AUDIO LABS

An Iterative Multichannel Subspace-Based Covariance Subtraction Method for Relative Transfer Function Estimation

Reza Varzandeh, Maja Taseska and Emanuël Habets

International Audio Laboratories Erlangen

March 1, 2017

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

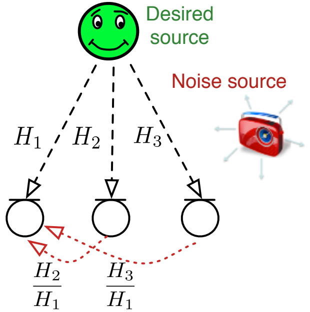
Fraunhofer IIS

Introduction and Motivation

Source extraction in noisy environments is ubiquitous in hands-free applications

To estimate the desired source we need to estimate the transfer functions H_m

To extract the desired source as received by the first microphones we only need to estimate H_m/H_1



- RTFs can be estimated from the data when the source is active
- We summarise state-of-the art estimators and propose an efficient iterative RTF estimator suitable for real-time applications

Content

Content

1. Signal Model and Source Extraction
2. Existing RTF Estimators
3. Proposed RTF Estimator
4. Performance Evaluation
5. Conclusions

Signal Model

- Desired speech and noise signals captured by M microphones
- STFT-domain signal at time n, frequency k:

$$\begin{aligned}\boldsymbol{y}(n,k) &= \boldsymbol{x}(n,k) + \boldsymbol{v}(n,k) \\ &= \boldsymbol{h}(n,k)\, S(n,k) + \boldsymbol{v}(n,k) \\ &= \boldsymbol{g}(n,k)\, X_1(n,k) + \boldsymbol{v}(n,k)\end{aligned}$$

- The RTF vector can be expressed in terms of the acoustic transfer functions $H_m(n,k)$:

$$\boldsymbol{g}(n,k) = \left[1,\ \frac{H_2(n,k)}{H_1(n,k)},\ \cdots,\ \frac{H_M(n,k)}{H_1(n,k)}\right]^{\mathrm{T}}$$

- The RTF vector is time-dependent to model source movements

Signal Model

- The power spectral density (PSD) matrices $\boldsymbol{\Phi}_{\boldsymbol{y}}$ and $\boldsymbol{\Phi}_{\boldsymbol{v}}$ are required for RTF estimation
- The PSD matrix of the received signal:

$$\boldsymbol{\Phi}_{\boldsymbol{y}}(n,k) = \boldsymbol{\Phi}_{\boldsymbol{x}}(n,k) + \boldsymbol{\Phi}_{\boldsymbol{v}}(n,k)$$

- The PSD matrix of the desired signal:

$$\boldsymbol{\Phi}_{\boldsymbol{x}}(n,k) = \phi_{x_1}(n,k)\, \boldsymbol{g}(n,k)\boldsymbol{g}^{\mathrm{H}}(n,k) \quad \text{with} \quad \phi_{x_1} = \mathrm{E}\left\{|X_1|^2\right\}$$

- The PSD matrix of the undesired signal, $\boldsymbol{\Phi}_{\boldsymbol{v}}$, can be estimated during speech absence, or using speech presence probability-controlled recursive averaging (Souden et al., 2011)

Source Extraction

- Estimate of the desired signal:

$$\begin{aligned}\widehat{X}_1(n,k) &= \boldsymbol{w}^{\mathrm{H}}(n,k)\,\boldsymbol{y}(n,k) \\ &= \boldsymbol{w}^{\mathrm{H}}(n,k)\left[\boldsymbol{g}(n,k)\,X_1(n,k) + \boldsymbol{v}(n,k)\right]\end{aligned}$$

- Distortionless response if $\boldsymbol{w}^{\mathrm{H}}\,\boldsymbol{g} = 1$
- Minimum Variance Distortionless Response (MVDR) filter:

$$\begin{aligned}\boldsymbol{w}(n,k) &= \arg\min_{\boldsymbol{w}}\ \boldsymbol{w}^{\mathrm{H}}\boldsymbol{\Phi}_{\boldsymbol{v}}(n,k)\boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{w}^{\mathrm{H}}\,\boldsymbol{g}(n,k) = 1 \\ &= \frac{\boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}(n,k)\,\boldsymbol{g}(n,k)}{\boldsymbol{g}(n,k)^{\mathrm{H}}\,\boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}(n,k)\,\boldsymbol{g}(n,k)}\end{aligned}$$

