Improved binaural speaker localization and separation robust to rotational head movement

Mehdi Zohourian, Gerald Enzner, Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany, Email: {first name.last name}@rub.de

Abstract

Adaptive binaural beamforming and source separation methods require information on the desired source locations relative to the current head position. The localization algorithm needs to be robust to unknown and dynamic source/receiver configurations and to adverse acoustic conditions. In this work we propose and compare binaural localization algorithms based on beamforming techniques. The algorithms use a joint ITD/ILD head model and do not require prior training. The methods are capable of accurate *direction-of-arrival* (DOA) estimation in the horizontal plane across a wide range of frequencies. In addition, we support our localization algorithms by means of a head tracker sensor to provide head movement information. Finally, we integrate the proposed localization algorithm in a generalized side-lobe canceller to separate concurrent speakers. We evaluate the performance of the proposed adaptive beamforming algorithm over different recording scenarios based on objective measurements and informal listening tests. Results demonstrate the efficiency of our algorithm.

Introduction

Binaural speech enhancement plays a major role in the performance of hearing aids. A common assumption in most of the binaural beamformer is that the target source is in front of the listener. Furthermore, ambient noise, interfering sources, and the listeners head movements degrade the performance of hearing aids considerably.

In this paper we propose an adaptive binaural beamformer to localize and separate concurrent speakers. The primary goal of this work is the investigation of binaural localization algorithm. The two binaural cues, namely *interaural time or phase difference* (ITD/IPD) and *interaural level difference* (ILD), are jointly employed in the proposed algorithms. The ILD plays an important role in DOA estimation especially at high frequencies and should be utilized in the localization algorithms. In free field conditions optimal approaches have been extensively used in localization. In the binaural context however, localization is performed based on heuristic combination of binaural cues. In [1] for example, the ILD cues were combined heuristically with IPD in the Fourier domain. In [2] the head model is employed to evaluate ITD cues while measured HRTFs are used for the evaluation of ILD. The authors in [3] incorporate a probabilistic model to train the interaction between ITD and ILD. In this study we contribute different localization methods that integrate both IPD and ILD in closed-form cost functions. The proposed localization methods enable the DOA estimation in a wide range of frequencies and do not require prior training of the interaction between ITD, ILD and source locations.

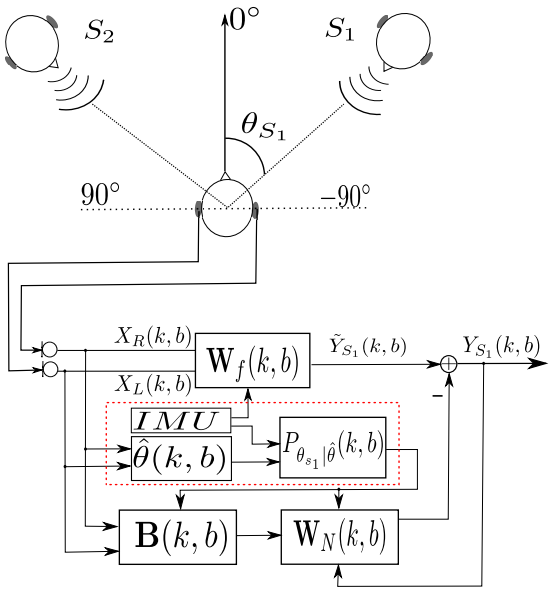


Figure 1: Binaural speaker localization and separation.

