Audio Engineering Society
Conference Paper
Presented at the International Conference on Audio for Virtual and Augmented Reality, 2020 August 17–19, Online

Anthropometric Features Estimation Using Integrated Sensors on a Headphone for HRTF Personalization

Md Tamzeed Islam*[1] and Ivan J. Tashev[2]

[1]*University of North Carolina at Chapel Hill*
[2]*Microsoft Research*

Correspondence should be addressed to Md Tamzeed Islam (tamzeed@cs.unc.edu)

ABSTRACT

Personalization of HRTF is essential for spatial sound rendering, for which a possible solution is based on one or more anthropological measures of the subject. Measuring these anthropometrics seamlessly, accurately and reliably is still a challenge. In this paper, we propose a system for obtaining anthropometric measurements, suitable for HRTF personalization, directly from a high-end headphone. The proposed system is multi-modal and leverages existing sensors to extract features related to listener's head dimensions. We propose three signal processing methodologies for three modalities of sensors and a fusion algorithm to aggregate these extracted features for a robust anthropometry estimation. To verify the design we use a data set, collected from 35 subjects. The proposed algorithm achieves a low error (RMSE) of $0.58 - 1.21$ cm for human anthropometry estimation.

1 Introduction

With the emergence of new generation devices, such as Augmented Reality (AR) and Virtual Reality (VR) glasses and corresponding software platforms, immersive experience in visual and audible media consumption has become a priority for users. Current generation headphones and computers support *spatial sound rendering* which encodes localization cues in audio to provide a perception in the listener that the sound is coming from a particular location. This has many applications ranging from gaming, live streaming, concerts, and VR interactions [5] where users get realistic experience of audio in terms of location of the sound sources. The spatial audio technologies impact even such everyday tasks as listening to stereo music, which, rendered properly, provides much better experience via externalization of the stereo sound.

The methodology behind spatialized sound rendering is to model the propagation of sound from source to the human ear. The sound wave is scattered around the human head, reflected by the shoulders and the torso, and additionally modified by the pinna on its path to enter the person's ear canal. These human anthropometry-induced changes to the sound signal can be modelled with a filtering function known as *head-related transfer function* (HRTF) [38]. Applying HRTF to monoaural sound and playing it back on headphone results into binaural [38] sound which allows us to synthesize sound from virtual source at any location around the user. Due to the difference in human anthropometric features, the HRTF is unique to each listener.

Common practice [35] in spatial sound rendering is to apply a generic HRTF set to encode the spatial cues in the audio. Because the HRTF is person dependent, applying a generic HRTF set leads to sub-optimal immersive sound experience. While measured HRTF for each person leads to improved perceptual quality of spatial sound, measuring HRTF [7] for every person is not feasible as it requires overly complex and costly equipment. This is why the research [36, 18] in this area is focused on personalization of the HRTF, most frequently using several anthropometric features of the individual. They can be obtained by direct measurements, or indirectly, by using a depth camera, for example.

In this paper, we propose a novel approach for obtaining the listener's anthropometrics such as head width, depth, circumeference and height for the needs of HRTF personalization. Our proposed system uses non-intrusive sensing with sensors available on high end headphones. The proposed approach can estimate the listeners anthropometrics in real time and provides

*Work on this project performed as an intern at Microsoft Research Labs, Redmond, WA.

seamless transition of the HRTF even when the headphone is passed from one listener to another.

The basic principle behind our system is multi-modal sensing with three different types of sensors: microphone, magnetometer and Inertial Measurement Unit (IMU) [22]. We exploit the basic principles of sound propagation around the head, magnetic field intensity attenuation with the distance, and acceleration caused by the human head movement, to extract features that are representative of human head dimensions. Each one of these modalities can be used separately for estimation of certain dimensions of the human head. We also propose a fusion algorithm to aggregate the extracted information from the three modalities and build a more robust and accurate estimator. To evaluate the feasibility and accuracy of these approaches we built an end-to-end system as is shown in Figure 1 and collected data from 35 subjects.

