EUSIPCO 2013 1569742579

ROBUST ITD ERROR ESTIMATION FOR CROSSTALK CANCELLATION SYSTEMS WITH A MICROPHONE-BASED HEAD-TRACKER

Yesenia Lacouture-Parodi and Emanuël A. P. Habets

International Audio Laboratories Erlangen†, Am Wolfsmantel 33, 91058 Erlangen, Germany

ABSTRACT

Previously, we proposed a crosstalk cancellation system with a microphone-based head-tracker. The head-tracker estimates the orientation of the head using the interaural time difference (ITD) error between the desired binaural signal and the signals at the microphones, which are placed near the listener's ears. In the presence of noise and multiple virtual sound sources, simple cross-correlation based techniques will fail to find the true ITD error value. In this study, we propose a method to estimate the ITD error from multiple competing sources in a noisy environment. We use relative transfer functions (RTFs) to estimate the ITD error in subbands and apply a weighting function based not only on the coherence, but also on the magnitude of the RTFs in each subband. Experiments show that in the presence of multiple sources and noise, the proposed weighting function yields better results compared to a weighting function that depends only on the coherence.

Index Terms— Interaural time differences, acoustic transfer functions ratio, time difference of arrival, crosstalk cancellation

1. INTRODUCTION

With binaural technology it is possible to simulate a virtual environment where the listener perceives an acoustic source located at a position where no physical source is located. To accurately reproduce binaural signals through loudspeakers, we need to compensate for the acoustic paths between the loudspeakers and the ears. This can be achieved by incorporating appropriate crosstalk cancellation filters (CCFs) into the reproduction chain. In general, the filters are dependent on the location of the listener such that head movements can easily impair the intended virtual sound experience.

In [1], we proposed a dynamic crosstalk cancellation system (CCS) that uses a microphone-based head-tracker (see Fig. 1). The orientation angle of the listener's head with respect to the loudspeakers is calculated by estimating the interaural time difference (ITD) error that is defined as the difference between the ITD of the desired binaural signal and the ITD of the signals captured by the microphones that are positioned close to the ears of the listener. In principle, the head orientation can be inferred from the acoustic transfer functions (ATFs) between the loudspeakers and the microphones. Unfortunately, the loudspeaker signals are often highly correlated, which makes the identification of the ATFs problematic. In [1, 2], it was shown that by minimizing the ITD error, we can accurately track head rotations and adapt the CCFs accordingly.

A method to estimate the ITD is the interaural cross-correlation (IACC) method, which estimates the ITD as the time at which the cross-correlation between left and right has its maximum value [3]. It is also possible to calculate the ITD based on the slope of the phase

† A joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS.

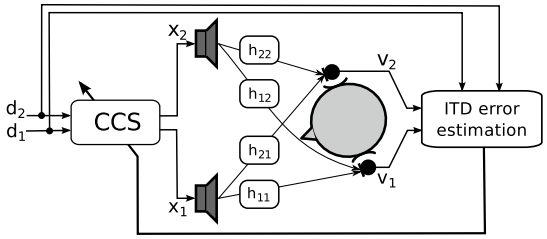


Fig. 1: Simplified diagram of the dynamic crosstalk cancellation system (CCS) with a head-tracker based on the microphone signals v_i (see [1] for a detailed description).

difference between the channels at low frequencies [4, 5]. However, this method has shown to be sensitive to errors in the phase of the signals, which are usually introduced by the CCFs [1]. Most commonly used methods assume a single source and no reflections. In the presence of multiple sources and reflections these methods can provide erroneous ITD estimates.

The ITD can be seen as a particular case of time difference of arrival (TDOA), where signals from only two sensors, namely the ears, are used. While for most sound source localization applications a TDOA accuracy of a few milliseconds might be sufficient, the required ITD accuracy for binaural reproduction is more critical given that the just noticeable audible difference (JNAD) of the ITD lies in the range of 10 to 20 μs [6].

In [7], Dvorkind and Gannot proposed to use the relative transfer function (RTF) to estimate the TDOA in noisy and reverberant environments. The TDOA is estimated as the time where the relative impulse response has its maximum. This approach shows to be robust to noise and reverberation. In the presence of multiple sources, the RTF will be a mixture of the ATFs of each source, and the estimated TDOA will not necessarily reflect the true location of any specific source. A common approach to cope with multiple sources, is to estimate TDOA in subbands [8–10]. However, most of the methods rely on the sparseness of the speech or assume mutually uncorrelated sound sources [8, 11].