Adaptive binaural beamformer

The block diagram of the proposed adaptive binaural beamformer considering the binaural signal model is shown in Fig. 1. In this scenario we consider binaural signals from two sources received by the front microphones of two BTE hearing aids. Using the convolution operator $*$ the received signal at each microphone m is written as

$$x_m(n) = \sum_{i=1}^{2} s_i(n) * h_{im}(n) + n_m(n) \qquad (1)$$

where $s_i(n)$ represents the i-th point source signal, $h_{im}(n)$ indicates a binaural room impulse response (BRIR) from the source i to the microphone m, $m \in \{L, R\}$, $n_m(n)$ is the noise at microphone m, and n is the sampling index. We analyse signals in the STFT domain and thus obtain

$$\begin{pmatrix} X_L(k,b) \\ X_R(k,b) \end{pmatrix} = \begin{pmatrix} H_{1L}(k,b) & H_{2L}(k,b) \\ H_{1R}(k,b) & H_{2R}(k,b) \end{pmatrix} \begin{pmatrix} S_1(k,b) \\ S_2(k,b) \end{pmatrix} + \begin{pmatrix} N_L(k,b) \\ N_R(k,b) \end{pmatrix}. \qquad (2)$$

Here, $\tilde{H}_{im}(k,b)$ are the transfer functions of both the left and right ears and (k,b) indicate frequency and frame indices. The received signals are processed through the proposed adaptive binaural beamformer. The beamformer is

an extension of [4] adapted to the binaural configuration using hearing aids. Our system consists of two parts: first a *generalized side-lobe canceller*(GSC) [5] with a beamformer $\mathbf{W}_f(k,b)$ looking towards the target, an adaptive blocking matrix $\mathbf{B}(k,b)$, and an adaptive noise canceler $\mathbf{W}_N(k,b)$. Secondly, a localization-tracking part comprises a binaural localization algorithm estimating the target DOA $\hat{\theta}(k,b)$, a head tracking sensor using an inertial measurement unit (IMU), and a target presence probability $P_{\theta_{s_i}|\hat{\theta}}(k,b)$ using a *Gaussian mixture model* (GMM). All of these components are used to extract the desired signal at each frequency bin.

Binaural speaker localization

In this section we introduce different binaural localization algorithms based on joint IPD/ILD information. In these approaches we derive different cost functions for localization that integrate IPD/ILD in a consistent fashion. Our approaches are based on *null-steering beamformer* (NSB) where the beamformer steers a null to each source position candidate and searches for the minimum value and the corresponding DOA . It has been shown for the free field scenario that *minimizing* the NSB output when microphone signals are compensated with phase only is equivalent to *maximizing* the output of the *steered response power* (SRP) [6]. We extend this idea to the binaural configuration considering joint IPD/ILD. The binaural cues are taken into account in the form of HRTFs $\hat{H}_L(\Omega_k,\theta)$ and $\hat{H}_R(\Omega_k,\theta)$ and derived from a spherical head model described in the following section.

Spherical head model

In [7] Brown and Duda propose a spherical head model for the approximation of the ITD the ILD cues. The head model is formed by cascading a first-order recursive head-shadow filter and a propagation delay. Taking the coordinate system in Fig. 1 into account, the HRTF for the right ear is expressed as:

$$\hat{H}_R(\omega,\theta) = \frac{1+j\frac{\omega}{2\omega_0}\gamma_R(\theta)}{1+j\frac{\omega}{2\omega_0}} e^{-j\omega\hat{\tau}_R(\theta)}. \tag{3}$$

In this equation we have $\omega_0 = c/a$, where c is the speed of sound, a is the radius of the head, and θ is the angle between the source and the right ear. $\gamma_R(\theta)$ and $\hat{\tau}_R(\theta)$ are two angle-dependent parameters that are defined as

$$\gamma_R(\theta) = 1.05 + 0.95\cos\left(\frac{6}{5}(\theta-\pi/2)\right) \tag{4}$$

$$\hat{\tau}_R(\theta) = \begin{cases} -\frac{a}{c}\sin(\theta) & \text{if} \quad -\pi/2 \le \theta < 0 \\ -\frac{a}{c}\theta & \text{if} \quad 0 \le \theta < \pi/2. \end{cases} \tag{5}$$

Then, for the left ear we have

$$\gamma_L(\theta) = 1.05 + 0.95\cos\left(\frac{6}{5}(\theta+\pi/2)\right) \tag{6}$$

$$\hat{\tau}_L(\theta) = \begin{cases} \frac{a}{c}(\theta) & \text{if} \quad -\pi/2 \le \theta < 0 \\ \frac{a}{c}\sin(\theta) & \text{if} \quad 0 \le \theta < \pi/2. \end{cases} \tag{7}$$

We integrate both IPD and ILD cues provided by the head model in different localization algorithms explained in the following sections.