2 Related Work

Usage of parametric models of head and torso for HRTF personalization is explored in [19, 6]. Authors in [21, 8] propose to use high resolution 3D head-scans for HRTF modeling. [7, 18, 36] model HRTF based on anthropometrical features. Recent works [26, 25, 28] use neural networks for HRTF estimation from anthropometric features. [14, 13] uses 3D head-scans for ITD modelling. These works require high resolution complete head-scan as input to estimate ITD of the user. Head shape and dimension estimation has been explored in computer vision research. In [**?**] the head dimensions are extracted from image data to use them for facial recognition. Ear detection and shape estimation from a side view image of the head has been explored in [33]. Recent works [29, 31] explore deep neural networks for facial landmark detection from image. All of these works rely on image for landmark detection and often have poor result for exact dimension estimation. In this paper, we deal with a novel problem as we do not have any visual information as input for anthropometric feature estimation.

3 Overview

In this section, we provide an overview of our proposed system for multi-modal human anthropometry estimation.

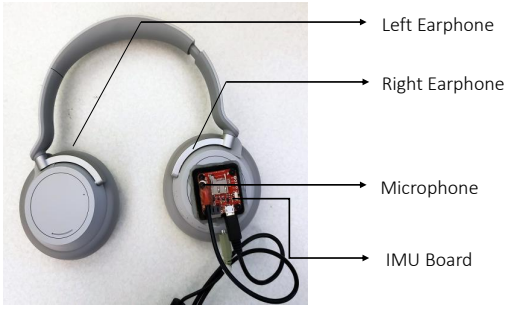


Fig. 1: Hardware Setup

3.1 Hardware Overview

Our experimental hardware is shown in Figure 1. On the right ear-cup of a Microsoft Surface headphone [3] we attached an IMU board [2] with integrated magnetometer, accelerometer and gyroscope. We also added an WM-60 omnidirectional microphone to have full end-to-end control of the entire hardware. The sampling rate of the IMU board is 100 Hz, the microphone signal is sampled at 48 KHz. It was connected to the computer via a USB interface.

3.2 System Overview

In Figure 2, we show the signal processing pipeline of our system. There are three main sensing modules: *a) Acoustic feature extraction, b) Magnetic feature extraction, c) Inertial feature extraction*. We also propose to use sensor fusion and combine them for obtaining more reliable measurements and better accuracy.

4 Acoustic Feature Extraction

The sound propagation time from source to receiver is a basic principle for distance estimation [10]. The main idea is to play a signal from one of the earphones and record it with the microphone on the other earphone. The time-delay between transmitted and received signals depends on the distance traveled. It is proportional to the relatively constant speed of sound and the half-circumference of the user's head.

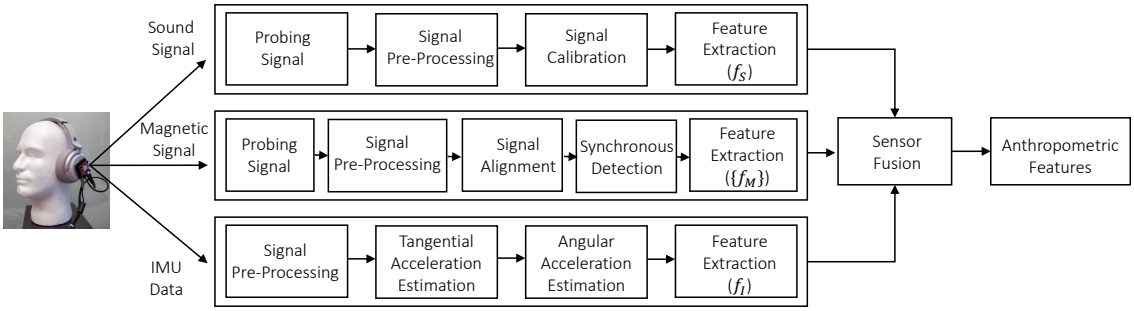


Fig. 2: System Pipeline

4.1 Probing Signal

To make the probing signal inaudible and considering the sampling rate of earphones we choose to use as probing signal a linear chirp of two seconds duration with $20-22$ KHz frequency range. The duration is chosen to be short to make our feature extraction faster. The signal is played first through the right earphone, where the microphone is placed. We refer to this transmitted signal as *calibration signal*, Tx_c. Then, after a short pause, the same chirp is played through the left earphone. We denote this as *response signal*, Tx_r. The two signals Tx_c and Tx_r are recorded with the microphone in a single channel and in one recording session. They are denoted as Rx_c and Rx_r, respectively.

