Framework for ITD-based sound source localization in complex acoustic environments

Tobias May and Nicolas Le Goff

January 30, 2014

1 Introduction

One of the most important physical cues that enables human sound source localization are the interaural time differences (ITDs) and interaural level differences (ILDs) between the left and the right ear signals. The purpose of this framework is to develop algorithms that are able to find the azimuth angle of multiple competing speech sources relative to an artificial head. The first two very basic algorithms only exploit estimates of the ITD.

The relation between a particular sound source azimuth and its corresponding ITD depends on the size and the geometry of the artificial head. Therefore, a mapping must be employed to relate the estimated ITD to its corresponding azimuth position. The mapping function is obtained during an initial *calibration* stage, in which the azimuth of a single sound source in anechoic conditions (no reflections) is systematically varied across the azimuth range of interest (e.g. $[-90\,°, 90\,°]$), and the resulting ITD is measured.

Given a binaural signal, the azimuth of a sound source is localized by performing the following two steps: first, the ITD is estimated by a correlation analysis. More specifically, the ITD is assumed to be reflected by the lag that corresponds to the most prominent peak in the cross-correlation function. In the second step, the observed ITD is related to its corresponding sound source azimuth by applying the mapping function that has been created during the calibration stage. So far, two basic sound source localization approaches are implemented:

1. The broadband approach (`estimate_Azimuth_Broadband.m`):

   Sound source localization is performed in the "audio domain" by computing the cross-correlation function over short time frames of 20 ms (`winSec = 20E-3`). The resulting 2-dimensional cross-correlation function `CCF`, which is a function of the number of time lags (ITDs) and the number of frames (`nLags x nFrames`), is mapped onto an azimuth grid (`nAzimuth x nFrames`). The required mapping function is computed by `calibrate_ITD_Broadband.m`. Then, this new CCF is

integrated across all time frames and the most prominent peaks are assumed to reflect the estimated sound source azimuth positions.

2. The subband approach (`estimate_Azimuth_Subband.m`:):

   A peripheral auditory processing stage is included, which decomposes the input signals into individual frequency channels using a gammatone filterbank (`gammaFB.m`). The center frequencies are equally spaced on the equivalent rectangular bandwidth (ERB)-rate scale between `fLowHz = 100` and `fHighHz = 12000` Hz. Sound source localization is performed in the "auditory domain" by computing the cross-correlation function for each subband over short windows of 20 ms (`winSec = 20E-3`). The resulting 3-dimensional cross-correlation function `CCF`, which is a function of the number of lags, subbands and frames (`nLags x nSubbands x nFrames`) is mapped onto an azimuth grid (`nAzimuth x nSubbands x nFrames`). The required mapping function is computed by `calibrate_ITD_Subband.m`. This mapping is performed for each subband separately. Then, this new CCF is integrated across all subbands and frames and the most prominent peaks are assumed to reflect the estimated sound source azimuth positions. This implementation is a simplified version of the algorithm described in Palomäki *et al.* (2004).

The estimation of the ITD, and accordingly, its sound source azimuth can be reliably obtained for one sound source in anechoic condition. However, in more complex acoustic environments, the direct sound from multiple sources that are positioned at different azimuth angles and room reflections overlap, thus changing the interaural time differences at the receiver. Therefore, more sophisticated localization approaches will be developed and implemented that also exploit the ILD cue. A general overview about binaural sound source localization techniques can be found in May *et al.* (2013), which is supplied in the literature folder. The modular structure allows to test different modifications and the evaluate the benefit of individual processing stages.

2 Experimental framework

Call the main script `localization_experiment.m` in order to perform one run of the localization experiment. During the experiment, a predefined number of speech sources are randomly positioned at unknown azimuth angles relative to the receiver. The amount of reverberation can be changed during each run of the localization experiment by selecting a particular set of binaural room impulse responses (BRIRs). The localization accuracy of both approaches is measured by computing the percentage of correctly localized sources (`pc1` and `pc2`) within a predefined error threshold, as well as the root mean square error of the correctly localized soures in comparison to their true azimuth (`rmse1` and `rmse2`). A summary of the settings that can be controlled in the header of the main script is given below:

1. Algorithm settings

Table 1: Acoustic properties of different BRIRs (Hummersone *et al.*, 2010).

Name	Room description	T_{60} (s)	DRR (dB)
`'SURREY_A'`	Anechoic condition	0.0	∞
`'SURREY_ROOM_A'`	Medium-sized office	0.32	6.09
`'SURREY_ROOM_B'`	Medium–small class room	0.47	5.31
`'SURREY_ROOM_C'`	Large cinema–style lecture theater	0.68	8.82
`'SURREY_ROOM_D'`	Medium–large sized seminar space	0.89	6.12

- `fs` controls the sampling frequency in Hz
- `fLowHz` and `fHighHz` define the lowest and the highest center frequency of the gammatone filterbank
- `winSec` determines the frame size for the cross-correlation analysis

2. Acoustic settings
   - `nSpeakers` specifies the number of competing speech sources that are randomly placed at unknown azimuth angles
   - `minDistance` defines the minimum angular distance between competing sources
   - `rooms` is a list of BRIRs that is used to spatialize the speech sources. An overview of all five rooms is given in Tab. 1. The acoustic properties of the different rooms are characterized by the reverberation time T_{60} and the direct-to-reverberant ratio (DRR).
3. Evaluation settings
   - `nMixtures` defines the number of binaural signals created for each room (`rooms`)
   - `thresDeg` controls the absolute error boundary in degree used to compute the percentage of correctly localized sound sources
   - `bVisualize` if `true`, the output of both localization approaches (`Broadband` and `Subband`) is plotted in two individual figures for each binaural signal. You can turn off the visualization by setting `bVisualize = false` for a faster execution of the experiment.

References

Hummersone, C., Mason, R., and Brookes, T. (**2010**). "Dynamic precedence effect modelling for source separation in reverberant environments", IEEE Trans. Audio, Speech, Lang. Process. **18**, 1867–1871.

May, T., van de Par, S., and Kohlrausch, A. (**2013**). "Binaural localization and detection of speakers in complex acoustic scenes", in *The technology of binaural listening*, edited by J. Blauert, 397–425 (Springer, Berlin–Heidelberg–New York NY).

Palomäki, K. J., Brown, G. J., and Wang, D. (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation", Speech Communication **43**, 361–789.