Personalization of Head-Related Transfer Functions with Limited Acoustic Measurements

Griffin D. Romigh
Dept. of Elec. & Comp. Eng.
Carnegie Mellon University
Pittsburgh, PA 15213
gromigh@ece.cmu.edu

Abstract

The main goal of this project was to take advantage of the spatial structure inherent within a set of HRTFs in order to provide a mechanism by which an entire set of individualized HRTFs could be estimated from a small set of measurement locations. Several methods were investigated for this purpose including missing feature techniques such as latent variable model decomposition and k-Nearest Neighbor averaging, as well as naïve methods such as spherical harmonic decomposition and linear interpolation. The most successful method investigated was a Linear Minimum Mean Square Error (LMMSE) estimation procedure which showed near complete reconstruction of the 232 HRTFs from a subset of less than 30 randomly distributed locations and promising results for as few as 20 locations. More practical measurement schemes like measurements only taken on the horizontal plane proved to be less successful than even distributions; however they still provided benefit over other rapid HRTF personalization techniques such as derivation from anthropometric measurements.

1 Introduction

Human listeners have the ability to determine the location of a sound source in three dimensions. This ability is largely due to robust cues based on the spatial separation of the two ears which causes a sound originating off to one side of the head to arrive at the near ear sooner and with greater amplitude then at the far ear. Thus creating an inter-aural time difference (ITD) and an inter-aural level difference (ILD). While the ITD and ILD dominate sound localization in most regards, these two cues alone fail to distinguish sound source positions which lie on contours of constant relative distance from the two ears; for instance a source directly in front of the listener and one directly above. In reality, these equal distance contours occur at every location in space and are referred to as cones of confusion. Sound source localization within one of these cones of confusion is accomplished using less robust spectral cues caused by the acoustic filtering properties of the listeners head, shoulders, and outer ear.

The concepts needed to accurately recreate this complete set of cues have been known for some time, and involve the characterization of the acoustic filtering effects of the listeners' anatomy. This characterization is accomplished by playing a known source from many locations in space and making recordings at the entrance of the listeners' two ear canals. For a given location in space the complex ratio of the Fourier transform of the recorded signal to that of the original signal is called the Head Related Transfer Function (HRTF), and can be used to recreate the entire set of cues needed for sound source localization. To adequately generate a 3-D representation anywhere in space however, HRTFs for more than 250 locations in space must be collected, and are generally only applicable for the person they were measured on. These two limitations make attaining high fidelity spatial audio difficult.

Several authors have shown that once an HRTF set is collected, individual filters can be represented well by weighted sums of a limited set of spectral basis vectors, and similarly that a single frequency value can be represented anywhere in space with a weighted sum of spatial basis vectors. These results prove that more compact and efficient representations of entire HRTF databases are feasible, but currently these studies provide no advantage over traditional techniques for acquiring the HRTF representation due to the fact that they are derived from HRTF sets acquired using traditional methods. The goal of the current project is to use the knowledge gained concerning low order representations of HRTF sets along with signal processing and machine learning techniques to investigate methods for generating adequate representations of an entire set of HRTFs from a limited set of easily attainable acoustic measurements.

2 Initial Method Evaluations

The HRTFs used in this study were gathered from the publically available CIPIC database [1] which consists of 200 sample Head Related Impulse Responses (HRIRs) captured at a sampling frequency of 44.1 kHz from 45 subjects at 1250 azimuth and elevation locations. A 200 point DTF was taken of each HRIR and the magnitude was kept as the reference HRTF signal. HRTF magnitudes for a small subset of 232 locations were kept from the available 1250 which corresponded to roughly 15 degrees of angular resolution in both azimuth and elevation.

For all of the techniques below a "one-out" training and testing procedure was used and results were averaged over the same five subjects for each condition. Relative performance was based on an **Average Spectral Distortion(SD)** measure presented in [3] as defined below in equation 1.

$$\overline{SD} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(20\log\left(\frac{|\hat{H}_k|}{|H_k|}\right)\right)^2} \qquad (1)$$

Above, H and $\hat{H}$ are the reference and modeled HRTFs at each of L locations respectively, all consisting of N frequency bins. It represents the mean RMS error of the log-magnitude spectra over an entire set of HRTFs.