4.2 Signal Pre-processing

Due to prior knowledge about the transmitted signal's frequency band we can apply a band-pass filter and remove the background noise. The frequency band $20-22$ kHz is relatively quiet and doesn't contain interfering sounds. We designed one high-pass and one low-pass finite impulse response (FIR) filters [9] using the Kaiser algorithm [24]. They were applied consecutively to the microphone signal, containing both Rx_c and Rx_r.

4.3 Signal Calibration

Because of delay caused by system calls and hardware pipeline [32] there is an additional *system-induced delay* between the transmitted and the received signal. Furthermore, this system-specific delay varies from recording to recording. To remove this random offset, we use a calibration step. The acoustical delay between the right earphone and the microphone, also placed on the right side, is small and constant, i.e. it doesn't depend on the listeners' anthropometrics. Assuming that the propagation time between Tx_c and Rx_c is close to zero, then the delay between these two signals is just the system induced delay. We estimate the system delay by first computing the cross-correlation function:

$$XCORR_{Tx_c,Rx_c}(K) = \frac{1}{N}\sum_{m=1}^{N-k-1} Tx_c[m] \times Rx_c[m+k] \quad (1)$$

Note that, we use unbiased cross-correlation [17] to reduce the variance of a standard cross-correlation function. Here, N is the signal length and k is the amount of shift or lag. The maximum value of $XCORR_{Tx_c,Rx_c}$ defines the required shift to align Tx_c and Rx_c, i.e. the shift in samples between the transmitted and received signal. This is the system-induced delay we denote as *calibration delay*, or D_c.

4.4 Feature Extraction

Because Rx_c and Rx_r are recorded in one iteration the estimated system delay is the same as between Tx_r and Rx_r. The delay D_{total} between the left earphone and the microphone is estimated using the cross-correlation function as in Equation 1. It is the summation of the system-induced calibration delay (D_c) and signal propagation delay ($D_{propagation}$), measured in samples. Therefore $D_{propagation}$ can be estimated as follows: $D_{propagation} = D_{total} - D_c$.

$D_{propagation}$ is the propagation delay between the left earphone and the microphone on the right earphone and encodes information about user's head circumference. It is expressed in samples and we can calculate the travelled distance as follows: $f_S = v \times$

$D_{propagation} \times T$. Here, f_S is the acoustic feature, i.e. the traveled distance of the sound signal, v is the speed of sound, T is the sampling period. f_S carries information about the head circumference of the listener. We can use non-linear curve-fitting [23] to estimate anthropometric parameter. f_S is also input for the fusion algorithm.

5 Magnetic Feature Extraction

Speakers, placed in the earphones, consist of constant magnet and a coil, connected to a diaphragm [1]. When a sound is played through the speaker, the coil generates variable magnetic field. The magnetic flux density is inversely proportional to the distance [30]. The magnetic sensor is mounted on the right earphone, i.e. at a constant distance to the sensor, while the distance to the left earphone depends on the head width of the listener.

5.1 Probing Signal

Here we also would like to have inaudible probing signal, aiming unobtrusive continuous sampling of the user's anthropometry. The magnetometer is sampled at 100 Hz, which limits us to infrasound below 20 Hz. We use a linear chirp of $8-10$ Hz frequency range as a probing signal. Similar to the acoustic signal, the magnetic probing signal also has a duration of two seconds. First, the signal is played through the right earphone, where the magnetometer is placed. This is denoted as *calibration signal*, Tx_c. Next the same signal is played through the left earphone and is denoted as the *response signal*, Tx_r. The received signals from the three axis magnetometer (Rx, Ry, and Rz) are comprised of the magnetic fields induced from both the calibration and response signals.

5.2 Signal Pre-processing

The magnetic field produced by headphone's speaker usually gets masked by noises and magnetic interference due to the presence of other ferromagnetic objects in the environment. We apply a band-pass filter to remove the interference outside of probing signal's frequency band. This step also takes care of earth's magnetism [15]. We designed one high-pass and one low-pass finite impulse response (FIR) filters using the same approach as with the audio feature and applied them consecutively to the magnetometer signals.

