Estimation of Relative Transfer Function in the presence of stationary noise based on segmental Power Spectral Density matrix subtraction

Xiaofei Li[1], Laurent Girin[1,2], Radu Horaud[1], Sharon Gannot[3]

[1] INRIA Grenoble Rhône-Alpes [2] GIPSA-Lab & Univ. Grenoble Alpes [3] Bar-Ilan University

{xiaofei.li, radu.horaud}@inria.fr laurent.girin@gipsa-lab.grenoble-inp.fr sharon.gannot@biu.ac.il

The scenario

- Microphone array with an arbitrary topology.
- Single static desired speech source & (directional) stationary noise.

Problem to be solved

- Estimate the relative transfer function (RTF) of the desired source.

Proposed method

- Segmental power spectral density (PSD) matrix subtraction method.
- Reducing the stationary noise component and Preserving non-stationary speech component.

Applications

- Sound source localization.

Problem Formulation

- Received signals in the short time Fourier transform (STFT) domain:

$$\mathbf{x}(l,\omega) = \mathbf{h}_s(\omega)s_s(l,\omega) + \mathbf{h}_i(\omega)s_i(l,\omega).$$

- $l = 1,\ldots,L$, $\omega = 0,\ldots\Omega-1$ - index of frame and frequency, respectively.
- $\mathbf{x}(l,\omega)$ - M-channel microphone signals.
- $s_s(l,\omega)$ - STFT spectrum of the desired speech source.
- $s_i(l,\omega)$ - STFT spectrum of the noise source.
- $\mathbf{h}_s(\omega)$ - (time-invariant) M-channel acoustic transfer functions (ATFs) of the desired source.
- $\mathbf{h}_i(\omega)$ - (time-invariant) M-channel ATFs of the noise source.

Segmental PSD Matrix Subtraction

Spectral Segment

- Segment as the concatenation of successive frames:

$$\mathbf{X}_{l'}(\omega) = [\mathbf{x}((l'-1)R+1,\omega),\ldots,\mathbf{x}((l'-1)R+W,\omega)].$$

- W frames.
- R - segment increment.
- $l' = 1\ldots L'$ - segment index.

PSD Matrix of Segment

- PSD matrix:

$$\mathbf{\Phi}_{l'}(\omega) = \mathbf{X}_{l'}(\omega)\mathbf{X}_{l'}^H(\omega) \approx \mathbf{h}_s(\omega)\mathbf{h}_s^H(\omega)\Phi_{l'}^s(\omega) + \mathbf{h}_i(\omega)\mathbf{h}_i^H(\omega)\Phi_{l'}^i(\omega),$$

- where

$$\Phi_{l'}^s(\omega) = \sum\nolimits_{l=(l'-1)R+1}^{(l'-1)R+W} |s_s(l,\omega)|^2$$

is the power summation of the desired source signal in the l'-th segment.

- The fluctuations of $\Phi_{l'}^s(\omega)$ are large because of the non-stationarity and sparsity of speech signals.
- $\Phi_{l'}^i(\omega)$ - power summation of the noise signal.
- $\Phi_{l'}^i(\omega)$ is the smoothed power spectrum using W frames, and has a small variance due to $s_i(t)$ stationarity.

Segmental PSD Matrix Subtraction

- PSD Matrix Subtraction:

$$\begin{aligned}&\mathbf{\Phi}_{l'_1}(\omega) - \mathbf{\Phi}_{l'_2}(\omega)\\ &= \mathbf{h}_s(\omega)\mathbf{h}_s^H(\omega)(\Phi_{l'_1}^s(\omega) - \Phi_{l'_2}^s(\omega)) + \mathbf{h}_i(\omega)\mathbf{h}_i^H(\omega)(\Phi_{l'_1}^i(\omega) - \Phi_{l'_2}^i(\omega)).\end{aligned}$$

- $|\Phi_{l'_1}^i(\omega) - \Phi_{l'_2}^i(\omega)| \ll |\Phi_{l'_1}^s(\omega) - \Phi_{l'_2}^s(\omega)|$.
- The PSD difference matrix matches the matrix spanned by $\mathbf{h}_s(\omega)$ well.

