Reverberant Sound Localization with a Robot Head Based on Direct-Path Relative Transfer Function

Xiaofei Li, Laurent Girin, Fabien Badeig, Radu Horaud

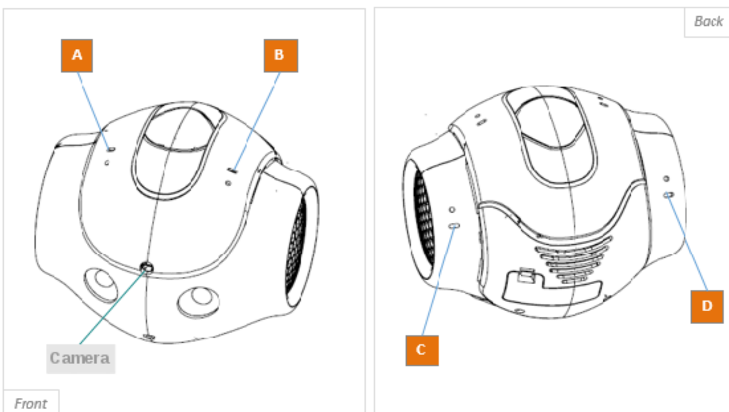
PERCEPTION Team, INRIA Grenoble Rhone-Alpes

October 12th, 2016

Sound Localization with a Robot Head

Considered Scenario

- Humanoid robot NAO (version 5)
- Speaker direction relative to the robot should be estimated



Microphone array (NAO robot)

Sound localization scene

Sound Localization with a Robot Head

Challenges

- Room reverberation
- Robot ego-noise and ambient noise

Proposed method

- Estimation of the Direct-Path Relative Transfer Function (DP-RTF)
- Sound source localization (DoA) calculated from DP-RTF
- Robustness towards noise increased by Spectral Subtraction

Microphone Signals

- Two-channel microphone signal:

$$x(n)=a(n)*s(n), \quad y(n)=b(n)*s(n)$$

  - $x(n)$, $y(n)$: microphone signals
  - $s(n)$: source signal
  - $a(b)$, $b(n)$: room impulse response including direct-path sound propagation *and* reflections.

    (The direct-path propagation indicates the sound direction.)

- Apply STFT to obtain the Convolutive Transfer Function (CTF):

$$x_{p,k} = a_{p,k} * s_{p,k}, \quad y_{p,k} = b_{p,k} * s_{p,k}$$

  - p, k: frame and frequency indices

Convolutive Transfer Function (CTF)

- **Problem:** Assumption of multiplicative transfer function

$$x_{p,k} = s_{p,k} a_k$$

not fulfilled if DFT size lower than room impulse response (RIR) length

- CTF needed in such cases given by the convolution

$$x_{p,k} = \sum_{p'=0}^{Q_k - 1} s_{p-p',k} a_{p',k} = s_{p,k} * a_{p,k},$$

  - Q_k depends the length of the RIR

Direct-Path Relative Transfer Function

- CTF $a_{p,k}$, with frame index $p=0,\ldots,Q-1$ is composed of
  - $a_{0,k}$: direct-path transfer function (at frame instance 0)
  - $a_{p,k}$,(unwanted) reverberation at frame instances $p=1,\ldots,Q-1$

- Direct-Path Relative Transfer Function (DP-RTF)
  - given by the ratio $\frac{b_{0,k}}{a_{0,k}}$
  - contains information about the source direction (by the phase difference for numerator and denominator)
  - robust to reverberation (since late reverberant part excluded)

DP-RTF Estimation

- Estimation from noise-free microphone signals
  - Two channel convolutive relation:

$$x_{p,k} * b_{p,k} = y_{p,k} * a_{p,k}$$

  - Division by $a_{0,k}$ and rearranging the terms leads to a set of linear equation:

$$y_{p,k} = \mathbf{z}_{p,k}' \, \mathbf{g}_k$$

with $\mathbf{z}_{p,k} = [x_{p,k}, \ldots, x_{p-Q+1,k}, y_{p-1,k}, \ldots, y_{p-Q+1,k}]'$,

$$\mathbf{g}_k = [b_{0,k}/a_{0,k}, \ldots, b_{Q-1,k}/a_{0,k}, -a_{1,k}/a_{0,k}, \ldots -a_{Q-1,k}/a_{0,k}]'.$$

  - Taking the expectation leads to an expression in terms of the cross- and auto power spectral density (PSD):

$$\varphi_{yy}(p,k) = \varphi_{zy}(p,k)' \, \mathbf{g}_k$$

  - At frequency k, DP-RTF $\frac{b_{0,k}}{a_{0,k}}$ is estimated by solving an overdetermined set of linear equations

Noisy Recordings

- DP-RTF estimation in the presence of noise
  - Noisy signal microphone signal:

