Approximation of Binaural Room Impulse Responses

Gavin Kearney†, Claire Masterson∗, Stephen Adams, Frank Boland

Department of Electronic and Electrical Engineering
Trinity College Dublin, Ireland

E-mail: †gpkearney@ee.tcd.ie  ∗mastercp@tcd.ie

Abstract — An approach to approximation of Binaural Room Impulse Responses (BRIRs) is presented. Here, a BRIR is considered to have two constituent parts: the anechoic Head Related Impulse Response (HRIR), convolved with the Room Impulse Response (RIR). Each part is approximated individually. The HRIR is partitioned as the convolution of a directional independent component and a directional dependent subsystem. The RIRs are split into their early and late reflection components, and an interpolation method is presented using Dynamic Time Warping (DTW).

I Introduction

In recent years, the ability to convincingly portray an acoustic scene via real-time binaural (two-ear) reproduction has gained increased popularity in applications such as videoconferencing, cinematic and gaming presentations and music production. However, such auditory scene synthesis demands that the ear signals presented over headphones must be identical, or as close as possible, to signals resulting from real world stimuli. This means that the cues for correct localisation (interaural time and level difference) as well as the directional dependent filtering of the head and pinnae must be perceptually correct. These cues are represented in Head Related Transfer Functions (HRTFs), or their time-domain equivalent Head Related Impulse Responses (HRIRs), which can be measured at the ear canals from each direction in a full 360° sphere with source stimuli at different distances from the head.

HRIR measurements are generally taken at best in 5° spatial increments and frequently at even lower spatial resolutions due to their tedious nature of their capture. Given that localisation blur for azimuthal (speech) sources to the front of the head is approximately ±1.5° [1] there are is clear need for HRIR interpolation. Furthermore, since such measurements are typically undertaken in anechoic environments there is also a requirement for the superposition of the acoustics of a desired auditory environment with the head related responses. For convincing auditory scene synthesis, it is not sufficient to merely convolve a static set of room impulse responses onto a HRIR dataset.

Moreover, the listener (and virtual source) movements should dictate the control of a dynamically changing RIR filter kernel.

It is quickly appreciated that convincing binaural synthesis represents both significant measurement and computational challenges. The aim of this paper is to outline several strategies to accommodate these. We will begin by first investigating the redundancies in binaural measurements which can be exploited for computational benefit. We will then focus on the interpolation and perceptually based synthesis of RIRs which will cater for reduced measurement and data storage.

II Binaural Impulse Responses

To create localized sound sources artificially for binaural playback it is necessary to convolve a mono source, \( x(n) \), with the left and right ear HRIRs, \( h_l(n) \) and \( h_r(n) \). A relative movement between the virtual sound source and the listener requires the filter coefficients to be altered and taking into account the considerable length of HRIRs there is an obvious computational strain which may result in sound artifacts occurring. This is a significant motivation to explore splitting the HRIR into two components, one which changes with the position of the sound source or the listener and one which is position independent. The order reduction that this technique achieves is also a useful base from which to implement HRIR interpolation.

Existing techniques used to split or factorise HRIRs typically involve simple averaging in the time or frequency domains [2, 3]. However, the concept that there is a directional independent
component, that can be extracted from a HRIR dataset has previously been considered by Moller in [4]. This paper suggests that directional independence starts a few millimeters outside the ear canal. Moller also shows that the directional dependence as far as 6 mm outside the entrance to the ear canal is relatively small.

a) Common Subsystem Algorithm

Here it is proposed that a set of HRIRs (denoted $h^\phi$) be simplified by factoring each filter into the convolution of a direction independent subsystem (denoted $f$) which is common to the whole set and a direction dependent residual (denoted $g^\phi$). The algorithm used in finding this common subsystem of a HRIR dataset is equivalent to finding the approximate greatest common divisor (AGCD) of the HRTF z-domain set. We formulate the task of finding the AGCD as a non linear optimisation problem:

$$\min_{f, g^\phi} \sum_{i=1}^{N} ||h^\phi - (f \ast g^\phi)||^2$$  \hspace{1cm} (1)

where $h^\phi = [h_0^\phi, ..., h_{m-1}^\phi]^T$, $g^\phi = [g_0^\phi, ..., g_{j-1}^\phi]^T$ and $f = [f_0, f_1, ..., f_{k-1}]^T$

Our proposed iterative least squares method is equivalent to the divisor quotient method described by Chin et al [5] wherein they provide a proof of convergence to a point on the mean square error surface with gradient zero, i.e. a local minimum or maximum. They however operate on small numbers of relatively low order polynomials (generally <10th order) while in this paper such methods are applied to polynomials of order 200 of which range from hundreds to thousands of polynomials. The Divisor-Quotient iteration is shown below. Here, $i=$ iteration count:

1. Guess $f_0$ ($i = 0$)

2. Solve for each residual, $g^\phi$, as follows:

$$g_{i+1}^\phi = F_i^\dagger h^\phi$$ \hspace{1cm} (2)

where $F_i$ is the convolution matrix formed from $f_i$ and $\dagger$ denotes the Moore-Penrose pseudoinverse.

