

SEGMENTATION AND ANALYSIS OF EARLY REFLECTIONS FROM A BINAURAL ROOM IMPULSE RESPONSE

Sampo Vesa and Tapio Lokki

Department of Media Technology,
Helsinki University of Technology
P.O. Box 5400

FI-02015, TKK, Finland

sampo.vesa@iki.fi, tapio.lokki@tkk.fi

ABSTRACT

In this paper, a novel method for analysis of binaural room impulse responses is presented. Individual reflections are localized in time and frequency from a measured binaural room impulse response based on the continuous cross-wavelet transform (XWT). The directions of the reflections are then analyzed based on KEMAR and CIPIC reference data lookup, and compared a previous approach. Analysis of the directions and arrival times of reflections allows detailed study of measured binaural room impulse responses. The reflections can also be resynthesized based on the continuous wavelet transform (CWT) and spread apart in time, resulting in a slow-motion room impulse response that can be useful for room acoustics design as well as in teaching room acoustics.

1. INTRODUCTION

Binaural room impulse responses (BRIRs) are encountered in many audio applications. The binaural impulse responses potentially hold a lot of useful information, which can not be directly seen in the time and frequency domain representations of the responses. One could ask, for example, where the individual reflections are located in time, what is their frequency content, and what directions are they coming from. Because a lot of binaural room impulse responses have been measured from concert halls, a method for extracting this kind of information could be useful.

The problem of time-localizing reflections from a binaural room impulse response has already been partly tackled by the authors [1]. The previously presented method allows the localization of reflections in time, but ignores the frequency dimension and does not allow estimating the direction-of-arrival. In the current work, the reflections are localized in both frequency and time using a segmentation algorithm borrowed from image processing. The segmentation algorithm is applied to the continuous cross-wavelet transform (XWT), which is basically the cross-spectrogram between two continuous wavelet transforms (CWT). When the time-frequency areas where each reflection resides are located, the azimuth angle of each reflection is estimated by comparing the interaural parameters of the reflection to a lookup table constructed from measured head-related transfer function (HRTF) data. Two HRTF databases are used: the KEMAR [2] and CIPIC [3] databases. The time-domain reflections can also be extracted by inverting the CWT at the time-frequency regions which the reflections occupy. These reconstructed reflections are used for slow-motion auralization [4]. Comparisons to a previous approach are also made. The approach is based on a method for time-localizing early reflec-

tions proposed by Kuster [5]. In addition, the azimuth angles of the reflections are estimated by calculating the cross correlation and mapping the lag of the maximum to azimuth angle.

The problem of detecting arrival times of room reflections has been investigated before by other authors. Kuster [5] used adaptive thresholding of the time-domain response to detect arrival times of early reflections from a single channel measured room impulse response. It was reported that the method can confidently detect 1–5 early reflections. Defrance et al. [6] reported a method for detecting the arrival times of reflections from a monaural room impulse response measured acoustically by firing a pistol. The direct sound part was used as an atom in a matching pursuit algorithm, so that the arrival times of sound rays were detected. However, the correspondence of the detected arrival times with the main early reflections in the room was not tested.

Previous approaches in localizing individual reflections in room impulse responses also include the work of Roper and Collins [7, 8], who applied microphone arrays to localize reflections in a listening room for purposes of room compensation in loudspeaker listening. The method involves the emitting of chirp pulses, matched filtering, matching pursuit, time-difference-of-arrival (TDOA) estimation, and the image source method. The combination of these methods permits localizing the listener, the sound sources, and the image sources (reflections) in a room. The results indicate that with this kind of a system it is possible to localize the first and second order image sources correctly. However, the method is not based on analysis of impulse responses measured the standard way. It also requires the acquisition of multichannel signals from a microphone array. Other approaches that utilize multichannel signals include the work of Gover et al. [9], Park and Rafaely [10], and Rafaely et al. [11]. The approach proposed in the current study differs from the previous approaches in the way that it is based on binaurally recorded impulse responses measured in the standard way using sweep or MLS signals.

Wavelet methods have been used in analysis of room impulse responses before. Loutridis [12] described how the continuous wavelet transform can be used for decomposing room and loudspeaker impulse responses, and for estimating modal frequencies and the reverberation time. Other audio applications of the continuous wavelet transform include noise reduction and signal compression [13], intermodulation effects analysis [14], sound synthesis [15] and sound signal modeling [16]. The wavelet decomposition has been used for approximating room impulse responses in simulations [17].

This paper is structured as follows. First, the wavelet analy-

sis method for binaural room impulse responses will be presented, along with the proposed reflection segmentation method. Then, the method for estimating the azimuth angles of the reflections is presented and evaluated. Finally, an application of the proposed reflection segmentation method to slow-motion auralization is discussed.

2. WAVELET ANALYSIS FOR EXTRACTING REFLECTIONS

In order to extract individual reflections from a binaural room impulse response, the reflections have to be first localized in time and frequency. Previous work of the authors concentrated on localizing the reflections in time using the continuous wavelet transform [1]. That work is extended in this section to include the frequency dimension as well.

2.1. Continuous wavelet transform

The continuous wavelet transform (CWT) for a discrete sequence $x(n)$ is defined by the equation (adapted from [18])

$$W_x(n, s) = \frac{1}{\sqrt{s}} \sum_{n'=0}^{N-1} x(n') \psi_0^* \left(\frac{n' - n}{s} \right) \quad (1)$$

where n is the discrete time index, N is the length of the discrete time series $x(n)$, s is the scale, n is the translation, $\psi_0(\eta)$ is a complex valued wavelet function (sometimes termed the mother wavelet). Complex conjugation is denoted by asterisk (*). The equation basically correlates scaled and translated wavelet functions with the input sequence in order to build a time-frequency representation of the sequence. In practice, the equation is implemented in the frequency domain. It should be noted that all quantities in this section are non-dimensional, unless indicated otherwise. For the sake of clarity, only discrete versions of the equations are given (a continuous-time treatment of the CWT can be found in e.g. [12]).

