Sampo Vesa and Tapio Lokki. 2006. Detection of room reflections from a binaural room impulse response. In: Proceedings of the 9th International Conference on Digital Audio Effects (DAFx 2006). Montreal, Canada. 18-20 September 2006, pages 215-220.

DETECTION OF ROOM REFLECTIONS FROM A BINAURAL ROOM IMPULSE RESPONSE

Sampo Vesa, Tapio Lokki

Telecommunications Software and Multimedia Laboratory
Helsinki University of Technology
sampo.vesa@tml.hut.fi

ABSTRACT

A novel analysis method for binaural room impulse responses (BRIRs) is presented. It is based on the analysis of ear canal signals with continuous wavelet transform (CWT). Then, the cross-wavelet transform (XWT) is used for detection of the direct sound and individual reflections from a BRIR. The new method seems to time-localize the reflections quite accurately. In addition, the proposed analysis method enables detailed study of the frequency content of the early reflections. The algorithm is tested with both measured and modeled impulse responses. A comparison with an FFT-based cross-spectrogram is made. The results show that XWT has potential in audio signal analysis.

1. INTRODUCTION

In many cases detailed time scale analysis of binaural impulse responses is needed. For example, a dense pattern of early reflections is usually associated with good concert hall acoustics and it would be great to be able to study these reflections individually from measured responses. One practical application—in which details of early reflections is needed—is the auralization in slow motion using measured binaural impulse responses [1]. When using simulated (binaural) impulse responses the auralization in slow motion is simple: the speed of sound is slowed down by a certain factor in the auralization. This allows perception of the time and direction of arrival of individual reflections. If measured binaural responses are to be slowed down, the situation becomes more complicated, since individual reflections should be very accurately localized in time in order to isolate them from the original response. This calls for a method that can time-localize the individual reflections as accurately as possible.

The problem studied in this paper is time-localization of early reflections from a binaural impulse response. It is assumed that if a reflection occurs there is correlation in short time window between left and right ear canal signals. Auditorily motivated approaches typically employ filter banks and calculation of cross-correlation between channels. Here, a more signal processing oriented approach is taken. An obvious method would be the frame-wise calculation of Fourier cross-spectrum between the left and right channels. A wavelet-based approach was hypothesized to be a better alternative, because of better time resolution compared to frame-based approaches. Therefore, both frame-based and wavelet-based approaches are evaluated in the study.

An introductory paper on wavelet analysis with applications to time series analysis [2] provided inspiration for using wavelet methods for localizing the reflections. Wavelet analysis, in the form of filter bank decomposition, has been used for approximating room impulse responses in simulations [3]. The continuous wavelet transform (CWT) has also been applied to audio signal processing previously for noise reduction and signal compression [4], intermodulation effects analysis [5], sound synthesis [6] and sound signal modeling [7]. The CWT has also been used for decomposition of room and loudspeaker impulse responses [8]. In the current work, the continuous, non-orthogonal and complex cross-wavelet transform (XWT) is used, because the interest is in the correlation between two time series, i.e., the left and right ear canal signals of a binaural impulse response. As far as we know, the cross-wavelet transform seems not to be applied to audio signal processing previously.

2. THE CONTINUOUS WAVELET TRANSFORM AND THE CROSS-WAVELET TRANSFORM

The CWT is a method that can be used for time series analysis. It gives a highly redundant time-frequency representation of a time series, being very different from the *discrete wavelet transform* (DWT), which gives a compact representation of the signal and is thus better suited for signal processing [2]. For time series analysis, the CWT is much more suitable. The DWT also has the disadvantage of an aperiodic shift in the time series giving a different wavelet spectrum [2].

The continuous wavelet transform for a signal $x(t)$ is defined as [9, 2, 8]:

$$W_x(t,s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t') \psi^* \left(\frac{t'-t}{s} \right) dt' \qquad (1)$$

where the asterisk $(*)$ indicates complex conjugation, t is time (*translation*), s is *scale* (dilation) and $\psi(t)$ is the *wavelet function*.

