THE INFLUENCE OF LOW ORDER REFLECTIONS ON THE INTERAURAL TIME DIFFERENCES IN CROSSTALK CANCELLATION SYSTEMS

Dimitrios Kosmidis, Yesenia Lacouture-Parodi, and Emanuël A. P. Habets

International Audio Laboratories Erlangen†

91058 Erlangen, Germany

Email: dimitrios.kosmidis@iis.fraunhofer.de, ylacoutu@ieee.org, emanuel.habets@audiolabs-erlangen.de

ABSTRACT

Reflections can play an important role in human perception of sound. While they can positively contribute to the perceived sound quality, they may also interfere with the reproduction of, for example, crosstalk cancelled binaural sounds through loudspeakers. In this paper, we study the influence of 1*st* and 2*nd* order reflections on a crosstalk-cancelled desktop reproduction system, through an analysis of the interaural time differences. The direct and reflected sounds are calculated using an image-source model. Meanwhile, the crosstalk cancellation filters are calculated assuming anechoic conditions, and therefore proper sound reproduction is now questionable. In this scenario, the reflections are found to introduce changes in the interaural phase differences and in the interaural group delay. These changes are analyzed and the possible effect on sound localization is investigated using a subjective localization experiment. The results indicated that for the studied setup, the localization accuracy was, practically, unaffected by the low order reflections.

Index Terms— Interaural time differences, crosstalk cancellation, interaural phase delay, interaural group delay, image model.

1. INTRODUCTION

Binaural reproduction is based on the assumption that by properly controlling the sound pressures at the ears of the listener, any virtual sound event can be simulated [1, 2]. In order to properly reproduce binaural signals through loudspeakers, however, one needs to compensate for the acoustic transfer functions (ATFs) between the loudspeakers and the contralateral ears. This can be achieved by incorporating crosstalk cancellation filters (CCFs) into the reproduction chain. Most crosstalk cancellation systems (CCSs) are designed using free-field (anechoic) measurements of the ATFs from the loudspeakers to the ears. When a CCS is used in a reverberant environment, however, there is a mismatch between the anechoic ATFs used to compute the CCFs and the ATFs in the reproduction environment

Reflections are generally known to affect the localization ability of humans [3, 4, 5]. The effect of a single wall or ground reflection was studied in [6], where they found that: a) the values of the binaural cues in the presence of reflections are extended beyond their anechoic range, and b) the reflections distort the interaural time differences (ITDs) at the ears of the listener depending on the time delay between the direct and reflected sounds. In the context of crosstalk cancellation, it has been shown that reflections can considerably degrade the crosstalk cancellation performance [7, 8], and affect the localization accuracy of virtual sources [8].

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

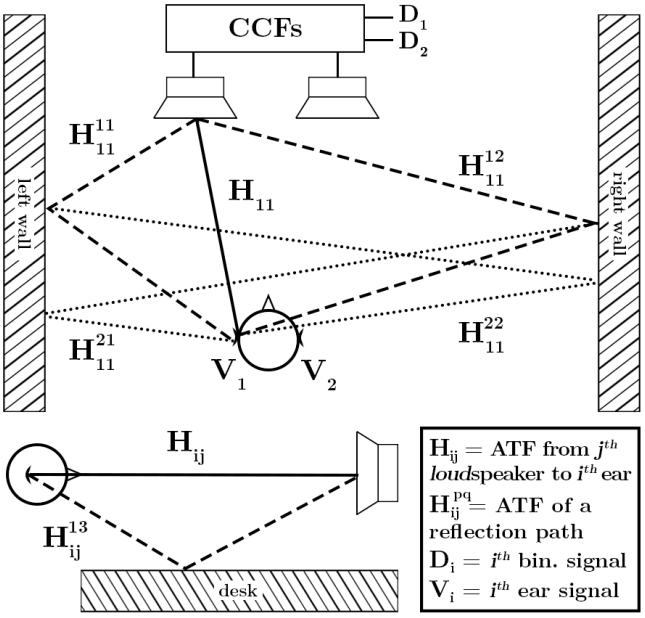

Fig. 1. Diagram of the desktop reproduction setup. Each loudspeaker gives rise to three 1*st* order and six 2*nd* order reflections. Thus, 18 reflections are calculated for each ear. Some of these reflections are shown in the above diagram with dashed lines.

