Local Relative Transfer Function for Sound Source Localization

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

[1]INRIA Grenoble Rhône-Alpes. {*firstname.lastname@inria.fr*}
[2]GIPSA-Lab & Univ. Grenoble Alpes
[3]Faculty of Engineering, Bar-Ilan University

September 1, 2015

Outline

Introduction

Task & The scenario

- Sound source localization.
- Microphone array with an arbitrary topology.
- Single static desired speech source.

Baseline method & Challenge

- Relative transfer function (RTF): as a funtion of direction of arrival.
- Challenge: It is hard to select a good reference channel in a complex acoustic environment.

Proposed method

- To avoid a potential bad unique reference channel, we propose
    - **local RTF** that takes local reference channel.
    - a biased local-RTF estimator and a unbiased estimator.

Problem formulation

In the STFT domain, the signals received by the M microphones are approximated as:

$$\mathbf{x}(\omega, l) \approx \mathbf{h}(\omega)s(\omega, l) + \mathbf{n}(\omega, l)$$

- ω and l are the indices of frequency-bin and time-frame.
- $s(\omega, l)$ is the source signal.
- $\mathbf{x}(\omega, l) = [x_1(\omega, l), \ldots, x_M(\omega, l)]^T$ is the sensor signal vector.
- $\mathbf{n}(\omega, l) = [n_1(\omega, l), \ldots, n_M(\omega, l)]^T$ is the sensor noise vector.
- $\mathbf{h}(\omega) = [h_1(\omega), \ldots, h_M(\omega)]^T$ is the acoustic transfer function (ATF) vector.

Relative transfer function

RTF Definition

ATF ratio $r_m(\omega) = \frac{h_m(\omega)}{h_1(\omega)}$, where the first channel is taken as the reference.

RTF Estimation

1. The cross-spectral method: $\hat{r}_m(\omega) = \frac{\hat{\Phi}_{x_m x_1}(\omega)}{\hat{\Phi}_{x_1 x_1}(\omega)}$.

   $\hat{\Phi}_{x_m x_1}(\omega)$ and $\hat{\Phi}_{x_1 x_1}(\omega)$ are the cross and auto-PSD of sensor signals.
2. An unbiased estimator based on the nonstationarity of speech [Gannot01][1].

In [Gannot01], it is proved that the RTF estimation error are *inversely proportional* to the SNR at the reference channel.

[1] S. Gannot, et al. "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Proc.*, vol. 49, no. 8, pp. 1614-1626, 2001.

Local relative transfer function: Definition 1

- We should select the channel with the highest SNR as the reference. However, it is hard to precisely estimate the SNR at each channel in a complex environment.

- As an alternative solution, we define *local-RTF*

$$a_m(\omega) = \frac{|h_m(\omega)|}{\| \mathbf{h}(\omega) \|} e^{j(\arg[h_m(\omega)] - \arg[h_{m-1}(\omega)])}$$

  where $\arg[\cdot]$ is the phase of complex number, $\| \cdot \|$ is the l_2-norm.
  - **Local phase difference** & **Normalized level**.
  - Avoid a potential bad global reference channel.

Local relative transfer function: Definition 2

The corresponding *local-RTF* vector is $\mathbf{a}(\omega) = [a_1(\omega), \ldots, a_M(\omega)]^T$.

- It is NOT an actual transfer function vector that can be directly used for beamforming.
- It is rather a robust feature expected to be appropriate for sound source localization due to its lower sensitivity to noise (compared to regular RTF vector).

Local relative transfer function: Biased estimator

The local-RTF of the m-th channel can be estimated by the *cross-spectral method*:

$$\hat{a}_m(\omega) = \frac{\sqrt{\hat{\Phi}_{x_m x_m}(\omega)}}{\sqrt{\sum_{m=1}^{M} \hat{\Phi}_{x_m x_m}(\omega)}} e^{j \arg[\hat{\Phi}_{x_m x_{m-1}}(\omega)]}.$$

- This estimator is biased, and in high SNR the bias is small.
- It is suitable for high SNR scenarios, due to the bias and low computational load.

Inspired by [Cohen04][2], we propose an unbiased local-RTF estimator.

[Cohen04] provides:

- $\hat{\rho}_m(\omega)$: an unbiased estimation of the ATF ratio $\rho_m(\omega) = \frac{h_m(\omega)}{h_{m-1}(\omega)}$.
- $\hat{\Phi}_{s_m s_m}(\omega, l)$: a PSD estimation of the image source $h_m(\omega)s(\omega, l)$.
- $\hat{\Phi}_{s_m s_m}(\omega) = \frac{1}{L}\sum_{l=1}^{L} \hat{\Phi}_{s_m s_m}(\omega, l)$: the frame-averaged power of the image source signal over frames.

[2] I. Cohen. "Relative transfer function identification using speech signals," *IEEE Trans. Speech and Audio Proc.*, vol. 12, no. 5, pp. 451-459, 2004.

Local relative transfer function: Unbiased estimator (2)

Based on $\hat{\rho}_m(\omega)$ and $\hat{\Phi}_{s_m s_m}(\omega)$, the local-RTF is estimated as

$$\hat{a}_m(\omega) = \frac{\sqrt{\hat{\Phi}_{s_m s_m}(\omega)}}{\sqrt{\sum_{m=1}^{M} \hat{\Phi}_{s_m s_m}(\omega)}} e^{j\arg[\hat{\rho}_m(\omega)]}$$

- The estimation error of this estimator depends on the estimate accuracy of $\hat{\rho}_m(\omega)$ and $\hat{\Phi}_{s_m s_m}(\omega)$. The detailed analysis can be found in [Cohen04].
- This unbiased estimator is more suitable for low SNRs.