For a discrete sequence, the continuous wavelet transform is defined as a convolution sum [2]:

$$W_x(n,s) = \frac{1}{\sqrt{s}} \sum_{n'=0}^{N-1} x(n') \psi^* \left(\frac{n'-n}{s} \right) \qquad (2)$$

For more efficient computation, the Fourier transform of Eq. (2) is used. Two complex wavelets are used in this study, the first one being the Morlet wavelet, defined as [2]:

$$\psi(t) = \pi^{-1/4} e^{j\omega_0 t} e^{-t^2/2} \qquad (3)$$

where t is a dimensionless time parameter and ω_0 is a non-dimensional frequency. Since the CWT can also be seen as a filter bank, ω_0 is also called the *center* or *oscillating frequency* of the wavelet. A wavelet with a better time resolution and poorer frequency resolution, compared to the Morlet wavelet, is the Paul wavelet [2]:

$$\psi(t) = \frac{2^m j^m m!}{\sqrt{\pi(2m)!}}(1-jt)^{-(m+1)} \quad (4)$$

where m is the order of the Paul wavelet.

Because the interest in audio signal analysis is on frequency, the scale s (non-dimensional) should be converted into frequency f (in Hz), using the following equation [8]:

$$f = \frac{f_s f_0}{s} \quad (5)$$

where f_s is the sampling frequency (in Hz) and $f_0 = \frac{\omega_0}{2\pi}$ is the non-dimensional wavelet center frequency. It should be noted that Eq. (5) only holds for the Morlet wavelet. When generalizing to all possible analyzing wavelets, f_0 can be seen as a proportionality constant which depends on the particular choice of wavelet base and order/center frequency. The relationships between scale and frequency for different wavelets can be found in [2].

The CWT enables also to study similarities of two signals in the same way as FFT-based cross-spectrum. The equivalent CWT-based tool is the *cross-wavelet transform* (XWT) [10], also known as *cross-wavelet spectrum*:

$$W_{xy}(t,s) = W_x(t,s)W_y^*(t,s) \quad (6)$$

Because the interest in this paper is on time-localizing reflections, we are only concentrated on the power of the XWT, which is used for all plots. The phase of the XWT could possible be used for calculating the azimuth angles of each individual reflection, but such a study is left for future work.

3. CROSS-WAVELET ANALYSIS OF BRIRS

The cross-wavelet transform gives information on the dependence between two signals as a function of time, similar to cross-correlogram (or the cross-spectrogram, i.e., the short-time cross-spectrum presented as a function of time). In our case these two signals are the left and right ear canal signals of a binaural impulse response. Therefore, the XWT should be useful in localizing individual reflections of a binaural room impulse response, which manifest themselves as correlation between left and right ear signals, the time lag of the correlation maximum being proportional to the azimuth angle of the reflection.

Fig. 1 compares the Fourier cross-spectrogram and cross-wavelet transforms calculated from a BRIR, measured in a small room (the listening room of the Laboratory of Acoustics and Audio Signal Processing at TKK, see Fig. 2). The Fourier spectrogram was calculated using a time-domain window length of 64 samples and a very small hop size of 2 samples. The time-domain windows were zero-padded to yield a 512-point FFT per frame. The cross-wavelet transforms were calculated using the Morlet and Paul wavelets, the scales ranging logarithmically (base 2) from 2 to 512 with 24 scales per octave, resulting in the total number of scales being 193. The frequency ranges covered were 83.4-21300 Hz for the Morlet wavelet and 61.7-15800 Hz for the Paul wavelet.

As can be seen in Fig. 1, the smaller details are given more emphasis in the cross-wavelet spectrograms. The difference in time and frequency resolutions of the Morlet and the Paul wavelets is also very evident. The logarithmic frequency resolution of the wavelet transforms is also clear in the figures. It is also clear how the time resolution of the wavelet transforms is much worse at low than high frequencies.