    $\hat{y}(n) = y(n) + v(n)$,
  - Source and noise signal are (assumed to be) uncorrelated.
  - PSD of noisy signal $\varphi_{\hat{y}\hat{y}}(p,k) = \varphi_{yy}(p,k)+\varphi_{vv}(p,k)$.
  - Clean PSDs can be obtained by Spectral Subtraction

$$\hat{\phi}_{yy}(p,k) \approx \hat{\phi}_{\tilde{y}\tilde{y}}(p,k) - \phi_{vv}(p,k)$$
$$\hat{\varphi}_{zy}(p,k) \approx \hat{\varphi}_{\tilde{z}\tilde{y}}(p,k) - \varphi_{wv}(p,k)$$

  - Estimation of noise PSDs $\phi_{vv}(p,k)$ and $\varphi_{wv}(p,k)$ easily obtained for stationary noise

Calculation of Sound Source Location

- DP-RTF feature vector **c**:
  - concatenates DP-RTFs across microphone pairs and frequencies.

- Calculation of sound direction **d**
  - Probablistic piecewise-linear regression $\mathbf{d} = f(\mathbf{c})$ [Deleforge et al., IEEE Trans. 2015].
  - The regression model f is learned from training data (feature-direction pairs) $\{\mathbf{c}_i,\mathbf{d}_i\}_{i=1,\ldots,I}$.

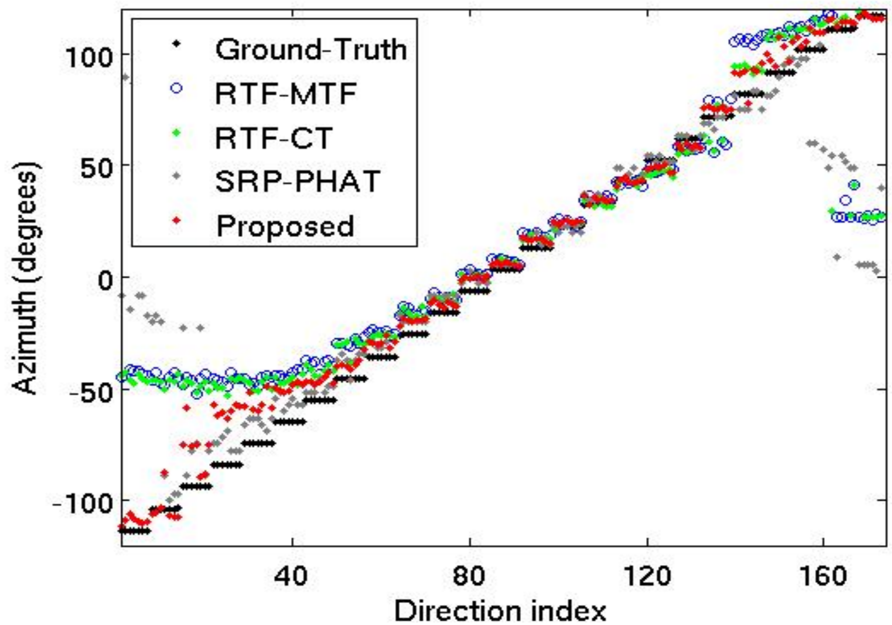
Experiments with the NAO Robot

- Experimental environments
  - Cafeteria, office, laboratory, and meeting room.
  - Reverberation time T_{60}: 0.24s, 0.47s, 0.52s, and 1.04s.
- Noise signals
  - Mainly the stationary fan-noise of robot head.
  - The signal-to-noise-ratio (SNR) is about 5 dB.
- Related methods
  - MTF-based RTF estimator (RTF-MTF) [Li et al., ICASSP 2015].
  - Coherence test (RTF-CT) [MOHAN et al., IEEE Trans. 2008].
  - SRP-PHAT [Do et al., ICASSP 2007].

Experiments with the NAO Robot

Results for laboratory room

- Azimuth angle from -120º to 120º (T60 of approx. 0.5s)



- Proposed method shows the best results
  - Related methods fail especially for large azimuths that are closer to the wall due to the strong reflections

Experiments with the NAO Robot

- Audio-visual: localize speaker position in the camera image
  - Metric: average absolute localization error in degrees
  - Azimuth (Azi.) and elevation (Ele.)

	Cafeteria		Office		Laboratory		Meeting Room	
	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.
RTF-MTF	0.45	1.57	0.62	2.14	1.44	2.31	1.87	3.66
RTF-CT	**0.44**	1.50	0.64	2.25	1.61	2.36	1.77	3.44
SRP-PHAT	0.77	1.95	1.03	2.80	1.41	3.33	2.04	3.52
Proposed	0.47	**1.47**	**0.55**	**1.87**	**0.82**	**1.84**	**0.95**	**2.12**

- The proposed localization method performs better, especially for high reverberation time.
- Azimuth results are better than elevation results since the coplanar microphone array has a low elevation resolution.

Conclusions

- A direct-path RTF estimator for sound source localization
- Robust to reverberation and noise.
- More details are available in the extended paper:

  X. Li et al., Estimation of the direct-path RTF for supervised sound-source localization, IEEE/ACM Trans. ASLP, 2016.

- In future studies, the extension to the multiple-speaker case could be investigated.