Sound source localization using local-RTF vector

- Concatenate the local-RTF vectors across frequencies: $\hat{\mathbf{a}} = [\hat{\mathbf{a}}^T(0), \ldots, \hat{\mathbf{a}}^T(\omega), \ldots, \hat{\mathbf{a}}^T(\Omega - 1)]^T$.
- Lookup table dataset: $\{\mathbf{a}_k, \mathbf{d}_k\}_{k=1}^K$. $\mathbf{a}_k$ and $\mathbf{d}_k$ denote the feature vector and source direction.
- Localization method
  - Lookup: find the I best directions $\{\mathbf{a}_{k_i}, \mathbf{d}_{k_i}\}_{i=1}^I$.
  - Interpolation: weighted mean

$$\hat{\mathbf{d}} = \frac{\sum_{i=1}^{I} \| \hat{\mathbf{a}} - \mathbf{a}_{k_i} \|^{-1} \mathbf{d}_{k_i}}{\sum_{i=1}^{I} \| \hat{\mathbf{a}} - \mathbf{a}_{k_i} \|^{-1}}$$

  where the reciprocal of the feature difference $\| \hat{\mathbf{a}} - \mathbf{a}_{k_i} \|^{-1}$ is taken as the weight.

Experiments: Audio-visual data set

- Audio-visual data set.
- **Lookup table**: 432 source directions in the camera field-of-view.
- **Test data**: the speech signal is emited from other 108 directions in the camera field-of-view.

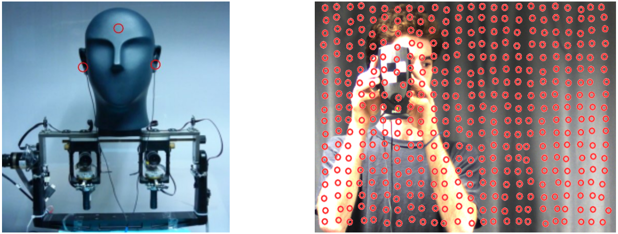

Figure: (left) Dummy head with four microphones (red circles) and cameras. (right) The lookup source directions.

Experiments: Noise and comparison method

Two types of **noise** are added into the test data with various SNRs.

- **Environmental noise** is recorded in a noisy office environment, includes people movements, devices, outside environment (passing cars, street noise), etc.
- **Directional WGN** is emitted by a loudspeaker with a direction beyond the camera field-of-view in the same noisy office.

Comparison method (Regular RTF): RTF with a unique reference derived from [Cohen04], using the reference channel with the highest input SNR[3].

[3] Note that the input SNR is computed using the estimated noise and speech power provided by [Cohen04].

Experiments: Results for environmental noise

Localization errors[4] for Biased estimator (Local-RTF 1), Unbiased estimator (Local-RTF 2) and the comparison method (Regular RTF). The bold values are the minimum error at each SNR.

SNR	Local-RTF 1		Local-RTF 2		Regular RTF	
(dB)	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.
10	**0.83**	0.51	0.85	**0.47**	0.96	0.76
5	**0.83**	0.56	0.86	**0.47**	0.95	0.82
0	**0.85**	0.62	0.89	**0.46**	1.02	0.74
-5	**1.00**	0.76	1.02	**0.51**	1.20	1.05
-10	1.53	1.22	**1.51**	**0.75**	1.79	1.30

- **Local-RTF 1 vs 2**: The biased estimator has comparable performance with the unbiased estimator in high SNRs, however larger elevation error in low SNRs.
- **Local-RTF 2 vs Regular RTF**: Regular RTF perform worse than the proposed, due to its imprecise input SNR estimation.

[4] The absolute angle error (in degrees) in azimuth (Azi.) and elevation (Ele.).

Experiments: Results for directional WGN

SNR (dB)	Local-RTF 1 Azi.	Local-RTF 1 Ele.	Local-RTF 2 Azi.	Local-RTF 2 Ele.	Regular RTF Azi.	Regular RTF Ele.
10	**0.80**	**0.49**	0.82	**0.49**	0.80	0.87
5	1.24	0.65	**0.80**	**0.54**	0.87	0.80
0	3.39	1.31	**0.91**	**0.56**	1.11	0.64
-5	8.33	2.74	1.40	0.77	**1.31**	**0.75**
-10	11.2	3.87	3.82	1.48	**1.64**	**1.00**

- **Local-RTF 1 vs 2**: Compared to the unbiased estimator, the biased estimator performs better slightly for 10 dB SNR, however deteriorates abruptly with the decreasing of SNR. Because the directional noise brings a larger estimation bias.
- **Local-RTF 2 vs Regular RTF**: Regular RTF performs better when the SNR is low (-5, -10 dB). This indicates that the highest SNR channel are correctly selected in Regular RTF. Because
  1. the noise directivity induces a large noise power difference among channels for low SNRs.
  2. the noise signal is relatively stationary.

Conclusions

- Local-RTF and two estimators are proposed.
- Experiments show that local-RTF is more robust than the regular RTF when the noise power cannot be precisely estimated.

Thank you very much!

Q & A

xiaofei.li@inria.fr