathbf{Z}_j = \mathbf{D} \sum_j \mathbf{Z}_j$$

Remark 2: TF makes universal dict. speaker specific

$$\widehat{\mathbf{V}}_m = \sum_j (\mathbf{H}_{mj} \mathbf{D}) \mathbf{Z}_j$$

Remark 3: EM-NMF with Universal Dictionary

- Unclear how to enforce sparsity in EM (to us)
- Left for future work

Remarks on Using a Universal Dictionary

Remark 1: Anechoic separation *cannot* work!

$$\widehat{\mathbf{V}}_m = \sum_j \mathbf{D}_j \mathbf{Z}_j \quad \rightarrow \quad \widehat{\mathbf{V}}_m = \sum_j \mathbf{D}\mathbf{Z}_j = \mathbf{D} \sum_j \mathbf{Z}_j$$

Remark 2: TF makes universal dict. speaker specific

$$\widehat{\mathbf{V}}_m = \sum_j (\mathbf{H}_{mj}\mathbf{D})\mathbf{Z}_j$$

Remark 3: EM-NMF with Universal Dictionary

- Unclear how to enforce sparsity in EM (to us)
- Left for future work

Results from Numerical Experiments

Conditions

# sources	2
# mics	3
STFT	2048

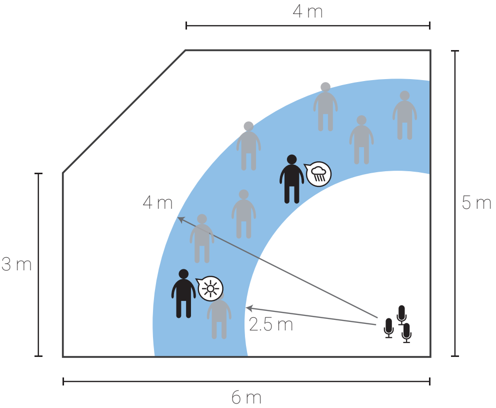
half-overlap, Hann win

Simulation with *pyroomacoustics*

T60 $\sim$ 100 ms

Baselines

- *Anechoic*
- *Learn* TF
- Ignore reverb

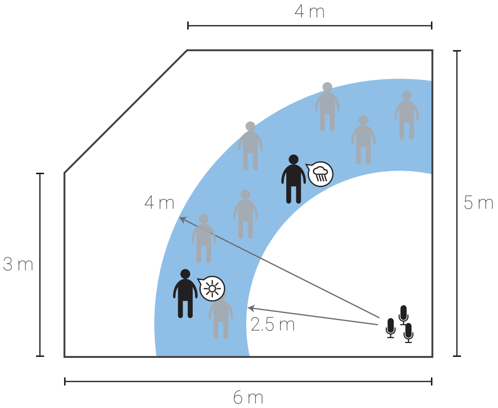


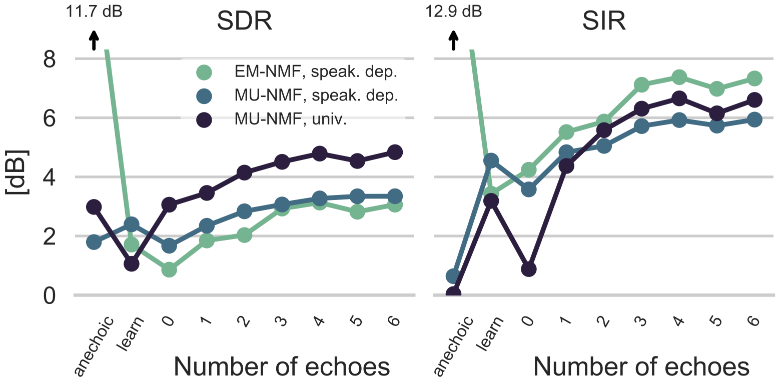
Conditions

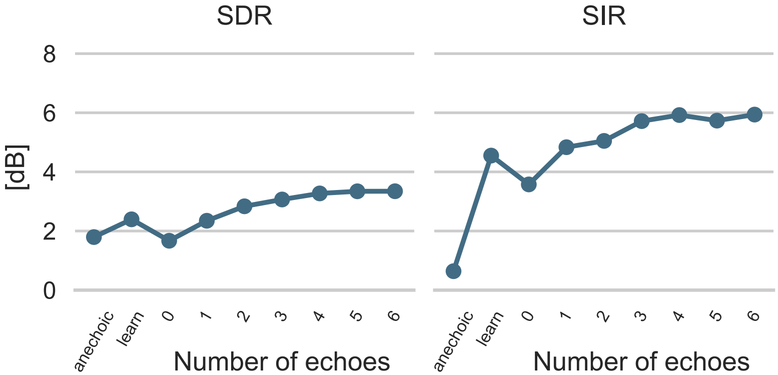
# sources	2
# mics	3
STFT	2048
half-overlap, Hann win	
Simulation with *pyroomacoustics*	
T60	$\sim$ 100 ms