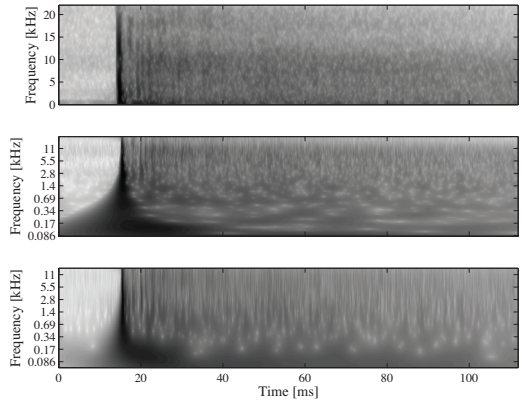

Figure 1: *Comparison between the Fourier cross-spectrogram (top panel) and the cross-wavelet spectrogram calculated using the Morlet (middle panel) and the Paul (bottom panel) wavelets.*

4. FINE-SCALE TIME-LOCALIZATION OF REFLECTIONS

Our hypothesis is that each reflection manifests itself as local maxima in the XWT on different scales at the time of the reflection, resulting in vertical stripes in the XWT plots (see Fig. 1). In order to use the XWT for time-localizing room reflections, we need to extract information on the time locations of the stripes from the XWT. An obvious way to do this is to "integrate out" the scale axis, i.e., to sum along the scales in the discrete case, yielding a one-dimensional signal, which is a function of time. Locations of the local maxima of this signal should correspond to the individual reflections. The sum across scale j_1 to j_2 is calculated from the cross-wavelet transform $W_{xy}(n,s)$ as:

$$W_{sum}(n) = \sum_{j=j_1}^{j_2} |W_{xy}(n,s_j)|^2 \quad (7)$$

As an alternative, the maximum across scale could be used:

$$W_{max}(n) = \max_{j\in[j_1,\, j_2]} \left\{|W_{xy}(n,s_j)|^2\right\} \quad (8)$$

Since it is expected that sum and/or maximum across scale could have some minor local maxima that do not correspond to reflections, Savitzky-Golay smoothing [11] was applied to them (using the function `sgolayfilt` in MATLAB). The filter order was 15 and the frame length was 23 samples. It turned out, that the maximum across scales was smooth enough as such and thus smoothing was only used for the sum across scale.

4.1. Choosing the set of scales

In contrast to orthogonal wavelet analysis, in non-orthogonal analysis the set of used scales can be chosen arbitrarily. The scales can be specified as fractional powers of two [2]:

$$s_i = s_0 2^{i\delta_i},\ i = 0, 1, \ldots, I \quad (9)$$

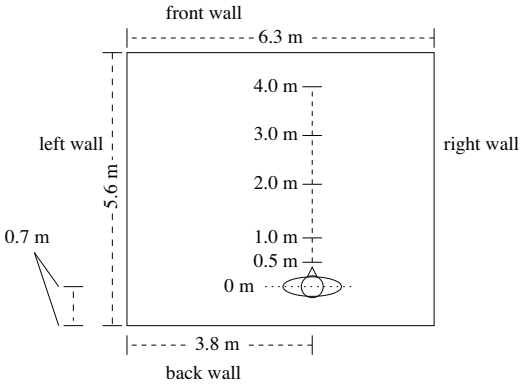

Figure 2: *Position of the binaural manikin in the room where the BRIRs were measured. The room height is 3.0 m and the listener and the source height was 1.5 m. The binaural room impulse responses were measured at source-to-receiver distances of 0.5, 1, 2, 3 and 4 meters.*

where s_0 is the smallest scale, I is the number of scales and δ_i controls the scale resolution ($D = 1/\delta_i$ gives the number of scales per octave).

5. EXPERIMENTS

As an example of the XWT analysis, both measured and modeled binaural impulse responses are analyzed and studied. First, the XWT is calculated for the entire measured binaural impulse response. This yields a matrix sized the length of the signal (in samples) times the number of scales the XWT was calculated at. The scale configuration is as described in Section 3. Since the largest scales convey little information about the time locations of the reflections, the XWTs were truncated in scales to the range $s \in [2,\ 64]$, which is 494-15800 Hz for the Paul wavelet and 667-21300 Hz for the Morlet wavelet. The range was chosen by hand so that the sum and maximum across scale seem to give a reasonable amount of detail, i.e., reasonable amount of local maxima. Including the frequencies below ∼500 Hz results in the sum and maximum across scale being overly smooth, and thus some reflections will not be detected.