In this work, the wavelet function used is the Morlet wavelet [18]

$$\psi_0(\eta) = \pi^{-1/4} e^{j\omega_0\eta} e^{-\eta^2/2} \quad (2)$$

where t is a dimensionless time parameter and ω_0 is a dimensionless oscillating period of the wavelet. The oscillating period determines the frequency resolution of the CWT [12] and here it is set to $\omega_0 = 6$, a value also used in [18].

It is often convenient to convert the scale to frequency in Hertz. For the Morlet wavelet, it proceeds as follows. First, the relationship between the scale s and the so called equivalent Fourier period λ is given as [19, 18]

$$\lambda = \frac{4\pi s}{\omega_0 + \sqrt{2 + \omega_0^2}}, \quad (3)$$

which is converted to frequency (in Hz) by

$$f = \frac{f_s}{\lambda} = f_s \cdot \left(\frac{\omega_0 + \sqrt{2 + \omega_0^2}}{4\pi s} \right) \quad (4)$$

where f_s is the sampling frequency in Hz.

When binaural signals are analyzed, the reflections can be localized based on the magnitude of the continuous cross-wavelet transform (XWT) (adapted from [20])

$$|W_{LR}(n, s)| = |W_L(n, s)W_R^*(n, s)| \quad (5)$$

where $W_L(n, s)$ and $W_R(n, s)$ are the CWTs of the left and right ear signals, respectively.

The set of scales s_j included in the transform can be conveniently described as power of two as (adapted from [18])

$$s_j = s_0 2^{j\delta_j}, \quad j = 0, 1, \dots, J \quad (6)$$

$$J = \frac{1}{\delta_j} \log_2 \left(\frac{s_{\max}}{s_0} \right) \quad (7)$$

where δ_j is the scale resolution, J is the total number of scales minus one, s_0 is the minimum scale, and s_{\max} is the maximum scale. In the wavelet transform computations in this work, the values used were $\delta_j = 1/32$, $J = 288$, $s_0 = 2$, and $s_{\max} = 1024$. This results in 289 different scales. However, scales corresponding to frequencies below approximately 300 Hz (scale indices j larger than 199) were discarded, because the interest was on reflections that are well-localized in time, and not on the room modes.

2.2. Segmenting the reflections

After calculating the XWT, the reflections are localized in time and frequency utilizing a segmentation procedure. Since the XWT can be seen as a gray-scale image, the watershed segmentation algorithm was chosen for segmenting [21]. The watershed algorithm is a basic algorithm for segmenting gray-scale images. The `watershed` function of the Image Processing Toolbox was used with default parameters for the segmentation in MATLAB.

Fig. 1 illustrates the process of segmentation. First, the base-2 logarithm of the magnitude (absolute value) of the XWT is taken, and the result is scaled so that the maximum is at zero (top panel in Fig. 1). A thresholding operation is then applied to discard the parts of the response that have small correlation between left and right ear signals, compared to the direct sound, which is usually where the maximum is located (middle panel in Fig. 1). The threshold values used in the experimental part of this work were between 9–14 decibels below the maximum. The thresholded XWT is then scaled between $[0, 1]$, the discarded parts are set to minus infinity, and the resulting image is then passed to the watershed algorithm, resulting in a matrix where the segments are marked with ascending numbers from left to right and top to bottom. The bottom panel of Fig. 1 illustrates this segmentation by marking each segmented reflection with a different shade of gray.

After the segmentation, some fine tuning is still needed. The segmentation may result in segments that are excessively large or small in terms of area on the time-frequency plane. Therefore, limits for acceptable segment area are set. In this work, the segments had to have an area in the range of 300–50000 “pixels” (when the XWT is seen as a 2D image). The top panel of Fig. 2 shows this final segmentation result, where the segment areas are in the aforementioned range. It can be seen that small segments present in some of the “holes” (in bottom panel of Fig. 1) have disappeared. In this case there were no excessively large segments.

One also needs to be aware that the direct sound is sometimes broken into multiple segments by the segmentation algorithm. Therefore, it may be necessary to manually choose to combine a few of the first segments. It is also possible that reflections are merged to one segment. Both of these problems happen at various parts of the response, due to reflections overlapping in time and frequency, which is why they can not be avoided completely.

The proposed method of segmenting reflections is compared to a previous monaural approach for time-locating early reflections proposed by Kuster [5]. This baseline method utilizes an adaptive

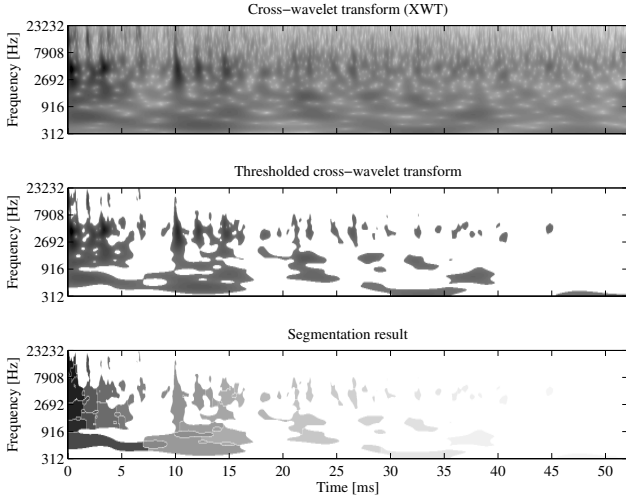


Figure 1: Segmentation of a BRIR. Top panel: the XWT of a bin-aural room impulse response. Middle panel: thresholded XWT. Bottom panel: segmentation result.