To properly reproduce binaural signals using a CCS, it is important to reproduce the correct ITDs at the ears of the listener [9]. Since the ITDs are affected by reflections, in this paper we examine the influence of low order reflections on the ITDs and on the ability to localize a virtual sound source. In the frequency domain, the ITDs correspond to interaural phase differences (IPDs) and interaural group delays (IGDs). The IPD is defined as the difference of the phase delays, and the IGD as the difference of the group delays, between the signals at the ears of the listener. The phase delay corresponds to a frequency-dependent time shift of the carrier, and the group delay to a frequency-dependent time shift of the envelope of a signal [2, 10]. Moreover, both the IPD and IGD are cues used for localization, especially at low frequencies, where the ITDs are known to be a dominant localization cue [11]. Another important factor is the audibility of the IPDs, which has been studied in [12, 13]. The lowest thresholds of audibility have been found to be in the low frequencies of up to 1 kHz, while it has been suggested that the IPDs mainly result to spatial changes of the sound [13]. In practice, these changes might be critical for the localization of the source.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

This paper is organized as follows: In Section 2, we present the virtual reproduction environment and the models used in the simulations. In Section 3, we analyze the IPD and IGD of the simulated signals at the ears of the listener. In Section 4, we study the effect on localization through a subjective sound localization experiment. In Section 5, we summarize the paper and draw some final conclusions.

2. VIRTUAL REPRODUCTION ENVIRONMENT

2.1. Signal Model

Figure 1 shows a diagram of a two-channel CCS in a typical desktop reproduction setup, where the listener is symmetrically positioned with respect to the loudspeakers. The input of the system are the binaural signals D_i that correspond to a virtual source at a position around the listener. These binaural signals are calculated by convolving a virtual source signal with HRTFs. In this study, these were HRTFs from a KEMAR manikin measured with a source at 2 m distance [14]. For the setup of Fig. 1, only 1*st* and 2*nd* order specular reflections were considered, while the reflecting surfaces (desk, left wall, right wall) were assumed to be of infinite extent.

In the absence of reflections, the sound signals that reach the listener can be defined in the frequency domain as:

$$\underbrace{\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}}_{\mathbf{v}} = \underbrace{\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} D_1 \\ D_2 \end{bmatrix}}_{\mathbf{d}}, \quad (1)$$

where V_i represents the sound pressure at the i-th ear of the listener, D_i the i-th binaural signal, H_{ij} the ATFs from the j-th loudspeaker to the i-th ear, and C_{ij} the corresponding CCFs. In (1) the dependence on the angular frequency ω is omitted for brevity. The CCFs are calculated here for an anechoic setup, where only the direct path ATFs (H_{ij}) are present. In such a scenario, perfect crosstalk cancellation is obtained when $\mathbf{v} = \mathbf{d}$, i.e., $\mathbf{H} \cdot \mathbf{C} = \mathbf{I}$. In this study, the CCFs were computed using the fast deconvolution method proposed in [15].

Equation (1) does not adequately describe the signals that the listener actually perceives in the non-anechoic reproduction setup shown in Fig. 1. In this scenario, the reflected path ATFs between the loudspeakers and the ears also need to be taken into account. This is done by defining $\widetilde{\mathbf{H}}$ as:

$$\widetilde{\mathbf{H}} = \begin{bmatrix} \widetilde{H}_{11} & \widetilde{H}_{12} \\ \widetilde{H}_{21} & \widetilde{H}_{22} \end{bmatrix} \quad \text{where} \quad \widetilde{H}_{ij} = H_{ij} + \sum_{p=1}^{3} \sum_{q=1}^{3} H_{ij}^{pq}. \quad (2)$$

$\widetilde{H}_{ij}$ represents an ATF including both the direct and reflected paths to the ears. Each loudspeaker (j) gives rise to three 1*st* order and six 2*nd* order reflections, which arrive at each ear (i). The index q corresponds to the surface on which a reflection has last impinged before reaching the listener, and p is used for indexing the reflections. For the calculation of the ATFs H_{ij} and H_{ij}^{pq}, the spherical head model of [16] was adopted, and modified to include the propagation delay.