2.1 Naïve Methods

In this context the term naïve is chosen to represent methods which do not require training data. Their predictions are based entirely on the available HRTF measurements for the test subject and can be thought of as a form of interpolation.

The simplest of these methods is a k-nearest neighbor linear interpolation, where k-nearest neighbors refer to the k measured HRTFs whose locations are closest in absolute angular distance to the location of the HRTF being predicted. The k HRTFs are then scaled and summed to produce the predicted HRTF where the scaling factors are defined as the ratio of the angular distance from that particular measurement location to the location being predicted divided by the sum of all k angular distances as seen in equation 2.

$$\hat{H}_m = \frac{1}{D}\sum_{i=1}^{k} d_i H_i \quad where \quad D = \sum_{i=1}^{k} d_i \qquad (2)$$

The second naïve method which was investigated was spherical harmonic decomposition. While the k-NN interpolation relied exclusively on local information, spherical harmonic decomposition uses information from all known measurement locations to determine the predicted HRTF. Spherical harmonics are continuous basis functions of angular position, and can be used to estimate the underlying function at positions other than those that were sampled. The HRTFs at the *L* measured locations are used to estimate spherical harmonic coefficients, $C_{pm,}$ which represent the underlying continuous HRTF function in spherical harmonics as shown in equation 3[2].

$$C^f_{pm} = \sum_{l=1}^{L} H^f(l)\, Y^{*f}_{pm}(l) \qquad (3)$$

The spherical harmonic coefficients are then used to estimate the continuous HRTF representation which can then be sampled at the missing locations as seen in equation 4.

$$\hat{H}^f(l) = \sum_{p=0}^{P} \sum_{m=-p}^{p} C^f_{pm}\, Y^f_{pm}(l) \qquad (4)$$

In this context *P* is the order of the spherical harmonic expansion and can be thought of as a measure of spatial variation (i.e. a higher order is capable of capturing more rapid spatial changes in the underlying function)[2].

2.2 Missing Feature Methods

In general, the term missing feature applies to any problem where portions of the data are unknown or corrupted and need to be filled with predicted or average values. In this way, an HRTF set lacking measurements at certain locations can be thought of as a missing feature problem.

A basic solution for a missing feature problem is the standard k-Nearest Neighbor (kNN) approach. As in the above naïve approach k of the “closest” known samples are averaged to predict the unknown values, however in this implementation the HRTFs available in the training set are evaluated and the k training HRTFs with the lowest spectral distortion at a certain location are averaged to obtain the prediction. In this way equation 2 holds for this formulation as well with all d_i’s equal to 1/k and the H_i’s are training HRTFs for that location. One possible outcome of this method if k is chosen to be one and the same training HRTF is chosen at every frequency and location corresponds to the scenario where a complete HRTF from a single training subject is used for the missing locations. This condition is reported as the “BestFit” training subject for use as a baseline performance condition since it represents the best non-individualized HRTF set available.

A more recent approach to a missing feature problem was proposed by Smaragdis et al. [4] who showed that a time sample of a spectrogram can be treated as a histogram generated by repeatedly drawing from a mixture multinomial distribution with time dependant mixture weights and source specific frequency multinomial bases. Since under previous assumptions an HRTF is essentially a space indexed spectrogram, the HRTF of a certain subject *s* at a given frequency *f* can be modeled as a histogram of N_d repeated draws from a mixture multinomial distribution with *subject* dependant mixture weights and a set of *spatial* multinomial bases as shown in equations 5 and 6.

$$P^f_s(l) = \sum_{z=1}^{Z} P^f_s(z)\, P^f(l|z) \qquad (5)$$

$$\hat{H}^f_s(l) = N_d P^f_s(l) \qquad (6)$$

The subject independent spatial bases can be trained using the EM algorithm from sample HRTFs. The expectation step at each location consisted of computing the *a posteriori* probability of the bases as in equation 7.

$$P^f_s(z|l) = \frac{P^f_s(z)\, P^f(l|z)}{\sum_{z'=1}^{Z} P^f_s(z')\, P^f(l|z')} \qquad (7)$$

The maximization step then consists of updating the bases and mixture weights as in equations 8 & 9.

$$P_s^f(z) = \frac{\sum_{l=1}^{L} P_s^f(z|l) H_s^f(l)}{\sum_{z'=1}^{Z}\sum_{l=1}^{L} P_s^f(z'|l) H_s^f(l)} \quad (8)$$

$$P^f(l|z) = \frac{\sum_{s=1}^{S} P_s^f(z|l) H_s^f(l)}{\sum_{l'=1}^{L}\sum_{s=1}^{S} P_s^f(z|l') H_s^f(l')} \quad (9)$$

The unknown HRTFs for the test subject can then be estimated via the above EM algorithm where equation 9 is skipped and the bases used are those obtained from the previous training. In this instance the *L* locations being used for the update will only consist of the measured HRTF locations.