thresholding, where the magnitude mean of the impulse response $h(n)$ at time index n is first calculated as (adapted from [5])

$$\mu(n) = \frac{1}{N_\mu} \sum_{m=n-N_\mu}^{n+N_\mu} |x(m)| \quad (8)$$

where $N_\mu = \lfloor (T_\mu f_s / 2) \rfloor$ is the number of samples corresponding to half of the length of the averaging filter of length T_μ seconds. Rounding to nearest integer is denoted by $\lfloor \cdot \rfloor$. Based on the local mean $\mu(n)$, a binary signal containing the reflection locations is calculated as (adapted from [5])

$$h_p(n) = \begin{cases} 0, & \forall h(n) < \epsilon \mu(n) \\ 1, & \forall h(n) \geq \epsilon \mu(n) \end{cases} \quad (9)$$

where ϵ is a parameter which defines the threshold. The lower the value of ϵ , the more sensitive the algorithm is to detect reflections. Because it was found that the algorithm finds multiple sequential peaks corresponding to a single reflection, a one-dimensional dilation operation is applied to $h_p(n)$. The dilation is performed with a structuring element [111111] so that peaks close to each other are combined to a contiguous sequence. The times of reflections are then taken to be the indices of the center points of contiguous sequences of ones in $h_p(n)$.

2.3. Reconstructing the reflections

Later in this study (see Sections 3 and 4), the time-domain reflections corresponding to the segmentation have to be recovered. Based on the segmentation of the XWT, each reflection can be reconstructed by using the wavelet transform reconstruction formula which reconstructs the time-domain signal as a sum of real parts of the wavelet transform $W_x(n, s_j)$ inside the bounding box of the segmented reflection in question, i.e., over a set of scales ranging from scales indices J_{\min} to J_{\max} and time indices N_{\min} to N_{\max}

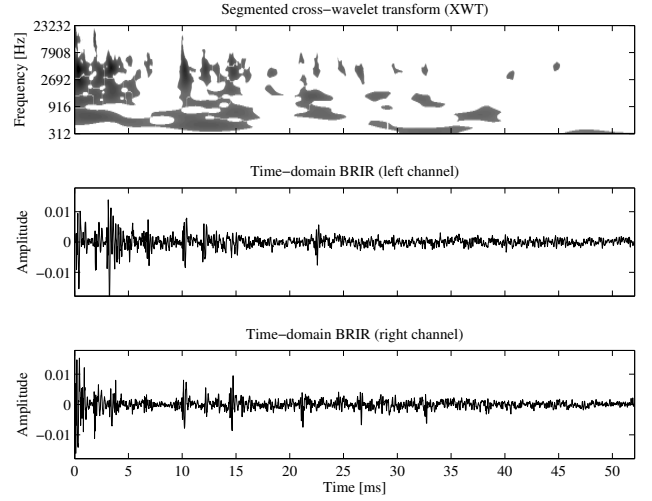


Figure 2: Segmented BRIR. Top panel: segmented XWT. Middle and bottom panels: left and right channels of the corresponding BRIR, respectively.

(adapted from [18])

$$x(n) = \frac{\delta j}{C_\delta \Psi_0(0)} \sum_{j=J_{\min}}^{J_{\max}} \frac{\Re\{W(n, s_j)\}}{\sqrt{s_j}}, \quad n \in [N_{\min}, N_{\max}] \quad (10)$$

where δj is the scale resolution ($\delta j = 1/32$ used in this work), C_δ is a reconstruction factor dependent on the wavelet function ($C_\delta = 0.776$ for the Morlet wavelet used here), and $\Psi_0(0)$ is a scaling factor ($\Psi_0(0) = \pi^{-1/4}$ for the Morlet wavelet). The scale is denoted by s_j . The reconstruction of Eq. (10) is applied to the CWTs of the left and right channel signals separately. The XWT is only used for the segmentation.

Fig. 3 shows an example of the reconstruction for a single reflection, which has its correlation peak between 14.5–15 ms. The top panel shows the original time-domain response during the time interval the reflection occupies. In the middle panel, the reflection reconstructed using Eq. (10) is depicted. Both the left and right ear signals resemble a sine-like waveform. It can be seen that this particular reflection is localized to the left, because the left channel waveform has a larger amplitude and the left channel waveform precedes the right channel waveform by approximately 0.7 ms. This can be seen as the time difference between the highest peak of the left channel signal close to 14 ms and the next peak of the right channel signal just after the 14.5 ms mark. The bottom panel shows the segmented part of the XWT, which specifies the bounding box inside which Eq. (10) is evaluated. The correlation peak is seen as the darkest area in the bottom panel around 14.5–15 ms. Even though the original signal in the top panel has a peak around 13.3 ms in the left channel, this is not the reflection that is segmented here, as the actual correlation peak is in a frequency range close to 1000 Hz (the darkest area in the XWT in bottom panel).

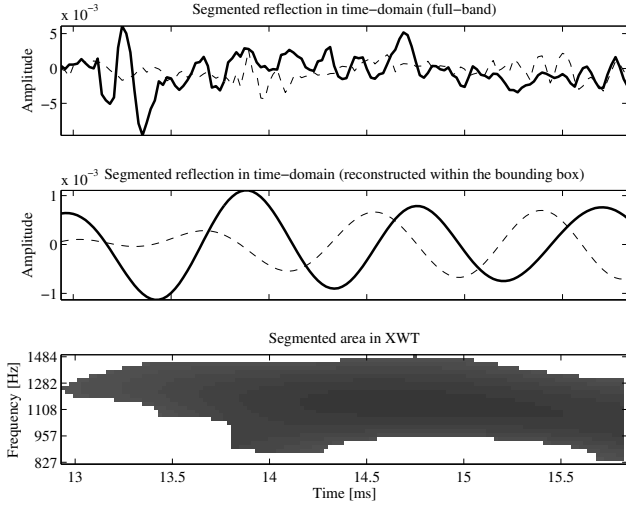


Figure 3: Reconstruction example of a single reflection. Top panel: the full-band reflection in the time domain. Middle panel: the reflection reconstructed using the bounding box. Bottom panel: the corresponding segmented area of the XWT. In the first two plots, the left and right channel signals are illustrated by bold and dashed lines, respectively.

