DYNAMIC TIME WARPING FOR ACOUSTIC RESPONSE INTERPOLATION: POSSIBILITIES AND LIMITATIONS

Gavin Kearney, Claire Masterson, Stephen Adams and Frank Boland

Department of Electronic and Electrical Engineering
Trinity College Dublin
email: gpkearney@ee.tcd.ie

ABSTRACT

We present a novel method for the interpolation of the direct and early reflection components in acoustic impulse responses. The method utilizes Dynamic Time Warping for automated temporal alignment of the sparse reflections. The method is evaluated using a simple acoustic model and its perceptual limitations are presented.

1. INTRODUCTION

Convolution reverberation has gained increasing popularity in recent years, due to the highly realistic timbre of convolved sources coupled with fast implementations and increasing computational flexibility. However, real-time convolution with finite room impulse responses (RIRs) for interactive audio-visual presentations is still a highly problematic area. Firstly, there is the drawback that a single convolution measurement reflects only a static source and receiver (microphone) position. This is also the case in the multichannel sense, since, for example a stereo capture of a RIR will still only reflect a static source position on playback. This is acceptable for the realm of music production, but not for convincing interactive real-time audio visual presentations, such as teleconferencing or gaming. In such cases the RIR needs to change with the movement of the listener and the source.

This would lead to more convincing auralization, but to achieve this, several simultaneous RIR measurements are required at spatially separated locations in a chosen room. The challenge then is to reduce the number of measurements for optimal storage and minimum reproduction effort whilst maintaining perceptually correct spatialisation. Thus, we investigate here the interpolation of RIRs to aid in measurement reduction of RIR datasets.

One method to interpolation, proposed by Haneda et al. [1] uses a common acoustical pole and residue model. They propose that variations in the RIR (or its frequency domain equivalent, the room transfer function) can be characterized by residue variations in the model with different source or receiver positions, and as such, interpolating between RIRs simplifies to interpolating residue functions. This approach seems to be effective for the low frequency component of the room transfer function. A more recent method, proposed by Huszty et.al [2], uses fuzzy modeling techniques for RIR interpolation. Their work recognizes that only the early reflections can be interpolated and for this, some type of temporal mapping is necessary. They propose a manual pairing of reflections to ensure the temporal mapping is accurate. In this paper, we too split the measured impulse responses into their early reflection and diffuse decay regions. However, as we shall demonstrate with our method, automatic temporal alignment is possible using a process known as Dynamic Time Warping (DTW).

1.1 Splitting the Impulse Response

First consider a RIR to have two significant parts: the direct sound with early reflections and the diffuse decay. Let h_i denote the room impulse response measured at position i in a 1-D microphone array. h_i can be split into two components, the early reflections, h_i^e and the diffuse decay h_i^l such that

$$h_i = \left[h_i^e[1:n_t] : h_i^l[(n_t+1):N] \right] \tag{1}$$

where n_t is the point of transition between the early and late reflections, and N is the total number of samples in the RIR.

The transition time, T_t, at which n_t occurs is of significant importance here. Existent measures of T_t are typically related to room volume and the density of reflections and a good summary can be found in [3]. Here we use the method suggested by Naylor and Rindel [4] where T_t is the time of arrival of the fourth order reflections in h_i. This can be computed from the mean-free path by

$$T_{ro} = \frac{4V}{cS}\left(O_e+1\right) \tag{2}$$

where V is the room volume, S is the surface area, c is the speed of sound and O_e is the reflection order [5]. We therefore take T_t as T_{ro} when O_e is equal to 4.

In this paper, we focus solely on the interpolation of the early reflections, and we consider that the tail can be synthesized effectively according to the method suggested by the authors in [6].

2. INTERPOLATION THROUGH DYNAMIC TIME WARPING

Let us consider the case where we have measured two RIRs h_1^e and h_3^e at points R1 and R3 as shown in Figure 1. We will now attempt to create a new interpolated (direct-sound and early reflections) RIR, $\hat{h}_2^e$, as if it were measured at position R2. Figure 2 shows an example of the first 20ms of two such measured RIRs.

