How to locate a sound source in a reverberant room?

Xiaofei Li and Radu Horaud

References

[Li2016a] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the Direct-Path Relative Transfer Function for Supervised Sound-Source Localization," IEEE Trans. on Audio, Speech and Lang. Proc., 24(11), Nov 2016 [link]

[Li2016b] X. Li, L. Girin, F. Badeig, and R. Horaud, "Reverberant Sound Localization with a Robot Head Based on Direct-Path Relative Transfer Function," IEEE/RSJ International Conference on Intelligent Robots and Systems, Oct 2016 [link]

[Li2015] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of Relative Transfer Function in the Presence of Stationary Noise Based on Segmental Power Spectral Density Matrix Subtraction," IEEE ICASSP 2015 [link]

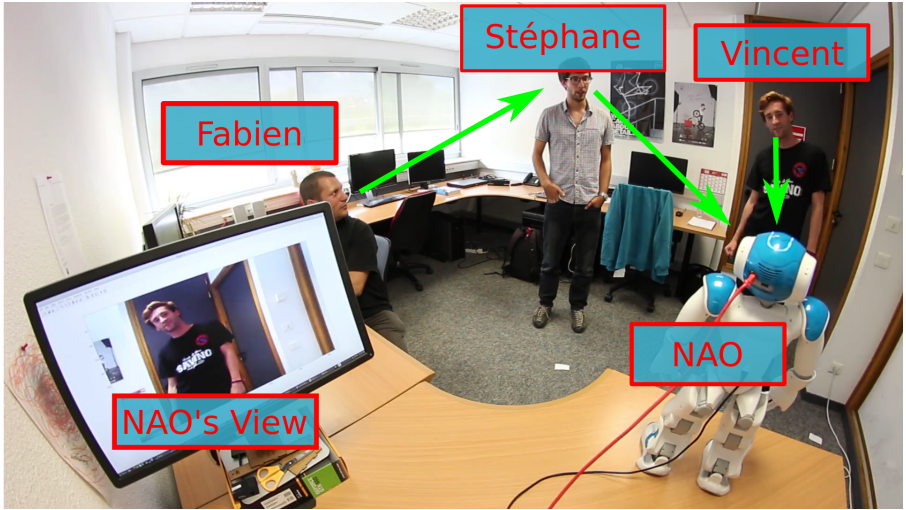
Dyadic Communication

Hand-held device (user controlled)

Table-top device (hands free)

Formal Meetings

Multi-Party Human-Human and Human-Robot Interaction



Dyadic communication, formal meetings and multi-party interaction

- Dyadic communication: source and microphone signals are quite similar, the user is in the loop.
- Formal meetings: the setup is quasi static over time.
- Human-to-robot and human-to-human interactions : complex and time-varying acoustic setups that are challenging the current SOA in sound-source localization.

Outline

- The short-time Fourier transform
- Audio-signal and acoustic models
- Dealing with room reverberation
- Estimating a sound's direct path

The short-time Fourier transform (STFT)

- Microphone signal, $x(t)$ and a window function $w(t)$, $t \in \mathbb{R}$
- Continuous-time STFT ($f \in \mathbb{R}$ is the frequency):

$$X(f, \tau) = \int_{t=-\infty}^{t=+\infty} x(t) w(t - \tau) \exp(-2\pi i f t) dt \in \mathbb{C}$$

- STFT coefficient at frame $p \in \mathbb{N}$, frequency $k \in \mathbb{N}$, window size N samples:

$$X_{p,k} = \frac{1}{N} \sum_{n=1}^{n=N} x(n) w(n - p) \exp(-2\pi j k n / N) \in \mathbb{C}$$

- The window is shifted along the signal with a discrete step of size L
- Speech signals are highly non stationary, hence N should be small to guarantee stationarity within a frame.
- For speech signals the size of w (or N) is 20 ms

Power spectral density (PSD)

- The PSD over P frames is defined by (where X^* is the complex conjugate of X):

$$\phi_{xx}(k) = \frac{1}{P} \sum_{p=1}^{P} X_{p,k} X_{p,k}^*$$

- The cross-PSD for two signals, x and y, is:

$$\phi_{xy}(k) = \frac{1}{P} \sum_{p=1}^{P} X_{p,k} Y_{p,k}^*$$

Acoustic model (I)

- Temporal representation:

$$x(t) = \underbrace{a(t) \star s(t)}_{\text{convolution}}$$

- where $s(t)$ is the speech source and $a(t)$ is the room impulse response (RIR)
- in a non-reverberant room, the size[1] of $a(t)$ is smaller than the size[2] of the STFT window, hence:

$$X_{p,k} = A_{p,k} S_{p,k} \in \mathbb{C}$$

- This is the *multiplicative* model, widely used in the audio signal processing literature.