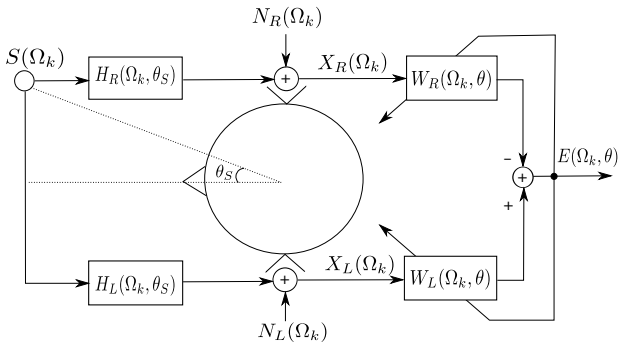


Figure 2: Binaural signal model and localization algorithm.

Cost functions for the localization

The binaural signal model followed by the binaural localization algorithm is shown in Fig. 2. Based on different types of filters we obtain cost functions for the localization which are introduced in the following sections:

NSB with signal equalization (NSB-SE)

We design the NSB filters such that each channel is equalized independently, i.e. $W_R(\Omega_k,\theta) = H_R^{-1}(\Omega_k,\theta)$, and $W_L(\Omega_k,\theta) = \hat{H}_L^{-1}(\Omega_k,\theta)$ and derive the cost function as

$$J(\Omega_k,\theta) = \left| \frac{X_L(\Omega_k)}{\hat{H}_L(\Omega_k,\theta)} - \frac{X_R(\Omega_k)}{\hat{H}_R(\Omega_k,\theta)} \right|^2 \tag{8}$$

where $\hat{H}_m$ $(m \in \{L,R\})$ is the HRTF for the angle θ, derived from the spherical head model. For simplicity, the time index b has been eliminated in the equation and $\Omega_k = 2\pi k f_S/M$. Expanding the cost function in (8) and exploiting the phase ϕ_m of the received signals we write

$$\begin{aligned} J(\Omega_k,\theta) = &\left| \frac{X_L(\Omega_k)}{\hat{H}_L(\Omega_k,\theta)} \right|^2 + \left| \frac{X_R(\Omega_k)}{H_R(\Omega_k,\theta)} \right|^2 \\ &- 2\Re\left\{ \frac{|X_L(\Omega_k)|}{|\hat{H}_L(\Omega_k,\theta)|} \frac{|X_R(\Omega_k)|}{|H_R(\Omega_k,\theta)|} e^{j\Delta\phi} \right\} \end{aligned} \tag{9}$$

where, $\Delta\phi = \big(\phi_{X_R}(\Omega_k) - \phi_{X_L}(\Omega_k) - \Omega_k(\hat{\tau}_R(\theta) - \hat{\tau}_L(\theta))\big)$ and $\Re(.)$ denotes the real part. We factorize (9) as

$$\begin{aligned} J(\Omega_k,\theta) = &\frac{|X_L(\Omega_k)||X_R(\Omega_k)|}{|\hat{H}_L(\Omega_k,\theta)||H_R(\Omega_k,\theta)|} \\ &\times \left(A(\Omega_k,\theta) + \frac{1}{A(\Omega_k,\theta)} - 2\cos(\Delta\phi) \right) \end{aligned} \tag{10}$$