For real-time applications, the RTF vector needs to be efficiently estimated online using the microphone signals $\boldsymbol{y}(n,k)$

Content

1. Signal Model and Source Extraction
2. Existing RTF Estimators
3. Proposed RTF Estimator
4. Performance Evaluation
5. Conclusions

Existing RTF Estimators

Method 1: Covariance Subtraction

- Recall the definition:

$$\boldsymbol{\Phi}_{\boldsymbol{x}}(n,k) = \phi_{x_1}(n,k)\boldsymbol{g}(n,k)\boldsymbol{g}^{\mathrm{H}}(n,k)$$

- The RTF can be obtained by

$$\boldsymbol{g}_{\mathrm{CS}}(n,k) = \frac{\boldsymbol{\Phi}_{\boldsymbol{x}}(n,k)\,\boldsymbol{e}_1}{\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{\Phi}_{\boldsymbol{x}}(n,k)\,\boldsymbol{e}_1} \quad \text{with} \quad \boldsymbol{e}_1 = [1,0,\ldots,0]^{\mathrm{T}}$$

- In practice $\boldsymbol{\Phi}_{\boldsymbol{x}}$ can be estimated using $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{x}} = \widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}} - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}$

Existing RTF Estimators

Method 2: Covariance Subtraction with EVD

- The RTF vector $\boldsymbol{g}$ is proportional to the principal eigenvector of $\boldsymbol{\Phi}_{\boldsymbol{x}}$
- An estimate of the RTF vector is given by the principal eigenvector $\boldsymbol{u}_{\max}$ of $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{x}} = \widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}} - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}$

$$\boldsymbol{g}_{\mathrm{CS-EVD}}(n,k) = \frac{\boldsymbol{u}_{\max}(n,k)}{\boldsymbol{e}_1^{\mathrm{T}}\,\boldsymbol{u}_{\max}(n,k)}$$

- The principal eigenvector of $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}} - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}$ provides better performance in spatial filtering than the column of $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}} - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}$ (Serizel et al., 2014)

R. Serizel *et al.*, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants", IEEE/ACM Transactions on ASLP, 2014

Existing RTF Estimators

Method 3: Covariance Whitening

- A generalized eigenvalue problem:

$$\underbrace{\left(\phi_{x_1} \boldsymbol{g}\boldsymbol{g}^{\mathrm{H}} + \boldsymbol{\Phi}_{\boldsymbol{v}}\right)}_{\boldsymbol{\Phi}_{\boldsymbol{y}}} \boldsymbol{u} = \lambda \boldsymbol{\Phi}_{\boldsymbol{v}} \boldsymbol{u}$$

- **In theory:** Only one eigenvalue $\lambda \neq 1$

- **In practice:** Use the principal eigenvector $\boldsymbol{u}_{\max}$ of $\boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}\boldsymbol{\Phi}_{\boldsymbol{y}}$

$$\boldsymbol{g}_{\mathrm{CW}}(n,k) = \frac{\widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}(n,k)\, \boldsymbol{u}_{\max}(n,k)}{\boldsymbol{e}_1^{\mathrm{T}} \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}(n,k)\, \boldsymbol{u}_{\max}(n,k)}$$

Existing RTF Estimators

Method 4: Covariance Whitening using PM

- Use power method to estimate the GEVD (Krueger et al., 2011)
- **Iteration matrix:** $\boldsymbol{A}_{\text{cw}}(n,k) = \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}^{-1}(n,k)\widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}}(n,k)$
- **Power iteration:** $\widehat{\boldsymbol{u}}_{\max}(n,k) = \frac{\boldsymbol{A}_{\text{cw}}(n,k)\widehat{\boldsymbol{u}}_{\max}(n-1,k)}{\|\boldsymbol{A}_{\text{cw}}(n,k)\widehat{\boldsymbol{u}}_{\max}(n-1,k)\|}$
- Compute the RTF vector:

$$\boldsymbol{g}_{\text{PM}-\text{CW}}(n,k) = \frac{\widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}(n,k)\,\widehat{\boldsymbol{u}}_{\max}(n,k)}{\boldsymbol{e}_1^{\text{T}}\widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}(n,k)\,\widehat{\boldsymbol{u}}_{\max}(n,k)}$$