For comparison, a standard FFT-based cross-spectrogram (calculated as described in Section 3) is analyzed in the 500-16000 Hz band, which is close to the band covered by the Paul wavelet. After calculating the XWT or FFT-based cross-spectrogram, the sum and maximum across scale/frequency is evaluated using Eqs. (7) and (8). The curve representing the sum across scale is then smoothed as described previously. Finally, all of the local maxima of the resulting curves are located in time.

For a quantitative validation of the time-localization method, the exact locations of the room reflections should be known. For modeled responses, the exact delays for each reflection are known precisely. The situation is more complicated with measured responses, when there is no exact information on the timing of the reflections. However, since the room dimensions are known in this case (see Fig. 2), the image-source method can be used to calculate the early reflections approximately. In this work, only the first and second order reflections were calculated.

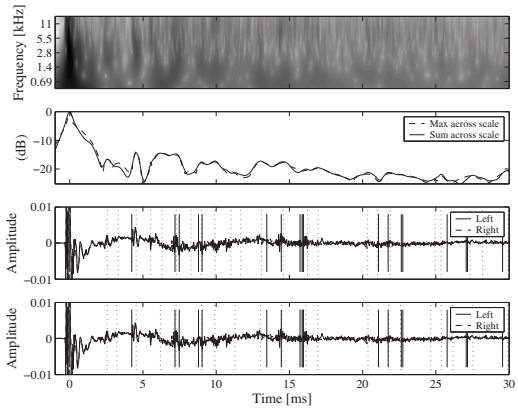

Figure 3: *Top panel: the XWT of a binaural impulse response measured in a standard listening room (source-to-receiver distance 0.5 m, azimuth angle 0°). The Paul wavelet was used in the analysis. Upper middle panel: maximum of the XWT across scale, versus time (dashed line) and sum of the XWT across scale, versus time (solid line), presented on a logarithmic scale (base 2). Lower middle panel: the binaural impulse response, with the localized reflections marked by dotted vertical lines (time-localization based on local maxima of the maximum across scale) and the first and second order reflections calculated from a shoebox model marked by solid vertical lines. Bottom panel: same as the lower middle panel, but with time-localization based on local maxima of the sum across scale.*

5.1. Qualitative evaluation

Figs. 3 and 4 present the XWT of the first 30 ms of a binaural impulse response measured in the standard listening room (refer to Section 3) with $T_{60} \approx 0.3$ s, its maximum and sum across scale, and the localized reflections based on both the maximum and sum across scales. In Fig. 3, the Paul wavelet was the analyzing wavelet, while in Fig. 4, the Morlet wavelet was used. A localized reflection is set to each local maximum of the maximum or sum across scale, and is indicated by dotted vertical lines. The first and second order reflections given by the image-source model [12] are shown by solid vertical lines.

The better time resolution of the Paul wavelet (Fig. 3) is evident when compared to the Morlet wavelet (Fig. 4). The FFT-based analysis of a measured response (Fig. 5) seems to have a time-localization accuracy comparable to the Paul wavelet. By visual inspection it is hard to tell which peaks in the time-domain impulse response are true reflections, but using the Paul wavelet or the FFT cross-spectrogram seems to locate many details from the response. Many of these details are likely caused by reflections. Some of the reflections given by the image-source model also seem to coincide with the localized reflections. The room model agrees well with visual inspection of the time-domain responses in the case of the almost simultaneous floor and ceiling reflections (at

Method	Avg. abs. error [ms]							
	back wall	front wall	left wall	right wall	floor	ceiling	1st ord.	1st & 2nd ord.
XWT, Paul wavelet (max)	0.24	0.51	0.29	0.23	0.063	0.21	0.26	0.33
XWT, Paul wavelet (sum)	0.30	0.39	0.36	0.27	0.076	0.29	0.28	0.32
XWT, Morlet wavelet (max)	0.39	0.52	0.67	0.20	0.12	0.27	0.36	0.63
XWT, Morlet wavelet (sum)	0.43	0.39	0.33	0.29	0.048	0.24	0.29	0.47
Fourier cross-spectrogram (max)	0.21	0.32	0.19	0.37	0.13	0.19	0.24	0.25
Fourier cross-spectrogram (sum)	0.37	0.43	0.32	0.44	0.087	0.16	0.30	0.32