Since the RIRs are recorded at different spatial locations their early component will contain sparse reflections occurring at different times in each impulse. Consequently, even at spatially close locations this spareness means that linear interpolation can result in significant smearing of reflections in the interpolated result, as shown in Figure 3. It is therefore necessary to align the signals in some way. One can apply a delay to one signal so that the direct paths align, but this does

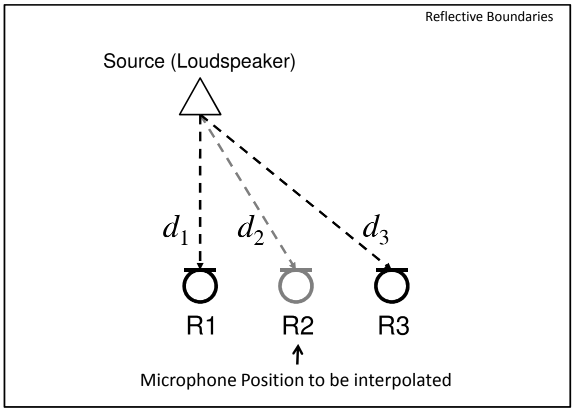

Figure 1: Interpolation between two microphones

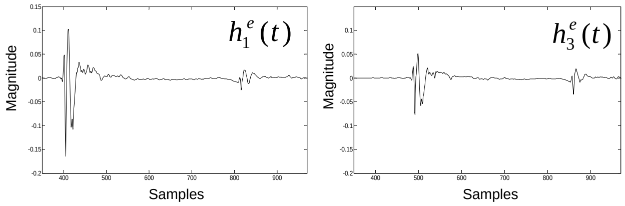

Figure 2: RIRs at positions 1 and 3

not guarantee that subsequent reflections will match up, due to different reflection path lengths at different positions in the room.

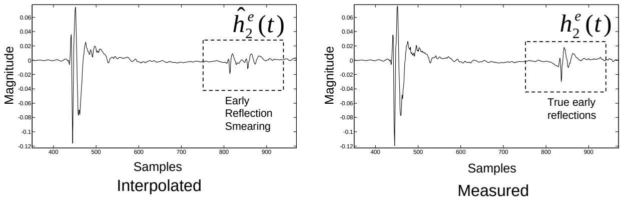

Figure 3: Comparison of impulse response created from linear interpolation between two RIRs to an actual impulse response measured at the position of interpolation.

It is therefore necessary to temporally align the main feature points of the impulses prior to interpolation. Dynamic Time Warping (DTW) [7] is a technique which allows us to do this. It stretches (warps) the signals non-linearly by repeating samples in each time series allowing us to 'line up' the early reflections. A warp vector is created for each time series which describes how the signals are stretched.

The warp vectors are formed by calculating a minimum distance warp path through an accumulated distance matrix. The 'distance' is the Euclidean distance between data point i in one time series and data point j in the other time series. The optimal warp path is then the path through the matrix with the minimum accumulated distance given by

$$D(W) = \sum_{k=1}^{k=K} D(w_{ki}, w_{kj}) \tag{3}$$

where $D(W)$ denotes the distance of the warp path and $D(w_{ki}, w_{kj})$ represent the distances between the sample indexes at the k^{th} element of the warp path. A trivial example of such a matrix with two time series (16 samples long) is shown in Figure 4.

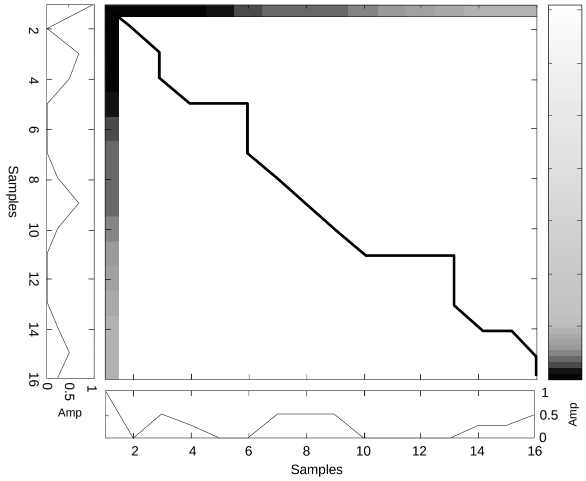

Figure 4: Accumulated Distance Matrix.