In this study, we propose to estimate the ITD at the input of the CCS and at the microphones using RTFs. To increase the robustness, the ITDs and ITD error are computed in subbands. Finally, the fullband ITD error is computed as a weighted sum of the subband ITD errors. In [12], the interaural coherence was used to select the interaural cues that are close to free-field binaural cues when several independent sources are concurrently active. In the case of crosstalk cancelled signals, there are subbands in which the signals are coherent due to insufficient channel separation, which can result in errors in the ITD error estimation. The channel separation is commonly defined as the magnitude ratio between the crosstalk and the direct signal. Therefore, a weighting function is proposed that depends on the interaural coherence and the RTF between the microphones.

2. SIGNAL MODEL

Fig. 1 shows a simplified diagram of the CCS proposed in [1]. The functions h_{ji} correspond to the acoustic impulse responses from the *i*th loudspeaker to the *j*th ear. The inputs to the CCS are the desired binaural signals d_i. The head orientation is calculated based on the estimated ITD error between the desired binaural signals d_i and the signals measured by the microphones v_i.

Let us define the binaural input signals to the CCS as

$$d_i(t) = a_i(t) * s(t) + q_i(t); \quad i \in \{1, 2\}, \tag{1}$$

where * denotes the convolution operation, $s(t)$ is the monophonic source signal, $a_i(t)$ corresponds to the head related impulse response (HRIR) of the desired virtual sound source to the *i*th ear and $q_i(t)$ is the ambient sound. The signals at the ears are thus defined as

$$v_i(t) = \tilde{a}_i(t) * s(t) + \tilde{q}_i(t) + n_i(t); \quad i \in \{1, 2\}, \tag{2}$$

where

$$\tilde{a}_i(t) = \sum_{k=1}^{2}\sum_{j=1}^{2} h_{ik}(t) * c_{kj}(t) * a_j(t), \tag{3}$$

$$\tilde{q}_i(t) = \sum_{k=1}^{2}\sum_{j=1}^{2} h_{ik}(t) * c_{kj}(t) * q_j(t), \tag{4}$$

c_{kj} corresponds to the CCF from the jth input signal to the kth loudspeaker and $n_i(t)$ is the noise at the *i*th microphone[1]. Perfect crosstalk cancellation is thus achieved when $v_i(t) - n_i(t) = d_i(t)$. In the case of multiple sources, the desired binaural signal will be a linear combination of the virtual sound sources $d'_i(t) = \sum_{l=1}^{L} a_{l,i}(t) * s_l(t) + q_i(t), i \in \{1, 2\}$, where L is the total number of sources. The signals at the ears become

$$v_i(t) = \sum_{l=1}^{L} \tilde{a}_{l,i}(t) * s_l(t) + \tilde{q}_i(t) + n_i(t); \quad i \in \{1, 2\}. \tag{5}$$

3. RELATIVE TRANSFER FUNCTION ESTIMATION

In [7] the authors proposed to estimate the TDOA between two spatially separated microphones using the RTF, which is defined as

$$\mathcal{R}_i(\omega) = \frac{B_i(\omega)}{B_1(\omega)}, \tag{6}$$

where $B_i(\omega)$ is the ATF from the desired signal to the *i*th microphone. Now, let $A_i(\omega)$ and $\tilde{A}_i(\omega)$ be, respectively, the transfer functions of the impulses $a_i(t)$ and $\tilde{a}_i(t)$ defined in (1) and (3). Since the ITD error depends on the ITD at the input of the CCS and the ITD at the microphones, we define the RTFs

$$\mathcal{R}_{\text{in}}(\omega) = \frac{A_2(\omega)}{A_1(\omega)} \quad \text{and} \quad \mathcal{R}_{\text{out}}(\omega) = \frac{\tilde{A}_2(\omega)}{\tilde{A}_1(\omega)}.$$