with $A(\Omega_k,\theta) = \frac{|X_L(\Omega_k)||\hat{H}_R(\Omega_k,\theta)|}{|X_R(\Omega_k)||\hat{H}_L(\Omega_k,\theta)|}$. For $A(\Omega_k,\theta) > 0$ the function $f(A) = A(\Omega_k,\theta) + \frac{1}{A(\Omega_k,\theta)}$ is always positive and attains its minimum value of $f(A) = 2$ for $A(\Omega_k,\theta) = 1$ and thus represents the effects of ILD deviations. Therefore, $\min\{f(A) - 2\cos(\Delta\phi)\} = 0$ when both the amplitudes and the phases match the head model. We can also ignore microphones signals $|X_L(\Omega_k)||X_R(\Omega_k)|$ which are independent of θ and write the cost function as

$$\begin{aligned} \widetilde{J}(\Omega_k,\theta) = &\frac{1}{|\hat{H}_L(\Omega_k,\theta)||\hat{H}_R(\Omega_k,\theta)|} \\ &\times \left(A(\Omega_k,\theta) + \frac{1}{A(\Omega_k,\theta)} - 2\cos(\Delta\phi) \right) \end{aligned} \tag{11}$$

and minimize it for each frequency across frontal angles.

NSB with cross-relation (NSB-CR)

In the second approach we design the NSB filters based on the cross relation technique, i.e. $W_R(\Omega_k,\theta) = \hat{H}_L(\Omega_k,\theta)$, and $W_L(\Omega_k,\theta) = \hat{H}_R(\Omega_k,\theta)$. Using the notation of the last section, we write the cost function for the NSB-CR method as

$$\Gamma(\Omega_k,\theta) = \left| X_L(\Omega_k)\hat{H}_R(\Omega_k,\theta) - X_R(\Omega_k)\hat{H}_L(\Omega_k,\theta) \right|^2. \tag{12}$$

The cost function (12) is expanded and factorized as

$$\Gamma(\Omega_k,\theta) = |X_L(\Omega_k)||X_R(\Omega_k)||\hat{H}_L(\Omega_k,\theta)||\hat{H}_R(\Omega_k,\theta)| \times \left(A(\Omega_k,\theta) + \frac{1}{A(\Omega_k,\theta)} - 2\cos(\Delta\phi) \right) \tag{13}$$

where again $\Delta\phi = \big(\phi_R(\Omega_k) - \phi_L(\Omega_k) - \Omega_k(\hat{\tau}_R(\theta) - \hat{\tau}_L(\theta)\big)$ and $A(\Omega_k,\theta) = \frac{|X_L(\Omega_k)||\hat{H}_R(\Omega_k,\theta)|}{|X_R(\Omega_k)||\hat{H}_L(\Omega_k,\theta)|}$. The microphone signals $|X_L(\Omega_k)||X_R(\Omega_k)|$ are independent of θ and are ignored. Then, we have

$$\widetilde{\Gamma}(\Omega_k,\theta) = |\hat{H}_L(\Omega_k,\theta)||\hat{H}_R(\Omega_k,\theta)| \times \left(A(\Omega_k,\theta) + \frac{1}{A(\Omega_k,\theta)} - 2\cos(\Delta\phi) \right). \tag{14}$$

The cost function is minimized when both the amplitude and the phase of the received signal match the binaural cues of the head model.

NSB with normalized signal equalization (NSB-SE/N)

The cost function (14) is similar to (11) with a difference in the gain functions. On the one hand the gain function $|\hat{H}_L(\Omega_k,\theta)||\hat{H}_R(\Omega_k,\theta)|$ amplifies the cost function for front angles and attenuates for the lateral angles. On the other hand due to the different factors in (14) and (11) the NSB-CR method localizes better than the NSB-SE method the lateral angles. Hence, the gain functions bias the cost functions regardless of the true position of sources. Consequently, in order to increase the robustness of the methods for all angles, and to decrease the computational complexity the gain functions are neglected. This results in the third approach with simplified closed-form cost function which we denote as NSB with normalized signal equalization (NSB-SE/N),

$$\widetilde{J}_N(\Omega_k,\theta) = A(\Omega_k,\theta) + \frac{1}{A(\Omega_k,\theta)} - 2\cos(\Delta\phi). \tag{15}$$