The warp path is subject to several constraints. First, the path must begin at the first sample of each signal ($h_1^e(1)$ and $h_3^e(1)$) and end at the last sample of each signal ($h_1^e(n_t)$ and $h_3^e(n_t)$), ensuring that each sample index of each signal is used during its formation. A continuity condition is also applied which ensures that the warp path only traverses through the matrix via adjacent cells. Furthermore the path must monotonically increase, in order to ensure that it never overlaps itself.

Thus, to obtain the interpolated impulse response $\hat{h}_2^e$, it is first necessary to apply DTW to h_1^e and h_3^e, which gives their warped versions h_{w1}^e and h_{w3}^e. This aligns the main feature points of both RIRs and allows for simple linear interpolation between them to obtain the magnitude of the unknown RIR, $\hat{h}_{w2}^e$. This is shown in Figure 5.

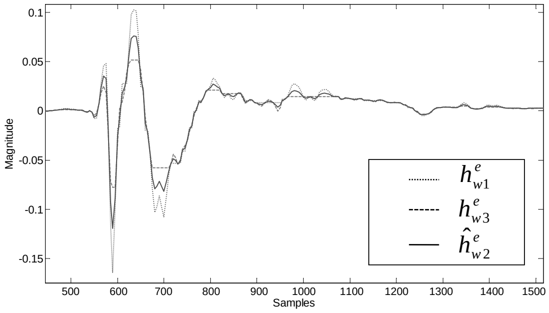

Figure 5: Warped RIRs at positions 1 and 2 and the interpolated warped RIR

The magnitude interpolation is weighted based on a ratio of the inverse distances between the source and the microphones, since sound pressure level is inversely proportional to the distance from the source.

$$\alpha = \frac{\frac{1}{d_3} - \frac{1}{d_2}}{\frac{1}{d_3} - \frac{1}{d_1}} \tag{4}$$

and hence,

$$\hat{h}^e_{w2} = \alpha h^e_{w1} + (1-\alpha) h^e_{w3} \tag{5}$$

Now the warp vectors, w_1 and w_3 that describe how h^e_1 and h^e_3 are mapped onto h^e_{w1} and h^e_{w3} by the DTW must be interpolated to obtain w_2. Again linear interpolation is used to accomplish this and the weights, β and $1-\beta$, are calculated based on the distances of the microphones to the source.

$$\beta = \frac{d_2 - d_3}{d_1 - d_3} \tag{6}$$

and hence,

$$w_{int} = \beta w_1 + (1-\beta) w_2 \tag{7}$$

The final step in the process is to map the warped interpolated vector back into the "unwarped" time domain using the interpolated warp vector. Figure 6 shows a comparison of an interpolated RIR $\hat{h}^e_2(1)$ and a real measured RIR $h^e_2(1)$ taken from the desired interpolated position. We notice that the temporal distortions that were present in the linear interpolation of Figure 3 are no longer present.

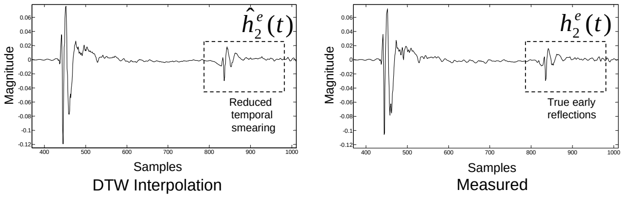

Figure 6: Comparison of impulse response created from DTW interpolation between two RIRs to an actual impulse response measured at the position of interpolation.

3. ANALYSIS OF INTERPOLATION METHOD

Since we are interested in applications to spatial auralization, we choose here to investigate the perceptual attributes of interpolated RIRs at reproduction. To this end, we simulate the capture and reproduction setup shown in Figure 7. The capture setup consists of an omnidirectional point source and two omnidirectional receivers enclosed in a reflective environment. We simulate RIR capture at the receiver points using the Image Method of Allen and Berkley [8].

In this setup, we will keep the source position S1 and receiver position R1 stationary and 2m apart. We will then vary the position of receiver R3 from 0 to 1m away from R1 in finite steps. At each step, we capture responses at R1 and R3 and interpolate another response $\hat{h}^e_2$ exactly in-between the receivers (at R2). By comparing this interpolated estimate to the true response at R2 $((h)^e_2)$, we can see the effect of increased microphone separation on the interpolation process. Furthermore, if we vary the reflection coefficient Γ of the room (where $\Gamma = 0$ represents free-field conditions and $\Gamma = 1$ represents completely reflective surfaces), we can study the effect of increasing the levels of early reflections on the interpolation process.