3. ESTIMATING THE AZIMUTH ANGLE OF THE REFLECTIONS

In order to study the segmented reflections in more detail, a method for localizing the azimuth angle of the reflections was implemented. The method is based on matching the estimated interaural time differences (ITDs) and interaural level differences (ILDs) computed from a reflection to reference values obtained from the KEMAR [2] and the CIPIC [3] HRTF databases.

3.1. Azimuth angle estimation method

First, reference values of interaural parameters are computed for each elevation and azimuth angle combination. Palomäki et al. [22] have reported that when learning sound source direction with a neural network, elevation angle estimation requires some head rotation information, which can be simulated by using localization cues from two azimuth angles (head rotations) simultaneously. The problem here may be even harder because the goal is to localize a very short sound, i.e. a single reflection, which also typically has a narrow bandwidth. The entire KEMAR and CIPIC data sets, with all elevations, are nevertheless used in order to avoid cone-of-confusion problems [23] which would happen if the elevation angle were forced to zero. The elevation angle is only used for matching with the reference HRTF data, since it can not accurately be estimated from the individual reflections.

A previously used way [24] of extracting ITD and ILD parameters from HRTFs for binaural localization of speech is to use a filterbank (gammatone, for example) and then feed it with HRTF-processed sinusoids having frequencies equal to the center frequencies of the filter bank. The means of the interaural cues calculated from the corresponding filter outputs are then used to construct a map which links the ILD and the ITD cues at each fre-

quency band to the azimuth angle. In this work, a white noise burst of 100 ms length is convolved with the HRIRs. The CWTs for the left and right channels are then calculated using the convolved burst as input. The reference ITD and ILD values are then calculated from the CWTs.

The ILD reference values for each frequency band f are calculated from the CWTs of the white noise bursts convolved with the KEMAR/CIPIC HRIRs in the time domain using the formula¹

$$ILD(f) = 20 \times \log_{10} \left(\frac{\sum_n |\Re \{W_R(n, f)\}|}{\sum_n |\Re \{W_L(n, f)\}|} \right) \quad (11)$$

where $W_L(n, f)$ and $W_R(n, f)$ represent the complex-valued CWTs of the white noise bursts convolved with the left and right ear HRIRs, respectively. The corresponding ITD reference values were calculated using the `delayestm.m` function [25], with $\Re \{W_L(n, f)\}$ and $\Re \{W_R(n, f)\}$ as the input signals. The aforementioned function basically calculates the cross-correlation and then refines the peak location estimate (which is the ITD) using upsampling. The function was modified so that the peak is only searched between lags of -1 ms and +1 ms, which is a realistic range for the ITD. The KEMAR database contains only one data set of the KEMAR dummy-head, while the CIPIC database has 45 subjects, including the KEMAR head. Therefore, for the CIPIC database, mean of the interaural parameters computed from all subjects was used as the reference data.

When the reference interaural parameters have been calculated, it is possible to localize the individual reflections by matching the ILD and ITD of each segmented reflection. The interaural parameters are calculated from the CWTs of the left and right channel inside the bounding box of the reflection in question, in the same manner as for the KEMAR/CIPIC HRIRs. Since the ILD is only useful at higher frequencies and the ITD at lower frequencies, an ITD/ILD crossover frequency of $f_c = 1.5$ kHz was found good for matching. The matching is done simply by comparing the ILD and ITD values at each frequency band of the reflection as defined by the bounding box, using

$$(\theta, \phi) = \arg \max_{(\theta, \phi)} \begin{cases} -\sum_{f_{\min}}^{f_{\max}} (ITD_{\text{ref}}(f, \theta, \phi) - ITD(f))^2, & f \leq f_c \\ -\sum_{f_{\min}}^{f_{\max}} (ILD_{\text{ref}}(f, \theta, \phi) - ILD(f))^2, & f > f_c \end{cases} \quad (12)$$

where $ITD_{\text{ref}}(f, \theta, \phi)$ and $ILD_{\text{ref}}(f, \theta, \phi)$ are the reference ITD and ILD values at frequency f for azimuth angle θ and elevation angle ϕ calculated from the KEMAR/CIPIC data. The elevation angle in the KEMAR data set ranges from -40° to $+90^\circ$ in 10° steps. The number of azimuth angles per elevation varies with the elevation angle, having a resolution of 5° for elevations from -20° to $+20^\circ$ and less at lower and higher elevations. In total the data set consists of 710 locations. In the CIPIC data set, the elevation angle has a resolution of 5.625° , ranging from -40° to $+230.625^\circ$. The azimuth angles sampled in CIPIC are -80° , -60° , -55° , -45° to $+45^\circ$ in 5° increments, $+55^\circ$, $+60^\circ$, and $+80^\circ$. The total number of locations in the data set is 1250.

In the baseline method, the lag corresponding to the maximum value of the cross-correlation between the signals within a 1.3 ms window, centered at each reflection detected using the method by Kuster [5] (as described in Sec. 2.2), is mapped to the azimuth

¹Conversion between scale and frequency is done with Eq. (4). Frequencies are used exclusively from now on.

angle using (adapted from [26])

$$\theta = \sin^{-1} \left(\frac{\tau_{\max} \cdot c}{d_{\text{head}}} \right) \quad (13)$$

where τ_{\max} is the lag of the cross-correlation maximum (in seconds), c is the speed of sound, and $d_{\text{head}} = 0.2$ m is the diameter of the head.