5.3 Signal Alignment

First we down-sample the transmitted stereo signal from 48 kHz sampling rate to 100 Hz sampling rate to match the sampling rate of the magnetometer. Then we sum the left and right channels to get a mono transmitted signal. The alignment of the transmitted and received signals is done by computing the cross-correlation function, finding its maximum, and shifting the received signals (Rx, Ry, and Rz) given number of samples. Note that in this case we compute the cross-correlation function on the entire signal, as the magnetic field propagates with the speed of light and the signals from left and right earphones are practically in phase.

5.4 Synchronous Detection

Even after the band-pass filtering the magnetometer signals are still very noisy for Rx. To suppress the non-correlated signals and increase the signal-to-noise ratio we use synchronous detection [11]. The idea of synchronous detection is to multiply the already prepared mono transmitted signal with the received signal to amplify the modulated segments of the received signal.

5.5 Feature Extraction

We use Root Mean Square (RMS) [12] to estimate the average magnetic flux density. This is done for two time intervals: during the right chirp and during the left chirp. In the process we combine the signals from the three axes (Rx, Ry, and Rz):

$$B(m_1,m_2) = \sqrt{\frac{1}{(m_2-m_1+1)}\sum_{k=m_1}^{m_2}\left(Rx(k)^2+Ry(k)^2+Rz(k)^2\right)}, \tag{2}$$

where m_1 and m_2 are the beginning and ending samples of the corresponding time interval. Lets denote them B_C and B_R for the right and left ear-cup signal. The magnetic sensors provide the magnetic flux density in Tesla ($\mathrm{N.A^{-1}.m^{-1}}$). It is inversely proportional to the distance, i.e., $B_C \propto \frac{1}{d_C}, B_R \propto \frac{1}{d_R}$ and $\frac{B_C}{B_R} = \frac{d_R}{d_C}$. Here d_C is the distance between the right earphone and the magnetic sensor, which is constant, and d_R is the distance between the left earphone and the magnetic sensor, which is proportional to the listener's head width. The ratio between the two densities is proportional to the head width, but for increased precision we

use a non-linear curve-fitting methodology to estimate this anthropometric parameter.

For the fusion, we extract four features ($\{f_M\}$: B_C, B_R, $\frac{B_C}{B_R}$, $B_C + B_R$) . We use the RMS values of both signals, the proportion between them, and the summation. These four features are fed into the fusion algorithm as input.

6 Inertial Feature Extraction

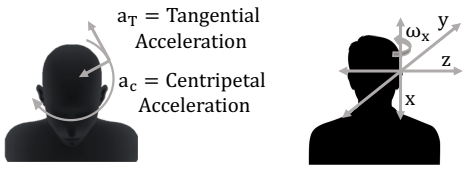
The third modality for feature extraction uses the signals from the accelerometer and gyroscope of the IMU. With these two sensors, we track human head rotation and exploit the relationship between linear acceleration and angular acceleration. Horizontal head movement can be modeled as a circular motion where the center of the circle is the center of the head and the radius is half the head width. For any circular motion, there are two types of accelerations associated with it: a) *Centripetal Acceleration:* The centripetal acceleration (a_c) [4] denotes the change in direction of the tangential velocity. b) *Tangential Acceleration:* Tangential acceleration (a_T) [27] defines the change in magnitude of the tangential velocity of an object. In Figure 3 (a) we show the two accelerations and their associated directions. There is another kind of acceleration related with circular movement. This is called angular acceleration (α) [27] that defines the change of angular velocity of the moving object, $\alpha = \Delta\omega/\Delta t$, where ω is the angular velocity. Tangential acceleration is related to the angular acceleration [27] as follows:

$$a_T = \alpha \times r. \tag{3}$$

Here r is the radius of the circular movement. From the accelerometer and the gyroscope we can calculate the tangential a_T and angular acceleration α, and to estimate half of user's head width r.

6.1 Signal Pre-processing

To denoise the signals from the sensors, we use a moving average window, normalized with the mean of the entire signal. In general such procedure is equivalent of a FIR filter, but is simpler to design and compute.



(a) (b)

Fig. 3: (a) Circular head movement (b) The reference axis for the head movement.