Following the above definition, the loudspeaker signals are obtained by passing the binaural signals through the CCFs. The signals at the ears of the listener are calculated as the superposition of the direct and reflected loudspeaker signals. The CCFs are still calculated using only the anechoic ATFs (H_{ij}), i.e., reflections are not taken into account. In other words, there is a mismatch between the CCFs and the actual ATFs between the loudspeakers and the ears of the listener. This mismatch can degrade the crosstalk cancellation performance and distort the sound perceived by the listener [7, 8].

2.2. Reflection Model

A modified version of the image method [17] was used to simulate the reflections, by considering a reflection factor that is dependent on frequency and source angle (see (3)). Each reflection corresponds to an image source, for which the position is calculated using the geometric reflection model found in [6]. Each image source reproduces an altered version of the signal of the corresponding loudspeaker, which is calculated in the frequency domain by multiplying the original signal with a spherical reflection factor Q. Meanwhile, the loudspeakers are modelled as point sources. Following the above, the signals at the left and right ear of the listener are calculated as the sum of the direct loudspeaker signals and the reflected image source signals. According to the acoustical model of reflections on natural surfaces, also found in [6], Q is defined as:

$$Q(\omega,\sigma) = R(\omega,\sigma) + [1 - R(\omega,\sigma)]\,F(w), \quad (3)$$

where R is the plane wave reflection coefficient, $F(w)$ is a "boundary loss factor", and w is the "numerical distance" [6]. As is shown in (3), Q is dependent on the angular frequency ω, as well as the so-called surface flow resistivity σ, which is a parameter that characterises the reflection of a surface. Using empirically measured σ values, one can calculate the specific acoustic impedances (Z_1, Z_2), and wave numbers (k_1, k_2) of the air and reflection surfaces, which are used in the calculation of R. The latter is also dependent on the position of the image source [6].

3. ANALYSIS OF THE IPD AND IGD

The virtual reproduction environment described in the previous section is used to investigate the behaviour of the signals at the ears of the listener. In the following analysis, the virtual sources are positioned only in the horizontal plane and the virtual source signal consists of an impulse. The loudspeakers are placed at 1 m in front of the listener, with spans of 30° or 60°. The desk, is placed 0.57 m below the head, and the distance to the left and right wall is, respectively, 1.5 and 3 m. The desk is assumed to be made of wood fiberboard with $\sigma = 1.5 \cdot 10^5$, and the walls are assumed to be made of plastered brick surfaces with $\sigma = 10^6$. These values were calculated according to the formulas and porosity values found in [18].

In the studied reproduction setup, the position of the listener, the head orientation, and the position of the loudspeakers (that give rise to reflections) are all fixed. This means that regardless of the virtual source angle, the reflections always arrive at the ears at the same time. The only thing that changes with every angle is the amplitude and phase of the signals that the loudspeakers reproduce, which, in turn, create different reflection signals. Additionally, the reflection signals in the case of 60° loudspeaker span, are stronger than in the 30° span, since the sources are closer to the reflecting walls and thus there is less propagation attenuation for the sound signals.

3.1. Analysis

In our analysis the IPD and IGD were calculated from the summation of the direct and reflected sound signals that arrive at the left and right ears. The IPD is defined as the difference of the phase delays between the signals, and the IGD as its derivative:

$$\mathrm{IPD}(\omega) = \left(-\frac{\phi_{\mathrm{left}}(\omega)}{\omega}\right) - \left(-\frac{\phi_{\mathrm{right}}(\omega)}{\omega}\right) \quad (4)$$

$$\mathrm{IGD}(\omega) = \left(-\frac{d\phi_{\mathrm{left}}(\omega)}{d\omega}\right) - \left(-\frac{d\phi_{\mathrm{right}}(\omega)}{d\omega}\right), \quad (5)$$

where $\phi_{\mathrm{i}}(\omega)$ is the phase of the corresponding aforementioned signals. These frequency-dependent cues are important for localization, and they also dependent on the angle of the virtual source.