Segment Classification

Large speech power spectrum difference $|\Phi_{l'_1}^s(\omega) - \Phi_{l'_2}^s(\omega)|$ is guaranteed by classifying segments into two classes $\mathbf{l}_1$ and $\mathbf{l}_2$ with high speech power and low speech power, respectively.

Power Spectrum Formulation

- The trace of the PSD matrix $\mathbf{\Phi}_{l'}(\omega)$:

$$\xi_{l'}(\omega) = \mathbf{h}_s^H(\omega)\mathbf{h}_s(\omega)\Phi_{l'}^s(\omega) + \mathbf{h}_i^H(\omega)\mathbf{h}_i(\omega)\Phi_{l'}^i(\omega),$$

- where, the power of the noise signal:

$$\xi_{l'}^i(\omega) = \mathbf{h}_i^H(\omega)\mathbf{h}_i(\omega)\sum\nolimits_{l=(l'-1)R+1}^{(l'-1)R+W} |s_i(l,\omega)|^2$$

obeys the Erlang distribution with the shape parameter W. Denote the cdf F.

Maximum and Minimum Statistics

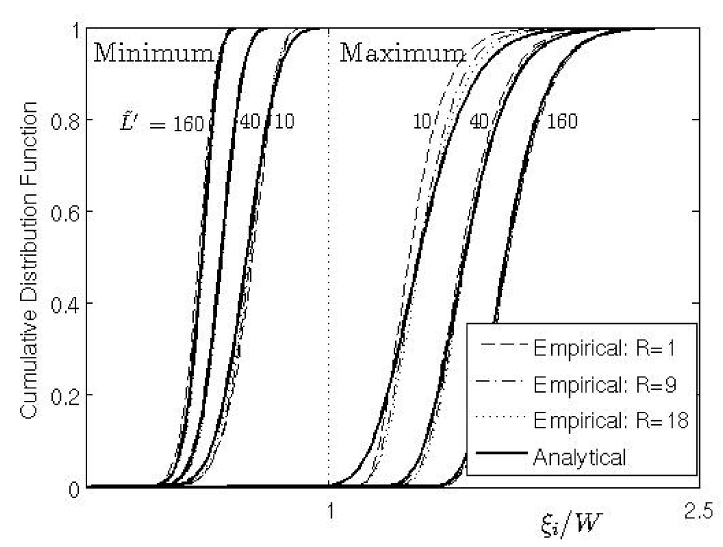
- Assuming adjacent segments are non-overlapping, the L' segments become independent and the pdfs of their minimum and maximum are:

$$f_{min}(\xi) = L'\cdot(1-F(\xi))^{L'-1}\cdot f(\xi), \quad f_{max}(\xi) = L'\cdot(F(\xi))^{L'-1}\cdot f(\xi).$$

- If overlap exists, $\xi_{l'}^i(\omega)$ becomes a correlated sequence.
- Simulations using a large dataset show that an approximate equivalent sequence length $\tilde{L}'$ is:

$$\tilde{L}' = \frac{L'R}{W}\cdot\left(1+\log\left(\frac{W}{R}\right)\right).$$

- Figure shows the cdf for $W = 18$, which demonstrates the applicability of the approximation.



Segment Classification

- Two classification threshold factors: maximum and minimum ratios

$$r_1 = \xi_{F_{max}(\xi)=0.95}/\bar{\xi}_{min}, \quad r_2 = \xi_{F_{max}(\xi)=0.5}/\bar{\xi}_{min}$$

- $F_{max}(\xi)$ - cdf of the maximum; and $\bar{\xi}_{min}$ - expectation of the minimum.
- Classification into two classes:

$$\mathbf{l}_1 = \{l' \mid \xi_{l'}(\omega) > r_1\cdot\min\{\xi_{l'}(\omega)\}\}, \quad \mathbf{l}_2 = \{l' \mid \xi_{l'}(\omega) < r_2\cdot\min\{\xi_{l'}(\omega)\}\}.$$

RTF Estimation

- The *global noise-free PSD matrix*: $\hat{\mathbf{\Phi}}(\omega) = \sum_{l'_1,l'_2\in\mathbf{l}_1}(\mathbf{\Phi}_{l'_1}(\omega) - \mathbf{\Phi}_{l'_2}(\omega))$.
- The principal eigenvector of $\hat{\mathbf{\Phi}}(\omega)$ is an unit-norm estimation of the RTF vector corresponding to $\mathbf{h}_s(\omega)$.