2.3 Minimum Mean Square Error Linear Estimation

While both forms of kNN described earlier are technically also linear estimations, a more common way of determining the scaling coefficients is to derive a closed form solution for them which minimizes some objective function. A frequent choice is the minimum mean square error solution which minimizes the average square difference between the actual and predicted values of the function being estimated. For known locations *k*, and missing locations *m*, the linear minimum mean square error solution is given in equation 10.

$$\hat{H}_m = \mu_m + C_{mk} C_{kk}^{-1} (\hat{H}_k - \mu_k) \quad (10)$$

Here C_{mk} is the cross covariance of the training HRTFs at a given frequency at missing and those at known locations; C_{kk} is the auto covariance at the known locations, and $\mu_{(.)}$ represent the corresponding means.

3 Initial Method Comparison

To save space, results for the variations of individual methods such as the various values of k in the k-NN approach are left out of the figures and discussions, and their best representative variation are used in all cross method comparisons. The final values for relevant methods and parameters are summarized in Table I.

TABLE I
FINAL PARAMETER VALUES

Method	Parameter	Value
Naïve k-NN	*k*	4
Sph. Harm.	*P*	7
k-NN	*k*	3
LVM	Z	30

The average spectral distortions over all of the predicted HRTFs for 5 subjects are plotted in Figure 1 versus the number of measure locations used. Spherical harmonic decomposition (SHD) is the only method which failed to reach the baseline best fit training subject condition. This large amount of error is likely due to the fact that SHD has an inherent tradeoff between reconstruction accuracy and necessary spatial sampling rate. In other words higher order modeling (larger P parameter) will yield better overall reconstruction of a function, however, the higher the order of the model the more spatial samples are needed to accurately calculate the spherical harmonic coefficients.

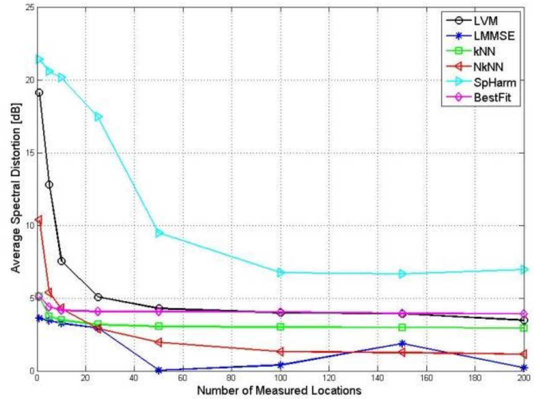

Fig. 1. Comparison of average spectral distortion for tested prediction methods as a function of the number of measurement locations.

K-NN proved to be one of the best performing methods for a low number of measurement locations, but failed to improve for high numbers of measurements. LVM and Naïve kNN also showed minimal improvement as the number of locations was increased, but Naïve kNN provided fairly strong results for when greater than 100 measurement locations were used. Linear minimum mean squared error estimation provided a clear improvement over all other methods and approached zero spectral distortion when around 50 measurements were used.

4 LMMSE Model Tuning

While LMMSE seemed to outperform other methods, it still suffered from discontinuities in resulting spectra for low numbers of sources as can be seen in the top row of Figure 2 and some degree of over fitting when more than 75 locations were used as shown by the small bump in Figure 1.

To help avoid over fitting of the training HRTFs in the LMMSE model, a Forward Model Selection algorithm was implemented based on the Akaike Information Criterion (AIC). AIC is a cost function which seeks to find the minimize model error while simultaneously penalizing models of increasing order (i.e. AIC will be lower for a model with less parameters if the two have similar amounts of error). AIC is calculated as in Equation 11 and indicates a better model with lower AIC scores [5].

$$AIC = 2k + n[\ln(RSS)] \tag{11}$$

Above, *RSS* refers to the sum of squared residuals (errors) , *k* is the number of measured locations, and *n* is the number of training samples.