Baselines

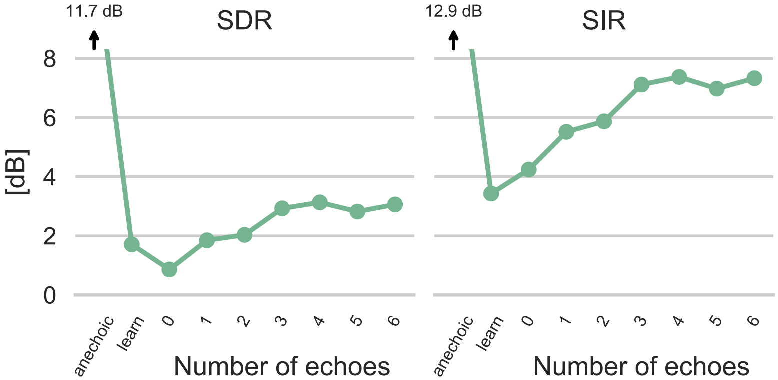
- *Anechoic*
- *Learn* TF
- Ignore reverb

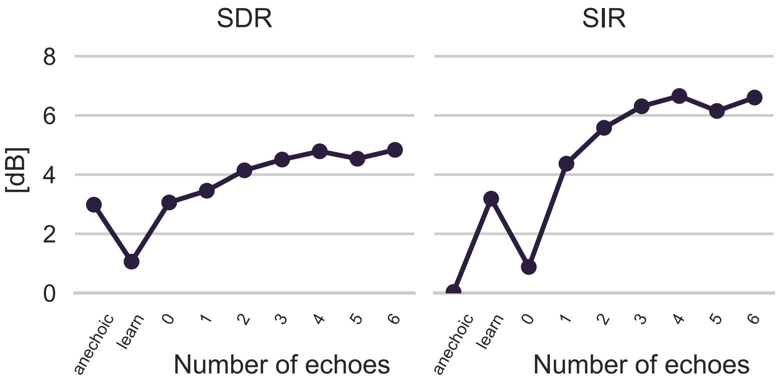



11.7 dB
SDR
12.9 dB
SIR
EM-NMF, speak. dep.
MU-NMF, speak. dep.
MU-NMF, univ.
[dB]
0
2
4
6
8
anechoic
learn
0
1
2
3
4
5
6
Number of echoes
anechoic
learn
0
1
2
3
4
5
6
Number of echoes


SDR
SIR
[dB]
8
6
4
2
0
anechoic
learn
0
1
2
3
4
5
6
Number of echoes

EM-NMF – Speaker Dependent




SDR
SIR
[dB]
8
6
4
2
0
anechoic
learn
0
1
2
3
4
5
6
Number of echoes
Number of echoes

Recall

$$C_{\mathrm{MU}}(\mathbf{Z}_j) = \sum_{mfn} d_{\mathrm{IS}}(V_m[f,n] \mid \widehat{V}_m[f,n]) + \gamma \sum_j \|\mathbf{Z}_j\|_1$$

			Number of echoes K						
	anechoic	learn	0	1	2	3	4	5	6
$\gamma =$	10	10^{-1}	10	10^{-4}	**0**	**0**	**0**	**0**	**0**

Table : Value of regularization parameter.

Partial RIR regularizes universal dictionary!

Recall

$$C_{\mathrm{MU}}(\mathbf{Z}_j) = \sum_{mfn} d_{\mathrm{IS}}(V_m[f,n] \mid \widehat{V}_m[f,n]) + \gamma \sum_j \|\mathbf{Z}_j\|_1$$

			Number of echoes K						
	anechoic	learn	0	1	2	3	4	5	6
$\gamma =$	10	10^{-1}	10	10^{-4}	**0**	**0**	**0**	**0**	**0**

Table : Value of regularization parameter.

Partial RIR regularizes universal dictionary!

Recall

$$C_{\mathrm{MU}}(\mathbf{Z}_j) = \sum_{mfn} d_{\mathrm{IS}}(V_m[f,n] \mid \widehat{V}_m[f,n]) + \gamma \sum_j \|\mathbf{Z}_j\|_1$$

			Number of echoes K						
	anechoic	learn	0	1	2	3	4	5	6
$\gamma =$	10	10^{-1}	10	10^{-4}	**0**	**0**	**0**	**0**	**0**

Table : Value of regularization parameter.

Partial RIR regularizes universal dictionary!

Conclusion

- Single echo improves performance
- Enables universal dictionary
- First few echoes most important

Future Work

- Compare to BSS
- Include (deeply) learnt models
- Underdetermined case

Thank you! Questions ?

Conclusion

- Single echo improves performance
- Enables universal dictionary
- First few echoes most important

Future Work

- Compare to BSS
- Include (deeply) learnt models
- Underdetermined case

Thank you! Questions ?

Conclusion

- Single echo improves performance
- Enables universal dictionary
- First few echoes most important

Future Work

- Compare to BSS
- Include (deeply) learnt models
- Underdetermined case

Thank you! Questions ?

Conclusion

- Single echo improves performance
- Enables universal dictionary
- First few echoes most important

Future Work

- Compare to BSS
- Include (deeply) learnt models
- Underdetermined case

Thank you! Questions ?

Conclusion

- Single echo improves performance
- Enables universal dictionary
- First few echoes most important

Future Work

- Compare to BSS
- Include (deeply) learnt models
- Underdetermined case

Thank you! Questions ?

Conclusion

- Single echo improves performance
- Enables universal dictionary
- First few echoes most important

Future Work

- Compare to BSS
- Include (deeply) learnt models
- Underdetermined case

Thank you! Questions ?

Conclusion

- Single echo improves performance
- Enables universal dictionary
- First few echoes most important

Future Work

- Compare to BSS
- Include (deeply) learnt models
- Underdetermined case

Thank you! Questions ?