[1] the temporal duration measured in number of samples

[2] the number of samples in a STFT frame

Binaural sound source localization in a non-reverberant room

- For a microphone pair:

$$X_{p,k} = A_{p,k} S_{p,k}$$
$$Y_{p,k} = B_{p,k} S_{p,k}$$

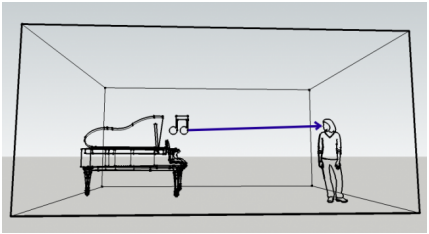
- The relative transfer function (RTF) contains localization information:

$$\frac{B_{p,k}}{A_{p,k}} = \frac{Y_{p,k}}{X_{p,k}} = \underbrace{\frac{|Y_{p,k}|}{|X_{p,k}|}}_{\text{ILD}} \exp(2\pi j k(\underbrace{\tau_2 - \tau_1}_{\text{ITD}}))$$

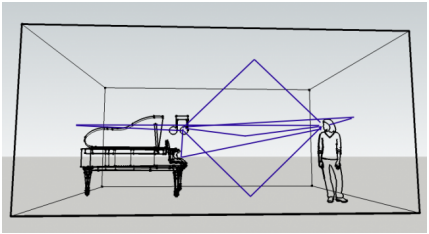
- $\text{ITD} = \tau_2 - \tau_1$: Interaural time difference between the two microphones
- ILD: Interaural level difference between the two microphones
- ITD and ILD are referred to as *binaural features* and contain sound direction information

What does the impulse response contain?

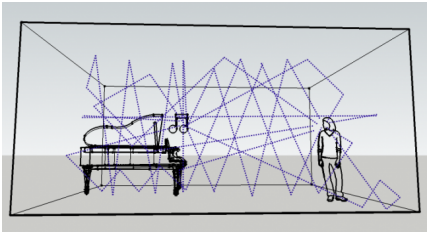
Direct path



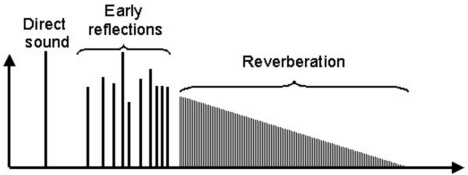
Early reflections



Reverberation



Room impulse response



Acoustic model (II)

- The multiplicative model is valid if the STFT frame length is long relative to the RIR length.
- But in practice:
  - STFT frame length: 20 ms (remember that speech is non-stationary)
  - RIR length : $>$ 100 ms (larger the room, longer the RIR)

Convolutive Model

- Convolution in the Fourier domain with $L \approx N$:

$$\begin{aligned} X_{p,k} &= A_{p,k} \star S_{p,k} \\ &= \sum_{q=0}^{q=Q-1} A_{q,k} S_{p-q,k} \\ &= \underbrace{A_{0,k} S_{p,k}}_{\text{direct path}} + \cdots + \underbrace{A_{q,k} S_{p-q,k} + \cdots + A_{Q-1,k} S_{p-Q+1,k}}_{\text{reflections}} \end{aligned}$$

- $A_{p,k}$ contains source-location information that was emitted at different "times", namely: p, $p-1$, $\ldots$, $p-(Q-1)$, along increasingly longer paths.
- Larger the room, better the separation between direct path and reflections.