The corresponding relative impulse responses are given by $r_{\text{in}}(t)$ and $r_{\text{out}}(t)$. Note that for the proposed application, the RTFs are equivalent to the interaural transfer functions. The ITD is defined as the time at which the impulses $r_{\text{in}}(t)$ and $r_{\text{out}}(t)$ have their maximum. In [7] it is proposed to estimate the RTFs by using the cross power spectral density (PSD) and the auto-PSD of the signals as follows:

$$\frac{\phi_{d_2d_1}(\omega)}{\phi_{d_1d_1}(\omega)} = \frac{A_2(\omega)A_1^*(\omega)\phi_{\text{ss}}}{A_1(\omega)A_1^*(\omega)\phi_{\text{ss}}} = \mathcal{R}_{\text{in}}(\omega), \tag{7}$$

where ϕ_{ss} is the auto-PSD of the source signal $s(t)$. Note that (7) holds for infinite observation windows, though in our application we have only access to short observation windows. As proposed in [7], we use the assumption that the desired signal is quasi-stationary[2] and uncorrelated with the ambient sound that is assumed to be stationary during the observation window. Exploiting these assumptions, we can write the estimate of the cross-PSD at the input as [7]

$$\hat{\phi}_{d_2d_1}(m, \omega) = \mathcal{R}_{\text{in}}(\omega)\hat{\phi}_{d_1d_1}(m, \omega) + \phi_{q_2}(\omega) + \xi(m, \omega), \tag{8}$$

for $m = 1 \dots M$, where M is the total number of frames, $\phi_{q_2}(\omega)$ is the PSD of the noise and $\xi(m, \omega) = \hat{\phi}_{q_2}(m, \omega) - \phi_{q_2}(\omega)$ is an error term that we seek to minimize in the least squares sense. Having the estimated PSDs for each frame, the least squares estimation of $\mathcal{R}_{\text{in}}$ is calculated as

$$\begin{bmatrix} \mathcal{R}_{\text{in}}(\omega) \\ \hat{\phi}_{q_2}(\omega) \end{bmatrix} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\hat{\boldsymbol{\phi}}_{d_2d_1}(\omega), \tag{9}$$

where

$$\mathbf{A} = \begin{bmatrix} \hat{\phi}_{d_1d_1}(1, \omega) & 1 \\ \vdots & \\ \hat{\phi}_{d_1d_1}(M, \omega) & 1 \end{bmatrix}, \quad \hat{\boldsymbol{\phi}}_{d_2d_1} = \begin{bmatrix} \hat{\phi}_{d_2d_1}(1, \omega) \\ \vdots \\ \hat{\phi}_{d_2d_1}(M, \omega) \end{bmatrix},$$

and the superscript H denotes the conjugate transpose. Finally, a similar procedure is used to estimate $\mathcal{R}_{\text{out}}$. Since we want to estimate the ITD error on a frame-by-frame basis, we used in this study the recursive implementation (RLS) of (9) proposed in [7].

4. ITD ERROR ESTIMATION

In principle, the ITD error could be directly estimated from the phase information of $\mathcal{R}_{\text{in}}(\omega)$ and $\mathcal{R}_{\text{out}}(\omega)$ [5]. For our application this is however sub-optimal, given that, even in ideal conditions, the CCFs introduce unavoidable distortions in the phase of the signal conveying large errors to the ITD error estimation [1]. Further, it is clear from (5) that in the presence of multiple sources, the relative impulse responses $r_{\text{in}}(t)$ and $r_{\text{out}}(t)$ will contain a mixture of the binaural cues from each source. Directly estimating the ITD error using these impulse responses will be prone to errors. To mitigate these problems, we propose to first compute an ITD error per subband and then compute a fullband ITD error as a weighted sum of the subband ITD errors. Note that according to the ITD error model presented in [1], the ITD error does not depend on the virtual sound sources but only on the head orientation.

Let $\mathcal{R}_{\text{in}}(m, k)$ and $\mathcal{R}_{\text{out}}(m, k)$ be the short-time frequency representation of the RTFs, where m is the time frame and k is the discrete frequency index. Let us now divide the spectrum into N subbands corresponding to frequency bins $k \in \{E_{b-1}, \dots, E_b - 1\}$, where $b = 1, \dots, N$ is the subband index. The subband ITD error at time frame m is defined as

$$\text{ITD}_{\text{error}}(m, b) = \arg\max_n \left\{r_{\text{in}}^{(m,b)}(n)\right\} - \arg\max_n \left\{r_{\text{out}}^{(m,b)}(n)\right\}, \tag{10}$$

[1] The noise at the microphone is a mixture of sensor and ambient noise.