3.2. Evaluation of azimuth angle estimation from segmented reflections

The azimuth angle estimation method was tested with four different responses — measured and simulated responses of two different listener/source configurations in a lecture hall with dimensions $12 \text{ m} \times 7.3 \text{ m} \times 2.6 \text{ m}$. The sound source was at height of 1.2 m and the height of the listener was 1.7 m. The lecture hall is illustrated in Fig. 4. The recordings were real-head recordings made using small electret microphones, and the simulated responses were produced by the image-source method [27] with reflections included up to the 4th order, as well as with first order edge diffraction and late reverberation modeling [28]. The time delays and arrival angles of reflections in the simulated responses were known a priori, which permits comparisons to the reflections segmented and analyzed by the proposed algorithm.

Figures 5–8 illustrate the reflections segmented by the proposed algorithm. The analysis is only performed up to 30 ms from the direct sound, which corresponds to 10.2 meters of sound propagation in the air when the speed of sound is 340 m/s. The reflections arriving later than 30 ms are much weaker and it is hard to tell whether they are just caused by statistical fluctuations or actual surface reflections in the room. The top panels of Figs. 5–8 plot the segmented reflections with asterisks (*), and the simulated reflections with crosses (×). For each of the four responses, the threshold value (see Sec. 2.2) is set to a value that results in the algorithm detecting as many reflections as were present in the simulated responses between 0–30 ms from the direct sound. This is necessary in order to have fair comparisons between the different methods. The azimuth angles are shown as a function of time, in order to show both the time and azimuth angle estimation performance of the proposed algorithm. The segmented reflections closest in time to each of the simulated reflections are joined by solid lines. The middle panels of Figs. 5–8 show the absolute time errors between each simulated reflection and the closest segmented reflections. The bottom panels show the corresponding azimuth angle errors.

Figures 5 and 6 show how reflections segmented from simulated room impulse responses are localized by the algorithm. From the top panels one can see that the algorithm finds reflections close to the simulated reflections. This is to be expected as there is a correspondence between the model used in the simulation and the impulse response that was analyzed, because the impulse response was generated from the model. From the time and azimuth error plots one can conclude that the time localization error is mostly < 1 ms, and the angle localization error $< 40^\circ$. However, there is a larger azimuth error for most reflections arriving after 20 ms and between 10–15 ms for the first and second receiver positions, respectively. For the first receiver position (Fig. 5) the large time errors for these reflections indicate that the reflections were actually missed by the algorithm, probably because they are too weak in amplitude. In the case of the second receiver position (Fig. 6) between 10–15 ms, there are modeled reflections close to each other

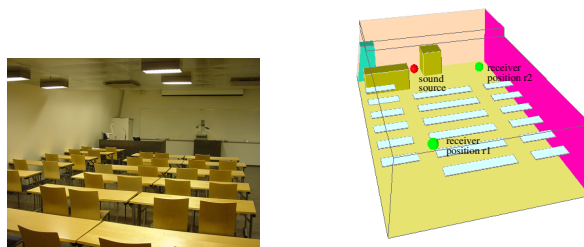


Figure 4: The measurement room. Left: a photo of the room. Right: a room model used for simulations with the source and receiver positions.

in time, and they come from several directions. Because the reflections in the simulated responses are wide-band, there is significant overlap in frequency as well and the correct angles can not be recovered. Instead, the reflections are localized in between the true angles.

Figures 7 and 8 are similar plots but now the impulse responses that were analyzed are measured responses. A few of the segmented reflections are relatively close to the simulated ones, but it is hard to tell whether or not this is a coincidence. Examples of this are the second and third reflections (after the direct sound) in Fig. 7, between 0–5 ms. In Fig. 9, an example of the performance of the baseline algorithm is shown for the measured response in the first receiver position. It is seen that the baseline algorithm can locate most of the reflections quite accurately in time, but there are difficulties in estimating the angle accurately, especially when there are many reflections close to each other in time.

Tables 1 and 2 summarize the time and azimuth errors for the proposed method when using the KEMAR and CIPIC databases with the proposed method, respectively. Table 3 presents the same information for the baseline method. The tables present the values of the threshold (or ϵ for the baseline algorithm, see Sec. 2.2), number of valid reflections (within ± 1 ms from the nearest modeled reflections), the number of simulated reflections (which equals the number of detected reflections in this evaluation), and the means and standard deviations of the time and azimuth errors. The errors are calculated by finding the nearest valid detected reflection for each simulated reflection, calculating the time and angle error, and then taking the mean and standard deviation of the errors. Tables 1 and 2 reveal that in terms of the azimuth angle, the proposed algorithm finds reflections closer in time to the simulated reflections than the simulated responses compared to the measured ones. The azimuth angle estimates seem to be of the same order (around 30°) for receiver position 1 for measured and simulated responses. In receiver position 2, there are differences of 12.1° and 18.8° between simulated and measured responses with the KEMAR and CIPIC databases, respectively. Overall, the errors in azimuth angle estimation are larger in position 2 compared to position 1. This may be due to there being more reflections coming from the sides in position 2, and the accuracy of the angle estimation decreasing when the reflections come from the sides. With the baseline method, the time errors are smaller for the simulated responses compared to the proposed method. The angle errors are of the same order, and position 2 has larger error than position 1.