Also shown in Figure 7 is the reproduction setup, which consists of stereophonic reproduction of the responses recorded at R1 and R2. Thus is simulated by convolution of the responses picked up at R1 and R2 with head related impulses responses (HRIRs) captured from a KEMAR binaural mannequin [9] i.e. for stereophonic reproduction of h^e_1 and h^e_2, the resultant binaural signals are

$$h_l(t) = h^e_1(t) \otimes k^{left}_1(t) + h^e_2(t) \otimes k^{left}_2(t) \tag{8}$$

$$h_r(t) = h^e_1(t) \otimes k^{right}_1(t) + h^e_2(t) \otimes k^{right}_2(t) \tag{9}$$

where k_1 and k_2 are the KEMAR binaural responses to loudspeakers 1 and 2.

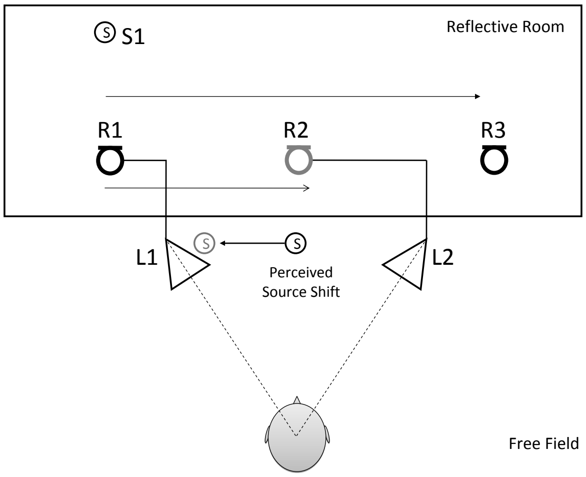

Figure 7: Simulated recording and reproduction setup utilizing image source method and KEMAR HRTFs.

Here, we investigate several binaural cues which aid localization and spatial impression. The first of these is the Interaural Cross Correlation function (IACF), a measure of the correlation between the received ear signals within the integration limits t_1 to t_2 as a function of the time delay τ, given by

$$IACF(\tau) = \frac{\int_{t_1}^{t_2} h_l(t) h_r(t+\tau) dt}{\sqrt{\int_{t_1}^{t_2} h_l^2(t) dt \int_{t_1}^{t_2} h_r^2(t) dt}} \tag{10}$$

The point at which this function yields its maximum is known as the Interaural Cross Correlation Coefficient (IACC), and is commonly used as a measure of the acoustic quality in concert halls [10]. We employ the IACC$_{E3}$ function here as defined by Beranek [11]. This is an average measure of the IACC in the 500, 1000 and 2000 octave bands within the first 80ms of the binaural responses.

The time delay at which the IACC is maximum is also representative of the reproduced source position. This is known as the Interaural Time Difference (ITD). ITD is presented here as an averaged value below 2kHz, since phase ambiguity occurs in the binaural measurements above this. Localisation accuracy is also determined by interaural level difference (ILD), given by

$$ILD = 20\log_{10}\frac{|H_l(f)|}{|H_r(f)|} \qquad (11)$$

where H_l and H_r are the Fourier domain representation of the ear responses. An averaged response above 2kHz is presented, since ILD values below this frequency range are extremely low.

4. RESULTS

The simulation setup of Figure 7 was implemented for three different size rooms, with dimensions listed in Table 1.

Room	L (m)	W (m)	H(m)
Small	5.85	4	2.5
Medium	8.7375	6	3.75
Large	11.65	8	5

Table 1: Room Dimensions used during simulation.

Errors in the interaural level difference due to the interpolation process are shown in Figure 8. It can be seen that the error is generally below 0.5dB in the medium and large rooms, with greatest errors of approximately 2dB occurring for the small room simulation (with large values of Γ). A distance error is present in the medium room at an approximately 40cm microphone spacing regardless of reflection level, and is attributed here to rounding differences in the interpolated and real result. In general, error is mainly a function of the reflection coefficient rather than of microphone spacing. This error is proportional to the density of reflections, and hence is greatest for small rooms. Although such errors are small, they will lead to high-frequency phantom source shifts above 2kHz.