Figure 2 shows the maximum IPDs (as perceived by the listener) as a function of the virtual source azimuth. The average thresholds

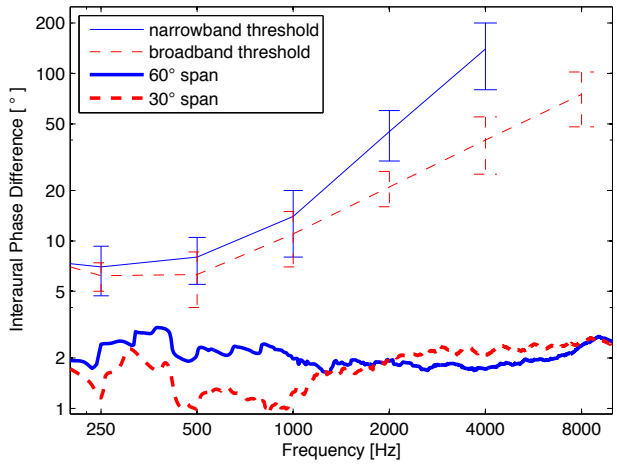

Fig. 2. Maximum absolute IPDs found between all possible azimuths. The average thresholds of audibility for headphone reproduction reported in [12, 13], are also plotted for comparison.

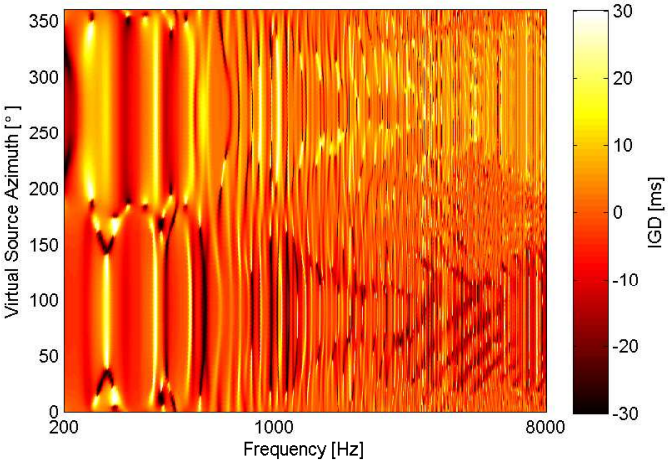

Fig. 3. IGD values as a function of frequency and virtual soure azimuth for 30° loudspeaker span.

of IPD audibility reported in [12, 13] for the headphone reproduction case are also plotted in the same graph for comparison. As can be seen, the IPDs that the reflections created are well below these mean thresholds of audibility. The calculated IGD values are shown in Fig. 3 as a function of frequency and virtual source azimuth for a loudspeaker span of 30°. A similar result was obtained for the span of 60°. A 'ringing' behaviour can be observed in Fig. 3 over all azimuths. This 'ringing' seems to be symmetric between left ($0° - 180°$) and right ($180° - 360°$) positions, especially for frequencies higher than 1 kHz. At lower frequencies, the 'ringing' exhibits wider peaks, as well as a more pronounced effect on the right-side positions. The latter can be explained by the fact that the left wall is simulated at a close distance (1.5 m), and, therefore, contributes to a larger-than-intended left ear signal for those right-side azimuths.

The IPDs that were found are small enough to point to the hypothesis that the phase changes due to reflections will not be detected by most of the listeners. On the other hand, the IGD values found are much larger than the minimum IGD values that can cause a lateral shift of the auditory image, as reported in [10]. The latter, along with the fact that the low frequencies are more critical for the ITDs to be used as localization cues [11], point to the need for further investigation. This is done next through a localization experiment.