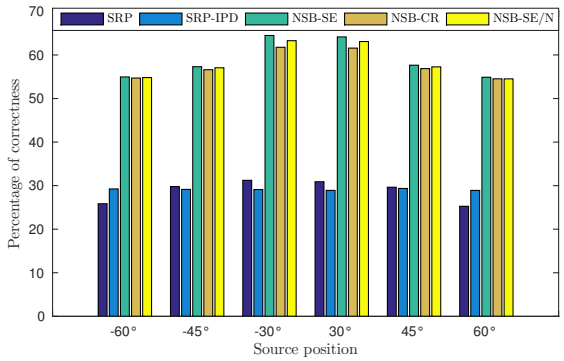
The advantage of this approach is that the magnitude of HRTFs are only used in ratio term and thus the stability of the cost function is improved.
Finally, we integrate each of the proposed localization algorithms in the adaptive binaural beamformer described in [4, 8].

Experimental results

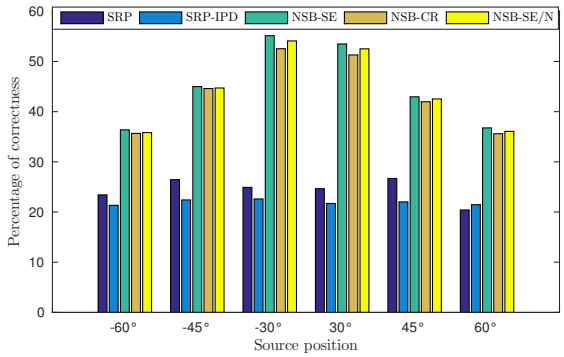
We used BTE hearing aid dummies attached to a dummy head. Loudspeakers playing male and female utterances were placed at a distance of 1 m from the dummy head. We conducted our experiments in an anechoic and a reverberated room with $T_{60} = 0.5$ s. Audio recordings were made at 48 kHz and later downsampled to 16 kHz.
The performance of the proposed localization algorithms is evaluated for the estimation of the position of two active sources and is presented in Fig. 3. We assume that one source is fixed at 0° and the other source is placed at the corresponding angles. We compare the proposed algorithms with the SRP algorithm with a free-field assumption, and the SRP algorithm incorporating IPD cues only (SRP-IPD). The results are expressed in terms of the percentage of correctness for different acoustic rooms. The percentage of correctness counts the number of correct estimation per frequency bin divided by the total number of bins in a given period of time. The absolute error threshold for the correct estimation is $|\theta_{thr}| = 5°$. According to this figure the proposed techniques that use joint IPD/ILD shows improved performance. Fig. 4 demonstrates the capability of the proposed methods for DOA estimation of a single source at $-60°$ at each time-frequency point. It is observed that the ambiguities in DOA estimation especially at high frequencies are remarkably resolved in the proposed algorithms that use both IPD and ILD cues.
The performance of the adaptive binaural beamformer for the separation of the desired speaker is evaluated for a dynamic recording with head movement in a low reverberated room. Two speakers are separated by 60° and the angular speed is $30°/s$ which represent a realistic scenario. One cycle of turns starts with one speaker in front of the head and ends when the other speaker is in front of the head. The total length of the recording is one minute. Experimental results are stated in terms of the perceptual evaluation of speech quality (PESQ) [9] and signal-to-interference ratio (SIR) [10] which are shown in Fig. 5. The results reflect that the beamformer which is controlled by the proposed NSB methods outperform the others in terms of quality and separation measurements. This confirms the idea of using joint IPD and ILD. The higher the accuracy in DOA estimation, the less distortion in the extracted target signal of the binaural beamformer is achieved.