6.2 Tangential Acceleration Estimation

From the accelerometer we receive accelerations along three axis: a_x, a_y, a_z. Due to the orientation of the IMU board in our prototype, shown in Figure 3 (b), the linear acceleration along y axis, a_y, contains the tangential acceleration. However, due to the tilt of human head, the linear acceleration also gets affected by the acceleration due to gravity. We use a dead reckoning module [34] to extract the tilt angles of the head from the gyroscope. Then we remove the effect of gravity by subtracting from the accelerometer readings: $a_y = a_y - gsin(\theta)$. Here θ is the tilt of head with respect to the reference axis.

6.3 Angular Acceleration Estimation

The gyroscope gives us the angular rotation, or velocity of the head movement, along three axis: w_x, w_y, w_z. The first order derivative yields the angular acceleration:

$$\alpha_{xt} = \frac{\omega_{xt} - \omega_{xt-1}}{\Delta t}. \tag{4}$$

Here, α_{xt} is the angular acceleration along x axis, ω_{xt} is the angular velocity we get from the gyroscope at timestamp t. Due to the orientation of the sensor, the horizontal movement is captured by the angular movement along x axis and we calculate α_x to get the angular acceleration of the head movement.

6.4 Feature Extraction

Once we calculated the tangential acceleration (a_T) and the angular acceleration (α_x), we use Equation 3 to calculate the radius of the circular movement. For additional stabilization of the estimation we apply another moving average window to estimate the feature (f_I), and use non-linear curve fitting on that to estimate the anthropometry – head width.

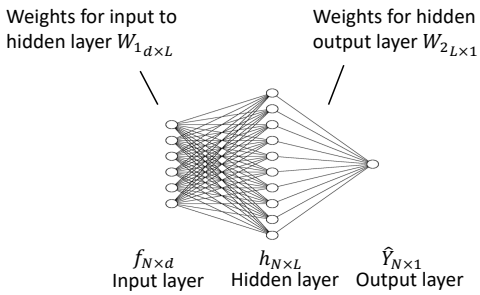


Fig. 4: Sensor Fusion Network

7 Sensor Fusion

Our proposed system aggregates extracted features from the three modalities $f = \{f_S, f_{M1}, f_{M2}, f_{M3}, f_{M4}, f_I\}$ and creates a late-fusion neural network for anthropometric feature estimation. The input feature vector contains information of all the modalities which make the representation robust as it does not rely on a single sensor. We use an Extreme Learning Machine (ELM) [20] in regression mode to estimate the desired anthropometry. The ELM network consists of one broad hidden layer of neurons. ELM has better generalization and performance with limited training data [20].

In Figure 4 is shown the architecture of the proposed sensor fusion network. Lets denote the input feature vector as f, the ground-truth value as Y, and the output of ELM as $\hat{Y}$. The input f has a dimension of $N \times d$ where N is the number of training samples, d is the dimension of feature vector, six in our case. In Figure 4 is also shown the single hidden layer ($h_{N\times L}$), where L is the number of neurons in the hidden layer. If we denote W_1 and W_2 to be the weights from input to hidden layer and from hidden layer to output layer, the output of the network is: $\hat{Y} = W_2 \times \sigma(W_1 \times f)$. Here, σ is the non-linearity function, typically sigmoid [16]. For ELM W_1 is randomly initialized. Then the one step learning estimates W_2 by least-squares fit [37] to match the output of the network $\hat{Y}$ to the ground truth value Y. In our system we train an individual ELM for each of the desired anthropometric features. For a dataset with N subjects, we train the network with $N-1$ subject's samples and test it with the subject's data that is not included in the training set. This *leave one out* methodology for evaluation ensures our network is not over-fitted to the training data and is generalizable to samples from completely unseen data during inference stage.

8 Data Collection

Using our prototype headphone we collected data from 35 subjects, 27 males and 8 females, age ranging from 20 to 59 years old. The data collection was done in a standard office room with other people present.