Table 1: *Average absolute errors for the detected reflections using sum and maximum across scale (measured responses).*

Method	Average absolute error [ms]
XWT, Paul wavelet (max)	0.10
XWT, Paul wavelet (sum)	0.19
XWT, Morlet wavelet (max)	0.28
XWT, Morlet wavelet (sum)	0.33
Fourier cross-spectrogram (max)	0.29
Fourier cross-spectrogram (sum)	0.35

Table 2: *Average absolute errors of reflections up to 30 ms using sum and maximum across scale (artificial response).*

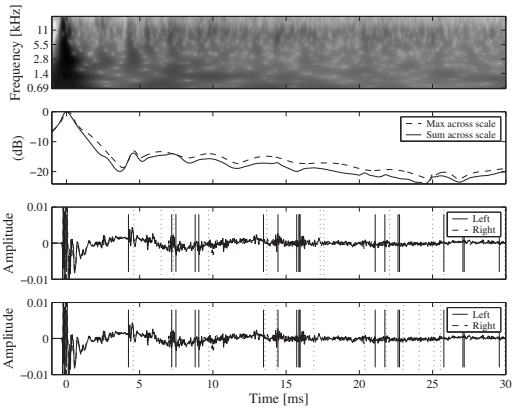

Figure 4: *Same as Fig. 3, but the Morlet wavelet was used in the analysis.*

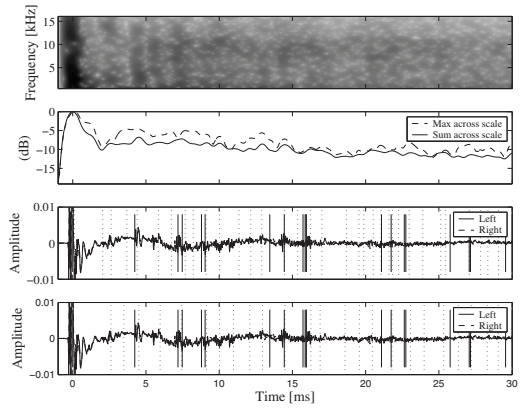

Figure 5: *Same as Fig. 3, but a standard FFT-based cross-spectrogram (with a hop size of one sample) was used.*

7.2 and 7.5 ms, respectively, according to the model). The analysis methods fuse the two reflections together, except with the combination of the FFT cross-spectrogram the maximum across frequency. The back wall reflection (at 4.2 ms) is correctly localized by both visual inspection and the analysis methods. However, room model may be slightly inaccurate, and therefore the three later reflections coming from the sides and the front wall may be inaccurate. By inspecting the two bottom panels of Fig. 3, it is very hard to tell, which of the "bumps" in the responses are actually due to the three later reflections. Besides the first and second order reflections from the walls, higher-order reflections and reflections from objects in the room appear in the response.

Fig. 6 presents the XWT analysis for an artificial modeled BRIR, using the Paul wavelet. The response is computed with the DIVA software [13] from a shoebox-shaped room. The reflections are modeled up to fifth order and each reflection is processed with the appropriate head-related transfer functions to get a binaural response. As can be seen in Fig. 6, the reflections contain energy over the whole frequency region and no background noise exists. Such facts make the artificial response easier for the proposed algorithm and the algorithm has localized most of the peaks correctly. This is also verified by comparing the localized reflections to the ground truth (solid vertical lines) in the two bottom panels of Fig. 6.

By looking at the top panels of Figs. 3, 4 and 6 it is seen that the XWT can also be used to analyze the frequency content of the early reflections[1]. As an example, the measured response (topmost panels of Figs. 3 and 4) contains a clear reflection (a darker area) right before the 5 ms time stamp. This reflection contains energy

[1] Because of its linear frequency resolution, the FFT cross-spectrogram (Fig. 5) is not so well suited for detailed visual analysis of the frequency content, at least at the low frequencies.

