An EM-Derived Approach to Blind HRTF Estimation

Eric Schwenker and **Griffin Romigh**
CS 229 Final Project
December 12th, 2014

Abstract

Current 3D audio technology used for virtual and augmented reality systems lack the immersive qualities of real acoustic spaces. This limitation is rooted in an inability to easily measure individualized head-related transfer functions (HRTFs) in a commercial setting. This study shows how the iterative construct of a joint-maximization EM algorithm can be applied to derive a novel method for cheap, "portable" HRTF estimation that eliminates both head-tracking and/or prior source location knowledge from the process.

1 Introduction

In natural listening environments, humans use a unique set of acoustic cues to localize incoming sounds. These localization cues, collectively represented by a head related transfer function (HRTF), describe the acoustic transformations caused by interactions of a sound wave with a listener's head, shoulders, and outer ears. In virtual audio applications, an individually measured HRTF is essential because it is used to give headphone-based sounds a realistic virtual spatial origin and immersive realism. Accordingly, if the HRTF is non-individualized or poorly estimated, it has the tendency to cause unwanted distortions and undesirable artifacts in the sound's perceived location when presented over headphones [1].

Traditionally, HRTFs are measured acoustically, based on binaural ("two-ear") recordings of a known test stimuli that are played from loudspeakers in 3D space. In general, the HRTF is a 2D continuous function defined over the unit sphere in 3D space, and a measurement for a given loudspeaker location represents a *sample* HRTF. It thus follows that a full representation of an HRTF obtained via conventional discrete acoustic measurements, requires an effective interpolation of collected *sample* HRTFs throughout all of space. Fortunately, with the development of a spherical harmonic (SH) based HRTF interpolation technique [2], a continuous individualized HRTF can be estimated using many spatially distributed samples.

While intepolation of a continuous HRTF is possible from discrete HRTF measurements, it requires knowledge of each *sample* HRTF's location. In typical laboratory setups, this information is provided by knowing a-priori a loudspeaker's location relative to a listener's fixed head position, or tracking a listener's moving head while keeping the loudspeakers fixed. Unfortunately, even though simple head tracking technology has recently become more cost effective, it still presents a significant financial investment for the average potential consumer of virtual audio.

A potential solution to this problem would be to try to estimate a continuous HRTF without head-tracking or loudspeaker arrays. Consider a *sample* HRTF collected without knowledge of the spatial origin of the recorded stimulus. This observation has some interesting consequences for HRTF estimation in context of machine learning paradigms because viewing this location as a latent variable converts HRTF estimation into an unsupervised learning problem. Here, the goal of estimating an HRTF could be approached as a two-step iterative machine learning algorithm: (1) binaural source localization based on knowledge of the underlying continuous HRTF and test stimulus, and (2) continuous HRTF interpolation from knowledge of several *sample* HRTFs and their locations. Figure 1 below illustrates the "chicken-versus-egg" nature of this two-step estimation problem.

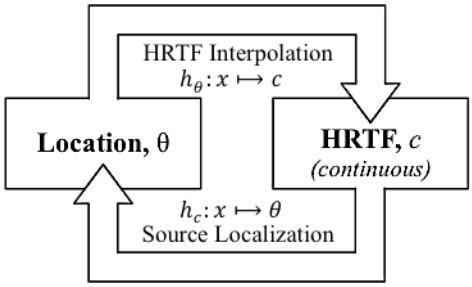

Figure 1: "Chicken-versus-egg" problem for HRTF estimation

This study presents a novel method for HRTF estimation derived from Expectation-Maximization (EM) theory, by collecting *sample* HRTF measurements on individuals varying their head rotation angle relative to a single fixed speaker without explicit knowledge of the source location. The goal is to achieve an accurate estimate of an individual's HRTF by eliminating any sort of head-tracking or prior source location knowledge from the process.

2 Data collection

Datasets of binaural recordings containing approximately 277 non-redundant measurements (evenly distributed over 3D space) were used in place of real-time measurements for the purpose of developing and refining the EM algorithm. Twelve separate datasets were collected in all, representing sets of binaural recordings collected for twelve different human subjects. Let $R_{\theta^{(i)}}$ represent a "clean" binaural recording at a given location $\theta^{(i)}$. "Clean" refers to the fact that these binaural recordings were collected in a quiet anechoic chamber. As a pre-processing step before initialization of the algorithm, the collection of binaural recordings were converted to *sample* HRTFs, $x^{(i)}$, computed as $x^{(i)}(\omega) = R_{\theta^{(i)}}(\omega)/s(\omega)$, where s was the source signal used in the recordings. Both $R_{\theta^{(i)}}$ and s were considered in the frequency domain so that a $x^{(i)}$ could be obtained through simple division.