Krueger et al., "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation", IEEE Transactions on ASLP, 2011

Existing RTF Estimators
Summary

- Covariance-Subtraction: g_{CS}
  - Computationally efficient
- Covariance-Subtraction with EVD: $g_{\mathrm{CS-EVD}}$
  - More accurate than g_{CS} (Serizel et al., 2014)
  - Requires EVD
- Covariance-Whitening: g_{CW}
  - More accurate than g_{CS} (Markovich-Golan et al., 2015)
  - Requires GEVD
- Covariance-Whitening with PM: $g_{\mathrm{PM-CW}}$ (Krueger et al., 2011)

S. Markovich-Golan *et al.*, "Performance analysis of the CS method for relative transfer function estimation and comparison to the CW method", IEEE Transactions on ASLP, 2015

Content

Proposed RTF Estimator

- Computing $\boldsymbol{g}_{\mathrm{PM-CW}}$ is less complex than computing $\boldsymbol{g}_{\mathrm{CW}}$
- It still involves the inversion of $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}$ to compute $\boldsymbol{A}_{\mathrm{cw}}$, and multiplication by $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}$ to obtain $\boldsymbol{g}_{\mathrm{PM-CW}}$ from the eigenvector $\boldsymbol{u}_{\max}$
- We propose to estimate $\boldsymbol{g}_{\mathrm{CS-EVD}}$ using the power method
  - **Iteration matrix:** $\boldsymbol{A}_{\mathrm{cs}}(n,k) = \widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}}(n,k) - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}(n,k)$
  - **Power iteration:** $\widehat{\boldsymbol{u}}_{\max}(n,k) = \frac{\boldsymbol{A}_{\mathrm{cs}}(n,k)\widehat{\boldsymbol{u}}_{\max}(n-1,k)}{\|\boldsymbol{A}_{\mathrm{cs}}(n,k)\widehat{\boldsymbol{u}}_{\max}(n-1,k)\|}$

$$\boldsymbol{g}_{\mathrm{PM-CS}}(n,k) = \frac{\widehat{\boldsymbol{u}}_{\max}(n,k)}{\boldsymbol{e}_1^{\mathrm{T}}\,\widehat{\boldsymbol{u}}_{\max}(n,k)}$$

Content

Experimental Setup

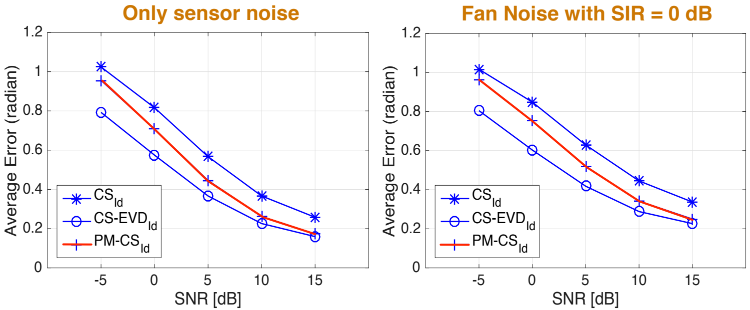
- Simulated room $4.5 \times 4 \times 3$ m^3, reverberation time $T_{60} = 0.3$ s
- Uniform 5-element linear array, inter-microphone distance 4 cm
- Microphone signals contain desired speech, directional interferer (fan noise), and sensor noise
  - signal-to-interference ratio (SIR): $\{0, \infty\}$ dB
  - signal-to-sensor noise ratios (SNRs): $[-5, 15]$ dB
- In all experiments, source-array distance was 1-1.2 m
- STFT frame-size is 128 ms, overlap 50%, sampling rate 16 kHz