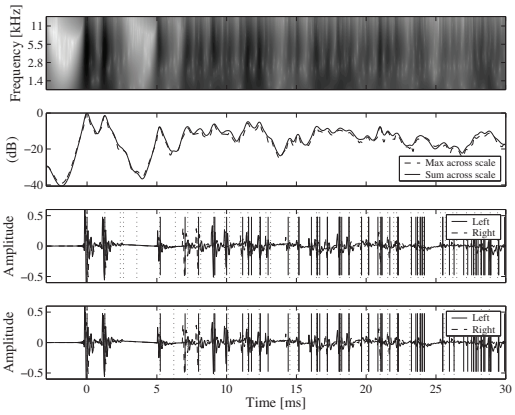

Figure 6: *Same as Fig. 3, but a modeled BRIR was analyzed. The vertical solid lines in the two bottom panels indicate the ground truth locations for the reflections.*

only at middle frequencies while next reflections a little bit later contain only low frequencies. From the artificial response analysis (Fig. 6) it can be seen that modeled reflections are all wideband reflections. Such analysis might be useful, e.g., in optimizing the loudspeaker placement in the home theaters, where some clear reflections typically occur.

5.2. Quantitative evaluation

Quantitative analysis of the accuracy of the time-localization is presented in Tables 1 and 2, where the average absolute errors of the time difference between the true reflection location and the nearest localized reflection are presented. This was done separately for the first and second order reflections in the case of the real responses, and for all of the reflections falling in the 30 ms time window following the direct sound in the case of the artificial response. For the measured responses (Table 1), the average error is taken as the arithmetic mean over the analysis results of set of binaural room impulse responses measured at source-to-receiver distances of 0.5, 1, 2, 3 and 4 meters (see Fig. 2). The average errors are presented for each of the six first order reflections separately. A total average for each analysis method is also given for the first order reflections and all calculated reflections, i.e., both first and second order reflections. Only the reflections falling within 30 ms of the direct sound are considered. A single artificial response was analyzed and the averages in Table 2 present the total average errors over all the reflections in the 30 ms time window following the direct sound, as calculated for each of the analysis methods.

By looking at Table 1 it seems that on average, the Paul wavelet and the FFT method perform equally well, while the Morlet wavelet performs slightly worse, on a real measured signal. The average absolute errors are always less than 1 ms. On average, the floor reflections are clearly located more accurately than the other reflections. Besides the floor reflection being a very strong one (it is actually superimposed with the ceiling reflection in this case), the theoretically calculated reflection is close to the true reflection. With the other first order reflections listed in Table 1, as well as the second order reflections, the theoretical time locations might not hold exactly due to inaccuracies in the measured distances (Fig. 2) used to calculate the theoretical locations. Because of this, the results presented in Table 1 give only an approximate sense of how the methods perform in real spaces, and compared to each other. The methods also differ in terms of the density of localized reflections. Some of the localized reflections might not actually correspond to a real reflection, but are just statistical fluctuations. This also affects the results of Table 1 so that a low absolute error may also be due to a localized reflection being close to the theoretical location just by chance.

Table 2 indicates that for the artificial response, the Paul wavelet performs significantly better than the Morlet wavelet or the FFT method. The average error of the Paul wavelet is almost half that of the Morlet wavelet or the FFT. The performance of the wavelet methods is also significantly better than that of with measured responses. With the artificial response, reflections up to fifth order are considered, not just the first and second order reflections. The FFT method seems to be unable to localize each of these reflections individually. However, with the real response (Fig. 5) it seems that the FFT method gives more detected reflections than the wavelet methods. Conclusions from the superiority of the any of the methods can not be therefore drawn based on this.

The maximum across scale/frequency seems to be superior to the sum across scale/frequency for each method listed in Tables 1 and 2. The only exception is with the Morlet wavelet for the measured responses. The superiority of the maximum might be explained so that when a reflection is present in the signal at a certain time point, there is typically a clear peak in the XWT/cross-spectrogram at some frequency at that point. When the maximum across scale/frequency is evaluated as a function of time, a peak in the curve results at the very same time point, and thus the reflection is localized accurately. The sum across scale/frequency averages this peak so that the time accuracy of localization is worse and some narrow-band reflections are easily missed completely. On the contrary, the maximum across scales also gives rise to many more localizations, many of which may be just due to statistical fluctuations. From Figs. 3, 4 and 6 it is evident that there are more localized reflections when using the maximum across scales.