[2] In [7] the assumption of quasi-stationarity is applied to speech, while for our application we need to extend this assumption to other type of sources.

where $r_{\text{in}}^{(m,b)}(n)$ and $r_{\text{out}}^{(m,b)}(n)$ are, respectively, the discrete time domain representations computed using $\mathcal{R}_{\text{in}}(m,k)$ and $\mathcal{R}_{\text{out}}(m,k)$ with $k \in \{E_{b-1}, \ldots, E_b - 1\}$.

Let us define the vector $\boldsymbol{\Delta}(m)$ which contains the estimated subband ITD errors at time frame m for all N subbands:

$$\boldsymbol{\Delta}(m) = [\, \text{ITD}_{\text{error}}(m,1), \ldots, \text{ITD}_{\text{error}}(m,N) \,]^T. \quad (11)$$

The fullband ITD error can be then defined as

$$\overline{\text{ITD}}_{\text{error}}(m) = \frac{\mathbf{w}^T(m)\boldsymbol{\Delta}(m)}{\|\mathbf{w}(m)\|^2}, \quad (12)$$

where $\mathbf{w}(m) = [W(m,1), \ldots, W(m,N)]^T$ is the weighting vector. Using the weighting vector a hard or soft decision can be realized. A hard decision can be realized by using weights that correspond to heavy-side step function. A soft-decision can be realized by using weights that correspond to a logistic sigmoid function. For any real number $y \in \mathbb{R}$, this function can be defined as

$$f(y) = \frac{1}{2} + \frac{1}{2}\tanh\left(\kappa\,(y-\epsilon)\right), \quad (13)$$

where the constant κ defines the slope of the transition of the logistic function and ϵ the transition point. Fig. 2 shows an example of this function for different values of κ and $\epsilon = 0.5$. It is clear that for $\kappa \to \infty$ (13) approximates the heavy-side step function which is equivalent to applying a threshold rule with a threshold value of ϵ.

4.1. Coherence-based weighting

In [12] it is suggested to use the magnitude-squared coherence (MSC) as an indicator of the validity of the estimated ITD. The MSC of the microphone signals is defined as:

$$C_{\text{out}}(m,k) = \frac{|\phi_{v_2v_1}(m,k)|^2}{\phi_{v_1v_1}(m,k)\phi_{v_2v_2}(m,k)}. \quad (14)$$

Finally, the weights $W(m,b)$ can be computed using (13) with

$$y = \overline{C}_{\text{out}}(k,b), \quad (15)$$

where $\overline{C}_{\text{out}}(m,b) = \frac{1}{E_b - E_{b-1}} \sum_{k=E_{b-1}}^{E_b-1} C_{\text{out}}(m,k)$.

4.2. Coherence and RTF based weighting

When a CCS is used, the subband microphone signals can be highly coherent when there is insufficient channel separation. In such a situation, the ITD estimates, and therefore ITD error, will be unreliable [7, 13]. It can be shown that the magnitude of the RTF depends on the channel separation. In this study, we therefore propose to make use of the coherence and the magnitude of the RTF ratio measured at the listener's location.

The normalized mean value of the magnitude of the RTFs at the microphones at time frame m and subband index b is[3]

$$\overline{\mathcal{R}}_{\text{out}}^{\infty}(m,b) = \frac{\overline{\mathcal{R}}_{\text{out}}(m,b)}{\|\overline{\boldsymbol{\mathcal{R}}}_{\text{out}}(m)\|_\infty}, \quad (16)$$

where $\overline{\mathcal{R}}_{\text{out}}(m,b) = \frac{1}{E_b - E_{b-1}} \sum_{k=E_{b-1}}^{E_b-1} |\mathcal{R}_{\text{out}}(m,k)|$ is the mean value of the magnitude of the RTF in subband b and $\overline{\boldsymbol{\mathcal{R}}}_{\text{out}}(m) =$ $[\, \overline{\mathcal{R}}_{\text{out}}(m,1), \ldots, \overline{\mathcal{R}}_{\text{out}}(m,N) \,]^T$.

