
© 2009 by authors
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 to 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 reflections 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 in 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 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 aurization 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])

\[
W_x(n,s) = \frac{1}{\sqrt{s}} \sum_{n'=-\infty}^{\infty} x(n') \psi^*_0 \left( \frac{n' - n}{s} \right)
\]  

(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, \( \psi_0(t) \) 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 equation. It is often convenient to convert the scale to frequency in Hertz. For the Morlet wavelet, it proceeds as follows. First, the relationless oscillating period of the wavelet. The oscillating period found in e.g. [12].

\[
\lambda = \frac{4\pi s}{\omega_0 + \sqrt{2 + \omega_0^2}}
\]  

(3)

which is converted to frequency (in Hz) by

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

(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])

\[
|W_{LR}(n,s)| = |W_L(n,s)W_R^*(n,s)|
\]  

(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])

\[
s_j = s_0 2^{j\delta_j}, \quad j = 0, 1, \ldots, J
\]  

(6)

\[
J = \left\lfloor \frac{\log_2 (s_{max}/s_0)}{\delta_j} \right\rfloor
\]  

(7)

where \( \delta_j \) 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 190) 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 on adaptive
thresholding, where the magnitude mean of the impulse response \( h(n) \) at time index \( n \) is first calculated as (adapted from [5])

\[
\mu(n) = \frac{1}{N_\mu} \sum_{m=n-N_\mu}^{n+N_\mu} |x(m)|
\]

where \( N_\mu = \lceil (T_s f_s/2) \rceil \) is the number of samples corresponding to half of the length of the averaging filter of length \( T_s \) seconds.

Rounding to nearest integer is denoted by \( \lceil \cdot \rceil \). Based on the local mean \( \mu(n) \), a binary signal containing the reflection locations is calculated as (adapted from [5])

\[
h_p(n) = \begin{cases} 
0, & \text{if } h(n) < \epsilon \mu(n) \\
1, & \text{if } h(n) \geq \epsilon \mu(n) 
\end{cases}
\]

where \( \epsilon \) is a parameter which defines the threshold. The lower the value of \( \epsilon \), 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 [1111111] 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 re-constructed by using the wavelet transform reconstruction formula which reconstructs the time-domain signal as a sum of real parts of the wavelet transform \( W_j(n, s_j) \) inside the bounding box of the segmented reflection in question, i.e., over a set of scales ranging from scales indices \( J_{\text{min}} \) to \( J_{\text{max}} \) and time indices \( N_{\text{min}} \) to \( N_{\text{max}} \). (adapted from [18])

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

where \( \delta_j \) is the scale resolution (\( \delta_j = 1/32 \) used in this work), \( C_\delta \) 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).
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. For the ITD, the elevation angle is only used for matching the reference values obtained from 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 $f_e = 1.5 \text{ kHz}$ 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

\[
\arg \max_{(\theta, \phi)} \left\{ -\sum_{f_{\max}}^{f_{\min}} (ITD_{\text{ref}}(f, \theta, \phi) - ITD(f))^2 : f \leq f_e \right\}
\]

\[
\arg \max_{(\theta, \phi)} \left\{ -\sum_{f_{\max}}^{f_{\min}} (ILD_{\text{ref}}(f, \theta, \phi) - ILD(f))^2 : f > f_e \right\}
\]

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 $\theta$ and elevation angle $\phi$ calculated from the KEMAR/CIPIC data. The elevation angle in the KEMAR data set ranges from $-40^\circ$ to $+90^\circ$ in $10^\circ$ steps. The number of azimuth angles per elevation varies with the elevation angle, having a resolution of $5^\circ$ for elevations from $-20^\circ$ to $-10^\circ$ 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 $5.625^\circ$, ranging from $-40^\circ$ to $+230.625^\circ$. The azimuth angles sampled in CIPIC are $-80^\circ$, $-60^\circ$, $-55^\circ$, $-45^\circ$ to $+45^\circ$ in $5^\circ$ 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 angle in the KEMAR data set ranges from $-80^\circ$ to $+80^\circ$.

\[1\] Conversion between scale and frequency is done with Eq. (4). Frequencies are used exclusively from now on.
angle using (adapted from [26])

\[ \theta = \sin^{-1}\left(\frac{\tau_{\text{max}} - c}{d_{\text{head}}}\right) \]

(13)

where \( \tau_{\text{max}} \) is the lag of the cross-correlation maximum (in seconds), \( c \) is the speed of sound, and \( d_{\text{head}} = 0.2 \text{ m} \) 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 m \( \times \) 7.3 m \( \times \) 2.6 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 the 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 (x). For each of the four responses, the threshold value (see Sec. 2.2) is set to a value that results in the segmented 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 \(<1\text{ ms}\), 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 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 \( \epsilon \) for the baseline algorithm, see Sec. 2.2), number of valid reflections (within \( \pm1\text{ ms} \) 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 from 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 ‘×’ 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.
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.

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 azimuthal 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, Teknikan Edistämissäätiö, and the Hecse Graduate School.

7. REFERENCES


Table 1: Accuracy of the segmentation (KEMAR)

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

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

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

Table 3: Accuracy of the segmentation (baseline)

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