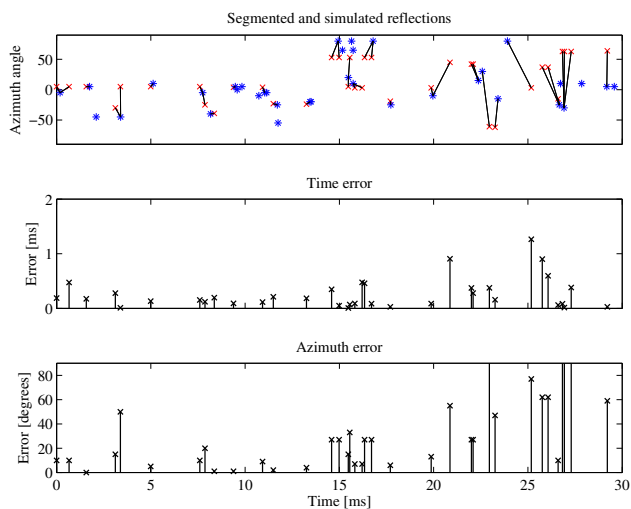


Figure 5: Segmented reflections and the localization errors (position 1, simulated response, CIPIC data). Top panel: Localized reflections. The crosses 'x' denote the reflection locations in the room model at the current position, and asterisks '*' mark the reflections as localized by the algorithm. The room model reflections that come from behind the listener are mirrored to the front. Middle panel: the absolute value of the time difference between each simulated reflection and the closest segmented reflection. Bottom panel: the absolute value of the azimuth angle difference between each simulated reflection and the closest segmented reflection.

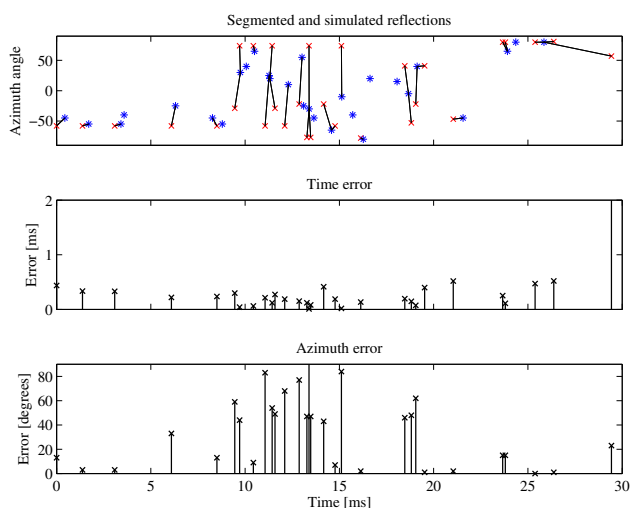


Figure 6: Same as Fig. 5 but for position 2, simulated response.

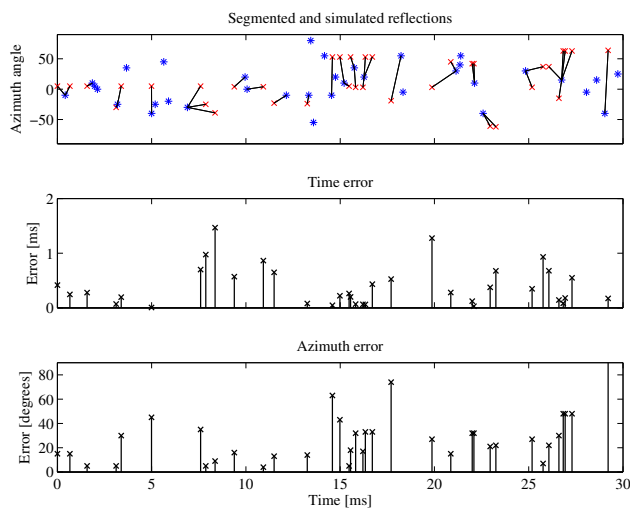


Figure 7: Same as Fig. 5 but for position 1, measured response.

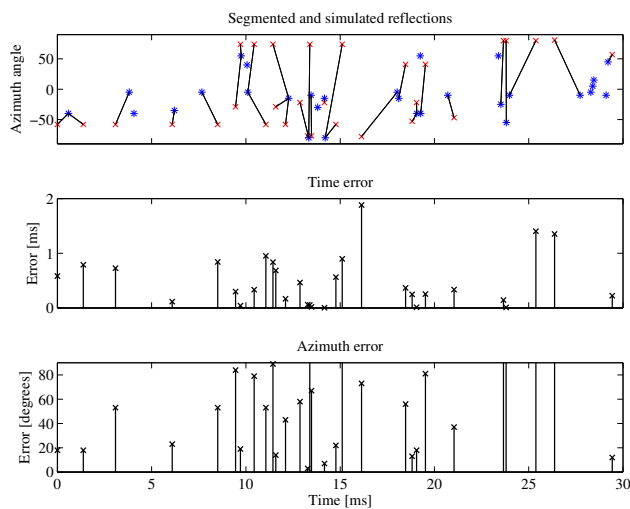


Figure 8: Same as Fig. 5 but for position 2, measured response.

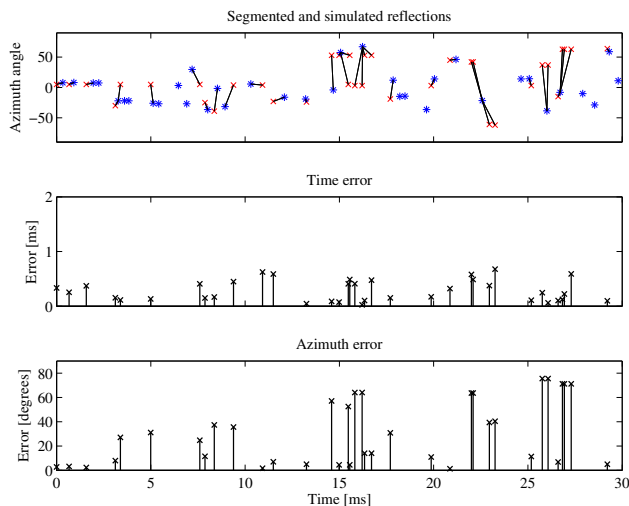


Figure 9: Same as Fig. 5 but for position 1, measured response, and using the baseline method.

4. APPLICATION: SLOW-MOTION AURALIZATION

The presented analysis method can be applied in slow-motion auralization of measured binaural room impulse responses. This auralization method makes it possible to hear the room reflections in detail by increasing the time delays between the reflections. The timing, frequency content, and direction of individual reflections can be heard in this slow-motion response [4].