Direct-Path Relative Transfer Function (DP-RTF)

- Convolutive model with two microphones:

$$X_{p,k} = A_{p,k} \star S_{p,k}$$
$$Y_{p,k} = B_{p,k} \star S_{p,k}$$

- which develops as:

$$X_{p,k} = A_{0,k}S_{p,k} + \cdots + A_{q,k}S_{p-q,k} + \cdots + A_{Q-1,k}S_{p-Q+1,k}$$
$$Y_{p,k} = B_{0,k}S_{p,k} + \cdots + B_{q,k}S_{p-q,k} + \cdots + B_{Q-1,k}S_{p-Q+1,k}$$

- If the time delay, between the direct-path wave and the first notable reflection, is large compared to N, the DP-RTF is given by:

$$G_{0,k} = \frac{B_{0,k}}{A_{0,k}} \quad \text{(denoted } d_k \text{ in eq. (8) of [Li2016a])}$$

Estimating the DP-RTF

- Convolutive model: $X_{p,k} = A_{p,k} \star S_{p,k}$ and $Y_{p,k} = B_{p,k} \star S_{p,k}$
- Convolve $A_{p,k}$ with the 2nd equation and use commutativity of convolution:

$$A_{p,k} \star Y_{p,k} = A_{p,k} \star B_{p,k} \star S_{p,k} = B_{p,k} \star X_{p,k}$$

- In vector form this writes as:

$$\boldsymbol{A}_{p,k}^{\top} \boldsymbol{Y}_{p,k} = \boldsymbol{B}_{p,k}^{\top} \boldsymbol{X}_{p,k}$$

- Divide with $A_{0,k}$ to obtain:

$$\begin{aligned} Y_{p,k} &= \boldsymbol{Z}_{p,k}^{\top} \boldsymbol{G}_k, \quad \text{with:} \\ \boldsymbol{G}_k^{\top} &= \left(\frac{B_{0,k}}{A_{0,k}}, \ldots, \frac{B_{Q-1,k}}{A_{0,k}}, -\frac{A_{k,1}}{A_{0,k}}, \ldots, -\frac{A_{Q-1,k}}{A_{0,k}} \right) \\ \boldsymbol{Z}_{p,k}^{\top} &= (X_{p,k}, \ldots, X_{p-Q+1,k}, Y_{p-1,k}, \ldots, Y_{p-Q+1,k}) \end{aligned}$$

DP-RTF estimation: the noise-free case

- By multiplying both sides of $Y_{p,k} = \boldsymbol{Z}_{p,k}^{\top}\boldsymbol{G}_k$ with $Y_{p,k}^*$ and taking the expectation, we obtain eq. (12) in [Li2016a]:

$$\phi_{yy}(p,k) = \phi_{zy}^{\top}(p,k)\boldsymbol{G}_k, \text{ with:}$$

$$\phi_{yy}(p,k) = E[Y_{p,k}Y_{p,k}^*]$$
$$\phi_{zy}(p,k) = \big(E[X_{p,k}Y_{p,k}^*], \ldots, E[X_{p-Q+1,k}Y_{p,k}^*], E[Y_{p-1,k}Y_{p,k}^*], \ldots, E[Y_{p-Q+1,k}Y_{p,k}^*]\big)$$

- Auto- and cross-PSD estimates, $\hat{\phi}_{yy}(p,k)$ and $\hat{\phi}_{zy}(p,k)$, are obtained by computing the expectations over the past D frames: $p, p-1, \ldots, p-D+1$.
- Using several consecutive frames one obtains a set of linear equations for each k:

$$\hat{\boldsymbol{\phi}}_{yy}(k) = \hat{\boldsymbol{\Phi}}_{zy}(k)\boldsymbol{G}_k$$

DP-RTF estimation: the noisy case

- The microphone signals are corrupted by additive noise:

$$X_{p,k} = A_{p,k} \star S_{p,k} + U_{p,k}$$
$$Y_{p,k} = B_{p,k} \star S_{p,k} + V_{p,k}$$