Azimuth [°]	**0**	**30**	**60**	**90**	**120**	**150**
binaural	64.2	50.8	26.4	16.0	23.2	36.2
direct	64.9	47.9	26.7	14.4	20.2	32.6
direct + reflected	54.5	49.2	25.0	14.4	24.1	34.2
overall (w/ FB)	*61.2*	*49.3*	*26.0*	*14.9*	*22.5*	*34.3*
overall (w/o FB)	*5.7*	*26.7*	*19.2*	*14.9*	*17.1*	*19.6*
Azimuth [°]	**180**	**210**	**240**	**270**	**300**	**330**
binaural	66.6	34.3	20.4	13.5	25.8	49.9
direct	71.9	26.7	23.5	14.8	25.6	50.5
direct + reflected	71.1	32.4	21.1	12.6	27.8	48.2
overall (w/ FB)	*69.9*	*31.1*	*21.7*	*13.7*	*26.4*	*49.5*
overall (w/o FB)	*7.4*	*19.7*	*17.1*	*13.7*	*20.5*	*26.4*

Table 1. Mean absolute localization error (in °) for different scenarios. The overall error is calculated with and without the front-back (FB) confusions. 0° correspond to a source directly in front of the listener, while azimuth increases counter-clockwise. The resolution of the answering interface was 10°.

4. LOCALIZATION EXPERIMENT

4.1. Design

An experiment was conducted to examine the localization accuracy in the presence of low order reflections. The experiment took place in a soundproofed listening booth in the facilities of Fraunhofer IIS. Sound reproduction was through the open-type STAX SR-Lambda Professional electrostatic headphones, connected to a STAX SRM-600 amplifier. The headphones simulated the loudspeaker reproduction conditions depicted in Fig. 1. In other words, the listeners were directly presented with a simulation of the ear signals that would occur under the depicted conditions.

Male speech, pink noise, and castanets, were rendered under three scenarios for comparison: a) plain binaural signals (*binaural*), b) crosstalk cancelled binaural signals containing only the direct signals of the loudspeakers (*direct*), and c) crosstalk cancelled binaural signals containing the direct and reflected signals of the loudspeakers (*direct + reflected*). For each of these 9 stimuli, a total of 12 azimuths ($0° : 30° : 330°$) were played. The order of presentation of the different stimuli was randomized across subjects using 9x9 Latin squares [19], while the order of each set of 12 azimuths was randomized in every test. The stimuli were also evaluated for reproduction conditions of 30° or 60° loudspeaker span.

Twelve volunteers between the ages of 23 and 41, with a median age of 28, participated in the experiment. The subjects were either students or researchers in the field of audio, some were aware of the purpose of the experiment, none had knowledge on the details, and almost none were familiar with localization experiments. The location was entered through a graphical user interface that displayed a picture of a head surrounded by azimuths distributed in 10° steps. The experiment consisted of a 15-minute training, and four main sessions of 10 minutes each. The training and main sessions were taken on different days, while each subject took two sessions per day. Two of the sessions included sounds under 30°, and the other two sounds under 60° loudspeaker span. The order of these sessions was randomized across subjects using 4x4 Latin squares.

4.2. Results

The results of the localization experiment are shown in Fig. 4, where the combined answers of all subjects (5184 in total) are plotted, grouped by sound, scenario, and loudspeaker span. The answers are distinguished into normal answers (blue 'x' markers) and front-back (FB) confusions (red '+' markers). An answer is regarded as