3 Joint maximization formulation

EM is a prevailing methodology for maximum likelihood parameter estimation in models with hidden or unobserved dependencies. To formalize the HRTF estimation problem in context of EM, let $\mathcal{X} = \left[x^{(1)}, \dots, x^{(m)}\right]^{\mathrm{T}}$ denote a space of m "unlabeled" *sample* HRTFs, and consider that if a set of known sound source location "labels" $\Theta = \left[\theta^{(1)}, \dots, \theta^{(m)}\right]^{\mathrm{T}}$ for $\mathcal{X}$ existed, finding the continuous HRTF parameters c, would become the standard continuous HRTF interpolation procedure. Likewise, having complete prior knowledge of the continuous HRTF, c, and $\mathcal{X}$ would trivialize the labeling of each recording, becoming binaural source localization. The proposed technique is an attempt at estimating the continuous HRTF,c, having a dataset $\mathcal{X}$, but no knowledge of Θ, hence the HRTFs collected are "blind" to location.

For this application, it is useful to begin the formulation of an EM as a coordinate ascent process. With this, the algorithm becomes a joint maximization procedure that iteratively maximizes a function, $\mathcal{F}(Q, c)$, where c is the parameter described above, and $Q_i(\theta^{(i)})$ is an arbitrary distribution over the unobserved variables, given the support $Q_i(\theta^{(i)}) = p\left(\theta^{(i)} \middle| x^{(i)}; c\right)$. The iterative procedure is carried out in two alternating maximization steps on function, $\mathcal{F}(Q, c)$, and proceeds as follows, repeating until convergence:

M-Step 1

For each i, set ${Q_i}^{(t)}$ to the Q_i that maximizes $\mathcal{F}\left(Q_i, c^{(t-1)}\right)$

$$ {Q_i}^{(t)} = \arg\max_{Q_i} \mathcal{F}\left(Q_i, c^{(t-1)}\right) $$

where $c^{(t-1)}$ is either the initial c_0, or the updated $c^{(t)}$ from previous iteration of M-Step 2.

M-Step 2
Set $c^{(t)}$ to the c that maximizes $\mathcal{F}\left(Q_i{}^{(t)}, c\right)$

$$c^{(t)} = \arg\max_{c \in \mathbb{C}} \mathcal{F}\left(Q_i{}^{(t)}, c\right)$$

where with this standard coordinate ascent view, the resulting algorithm is maximizing $\mathcal{F}$ in a process akin to maximizing a tight lower bound to the true likelihood surface [3].

In this study, a broader view of the traditional EM algorithm is adopted, in which focus is placed on the strategy and overall success of the individual M-Steps rather than successful attainment of the optimal joint distribution, $\mathcal{F}$. This choice makes the formulation of the algorithm more amenable to the existing body of work on HRTF interpolation and binaural source localization, as practical constructs established from within existing research can be used to approximate each M-Step. This is the topic of the proceeding section.

4 EM-derived approach

To complete the formulation of the EM-derived approach, consider that M-Step 2 and the process of continuous HRTF interpolation share a common objective: find a c, (continuous HRTF representation) which optimizes the probability of a source location corresponding to a given *sample* HRTF. Accordingly, it is logical to assume that an effective HRTF interpolation strategy provides a sufficient proxy to the closed-form solution for the update of parameter c. Under this assumption, M-Step 2 can be viewed as a module that performs some sort of HRTF interpolation procedure given the information from M-Step 1. Using similar logic and considering a greedy distribution, $Q_i\left(\theta^{(i)}\right)$, (constrained so as to assign zero probability to all but one value of $\theta^{(i)}$, so that $Q_i\left(\theta^{(i)}\right) = p\left(\theta^{(i)} \middle| x^{(i)}; c\right) = 1\left\{\theta^{(i)} = \arg\max_{\theta} f_c\left(\theta^{(i)}, x^{(i)}\right)\right\}$) [4], both M-Step 1 and binaural localization models function as the machinery for identifying a $\theta^{(i)}$ that maximizes $Q_i(\theta^{(i)})$. As such, the joint maximization of $\mathcal{F}$ can be recast according to the module convention outlined above as:

Module 1: Binaural Source Localization
For each i, set $\theta_i^{(t)}$ to the $\theta^{(i)}$ that maximizes $f_{c^{(t-1)}}\left(\theta^{(i)}, x^{(i)}\right)$

$$\theta_i^{(t)} = \arg\max_{\theta} f_{c^{(t-1)}}\left(\theta^{(i)}, x^{(i)}\right) \approx \mathcal{F}\left(Q_i{}^{(t)}, c^{(t-1)}\right)$$

where $c^{(t-1)}$ is either the initial c_0, or the updated $c^{(t)}$ from a previous iteration of Module 2, and $\mathcal{F}\left(Q_i{}^{(t)}, c^{(t-1)}\right)$ represents the resulting maximized function from M-Step 1. Note that the initial c_0 represents the average continuous HRTF from an existing database.

