Automated HRTF Individualisation Based on Localization Errors in 3D Space

Robin Yonge

August 2017

Background

TODO: sort out subsubsection topics to make it more elegant, fill out notes'd bits, maybe refer to this bit as introduction or motivation? mention my interest in the overall UX of the calibration process

How Humans Localise Sound/What do I mean by spatial audio?

add in a bit about this, citing blauert, ez

When HRTFs are applied to an arbitrary signal and presented to the listeners two ears through headphones, he or she hears a virtual target that appears to originate from the location of the original sound source ?e.g., Wight-man and Kistler, 1989b; Bronkhorst, 1995; Mller et al., 1996?.

users inability to localise using someone else's hrtfs in: **?**

The Present Importance of Spatial Audio etc

Spatial audio has never been more important. Though there has been a steady stream of interest in applications of spatial audio in fields like defence - primarily applied to Virtual Auditory Displays (VADs)(**?**) - virtual, augmented, and mixed reality form a large component of the current technological zeitgeist. One major stated goal of these technologies is that of immersion. This doesn't have to mean that the user feels as if they have been transported to somewhere completely new, they just have to believe in the virtual elements of the experience. No matter whether the intended application is entertainment, productivity, or assistance (unsure about this line), the end user must be deceived into believing in what they are experiencing.

Audio has parity with visuals here, as just small errors in either can irreparably break immersion(**?**). listeners often complain that auditory events are spatially diffuse, and listeners often make incorrect judg- ments of the source locations ?Wenzel et al., 1993; Mller et al., 1996?. [something about binaural audio here??? Explain meaning etc??? hrtfs as the principal determinant of sound localisation blah Reproducing spatial audio convincingly involves a number of factors, including reflections and occlusion caused by the room and the objects in it. This project however, will focus entirely on the effect the anthropometry of the listener has on the audio signal - the attenuation of the sound caused by the various body parts that they sound waves come into contact with. [also mention ITD and IID]

Recreating Spatial Audio/Representing this change?? idk

In most applications involving spatial audio, representing this attenuation is done using Head-Related Transfer Functions, or HRTFs, derived from their time-domain counterparts Head-Related Impulse Responses, or HRIRs. HRIR measurements are taken by placing microphones in the ears of a participant (human or mannequin) and measuring the impulse response resulting when a tone is played from a loudspeaker (**?**). This measurement process should be repeated for as many positions/points of origin around the participant as possible in order to maximise coverage and provide the greatest amount of information

when it comes to using the information to process audio. This process is incredibly labour-intensive, requires specialist equipment, and can take hours to perform. As a result, there are few organisations capable of performing these measurements, and generating a set of HRTFs for most people is impractical at best. [maybe a section on databases]There are a few organisations that have assembled databases of HRTFs, that involve measurements from a range of participants. The two main differences in these databases are the number of source positions, and the number of participants.

Source Positions:

The number of source positions varies from database to database [in the case of CIPIC it is every L degrees from N to M, in the case of ARI, it is every Y degree from X to Z, etc] [add diagrams!]

Subjects:

These databases may contain anything from data from a single mannequin in the case of the MIT KEMAR set (**?**), to the CIPIC database's 45 subjects (**?**), up to the 110-subjects-and-growing ARI HRTF database (**?**).

[table of databases by subjects and participants maybe?]

The Problem

Because of the aforementioned difficulty in measuring HRTFs, data from these databases is commonly used in attempts to implement spatial audio solutions. In the simplest implementations, the audio sample is convolved with the HRIR, producing audio that appears to come, convincingly or otherwise, from the position in 3D space that the HRIR was originally measured from.

The problem with using this data in any spatial audio implementations that are to be used in applications for the consumption of a wide range of end users, is that HRTF data is incredibly specific to the person the measurements have been taken from. Just small differences in the anthropometry of the measured participant and the end user can compromise the efficacy of the HRTF used(**?**). However, when one tries instead to use a generalised HRTF - derived from the average of a set of measurements, or from a mannequin like the KEMAR(**?**) - the processed audio becomes too average and the same problems arise as using an HRTF from a single human. When using HRTFs that are not well matched to the user, front/back and elevation confusion is very common (**?**).

Re-mention that VR/AR is popular, and that audio is integral to the progression of the technology!!!!.

It follows, then, that in a system that implements binaural audio, the audio for a user would be processed using a set of HRTFs could be used to recreate audio in such a manner that the user would be able to accurately localise the source of a sound. As we have already established, the traditional method of measuring HRTFs is impractical for the vast majority of users, which leaves us at something of an impasse. We need a method for producing individualised

sets of HRTFs with minimal specialist equipment, an easy user experience that does not require expert knowledge, as small a time investment as possible.