3. Solve for $f_{i+1}$ using

$$f_{i+1} = \begin{pmatrix} G_{i+1}^1 \\ \vdots \\ G_{i+1}^N \end{pmatrix}^\dagger \begin{pmatrix} h_1 \\ \vdots \\ h_N \end{pmatrix}$$

where $G_{i+1}^\phi$ is the convolution matrix formed from $g_{i+1}^\phi$

4. Set $i = i + 1$ and repeat steps 2 and 3 until there is convergence.

b) Results of Tests on CIPIC Database

The CIPIC database [6] is a public domain HRIR database which consists of 1250 HRIR measurements for each of 45 subjects. Each 200 sample long HRIR is measured at a location on a sphere of radius one meter centered on the subject head and is sampled at 44.1kHz. The results displayed below are using the left ear HRIRs from subjects 3 and 21. Subjects 3 are human subjects while subject 21 is the KEMAR dummy head with large pinnae.

Figures 1 and 2 show examples where a 100 sample long common subsystem has been extracted from the whole dataset (1250 measurements). There are three different initial guesses used in these examples; all ones, an average of all the HRIRs, and a random vector. The same random vector is used in each case. The HRIRs shown in these examples are from the $(0^\phi, 0^\phi)$ position i.e. straight ahead of the subject. Figure 1a compares an original HRIR for subject 21 to reconvolved versions of the same HRIR originating from three different initial guesses of the common subsystem. Figure 1b shows the frequency domain comparison. It is evident that there is negligible difference between the original and the reconvolved versions for each initial condition. Figures 2a and

![Figure 1: Subject 21. Comparison of original HRIR to reconvolved HRIRs with different initial $f_0$ guesses](image-url)
For this to be effective, the characteristics of the room response should change with source/listener position and orientation. To this end we focus on the synthesis of large datasets of RIRs utilized in such auralization processes, that have perceptual equivalence to real-world measurements. The motivation is to reduce the effort involved in measuring and storing data pertaining to acoustic fields, and this work is seen as a precursor to applicable spatialisation techniques for binaural auralization. These include higher order Ambisonics [7], as well as the recently suggested Virtual Wave Field Synthesis system [8], both of which benefit from measurement of real acoustic fields for convincing auralization.

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

$$h_i = [h_i^e[1 : n_t] : h_i^l[(n_t + 1) : N]]$$

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 [9]. Here we use the method suggested by Naylor and Rindel [10] 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

\[ T_{ro} = 4V \frac{c}{S} (O_e + 1) \]  

(5)

where \( V \) is the room volume, \( S \) is the surface area, \( c \) is the speed of sound and \( O_e \) is the reflection order [11]. 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 decorrelation method suggested by the authors in [12].

a) Interpolation of early reflections through Dynamic Time Warping

Let us consider the case where we have measured two RIRs \( h_e^1 \) and \( h_e^3 \) at points R1 and R3 as shown in Figure 4. We will now attempt to create a new interpolated (direct-sound and early reflections) RIR, \( \hat{h}_e^2 \), as if it were measured at position R2. Figure 5 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 sparseness means that linear interpolation can result in significant smearing of reflections in the interpolated result, as shown in Figure 6. 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 not guarantee that subsequent reflections will match up, due to different reflection path lengths at different positions in the room.

It is therefore necessary to temporally align the early reflections of the impulses prior to interpolation. Dynamic Time Warping (DTW) [13] 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 main feature points. 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
\[ D(W) = \sum_{k=1}^{K} D(w_{k1}, w_{kj}) \] (6)
where \( D(W) \) denotes the distance of the warp path and \( D(w_{k1}, w_{kj}) \) 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 7.

![Fig. 7: Accumulated Distance Matrix.](image)

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 (\( h_1^e(n_1) \) and \( h_3^e(n_1) \)), 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 \( h_2^e \), 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 warped versions of the direct sound and early reflections of both RIRs and allows for simple linear interpolation between them to obtain the magnitude of the unknown RIR, \( h_{w2}^e \). This is shown in Figure 8.

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.
\[ \alpha = \frac{1}{d_3} - \frac{1}{d_2} \] (7)
and hence,
\[ h_{w2}^e = \alpha h_{w1}^e + (1 - \alpha) h_{w3}^e \] (8)

Now the warp vectors, \( w_1 \) and \( w_3 \) that describe how \( h_1^e \) and \( h_3^e \) are mapped onto \( h_{w1}^e \) and \( h_{w3}^e \) by the DTW must be interpolated to obtain \( w_2 \). Again linear interpolation is used to accomplish this and the weights, \( \beta \) and \( 1 - \beta \), are calculated based on the distances of the microphones to the source.
\[ \beta = \frac{d_2 - d_3}{d_1 - d_3} \] (9)
and hence,
\[ w_{int} = \beta w_1 + (1 - \beta) w_2 \] (10)

The final step in the process is to map the warped interpolated vector back into the “un-warped” time domain using the interpolated warp vector. Figure 9 shows a comparison of an interpolated RIR \( h_2^e(1) \) and a real measured RIR \( h_2^e(1) \) taken from the desired interpolated position. We notice that the temporal distortions that were present in the linear interpolation of Figure 6 are no longer present.

IV Conclusions

The process of producing real-time auralization for binaural room synthesis is a challenging one, which not only requires considered measurement processes but also computationally expensive rendering implementations. We have presented in this paper several techniques which can be used reduce such complexities. In a real time interactive environment the extraction of a common subsystem should ease the computational strain as relative movements between the source and listener require fast HRIR changes. There are also possible applications of this method to the spatial interpolation of HRIRs and to HRIR classification. Further work is needed to examine the convergence of the common subsystem algorithm.

The performance of DTW interpolation as a precursor to spatialisation systems such as higher order ambisonics and wave field synthesis also needs to be studied, as well as the application of the method to interpolate across different spatial positions and room acoustic parameters.
Fig. 9: Comparison of impulse response created from DTW interpolation between two RIRs to an actual impulse response measured at the position of interpolation.

V 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