Table 1: Anthropometric Feature Statistics

Feature	Mean (cm)	Maximum (cm)	Minimum (cm)
Head width	16.1	18	15
Head height	22.1	27.5	19.2
Head depth	18.9	21.4	17
Head circumference	57.6	63.2	50

For ground truth we have measured the subject's anthropometric features (head width, head height, head depth, head circumference) using a measuring tape and a caliper. Comparing to previous works [36], we find that our dataset has adequate diversity in terms of distribution. Statistics about all the anthropometric features are presented in Table 1. For each of acoustic and magnetic features we took 10 measurements, totaling 350 for each modality. For the inertial feature the subjects were instructed to move their head left and right with commands through the headphones for 5 seconds.

9 Evaluation

In this section we present each individual modality and the fusion algorithm's performance. For evaluation metric we use Root mean square error (RMSE) and Pearson Correlation Coefficient. Here, we report the performance of individual sensing modalities and fusion algorithm's performance in anthropometry estimation.

9.1 Root Mean Square Error (RMSE)

In Figure 5a are shown the RMSE for the individual features and the fusion. For head width estimation our proposed fusion-based model achieves the lowest RMSE of 0.58 cm. From the individual features *Inertial feature* achieves the lowest RMSE of 0.61 cm. This feature implicitly estimates the head width and is less prone to environmental noise. The next best performance is from the *Magnetic feature*, which measures exactly this parameter.

For head height estimation our proposed fusion algorithm has a RMSE of 1.21 cm. On the other hand, acoustic feature has a RMSE of 1.54 cm which is the lowest among the individual features. The magnetic

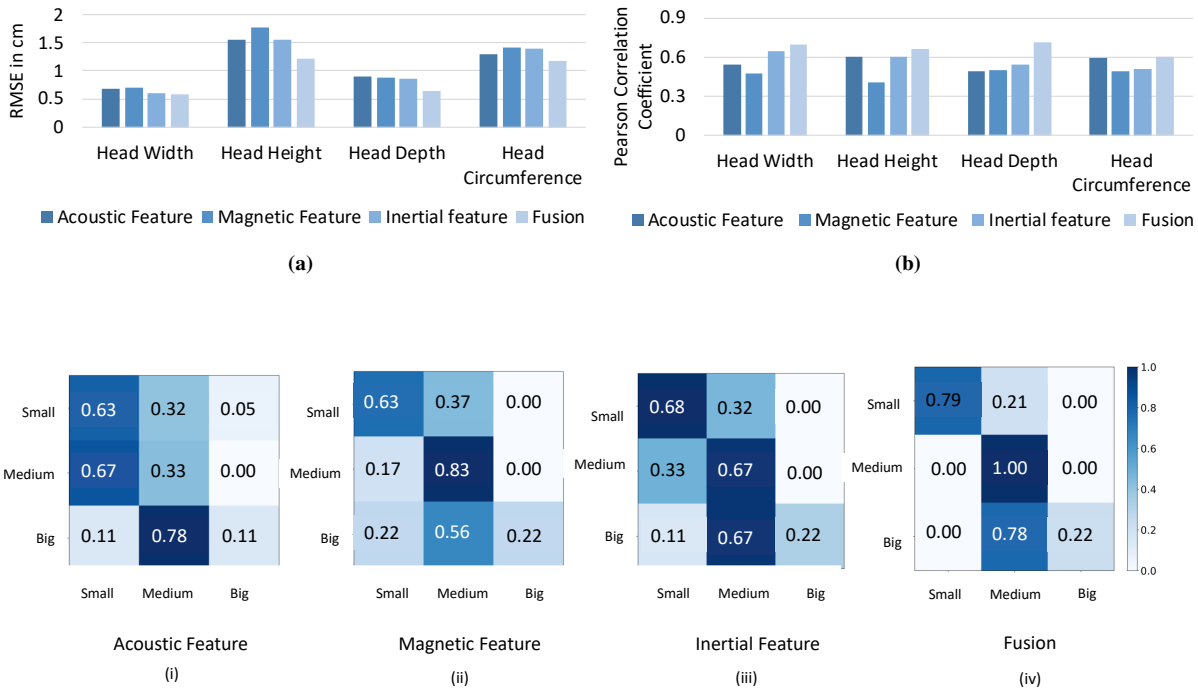


Fig. 5: (a) RMSE and (b) Pearson Correlation Coefficient of different methods for anthropometric feature estimation. (c) Confusion matrices for individual sensing modalities and fusion based cluster analysis.

and inertial features have 1.78 and 1.56 cm RMSE, respectively. There is no modality in our system that actively senses head height. All the modalities rely on the correlation between head height and width to estimate the value of head height. Therefore the error is larger in this case for both individual feature and fusion.