The function $f_{c^{(t-1)}}\left(\theta^{(i)}, x^{(i)}\right)$ computes the similarity between the spectra of $c^{(t-1)}$ at $\theta^{(i)}$ with $x^{(i)}$, thus for each $x^{(i)}$, the localizer tries to find a $\theta^{(i)}$ that maximizes their similarity. Note that with this modular formulation, any binaural localization algorithm can be substituted and tested in Module 1 as long as it follows the same general construct as given above.

Module 2: Continuous HRTF Interpolation
Set $c^{(t)}$ to the c that minimizes $f_{\theta^{(t)}}\left(c, x^{(i)}\right)$

$$c^{(t)} = \arg\min_{c \in \mathbb{C}} f_{\theta^{(t)}}\left(c, x^{(i)}\right) \approx \mathcal{F}\left(Q_i{}^{(t)}, c^{(t)}\right)$$

where $f_{\theta^{(t)}}\left(c, x^{(i)}\right)$ is a function that computes the mean square estimate for c, given the *sample* HRTF measurement $x^{(i)}$.

Alternatively, $c^{(t)}$ can be expressed as the least squares solution

$$c^{(t)} = (\mathrm{Y}^{\mathrm{T}}\mathrm{Y})^{-1}\mathrm{Y}^{\mathrm{T}}\Theta^{(t)}$$

where Y represents a matrix of spherical harmonic (SH) basis functions (see [2] for more details), and $\Theta^{(t)}$ represents a vector of $\theta_i^{(t)}$ for all i from Module 1. Remember, $\theta_i^{(t)}$ has an x dependence (from Module 1), thus the alternative view is not an attempt to eliminate an x dependence, but rather, it is motivated by the discussion of the results. The basis functions contained in Y are indexed according to their order, a constant that determines the rate of spatial change of the basis function over the sphere, and this is presumed to have an effect on the interpolation scheme (and therefore the proceeding localization). Again, with this design, other HRTF interpolation schemes can be substituted and tested in Module 2 without forcing a re-derivation of the entire construct. Note that it is assumed that $c^{(t)}$ is a suitable approximation for the result of M-Step 2.

5 Results and discussion

Twelve separate ground truth HRTFs representing twelve individual subjects by proxy were used for each experiment; these HRTFs were custom measured on real human participants. The metrics used to define the success of the algorithm are expressed according to the purpose assigned to each module. Moreover, the purpose of the binaural localizer in Module 1 is to find a $\theta^{(i)}$ that maximizes the similarity of the spectra under comparison, which is analogous to minimizing the localization error for each given $x^{(i)}$. This localization error across all subjects (for sets of random locations not contained in the training set) is measured with average angular error and is given in Figure 2a below as a function of the number of measurements (number of discrete binaural recordings) used in estimating the continuous HRTF parameter, c.

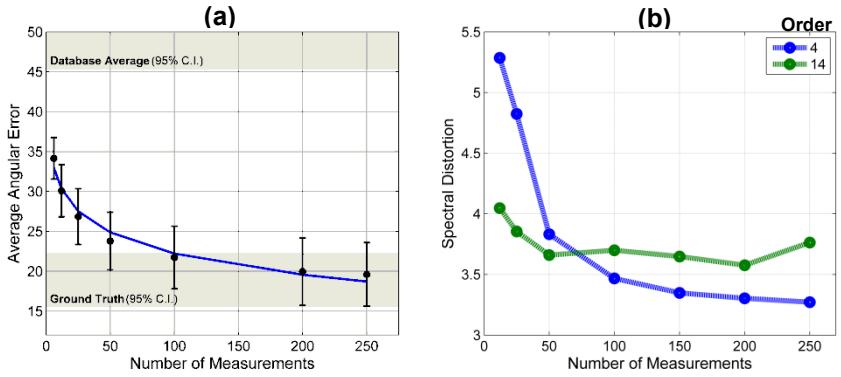

Figure 2: (a) Average angular error (simulated localization with final *c* estimate) vs. number of training locations (b) Spectral distortion vs number of training locations as a function of SH interpolation order

The results of Figure 2a indicate that the average angular error in the simulated localization (over "unseen" sample locations) decreases rapidly over small measurement set sizes. This means that the localization benefit obtained from a small measurement set size is well generalizable over all relevant locations, and that it is likely unnecessary to consider any measurement set sizes over 100 as the localization benefit for those sizes appears negligible.