Errors in the ITD, shown in Figure 9, again occur mainly due to the level of the reflections, and are largely independent of the spatial separation of the microphones.

The best ITD results are obtained for the large room, which has the smallest region of error. In the small room, the interpolation does not significantly affect the ITD until the reflection coefficient is above 0.5, with a large ITD error occurring for 1m separation and maximum Γ. This peak occurs at approximately 0.3ms, which considering the interaural cross head delay is in the range of 1ms, will lead to significant localisation error. We note that this represents an extreme case, since reflection coefficients of surfaces in real rooms rarely equal 1.

Errors in the apparent source width are shown in Figure 10. These errors are again mainly a function of reflection coefficient, and are low with values below 0.2. The small room again represents the poorest case, with high values of reflection coefficient causing 35% error in the perceived source width.

The performance of the algorithm in the small room demonstrates a necessity for the reflections to be sufficiently sparse in order for the interpolation to be effective. The density of the reflections results in a feature misalignment in the algorithm. Such errors in the early reflection interpolation lead to comb-filtering effects that differ from those caused by the true response, and as a result, we see errors in ILD. Sparseness in the acoustic response can be increased if a more directional receiver pattern is used, such as a cardioid or super-cardioid. However, the interpolation method needs to be adjusted to accommodate microphone directionality. Furthermore, a more accurate measure of the transition region between early and late reflections would ensure that dense reflection regions are not included in the interpolation process.

5. CONCLUSIONS

We have presented a method of interpolation of room acoustic responses using dynamic time warping to align the main feature points of the signals prior to interpolation. The method was shown to avoid the smearing distortions that occur in linear interpolation processes. A perceptual based evaluation of the interpolated impulse responses was presented, and within the scope of this study, the method was found to work well for medium to large rooms with effective interpolation up to approximately 0.5m spacing under highly reflective conditions. The studies presented here indicate that the method will work well for other rooms of larger dimensions. However, the method in its current form is not suited for interpolation in small rooms, since the early reflections are not as sparse. Further work will involve improving the algorithm to incorporate microphone directivity as well as investigation of measurements in real acoustic environments.

6. ACKNOWLEDGMENTS

Gavin Kearney acknowledges the support of Science Foundation Ireland. Claire Masterson and Stephen Adams acknowledge the support of the Irish Research Council for Science and Technology.

REFERENCES

[1] Y. Haneda, Y. Kaneda, and N. Kitawaki. Common-acoustical-pole and residue model and its application to spatial interpolation and extrapolation of a room transfer function. *Speech and Audio Processing, IEEE Transactions on*, 7(6):709–717, 1999.

[2] C. Huszty, B. Németh, P. Baranyi, and F. Augusztinovicz. Measurement-based fuzzy interpolation of room impulse responses. *The Journal of the Acoustical Society of America*, 123(5):3771, 2008.

[3] K. Meesawat and F. Bajers. An investigation on the transition from early reflections to a reverberation tail in a BRIR. In *Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan*, July 2002.

[4] G. Naylor and J. H. Rindel. Predicting room acoustical behaviour with the ODEAN computer model. In *124th Meeting of Acoustical Society of America*, November 1992.

[5] H. Kuttruff. *Room Acoustics*. Appl. Science Publ., London, 1979.

[6] C. Masterson, G. Kearney and F. Boland. Acoustic impulse response interpolation using Dynamic Time Warping. In *Audio Engineering Society 35th International Conference*, February 2009.

[7] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, Feb 1978.

[8] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *JASA*, 65:943–950, 1979.

[9] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102, 2001.

[10] T. Okano, L. L. Beranek, and T. Hidaka. Relations among interaural cross-correlation coefficient (IACC), lateral fraction (LF), and apparent source width (ASW) in concert halls. *The Journal of the Acoustical Society of America*, 104(1):255–265, 1998.

[11] L. Beranek. *Concert and opera halls: How they sound.* Acoustical Society of America, 1996.