For head depth estimation, fusion has a RMSE of 0.64 cm which is lower than the individual feature based estimations. The performance of the individual features is quite similar in this case. Among the individual features, inertial feature has the minimum RMSE. The acoustic, magnetic and inertial features have 0.89 cm, 0.88 cm, 0.85 cm RMSE, respectively.

In case of head circumference estimation, the fusion based approach is again superior to the individual feature based estimations. Fusion has an RMSE of 1.18 cm for head circumference estimation. On the other hand, acoustic feature's RMSE is 1.29 cm which is lower than other modalities, as acoustic feature is extracted by sound signal's propagation around the head. The maximum value of head circumference is 63.2 cm which is more than 3 times than the maximum value of head width. Therefore, the relative error is actually lowest in this case. In all cases the fusion has lower RMSE than single modality based anthropometry estimation. This proves the advantage of fusion approach over single modality sensing.

9.2 Pearson Correlation Coefficient

In Figure 5b we find that for head width estimation, inertial feature has the maximum coefficient of 0.65 among the individual feature based algorithms. This is consistent with the analysis based on RMSE as inertial feature has the lowest RMSE for head width estimation. Our proposed fusion algorithm has a coefficient of 0.7, which is the highest among all the modalities. This shows that fusion algorithm's estimation has similar trend with the ground truth value. For head height estimation, fusion has the highest Pearson correlation coefficient of 0.66, whereas the acoustic, magnetic and inertial features have 0.61, 0.4, 0.6 coefficient, respectively. Therefore, acoustic feature has the maximum Pearson correlation coefficient from individual features. This is also reflective of the analysis

with RMSE value, as acoustic feature has the lowest RMSE among the individual features. Consistent with the RMSE analysis, the fusion based approach has the best performance for head depth estimation as well with Pearson correlation coefficient of 0.72. The inertial feature has the maximum Pearson correlation coefficient among the individual modalities. The acoustic, magnetic and inertial features have 0.49, 0.5, 0.55 coefficient, respectively. For head circumference, the fusion and acoustic feature have 0.61 and 0.59 Pearson correlation coefficient, respectively. On the other hand, magnetic and inertial features suffer in this case and have coefficient of only 0.49 and 0.51. These two modalities sense head width only and it is evident in their performance drop for head circumference estimation.

9.3 Clustering Based Analysis

In this experiment we evaluate the algorithms' performance in terms of clustering heads with respect to anthropometry. We divide listeners heads into three groups based on the ground-truth head width: a) Small (15–16 cm), b) Medium (16–17 cm), c) Big (>17 cm).

In Figure 5c we show the confusion matrices for individual features and for the fusion algorithm. From Figure 5c (i) we see for acoustic feature based clustering, the major misclassifications occur between *medium* and *big* classes. The overall weighted accuracy for acoustic feature based clustering is 35%. In Figure 5c (ii) we see an improvement in accuracy using magnetic feature for clustering. The accuracy in this case is 53%. For inertial feature based clustering (Figure 5c (iii)) the accuracy goes up to 59%. Finally our fusion approach has a weighted accuracy of 67% in terms of correct cluster prediction (Figure 5c (iv)). Note that the primary goal of this analysis is to evaluate the confusion matrix rather than the accuracy. From confusion matrices we see that all the approaches struggle with big heads. The acoustic feature and fusion algorithms are better than the rest as it classifies 78% of the big heads as medium, i.e. the nearest cluster. For medium heads fusion outperforms the rest by predicting every one of them correctly. For small heads the individual feature-based clustering have similar results, but our fusion based approach has the maximum accuracy of 79%. Also, fusion is the only algorithm with all the misclassifications in the nearest cluster.

10 Conclusion

In this paper, we propose a system for human anthropometry estimation using existing sensors on current high-end headphones for the need of HRTF personalization. Our system allows seamless, real-time HRTF personalization, which is scalable and feasible for practical usage. We propose three algorithms for three different sensing modalities to extract features representative of listener's head shape. We also propose a fusion algorithm to create a multi-modal anthropometry estimator. We evaluate our system with collected data using our prototype hardware. We find that our algorithm has low error (RMSE) of $0.58 - 1.21$ cm for estimation of the parameters needed for HRTF personalization.