Finally, the weights $W(m,b)$ are obtained using (13) with

$$y = \overline{\mathcal{R}}_{\text{out}}^{\infty}(m,b)\,\overline{C}_{\text{out}}(m,b). \quad (17)$$

Using these weights, the fullband ITD error will depend mostly on the subband ITD errors for which the microphone signals are coherent and $\overline{\mathcal{R}}_{\text{out}}(m,b)$ is close to the maximum value in $\overline{\boldsymbol{\mathcal{R}}}_{\text{out}}(m)$.

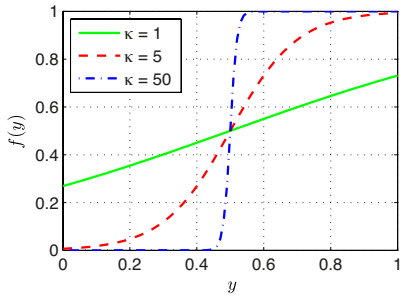


Fig. 2: Example of the logistic function (13) for different values of κ when $\epsilon = 0.5$.

5. EXPERIMENTAL RESULTS

We simulated a two-channel CCS as depicted in Fig. 1. The loudspeakers were placed symmetrically with respect to the center of the head of the listener at a distance of 1.2 m and the span angle between them was set to 30°. The CCFs were calculated for a symmetric position using the fast deconvolution method [14]. We simulated a listener rotating his head from right to left at constant speed of 4 degrees per second.

Three virtual sound scenarios were simulated with different number of virtual sources. The first scenario consisted of a single virtual sound source (a female singer) located at 45°. The second scenario consisted of two virtual sources overlapping in time: the same female singer located at 45° and a saxophone located at −30°. The third scenario consisted of the previous scenario plus a crowd sound located at 0°, i.e. three virtual sound sources overlapping in time. All angles were relative to the median plane and negative angles denoted sources at the right side of the listener. All virtual sources were placed at $r_s = 0.75$ m from the listener and had a duration of 32 s, with a sampling frequency of 48 kHz. The HRIRs used for the CCS and the virtual sources were calculated using the spherical head model [15]. Sensor noise at the microphones was simulated using white Gaussian noise and the signal-to-noise ratio was set to 25 dB.

The process was done on a frame-by-frame basis, with a frame size of $M = 2048$ samples, i.e. 42.7 ms. For each frame the PSDs were estimated using the Welch method with a Hamming window of the same length and 50% overlap. We divided the spectrum into $N = 15$ subbands uniformly spaced between 400 Hz and 6000 Hz on an ERB scale. The stepsize for the recursive implementation of (9) was set to 0.85. The transition point ϵ was set to 0.6, which was found to give the best empirical results. Note that since both coherence and RTFs are used in the proposed ITD error estimation method, a large value of ϵ results in a much stricter decision rule. This will give rise to a larger number of subbands being disregarded, reducing in that way the tracking ability of the algorithm. In the experiments, we also compared the performance of the proposed ITD error estimation using the coherence-based weighting function (15)

[3] The RTFs are normalized with the ∞-norm such that a constant transition point ϵ can be used.

in (13). The transition point was set to $\epsilon_c = 0.9$ for this case, which was found to give the best empirical results with respect to estimation error distributions and median values. The slope κ of the weighting function was varied.

As a reference, we used the ITD error model proposed in [1], namely

$$\mathrm{ITD}_{\mathrm{error}}^{\mathrm{model}}(m) \approx \left(\tau_{22}^{(\mathrm{CCS})} - \tau_{11}^{(\mathrm{CCS})}\right) - \left(\tau_{22}(m) - \tau_{11}(m)\right) \quad (18)$$

where $\tau_{ii}(m)$, $i \in \{1,2\}$, are the delays of the direct paths between the ith loudspeaker to the ith ear corresponding to the head orientation at time frame m and $\tau_{ii}^{(\mathrm{CCS})}$, $i \in \{1,2\}$, are the delays corresponding to the ATFs used to calculate the CCFs. These delays are modeled as [1]