Literature Review

Modifying HRTFs/Overview

This idea of generating individualised sets of HRTFs without having to perform the complex measurements that would usually be required has existed since the 1990s (**?**). The ideal scenario for commonplace spatial audio involves every user having access to an HRTF set that works for them. If traditional methods of measurements are impractical, then alternatives are necessary.

Methods

Investigations into HRTF individualisation have been done using a range methodologies, some involving just simple selection tasks and others complex tuning - adjusting multiple parameters against listening tests. Often these methods hinge on a specific model that is used to decompose the HRTF into individual parameters that can be manipulated independently in order to achieve meaningful control over the customisation process. In some cases these models also seek to make clear the relationship between the features of the HRTF and the features of the user - the morphological properties of the measuree being the primary determinant of generated HRIR this seems a logical approach. In these next few sections I will cover the main of the approaches that have been investigated to date, as well as their efficacy and why they are or are not well suited to this project. We will see that there is a definite overlap between these methods, leading to the idea that perhaps in a more comprehensive but laborious model for HRTF individualisation, a combination of these techniques might be used (**?**).

Database Matching

Database matching is often incorporated into other models for HRTF individualisation, and is somewhat self-explanatory. It is based on the predicate that within a database of a given size, there must be a set of HRTF measurements that have been taken from a participant with similar anthropometric features as a given user. This technique has been used in a range of studies on spatial audio, both as part of a wider study on binaural audio and localisation, (**?**) and as the sole focus of the study (**?**). As in both of these papers from Zotkin et al, many attempts to match participants with closely matching HRTF sets use measurements of the user's anthropometry, which they will then try to match to the anthropometric measurements taken in the process of assembling the database.

This can work reasonably well, assuming the database used contains measurements from a great enough range of people. The CIPIC database contains anthropometric measurements for all 45 of its participants (**?**), while the ARI database comes with measurements for 50 of its participants (**?**). Problems with

this method can of course arise when the database does not include measurements from a participant with a morphology that does not closely match those of the user. The second problem with this method is more of an issue when considering this method in terms of what this project hopes to achieve. Given the my requirements for the customisation method I intend to design, a method that requires precise measurements that would be difficult to perform at home is not ideal. A method that requires precise measurements to be taken has two problems, both in the difficulty of performing the measurements, and the effort that such an act involves, does not satisfy any of my self-imposed standards for user experience.

An alternative method for matching users to their closest-matching HRTF set could be based on subjective listening tests. Playing a user a sample, filtered using an HRTF taken from a database, and asking them to indicate where they believed the sound came from. This process can be repeated for as many examples as are contained in the database, and the one that results in the least incorrect localisation attempts chosen. The problems with this method are again clear, in that the labour required to search all the entries in a database is more than anyone but the most die-hard users are likely to pursue [citation for some human-computer interaction junk about how much effort people will put in?]. Improvements are made on these kinds of subjective listening tests, however, in attempts to match users to more appropriate HRTFs through the clustering of similar sets.

Clustering

Clustering involves collating a database of HRTF sets measured from different participants, and then sorting these into orthogonal groups based on a specific feature. Fahn and Lo in their 2003 paper (**?**) grouped HRTFs based on the power cepstra of each HRTF set. They then used a modified version of the LBG algorithm to form 6 different clusters. Other studies, such as a 2013 paper by Xie et al **?** found a total of 7 clusters were required. Either way, the idea is to group HRTFs into groups - or clusters - where each HRTF is similar enough to the others in the cluster, but where the differences between each cluster are sufficiently great. You can then take the central example from each cluster, the HRTF that best represents that cluster or that represents the average, and provide to the end user the example from this set of 6 or 7 that best matches them.

Given that clustering is meant to make simpler the process of matching a user with a more personal HRTF set, trying to match users by anthropometry again would be nonsensical. Instead, subjective listening tests are used more often **?** (**?**). Using this method, the comparative efficacy of subejctive tests in this instance is clear versus raw database matching. As opposed to subjecting an undending barrage of tests against 45 or more (as a slightly facetious example), the listener has to compare between 6 or 7. However, the resulting localisation is going to be less precise, given the inherently more generalised approach. The increase in user-friendliness it interesting, though. In lighter-weight applications of VR/AR, perhaps for example on mobile devices, this approach could work. Giving interested users the option to choose between a subset of sufficiently disparate HRTFs, adding a little lightweight customisation.

(**?**) (**?**)

Frequency Scaling

Another methodology is based upon scaling in frequency entire HRTFs or elements of the HRTF. A method that was investigated early on in attempts to devise individualisation methods, it is one that lost out to cluster/database matching methods in the longer run.