6. CONCLUSIONS

The cross-wavelet transform is applied for time-localizing reflections from binaural room impulse responses. The method performs accurately for artificial responses and finds almost all early reflections. Locations of the first and second order reflections were calculated for the room and a comparison to the reflections found by the proposed method was made. At least the strongest of the six first order reflections were localized quite accurately. However, the performance with real measured responses is hard to evaluate, since the exact locations of all of the reflections are not known. The proposed XWT with Morlet and Paul wavelets was also compared to the conventional FFT cross-spectrum. The Paul wavelet has the best time resolution, and thus seems to be the most accurate one of the tested methods, at least in the case of the artificially generated room response. For the measured response, the FFT cross-spectrogram method gave equally good results.

Future work should concentrate on improving the accuracy of the method. The algorithm should be made robust so that no reflections are missed and no false detections are made either. The

spectral content of the reflections should also be taken into account – there might be need to differentiate between low and high frequency reflections, as well as narrowband and wideband reflections. More advanced algorithms for detecting the reflections could be developed. Worth of investigation is also how the direction, i.e., azimuth angle, of the individual reflections could be calculated. One possibility is the use of the phase of the XWT for azimuth localization.

7. ACKNOWLEDGMENT

A freely available MATLAB toolbox for calculating the cross-wavelet transform and wavelet coherence was used [14]. This work was partly funded by the HeCSE graduate school.

8. REFERENCES

[1] T. Lokki, "Auralization of simulated impulse responses in slow motion," in *118th Conv. Audio Eng. Soc.,* Barcelona, Spain, May 2005, paper no. 6500.

[2] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin Am. Meteorological Soc.*, vol. 79, no. 1, pp. 61–78, 1998.

[3] M. Schönle, N. Fliege, and U. Zölzer, "Parametric approximation of room impulse responses based on wavelet decomposition," in *Proc. IEEE Workshop Appl. of Dig. Sig. Proc. to Audio and Acoust.,* New Palz, NY, Oct. 1993, pp. 68–71.

[4] P. J. Wolfe and S. J. Godsill, "Audio signal processing using complex wavelets," in *114th Conv. Audio Eng. Soc.,* Amsterdam, The Netherlands, Mar. 2003.

[5] J. R. Beltrán, J. P. de León, and E. Estopiñán, "Intermodulation effects analysis using complex bandpass filterbanks," in *Proc. Int. Conf. on Digital Audio Effects (DAFx-05),* Madrid, Spain, Sept. 2005, pp. 149–154.

[6] J. R. Beltrán and F. Beltrán, "Additive synthesis based on the continuous wavelet transform: a sinusoidal plus transient mode," in *Proc. Int. Conf. on Digital Audio Effects (DAFx-03),* London, UK, Sept. 2003.

[7] P. Guillemain and R. Kronland-Martinet, "Characterization of acoustic signals through continuous linear time-frequency representations," *Proc. IEEE*, vol. 84, no. 4, pp. 561–585, 1996.

[8] S. J. Loutridis, "Decomposition of impulse responses using complex wavelets," *J. Audio Eng. Soc.*, vol. 53, no. 9, pp. 796–811, 2005.

[9] A. Cohen and J. Kovacevic, "Wavelets: The mathematical background," *Proc. IEEE*, vol. 84, no. 4, pp. 514–522, 1996.

[10] 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.

[11] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992, ch. 14.8.

[12] 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, 1979.

[13] T. Lokki, "Physically-based auralization – design, implementation and evaluation," Ph.D. dissertation, Helsinki University of Technology, Espoo, Finland, 2002, iSBN 951-22-6157-X.

[14] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Cross wavelet and wavelet coherence MATLAB toolbox," 2006, [Online] http://www.pol.ac.uk/home/research/waveletcoherence/.