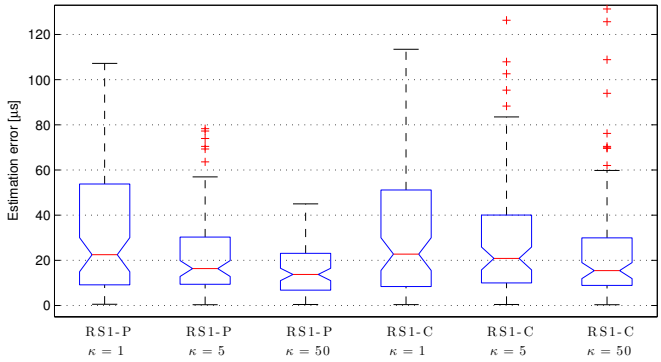
$$\tau_{22} - \tau_{11} = \quad (19)$$
$$\frac{r_\mathrm{s}}{c}\begin{cases} \theta_e - \frac{\theta_\mathrm{s}}{2} - \alpha + \Gamma\left(\frac{\theta_\mathrm{s}}{2} - \alpha\right) & -\frac{\theta_\mathrm{s}}{2} \le \alpha \le -\theta_{\mathrm{lim}} \\ -2\,\alpha & -\theta_{\mathrm{lim}} \le \alpha \le \theta_{\mathrm{lim}} \\ -\theta_e + \frac{\theta_\mathrm{s}}{2} - \alpha - \Gamma\left(\frac{\theta_\mathrm{s}}{2} + \alpha\right) & \theta_{\mathrm{lim}} \le \alpha \le \frac{\theta_\mathrm{s}}{2} \end{cases},$$

where θ_s is the span angle between the loudspeakers, α is the head orientation angle with respect to the middle point between loudspeakers, θ_e is the angle of the ears with respect to the median plane, $\theta_{\mathrm{lim}} = \theta_e - \theta_0 - \frac{\theta_\mathrm{s}}{2}$ with $\theta_0 = \cos^{-1}(1/\rho)$, where $\rho = r_\mathrm{s}/r$ is the normalized distance of the loudspeakers (r_s) with respect to the radius of the sphere (r), and $\Gamma(\Theta) = -\theta_0 + \sqrt{\rho^2 - 1} - \sqrt{\rho^2 - 2\rho\cos(\theta_e - \Theta) + 1}$. In this study we focused only on the estimated ITD error, thus the CCFs were not updated according to the estimated head orientation. We set $\tau_{11}^{\mathrm{CCS}} = \tau_{22}^{\mathrm{CCS}}$ for all m, i.e. $\mathrm{ITD}_{\mathrm{error}}^{\mathrm{model}}(m) \approx \tau_{11}(m) - \tau_{22}(m)$. The estimation error is defined as

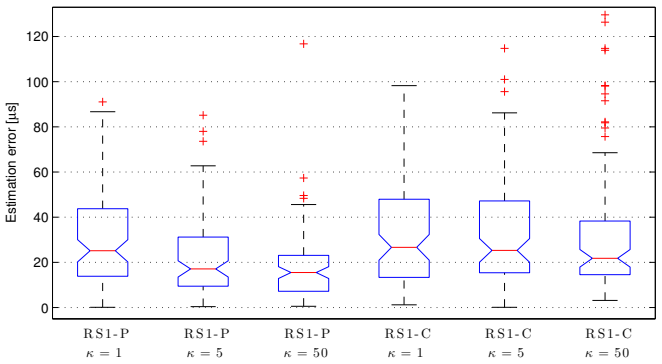
$$e(m) = \left|\mathrm{ITD}_{\mathrm{error}}^{\mathrm{model}}(m) - \overline{\mathrm{ITD}}_{\mathrm{error}}(m)\right|. \quad (20)$$

First we study the distribution of the estimation errors for the three aforementioned scenarios, which are presented in Fig. 3. The proposed weighting function (17) (RS1-P) and coherence-based weighting function (RS1-C) are compared for different values of κ. In general, for RS1-P, the dispersion of the errors decreases with increasing values of κ, while for RS1-C a decrease in dispersion is only observed in the single source scenario. The estimation errors with the coherence-based weighting function exhibit a rather large variance and number of outliers for all scenarios. The performance of RS1-C with $\kappa = 50$ in the multiple source scenarios (Figs. 3(b) and 3(c)) is comparable to the performance obtained with proposed method RS1-P when $\kappa = 1$. In the three sources scenario (Fig. 3(c)), the distribution of the estimation errors of RS1-P with $\kappa = 50$ is skewed towards 10 μs, whereas for RS1-P with $\kappa \in \{1,5\}$ and RS1-C $\forall\kappa$, the distributions are skewed towards 40 μs.