Some notable examples of studies into this technique include one by Middlebrooks **?**. In this study, they used Directional Transfer Functions (DTFs) which are processed HRTFs with the source location information isolated **?**. Initially finding that spectral features from one participant's DTF could be aligned with those of another by scaling. In further investigation participants used DTFs from the other participants, which were then scaled by a range of different factors based on comparisons in the two participants anthropometry - primarily the size of the head, and pinnae. This study then compared the participant's ability to localise sounds convolved with another's DTF against localisation when using the scaled DTFs and found a roughly 50% increase in accuracy with the most effective scale factor.

Another method investigated by Tan et al **?**, involved building a tool that allowed users to manipulate the scaling of an HRTF themselves. Given that front/back and elevation confusion is most common when using non-individual HRTFs, they opted to provide options to add a bias towards the front/back, as well as another parameter to tweak how elevation was perceived. Their results showed an improvement over the non-individualised HRTFs, but not a substantial enough one. Given the simplicity of adjusting a mere two parameters this approach could have been very convenient. But the lack of an impressive improvement in localisation makes it a less tenable solution than some of the others explored, and it is overshadowed by later methods.

Structural Models

Structural models appear to be the most commonly studied models for understanding HRTFs as well as for attempting to synthesise individual sets or customise generalised sets[I have roughly a billion citations for this, should I just plonk them all here?]. Because HRTFs and HRIRs represent the affects on the sound signal/wave of the features of a human's body, then one should be able to extract and isolate the discrete elements an HRTF that relate to the individual body parts. A German researcher by the name of Klaus Genuit first proposed a model for understanding HRTFs as a series of filters that each represented the effects of a certain anatomical feature (**?**). The idea of a structural model, or of HRTF individualisation based on a user's anthropometry is pervasive, and many other methods incorporate elements from it. For example the aforementioned 2003 study by Zotkin, Duraiswami, Davis, and Hwang (**?**) used anthropometric measurements to match a user to closely-matching set of HRTFs from the CIPIC database. Similarly, later studies centered on Principal Components Analysis - something I will go into in more detail later - look at the relationship between principal components (PCs) and morphological features.

1998 work by Philip Brown and Richard Duda **?** (itself based on a 1996 paper **?**) looked primarily at HRIRs, focusing on the additional temporal information that the frequency-domain HRTFs lacked. The decision to focus on the time domain was to allow them to identify the characteristics of HRTFs that are the

result of the different paths to the inner ear that the sound waves took, over time. This study involved only a small number of participants, and so whether or not the synthesised HRTFs produced with this model could replace measured ones is left to more comprehensive studies.

In a 2001 study Algazi, Duda, Morrison, and Thompson attempted to produce an approximated HRTF from the isolated responses of different structural components (**?**). As with other studies, the synthesis was performed based on anthropometric measurements of the subjects, and the final HRTF composite - made up of the responses of each structural component. This approach was evaluated using a composite HRTF vs a measured HRTF, and when viewed spectrally the two had significant similarities. The study did not go as far as to perform subjective/psychoacoustic tests, however. A similar study in 2003 by Raykar and Duraiswami (**?**) aimed to decompose the HRTF into a set of significant features that are integral to the localisation of sound sources. Their results were promising, developing an algorithm to decompose a given HRTF, and testing it successfully on every participant in the CIPIC database.

This model often provides promising results, and can be well-suited for applications where a high level of localisation accuracy is required, but a full measurement session is out of the question. The main problem with this method being the precision that is required for the measurements. If an ideal implementation for widespread consumer use relies on a simple calibration process, detailed measurements it becomes more difficult to fulfil that requirement. Investigations have been made into the use of computer vision to automate the measurement process and eliminate human error(**?**)[find citation again], but unless this process can be distilled into something simple that requires minimal additional hardware (perhaps a smartphone camera), then it is sub-optimal for widespread use.

More recent studies have tried to combine this approach with others, using a combination of PCA and a reduced number of anthropometric measurements (**?**). Or a similarly reduced number of measurements to match a set of HRTFs to a subject, then modify them to improve their performance (**?**). Even trying combination of structural data and a radial basis function (RBF) nerual network (**?**).

Principal Components Analysis

Understanding PCA

PCA and HRTFs/HRIR

test citation (Hölzl, 2012)

Search Methods

Simulated Annealing

Method

Analysis

Discussion

Bibliography

Josef Hölzl. *An initial Investigation into HRTF Adaptation using PCA IEM Project Thesis*. PhD thesis, Graz University of Technology, 2012. URL `https://github.com/jhoelzl/HRTF-Individualization`.