- Both noise signals are assumed to be wide-sense stationary (WSS) and uncorrelated with the speech signal.
- Inter-frame spectral subtraction can be used between frames with high-speech and low-speech power, respectively. Please consult [Li2015] and Section IV of [Li2016a].
- By interchanging the two microphones, we obtain two estimates of the DP-RTF $\frac{B_{0,k}}{A_{0,k}}$ and $\frac{A_{0,k}}{B_{0,k}}$, namely $\hat{\boldsymbol{G}}_{0,k}$ and $\hat{\boldsymbol{G}}'_{0,k}$. Finally the DP-RTF is estimated with:

$$\hat{C}_k = 0.5(\hat{\boldsymbol{G}}_{0,k} + \hat{\boldsymbol{G}}'^{-1}_{0,k})$$

DP-RTF and HRTF

- When the microphones are embedded in a "dummy" head, e.g. a robot, the DP-RTF corresponds to the head-related transfer function (HRTF) associated with a binaural microphone pair.
- Otherwise said, in a noise-free and anechoic room, the DP-RTF is equivalent with the HRTF
- There is no analytical formula between the HRTF/DP-RTF and the direction of arrival (DOA) of a sound source.
- Regression methods can be used to learn this high-dimensional to low-dimensional mapping

Supervised Sound-source localization

- Training dataset of pairs $D = \{\hat{\boldsymbol{C}}_i, \boldsymbol{q}_i\}_{i=1}^{I}$, DP-RTF $\hat{\boldsymbol{C}}$ and DOA $\boldsymbol{q}$:

$$\hat{\boldsymbol{C}}_i = \begin{pmatrix} \hat{C}_{1,i} \\ \vdots \\ \hat{C}_{k,i} \\ \vdots \\ \hat{C}_{K,i} \end{pmatrix} \in \mathbb{C}^K \iff \boldsymbol{q}_i = \underbrace{\begin{pmatrix} x_i \\ y_i \end{pmatrix} \in \mathbb{R}^2}_{\text{DOA}}$$

- Predicting a DOA from a DP-RTF:

$$\hat{\boldsymbol{C}} \Longrightarrow \boldsymbol{q} \tag{1}$$

Example of training/testing

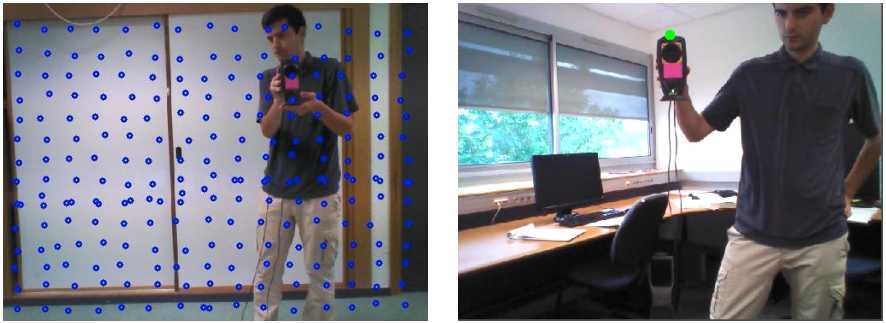


Figure: Left: an *audio-visual* training dataset contains sound sources emitted by a loud-speaker that correspond to sound directions, i.e. image locations marked as blue circles. Right: The sound emitted by the loud-speaker is localized in the image plane with a green circle

Discussion

- The learning method is room-independent, but *robot-head dependent*, the DP-RTF still depends on the head shape, material, microphone locations, number of microphones, etc.
- In the case of a robot head, DP-RTF may well be viewed as the *head-related transfer function* (HRTF).
- DP-RTF estimation in the presence of stationary noise uses *inter-frame spectral subtraction* based on estimated the auto- and cross power spectral densities.
- Extensions to multiple static and moving sound sources are available
- One possible metric is the *direct to reverberation ratio* (DRR) which is the energy ratio between direct path and reflections. The direct-path energy will become smaller (relative to reflections) when the speaker gets farther from the microphones, or when the speaker does not face toward the microphone.

Additional references

- A. Deleforge, R. Horaud, Y. Schechner, and L. Girin, "Co-Localization of Audio Sources in Images Using Binaural Features and Locally-Linear Regression," *IEEE TASLP*, 2015. [link]
- X. Li, L. Girin, R. Horaud and S. Gannot, "Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization", *IEEE TASLP*, 2017. [link]
- X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud, "Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environment", *IEEE J-STSP*, 2019. [link]