To evaluate the performance of Module 2, consider that a continuous HRTF that is sufficiently interpolated yields minimal spectral distortion when compared to the *sample* HRTF from which it is constructed. Figure 2b above highlights how the choice of order in the continuous HRTF interpolation (Module 2) affects spectral distortion, again is given as a function of the number of measurements used in the parameter estimation. The results show that for a small measurement set size (<50 locations), a high (14th order) representation produces a less distorted estimation of ground truth. Since the higher orders in the spherical harmonic interpolation strategy represent a projection onto a basis with a large frequency over space, it is plausible that when given a sizable measurement set size (>50), higher orders begin over fitting the data. All things considered, the lower (4th) order representation appears to benefit most from the design of the algorithm and presents an interesting point of discussion, as Romigh *et. al* [2] find that a 4th order SH representation achieved localization accuracy at a level of performance comparable

to a fully individualized HRTF, despite the fact that a low order representation induces a significant amount of both spectral and spatial smoothing. The spectral smoothing present in a 4th order estimate can be visualized in Figure 3

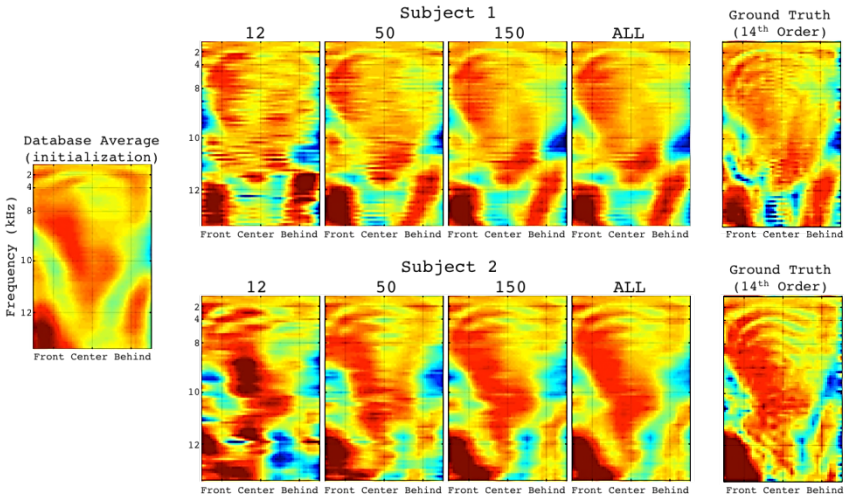

Figure 3: 4th order HRTF magnitudes (in dB) plotted as a function of angle along the median plane (comparison between 2 different subjects for 4 different measurement set sizes)

Figure 3 shows the dramatic progression of the estimation procedure as more locations are used in formation of the estimation. Each time the algorithm is run, it starts with the same initial guess for c_0(the database average), given on the far left and as is shown, does begin to capture the important features of an individual's ground truth 14th order HRTF that was custom measured in an anechoic facility. As a final point, it's important to recognize that the results presented here were exploratory in scope and that these insights into ideal measurement set sizes and interpolation order, will serve to help guide the algorithm towards a more refined design.

6 Conclusion and future work

This study used the iterative construct of a joint-maximization EM algorithm to derive a novel method for HRTF estimation that eliminates both head-tracking and/or prior positional source location knowledge from the process. Keeping to the practical problem, assumptions were made which generalized each M-Step into a more modular form and did in fact eliminate some of the rigor built into the strict EM formulation; however, the overall method was successful in its estimation of a continuous HRTF across a database of twelve subjects, as both the simulated average angular error on testing data and the spectral distortion improved over the course of iteration. To ensure robustness for measurements taken in everyday listening environments (the eventual objective), the necessary next steps involve consideration of non-anechoic and noisy recordings to see how the proposed method handles more realistic input data. Furthermore, note that currently, the testing locations are evenly distributed over 3D space. Randomizing or perhaps developing realistic paths (representing the motions of a person using the technique) would present an interesting follow up study, as well. Finally, it is essential to begin formulating a plan for perceptual testing of the estimated continuous HRTF structures, to ensure that the metrics defined function as suitable maximum likelihood estimators.

Authors' contributions to manuscript
G.R. advised E.S. // E.S. responsible for coding and write up. // HRTF interpolation step based on G.R.'s PhD thesis [2].

References

[1] W.M. Hartmann and A.Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, pp. 3678–3688, 1996.
[2] G. D. Romigh, D. S. Brungart, and R. M. Stern, "A continuous hrtf representation for modeling and estimation," 2012.
[3] A.Y. Ng, "The EM Algorithm," *CS229 Course Notes,* Fall 2014.
[4] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants." In: M. Jordan, editor, *Learning in Graphical Models*, pp 355-368. 1998.