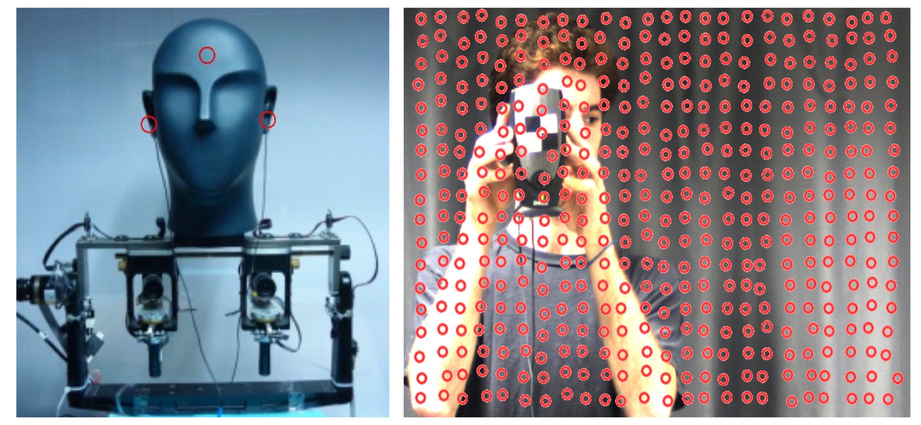
Experiments: Application to Sound Source Localization

Sound Source Localization Principle

- Supervised "look-up table" approach.
- Feature vector: $\mathbf{h} = [\mathbf{h}^T(0),\ldots,\mathbf{h}^T(\Omega-1)]^T$.
- A pre-trained dictionary $\{\mathbf{h}_k,\mathbf{p}_k\}, k = 1,\ldots,K$ of pairs of feature vectors and source directions, for a given room and a given microphone constellation.
- In the test stage select the best fitting feature vector:

$$\hat{\mathbf{p}} = \mathbf{p}_{k_0} \quad \text{with} \quad k_0 = \underset{k\in[1,K]}{\operatorname{argmin}} \parallel \tilde{\mathbf{h}} - \mathbf{h}_k \parallel.$$

The Dataset: Audio-Visual Alignment



- Four microphones: left/right and front/back.
- Lookup table: 1s white-noise signal is emitted from 24×18 (azimuth and elevation) directions.
- Test data: 108 speech signals are emitted from 108 directions.
- Directional noise: White Gaussian noise (WGN) and babble noise are separately emitted from different directions outside the camera field-of-view.

Results

- Comparison methods: non-stationarity (NS) method [Gannot et al., 2001], speech presence probability (SPP) method [Cohen, 2004].
- Performance metric: average absolute localization error (in degrees).

	WGN			babble		
SNR(dB)	NS	SPP	Prop.	NS	SPP	Prop.
10	1.51	1.35	1.21	1.47	1.31	1.24
5	1.58	1.34	1.27	1.77	1.58	1.56
0	2.14	1.65	1.30	2.40	2.55	2.47
-5	4.61	2.79	1.77	-	-	-
-10	9.20	6.64	2.62	-	-	-

- Achieving the best performance for WGN.
- Advantages: 1) only the segments containing speech are selected; 2) the noise PSD matrix is accurately subtracted; 3) the eigenvalue decomposition is an optimization criterion that considers all channels simultaneously.
- Performance degrade for babble noise due to its non-stationarity.

Summary

- A RTF identification method based on segmental PSD matrix subtraction.
- A classification between speech and noise segments and noise-only-segment based on maximum and minimum statistics.
- Outperforms commonly used methods when the noise is stationary.
- Can be extended to the case of multiple speakers (future work).