The Forward Model Selection Algorithm is a technique used to find approximate best models. Finding the best model with N measurement locations becomes intractable by brute force methods after N is greater than 3 or 4, so the Forward Model Selection Algorithm approximates this by finding the best N location model which includes the N-1 locations from the best N-1 location model. Using this strategy a good solution can be found in a far less number of computations.

When the Forward Model Selection Algorithm was implemented using AIC as its objective function for determining which of the k possible known locations should be used for prediction, the large spectral discontinuities for low measurement numbers went away as can be seen in the median plane plots in Figure 2. With this additional step, adequate prediction fell from around 50 measurement locations to below 30.

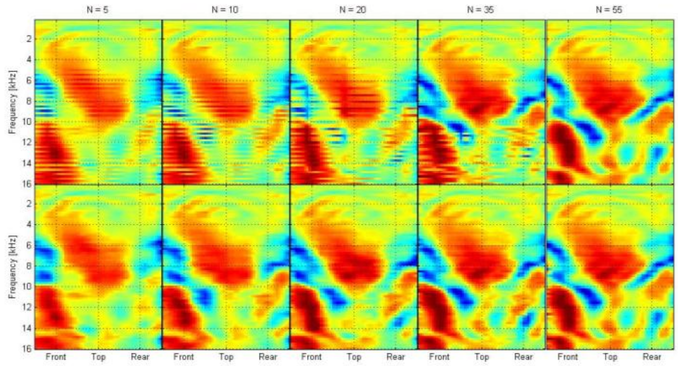

Fig. 2. Predicted HRTFs from the median plane using the LMMSE technique (TOP ROW) and LMMSE technique with forward model selection (BOTTOM ROW).

5 Practical Measurement Locations

While a general decrease in the number of measurement locations needed to accurately predict an individualized HRTF is useful, the most useful reductions in measurement locations would be ones which restricted measurement locations to those on single directional planes, such as the horizontal plane, or hemi-planes, such as the front half of the horizontal and median planes. Results for the LMMSE method for several of these setups are displayed in Figure 3. It can be seen that in general roughly equally distributed measurement locations out perform the same number of locations restricted plane locations, however the two conditions which feature sources on the median plain fair better than their equally distributed counterparts.

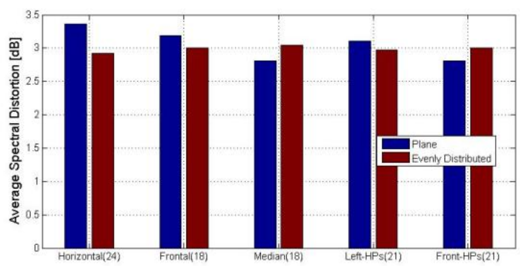

Fig. 3. Average spectral distortion for LMMSE method for practical measurement locations along principle planes and hemi-planes, number of measurement locations indicated in parentheses.

6 Conclusion

Several of the techniques developed show promise for providing personalization of a set of HRTFs. More so then the rest, Linear Minimum Mean Square Error estimation proved to be very reliable for setups including as little as 12% of the original measurement locations. Locations on the median plane also show a higher than average performance versus more distributed locations, which indicates highly practical implementations with low numbers of measurement locations are also feasible.

References

[1] Algazi, V. R., and R. O. Duda. "THE CIPIC HRTF DATABASE." *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2001).

[2] Evans, Micheal J. "Analyzing Head-related Transfer Function Measurements Using Surface Spherical Harmonics." *J. Acoust. Soc. Am.* 104.2400 (1998).

[3] Grindlay, Graham, and M. A.O Vasilescu. "A MULTILINEAR APPROACH TO HRTF PERSONALIZATION." *ICASSP* (2007).

[4] Smaragdis, Paris, Bhiksha Raj, and Madhusudana Shashanka. "Missing Data Imputation for Time-Frequency Representations of Audio Signals." *JOURNAL OF SIGNAL PROCESSING SYSTEMS* (2010).

[5] H. Bozdogan, Akaike's Information Criterion and Recent Developments in Information Complexity, Journal of Mathematical Psychology, 44 (2000).