Conclusion

We presented binaural localization approaches based on the combination of IPD and ILD. Conventional binaural localization methods often use heuristic measurements to combine binaural cues since the interaction of IPD/ILD is unknown. Not much work, however, has been done for the use of optimization criteria in binaural localization. Therefore, in this paper we contribute novel localization algorithms that combine the IPD and ILD cues in simple closed-form cost functions. The localization methods are robust against the presence of multiple speakers and a limited amount of reverberation. Our algorithms enable the narrowband DOA estimation regardless of prior



Anechoic



$T_{60} = 0.5$

Figure 3: The comparison between the localization methods for the estimation of two sources in different reverberation rooms. (One source is fixed at 0° and the other source is placed at the corresponding angles).

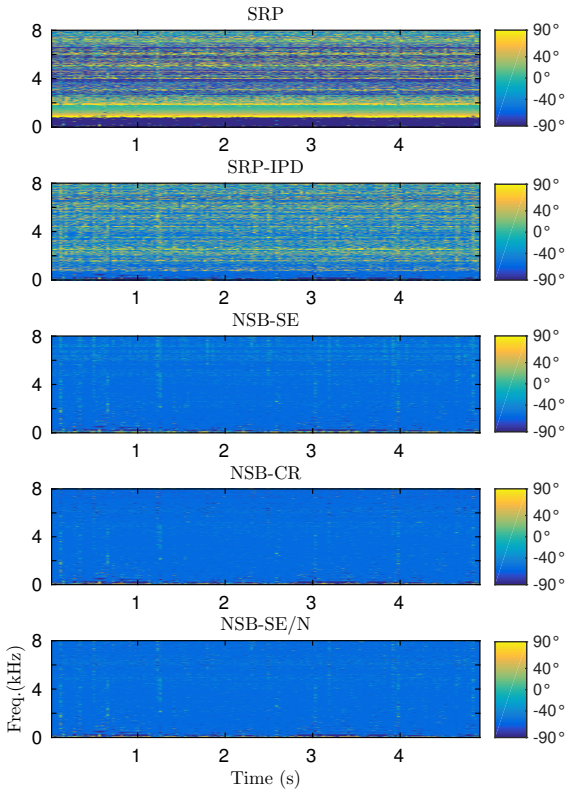


Figure 4: DOA estimation at each time-frequency bin for a source located at −60°.

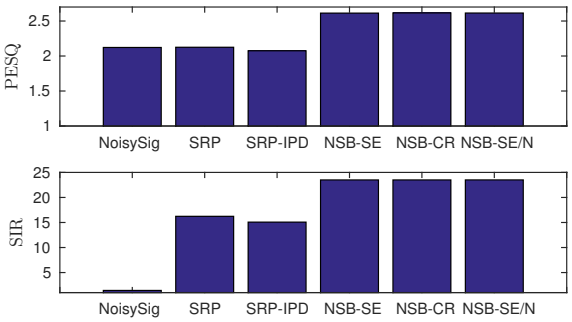


Figure 5: The performance of the adaptive binaural beamformer controlled by different localization methods.

training of source/receiver configuration. The localization algorithm is then integrated in the adaptive binaural beamformer [4, 8] to extract the target signal. Results corroborate the superiority of our system for both localization and separation versus other methods.

References

[1] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio Speech Language Process.*, vol. 18, no. 1, Jan 2010.

[2] I. Merks, G. Enzner, and T. Zhang, "Sound source localization with binaural hearing aids using adaptive blind channel identification," in *IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, May 2013.

[3] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 1, Jan 2011.

[4] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 7, Sept 2011.

[5] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, Jan 1982.

[6] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[7] C.P. Brown and R.O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, Sep 1998.

[8] M. Zohourian, A. Archer-Boyd, and R. Martin, "Multichannel speaker localization and separation using a model-based GSC and an inertial measurement unit," in *IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, April 2015.

[9] A.W Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, 2001, vol. 2.

[10] C. Févotte, R.I. Gribonval, E. Vincent, et al., "BSS_EVAL toolbox user guide–revision 2.0," 2005.