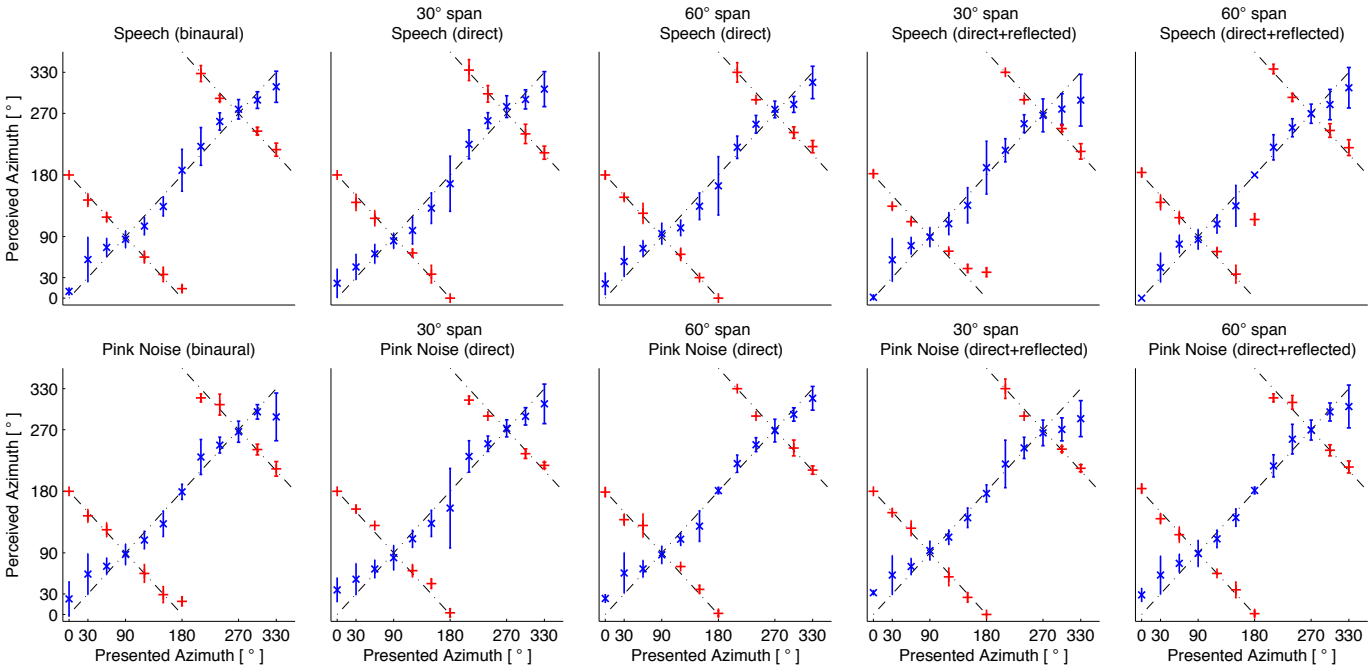

Fig. 4. Combined answers of all the subjects of the localization experiment, grouped by sound, scenario, and loudspeaker span. For each of the presented azimuths, the mean perceived azimuth and its deviation are plotted. The answers are separated into normal answers (blue 'x' markers) and front-back confused answers (red '+' markers). The results of the castanets are omitted, but were found to be very similar.

a FB confusion if its azimuth corresponds to the symmetric position of the correct answer's azimuth in the opposite hemisphere (front or back) and up to $\pm 20^\circ$ from it. Taking into account the considerations mentioned in [20], an exception is made for the azimuths up to $\pm 15^\circ$ from the 90° and 270° positions, where incorrect answers are considered as normal errors.

The plots of Fig. 4 did not reveal any significant differences in the localization performance between the various cases. The results were, therefore, further analysed with the help of statistical analysis. A multi-factor analysis of variance (ANOVA) was carried out, with the absolute localization error as the dependent variable, and the sound, scenario, loudspeaker span, presented azimuth, and their 2- and 3-way combinations as independent variables. The only parameters that were found to have a significant effect (at the 0.05 level) on the absolute localization error were: a) the loudspeaker span ($F_1 = 5.871$, $p = 0.015$), and b) the presented azimuth ($F_{11} = 74.749$, $p < 0.001$). In the pairwise comparisons, the 60° loudspeaker span showed a mean absolute error 3° smaller than the 30° span ($p = 0.015$), which is for most applications insignificant. Regarding the presented azimuth, 55/66 pairwise comparisons showed a statistically significant difference in the mean absolute error ($p \leq 0.024$). This mean absolute error is shown in Table 1 for each presented azimuth. It can be seen that the biggest errors appear for the 0° and 180° positions when the FB confusions are included, but when the FB confusions are removed from the answers, the biggest errors are noted for azimuths between the center (0°, 180°) and side positions (90°, 270°). It is also noted that most of the FB confusions were observed in the 0° and 180° azimuths.

The analysis of the results suggests that the localization performance of the listeners was not affected by the sound (speech, pink noise, castanets), scenario (binaural, direct, direct+reflections), or loudspeaker span. The only factor that actually influenced localization was the presented azimuth. In fact, the effect of the source position on the localization performance (Table 1) matched the behaviour that is expected from the natural limitations of human hearing [2]. This result also confirms the observed similarity of the plots of Fig. 4, where small differences between the plots can now be attributed to random errors. Finally, the FB confusions did not seem to increase with the addition of reflections. In fact, they constituted $\sim$17% of the answers in every scenario, and therefore, they can only be attributed to the limitations of headphone reproduction and the use non-individualized HRTFs.