4.1. Constructing the slow motion impulse response

After each reflection has been reconstructed, the slow motion impulse response is constructed. For this reconstruction, the exact time indices of the reflections have to be known and the time differences between each reflection and the direct sound have to be increased by multiplying the delays with a constant factor K .

Each reflection is localized in time by summing the scale axis out of the absolute value of the XWT of each segmented reflection inside its bounding box, and adding the maximum location to the left edge of the bounding box. The time index of the direct sound, which is assumed to be the global maximum of the XWT (which might not hold in all cases), is then subtracted from the time location calculated before. The result is the time difference of the reflection relative to the direct sound.

After segmenting, time-localizing, and reconstructing each of the reflections, the slow motion response can be constructed. Different factors K can be chosen to hear different aspects of the impulse response. Typical values could be $K \in \{10, 50, 100, 150\}$. The reconstructed time-domain reflections are placed in the slow-motion response so that the time differences of the reflections relative to the direct sound are multiplied by K and the samples of time-domain reflections at the maxima of the corresponding XWTs as described before are placed to that exact time index. Fig. 10 shows examples of responses of measured and simulated BRIR at listening position 1 in the lecture room, slowed down using the proposed method with $K = 100$. One can see that there is some correspondence between the responses at the first reflections.

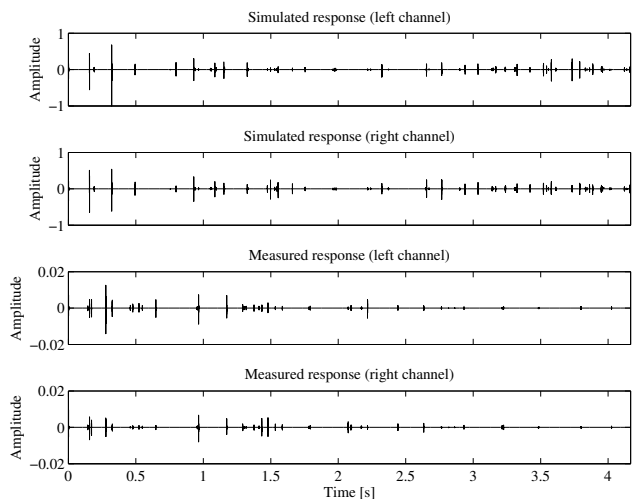


Figure 10: An example of slowing down a simulated and a measured response ($K = 100$, position 1).

5. DISCUSSION AND CONCLUSIONS

A method for segmenting and analyzing reflections from a binaural room impulse response was presented, and evaluated. It is shown that the method can segment reflections from simulated reflections quite accurately, but the estimates of the azimuth angles of the reflections are not very accurate compared to the ground truth from the room model. With measured responses, there is an even larger discrepancy. However, the room model does not match reality exactly, which probably explains many of the differences. In measured responses there is also measurement noise and diffraction, which makes reliable estimation of the azimuth angles of the reflections difficult. Furthermore, the room used for the evaluation is small, which results in the reflections being closer to each other in time compared to concert halls, for example. Future work includes improvements of the azimuth angle estimation method, and investigation into the possibilities of estimating the elevation angle. Different segmentation methods could also be tried.

The present study raises questions on where are the limits of analyzing reflections from a binaural room impulse response, especially from measured data. It is clear that it is possible to segment the reflections only up to the mixing time [29], where the sound field becomes more or less diffuse. Even before the mixing time, the reflection density keeps increasing and there is more and more overlap between the reflections, both in time and in frequency. The overlap causes problems in both segmentation and direction-of-arrival estimation.

6. ACKNOWLEDGMENTS

The authors wish to thank Dr. Kalle Palomäki for comments on the manuscript. A freely available MATLAB toolbox by Grinsted et al. was used for calculating the wavelet transforms [30]. A function for delay estimation by Kevin D. Donohue [25] was used in estimating the ITDs. The research leading to these results has received funding from the Academy of Finland, project no. [119092] and the European Research Council under the European

Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636]. The first author has also received funding from the Nokia Foundation, Tekniikan Edistämissäätiö, and the Hecse Graduate School.