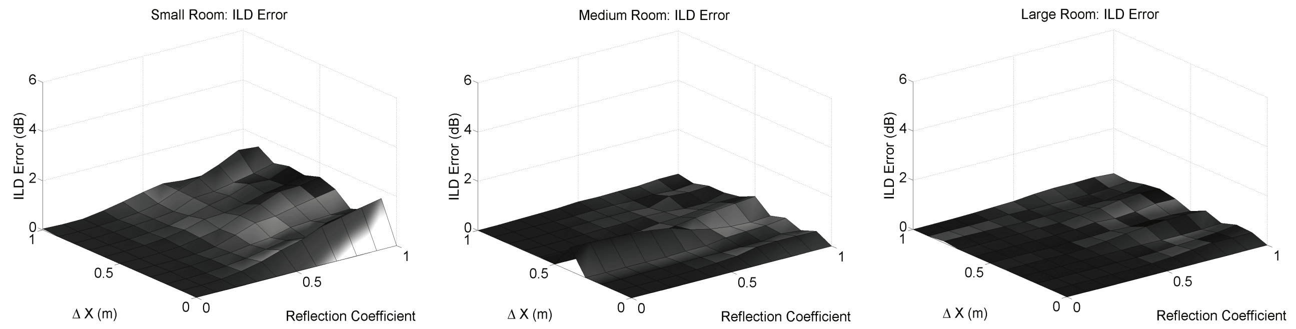

Figure 8: Interaural level difference error for small, medium and large rooms.

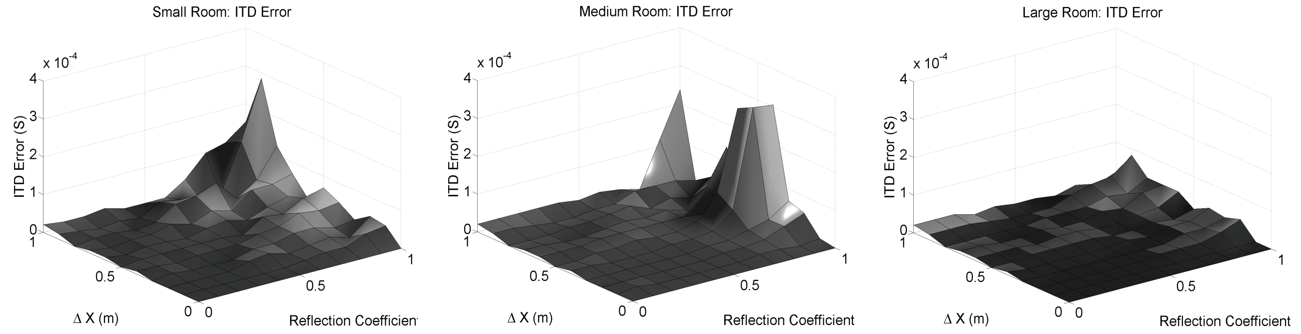

Figure 9: Interaural time difference error for small, medium and large rooms.

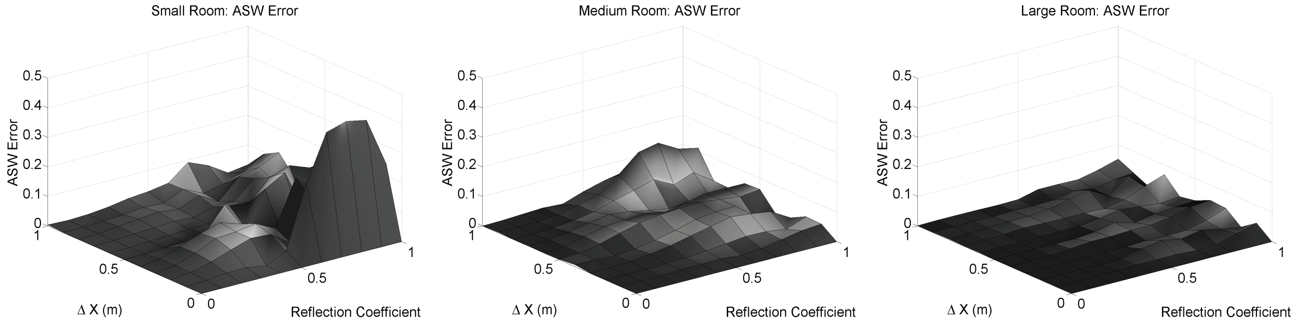

Figure 10: Apparent source width error for small, medium and large rooms.