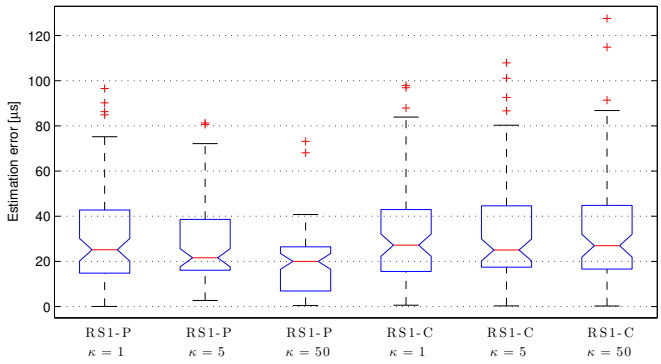
Fig. 4(a) shows the estimated ITD error as function of time for the scenario with three virtual sources. The ITD errors are calculated using the model (18), using the proposed weighting function RS1-P for different values of κ, and using the coherence-based weighting function RS1-C with $\kappa = 50$. Only the last 20 s are presented to get a better overview. The binaural signal as a function of time is shown in the lower panels. We can see that in general, the proposed weighting function approximates better the ITD error model, while the coherence-only method shows a large variability in ITD error estimation. As expected, large deviations are observed after silences in the input signal with RS1-P. On the other hand, errors with RS1-C are clearly independent of the energy of the input signal. The latter can be better seen by zooming in the estimation error results. Fig. 4(b) shows 5 s of the estimation error $e(m)$ as a function of time. With the RS1-C approach, large variations in the estimation error are obtained but no clear correlation with the fluctuations of the input signal is observed. This suggests that the outliers observed in Fig. 3 for the RS1-P can be attributed to errors due to silences in the signal, while outliers for the RS1-C do not necessarily reflect a dependence on the temporal fluctuations of the signal. For this fragment of the sound, the RS1-P with $\kappa = 50$ shows errors in the range of 20 μs, which are in the range of the JNAD of the ITD, while for $\kappa \in \{1,5\}$, the estimation errors are in the range of 40 μs.



(a) One virtual source.

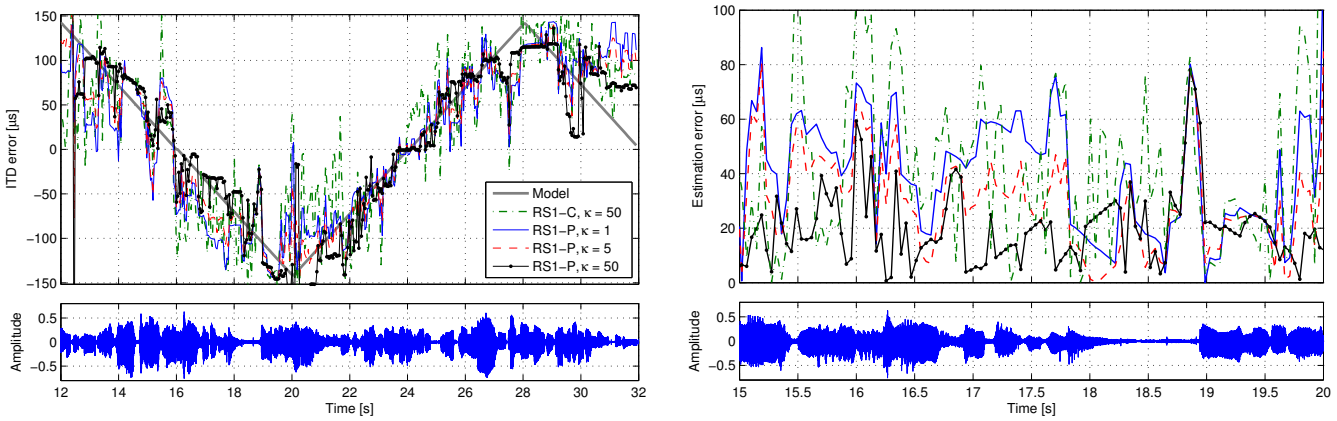


(b) Two virtual sources.