7. REFERENCES

- [1] S. Vesa and T. Lokki, "Detection of room reflections from a binaural room impulse response," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada, September 2006, pp. 215–220.
- [2] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," Tech. Rep., MIT Media Lab Perceptual Computing, 1994, Available at <http://sound.media.mit.edu/resources/KEMAR.html>. Visited May 25, 2009.
- [3] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2001)*, New Paltz, NY, USA, October 2001, pp. 99–102, Available at <http://interface.cipic.ucdavis.edu/>. Visited May 25, 2009.
- [4] T. Lokki, "Auralization of simulated impulse responses in slow motion," in *Proceedings of the AES 118th Convention*, Barcelona, Spain, May 2005, Paper no. 6500.
- [5] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 982–993, 2008.
- [6] G. Defrance, L. Daudet, and J.-D. Polack, "Detecting arrivals within room impulse responses using matching pursuit," in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, September 2008, pp. 297–300.
- [7] S. Roper and T. Collins, "The localisation of a sound source in a reverberant room using arrays of microphones," in *Proceedings of the AES 31st International Conference*, London, United Kingdom, June 2007.
- [8] S. Roper and T. Collins, "A sound sources and reflections localization method for reverberant rooms using arrays of microphones," in *Proceedings of the AES 32nd International Conference: DSP for Loudspeakers*, Hillerød, Denmark, September 2007.
- [9] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2138–2148, 2004.
- [10] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3094–3103, 2005.
- [11] B. Rafaely, I. Balmages, and L. Eger, "High-resolution plane-wave decomposition in an auditorium using a dual-radius scanning spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 122, no. 5, pp. 2661–2668, 2007.
- [12] S. J. Loutridis, "Decomposition of impulse responses using complex wavelets," *Journal of the Audio Engineering Society*, vol. 53, no. 9, pp. 796–811, 2005.
- [13] P. J. Wolfe and S. J. Godsill, "Audio signal processing using complex wavelets," in *Proceedings of the 114th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, March 2003, Paper no. 5829.
- [14] J. R. Beltrán, J. P. de León, and E. Estopiñán, "Intermodulation effects analysis using complex bandpass filterbanks," in *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, Madrid, Spain, September 2005, pp. 149–154.
- [15] J. R. Beltrán and F. Beltrán, "Additive synthesis based on the continuous wavelet transform: a sinusoidal plus transient model," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 2003.
- [16] P. Guillemain and R. Kronland-Martinet, "Characterization of acoustic signals through continuous linear time-frequency representations," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 561–585, 1996.
- [17] M. Schönle, N. Fliege, and U. Zölzer, "Parametric approximation of room impulse responses based on wavelet decomposition," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 1993, pp. 68–71.
- [18] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998.
- [19] S. D. Meyers, B. G. Kelly, and J. J. O'Brien, "An introduction to wavelet analysis in oceanography and meteorology: With application to the dispersion of Yanai waves," *Monthly Weather Review*, vol. 121, no. 10, pp. 2858–2866, 1993.
- [20] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlinear Processes in Geophysics*, vol. 11, pp. 561–566, 2004.
- [21] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice Hall, 2001.
- [22] K. Palomäki, V. Pulkki, and M. Karjalainen, "Neural network approach to analyze spatial sound," in *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, March 1999, pp. 233–245.
- [23] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, The MIT Press, October 1996.
- [24] M. Takanen and G. Lorho, "A binaural auditory model for the evaluation of reproduced stereophonic sound," in *Proceedings of the 124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, May 2008, Paper no. 7371.
- [25] K. D. Donohue, "Function for estimating time delay," Available at <http://www.engr.uky.edu/~donohue/audio/Arrays/delayesttm.m>. Visited May 25, 2009.

- [26] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, England, September 2003, pp. 209–213.
- [27] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.
- [28] T. Lokki, *Physically-based auralization — design, implementation, and evaluation*, Ph.D. thesis, Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory, report TML-A5, 2002.
- [29] T. Hidaka, Y. Yamada, and T. Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses," *The Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 326–332, 2007.
- [30] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Cross wavelet and wavelet coherence MATLAB toolbox," Available at <http://www.pol.ac.uk/home/research/waveletcoherence/>. Visited May 25, 2009.
- [31] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
- [32] A. Cohen and J. Kovacevic, "Wavelets: the mathematical background," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 514–522, 1996.
- [33] T. Lokki, *Physically-Based Auralization - Design, Implementation and Evaluation*, Ph.D. thesis, Helsinki University of Technology, Espoo, Finland, 2002, ISBN 951-22-6157-X.
- [34] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C: the art of scientific computing*, chapter 14.8, Cambridge University Press, 1992.
- [35] S. Vesa and T. Lokki, "An eyes-free user interface controlled by finger snaps," in *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, Madrid, Spain, September 2005, pp. 262–265.
- [36] T. Lokki and H. Järveläinen, "Subjective evaluation of auralization of physics-based room acoustic modeling," in *Proceedings of the 7th International Conference on Auditory Display (ICAD 2001)*, Espoo, Finland, August 2001, pp. 26–31.
- [37] T. Lokki and V. Pulkki, "Evaluation of geometry-based parametric auralization," in *Proceedings of the Audio Engineering Society 22nd International Conference: Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, June 2002, pp. 367–376.
- [38] M. B. Gardner, "Distance estimation of 0° or apparent 0° -oriented speech signals in anechoic space," *The Journal of the Acoustical Society of America*, vol. 45, no. 1, pp. 47–53, 1969.
- [39] M. Gröhn, T. Lokki, and T. Takala, "Static and dynamic sound source localization in a virtual room," in *Proceedings of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, June 2002, pp. 337–344.
- [40] M. Gröhn, T. Lokki, and T. Takala, "Localizing sound sources in a cave-like virtual environment with loudspeaker array reproduction," *Presence: Teleoperators and Virtual Environments*, vol. 16, no. 2, pp. 157–171, 2007.

Table 1: Accuracy of the segmentation (KEMAR)

Position	Thr.	# valid/det	Mean time err. [ms]	Std. time err. [ms]	Mean azi. err. [°]	Std. azi. err. [°]
1 (mod.)	-13.74	35/36	0.23	0.23	33.2	35.1
2 (mod.)	-13.30	29/30	0.23	0.15	42.1	32.9
1 (meas.)	-12.95	34/36	0.34	0.28	30.7	23.1
2 (meas.)	-9.65	27/30	0.37	0.32	54.2	47.8

Table 2: Accuracy of the segmentation (avg. of the 45 subjects of CIPIC)

Position	Thr.	# valid/det	Mean time err. [ms]	Std. time err. [ms]	Mean azi. err. [°]	Std. azi. err. [°]
1 (mod.)	-13.74	35/36	0.23	0.23	29.9	29.2
2 (mod.)	-13.30	29/30	0.23	0.15	35.6	30.4
1 (meas.)	-12.95	34/36	0.34	0.28	28.7	21.7
2 (meas.)	-9.65	27/30	0.37	0.32	54.4	43.9

Table 3: Accuracy of the segmentation (baseline)

Position	ϵ	# valid/det	Mean time err. [ms]	Std. time err. [ms]	Mean azi. err. [°]	Std. azi. err. [°]
1 (mod.)	2.27	36/36	0.19	0.16	32.9	34.3
2 (mod.)	1.53	30/30	0.21	0.16	38.6	36.0
1 (meas.)	2.25	36/36	0.28	0.19	30.8	26.7
2 (meas.)	2.50	29/30	0.22	0.16	52.2	34.5