Noise PSD matrix:

1. Estimated in advance during speech absence (denoted by ”Id”)
2. Estimated using speech presence probability-based framework

Results: Distance Measure

Hermitian Angle: $\Theta(n,k) = \arccos \frac{|\boldsymbol{g}^{\mathrm{H}}(n,k)\,\widehat{\boldsymbol{g}}(n,k)|}{\|\boldsymbol{g}(n,k)\|\,\|\widehat{\boldsymbol{g}}(n,k)\|}$

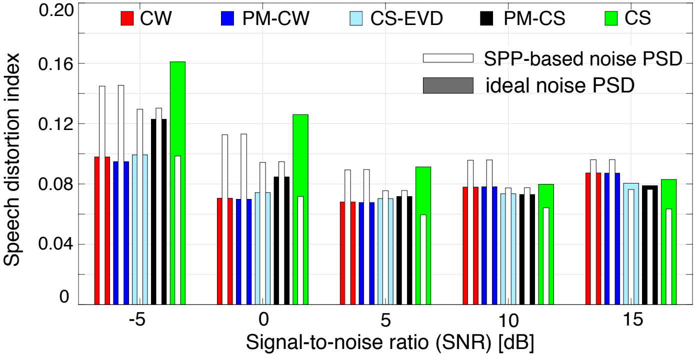


- Averaged $\Theta(n,k)$ over time segment of 15 s for all n and k
- CS-EVD outperforms CS and the error of the proposed PM-CS lies between the two methods

Results: Source Extraction Using MVDR

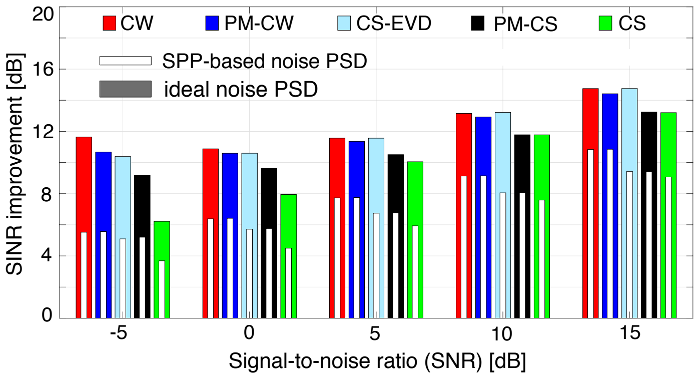
- MVDR filters using different RTF estimates
- Objective quality evaluation:
  - Speech distortion (SD) index
  - Signal to interference-plus-noise ratio (SINR) improvement compared to the reference microphone
- The measures are computed for non-overlapping 30 ms frames and are then averaged over all frames (15 seconds)

Speech distortion (fan noise with 0 dB SIR)



- Ideal noise PSD matrix: The proposed PM-CS causes similar or larger SD than the CS-EVD, but smaller than the CS
- Estimated noise PSD matrix: The distortion of PM-CS and CS-EVD is comparable
- Estimated noise PSD matrix: PM-CS causes lower SD than CW and PM-CW

SINR improvement (fan noise with 0 dB SIR)



- CS provides less SINR improvement than the alternatives which is consistent with (Markovich-Golan et al., 2015)
- Estimated noise PSD matrix: The proposed PM-CS has similar SINR improvement than CS-EVD

Content

1. Signal Model and Source Extraction
2. Existing RTF Estimators
3. Proposed RTF Estimator
4. Performance Evaluation
5. Conclusions

Conclusions

- Motivated by the advantage of $\boldsymbol{g}_{\mathrm{CS-EVD}}$ compared to $\boldsymbol{g}_{\mathrm{CS}}$, we proposed an iterative estimator to reduce the complexity
- Although the proposed PM-CS estimator has a greater computationally complexity than the CS estimator, it is less complex than the PM-CW estimator
- When the noise statistics are estimated, the performance of the proposed estimator is comparable to the CS-EVD estimator

AUDIO LABS

Thank you for your attention.