(c) Three virtual sources.

Fig. 3: Boxplots of the estimation error for the different approaches evaluated in three scenarios. The whiskers correspond to the lowest value within 1.5 interquartile range (IQR) of the lower quartile, and the highest value within 1.5 IQR of the upper quartile.



(a) Estimated ITD error between 12 and 32 seconds

(b) Estimation error $e(m)$ between 15 and 20 seconds

Fig. 4: Results obtained using (18), the proposed weighting function (RS1-P), and the coherence-based weighting function (RS1-C) for different values of κ. The left binaural signal is shown in the lower panel.

6. DISCUSSION

We presented a new approach to estimate the ITD error that is used to determine the orientation of the listener's head in a recently proposed CCS with microphone-based head-tracker [1]. We proposed to estimate the subband ITD error based on the RTFs estimation method in [7]. A fullband ITD error is calculated as a weighted sum of the subband ITD errors. The weighting function is based on the coherence and the energy of the RTFs in each subband. The proposed method was evaluated for a CCS reproducing multiple virtual sound sources simultaneously and in the presence of noise. The ITD error model presented in [1] was used as a reference. The performance of the proposed method was assessed for different settings of the weighting function and compared with a coherence-based weighting function. Experimental results show that in the presence of multiple virtual sound sources and noise, it is possible to estimate the ITD error accurately by using a subband analysis of the RTFs. Compared to the approach of selecting the ITDs based only on the coherence, the proposed weighting function showed to be more robust to signal changes introduced by the CCS and sensor noise. As expected, deviations were still observed at time frames where the energy of the input signal was rather low. As shown in [2], these deviations can be mitigated by using a tracking algorithm that accounts for silences in the input signal.

7. REFERENCES

[1] Y. Lacouture-Parodi and E. A. P. Habets, "Crosstalk cancellation system using a head tracker based on interaural time differences," in *Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2012.

[2] Y. Lacouture-Parodi and E. A. P. Habets, "Application of particle filtering to an interaural time difference based head tracker for crosstalk cancellation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[3] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction.," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, Mar. 1992.

[4] C. Tournery and C. Faller, "Improved time delay analysis/synthesis for parametric stereo audio coding," in *120th AES Conv.*, May 20 - 23 2005.

[5] J.M. Yang and H G. Kang, "Two-stage source tracking method using a multiple linear regression model in the expanded phase domain," *EURASIP J. on Advances in Sig. Proc.*, vol. 2012, no. 1, pp. 5, 2012.

[6] R. G. Klump and H. R. Eady, "Some measurement of interaural time difference thresholds," *J. Acoust. Soc. Am.*, vol. 28, no. 5, pp. 859 – 860, September 1956.

[7] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[8] M. Baeck and U. Zölzer, "Real-time implementation of a source separation algorithm," in *Int. Conf. on Digital Audio Effects (DAFx-03)*, London, September 2003.

[9] S Wang, D Sen, and W Lu, "Subband analysis of time delay estimation in STFT domain," *Proc. of the 11th Australian Int. Conf. on Speech Science and Technology*, pp. 211–215, 2006.

[10] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Proc.*, vol. 92, pp. 1950 – 1960, 2012.

[11] A Lombard, Y. Zheng, H. Buchner, and W Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 19, no. 6, pp. 1490–1503, 2011.

[12] C. Faller and J. Merimaa, "Source localization in complex listening situations: selection of binaural cues based on interaural coherence.," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[13] Y. Lacouture-Parodi and P. Rubak, "Sweet spot size in virtual sound reproduction: a temporal analysis," in *Principles and App. of Spatial Hearing*. World Scientific, Singapore, February 2011, ISBN: 978-981-4313-87-2.

[14] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multi-channel systems using regularization," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, no. 2, pp. 189–195, 1998.

[15] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058, Jan. 1998.