5. DISCUSSION

The presence of reflections is generally known to interfere with sound reproduction and perception. More specifically, reflections can deteriorate the performance of a CCS [7], and/or the sound localization accuracy [8]. In this paper, we investigated this matter further by analysing the ITDs and evaluating the localization ability of listeners in the presence of 1*st* and 2*nd* order reflections using a headphone-simulated desktop reproduction setup. In our study, we observed that the reflections created only small changes in the IPDs of the signals that arrive at the ears of the listener. This suggested that the localization ability would not be disturbed by the presence of reflections. On the other hand, reflections were found to cause a big variation of the IGD, which could affect the sound localization accuracy. This was further investigated through a subjective localization experiment. Contrary to [8], the results of the experiment showed that the reflections did not have a significant influence on the localization accuracy, which also hints to a well-behaved performance of the CCS in a moderately reverberant environment.

The effect of reflections on binaural reproduction through loudspeakers (real or simulated) in different environments is a topic for future research. In addition to crosstalk cancellation performance and sound localization, the evaluation of sound attributes, such as the perceived width of the sound source, should be investigated.

6. REFERENCES

[1] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, pp. 171–218, 1992.

[2] J. Blauert, *Spatial Hearing*, Hirzel-Verlag, 3rd edition, 2001.

[3] W. M. Hartmann, "Localization of sound in rooms," *J. Acoust. Soc. Am.*, vol. 74, no. 5, pp. 1380–1391, 1983.

[4] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms, ii: the effects of a single reflecting surface," *J. Acoust. Soc. Am.*, vol. 78, no. 2, pp. 524–533, 1985.

[5] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, 1999.

[6] B. Gourévitch and R. Brette, "The impact of early reflections on binaural cues," *J. Acoust. Soc. Am.*, vol. 132, no. 1, pp. 9–27, 2012.

[7] D. B. Ward, "On the performance of acoustic crosstalk cancellation in a reverberant environment," *J. Acoust. Soc. Am.*, vol. 110, pp. 1195–1198, 2001.

[8] A. Sæbø, *Influence of reflections on crosstalk cancelled playback of binaural sound*, Ph.D. thesis, Norwegian University of Science and Technology (NTNU), 2001.

[9] Y. Lacouture-Parodi and P. Rubak, "Sweet spot size in virtual sound reproduction: a temporal analysis," in *Principles and App. of Spatial Hearing*, Y. Suzuki, D. Brungart, Y. Iwaya, K. Lida, D. Cabrera, and H. Kato, Eds. World Scientific, Singapore, February 2011, ISBN: 978-981-4313-87-2.

[10] H. Møller, P. Minnaar, S. K. Olesen, F. Christensen, and J. Plogsties, "On the audibility of all-pass phase in electroacoustical transfer functions," *J. Audio Eng. Soc*, vol. 55, no. 3, pp. 113–134, 2007.

[11] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661, 1992.

[12] G. Martin and S. Choisel, "Audibility of phase response differences in a stereo playback system. Part 1: Headphone reproduction of wide-band stimuli," in *Proc. 124th AES Convention*, Amsterdam, The Netherlands, May 2008.

[13] S. Choisel and G. Martin, "Audibility of phase response differences in a stereo playback system. Part 2: Narrow-band stimuli in headphones and loudspeakers," in *Proc. 125th AES Convention*, San Francisco, CA, USA, October 2008.

[14] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *Proc. 130th AES Convention*, May 2011.

[15] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 189–194, 1998.

[16] 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.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[18] T. J. Cox and P. D'Antonio, *Acoustic Absorbers and Diffusers: Theory, design and application*, Taylor & Francis, Second edition, 2009.

[19] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*, John Wiley & Sons, Inc., 2006.

[20] R. L. Martin, K. I. McAnally, and M. A. Senova, "Free-field equivalent localization of virtual audio," *J. Audio Eng. Soc*, vol. 49, no. 1/2, pp. 14–22, 2001.