References

[1] How do speakers work? https://bit.ly/32ZJKVH.

[2] SparkFun 9DoF Razor IMU M0. https://www.sparkfun.com/products/14001.

[3] Surface Headphones. https://bit.ly/2TNGHfl.

[4] Estimating the angular velocity of a rigid body moving in the plane from tangential and centripetal acceleration measurements. *Multibody System Dynamics*, 2008.

[5] V. R. Algazi and R. O. Duda. Immersive spatial sound for mobile multimedia. In *Seventh IEEE International Symposium on Multimedia (ISM'05)*, pages 8–pp. IEEE, 2005.

[6] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112, 2002.

[7] P. Bilinski, J. Ahrens, M. R. Thomas, I. J. Tashev, and J. C. Platt. HRTF magnitude synthesis via sparse representation of anthropometric features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4468–4472. IEEE, 2014.

[8] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl. A

cross-evaluated database of measured and simulated hrtfs including 3d head meshes, anthropometric features, and headphone impulse responses. *Journal of the Audio Engineering Society*, 67(9):705–718, 2019.

[9] A. E. Cetin, O. N. Gerek, and Y. Yardimci. Equiripple FIR filter design by the FFT algorithm. *IEEE Signal Processing Magazine*, 1997.

[10] D. de Godoy, B. Islam, S. Xia, M. T. Islam, R. Chandrasekaran, Y.-C. Chen, S. Nirjon, P. R. Kinget, and X. Jiang. Paws: A wearable acoustic system for pedestrian safety. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 237–248. IEEE, 2018.

[11] C. B. Fisher and S. T. Fisher. Synchronous detection with sampling, Feb. 24 1981. US Patent 4,253,066.

[12] T. Y. Fukuda, J. O. Echeimberg, J. E. Pompeu, P. R. G. Lucareli, S. Garbelotti, R. O. Gimenes, and A. Apolinário. Root mean square value of the electromyographic signal in the isometric torque of the quadriceps, hamstrings and brachial biceps muscles in female subjects. *J Appl Res*, 10(1):32–39, 2010.

[13] H. Gamper, M. R. Thomas, and I. J. Tashev. Anthropometric parameterisation of a spherical scatterer ITD model with arbitrary ear angles. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2015.

[14] H. Gamper, M. R. Thomas, and I. J. Tashev. Estimation of multipath propagation delays and interaural time differences from 3-D head scans. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 499–503. IEEE, 2015.

[15] G. A. Glatzmaier and P. H. Roberts. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal.

[16] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[17] R. M. Gray and L. D. Davisson. *An introduction to statistical signal processing*. Cambridge University Press, 2004.

[18] F. Grijalva, L. Martini, S. Goldenstein, and D. Florencio. Anthropometric-based customization of head-related transfer functions using isomap in the horizontal plane. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[19] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki. Common-acoustical-pole and zero modeling of head-related transfer functions. *IEEE Transactions on speech and audio processing*, 7(2):188–196, 1999.

[20] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):513–529, 2011.

[21] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. Van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe. Creating the Sydney York morphological and acoustic recordings of ears database. *IEEE Transactions on Multimedia*, 16(1), 2013.

[22] J. E. Kain and C. Yates. Airborne imaging system using global positioning system (GPS) and inertial measurement unit (IMU) data, Apr. 13 1999. US Patent 5,894,323.

[23] O. J. Karst. Linear curve fitting using least deviations. *Journal of the American Statistical Association*, 53(281):118–132, 1958.

[24] J. Knowles and E. Olcayto. Coefficient accuracy and digital filter response. *IEEE Transactions on Circuit Theory*, 15(1):31–41, 1968.

[25] G. Lee and H. Kim. Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear. *Applied Sciences*, 8(11):2180, 2018.

[26] G. W. Lee, J. H. Lee, S. J. Kim, and H. K. Kim. Directional audio rendering using a neural network based personalized HRTF. *Proc. Interspeech 2019*, pages 2364–2365, 2019.

[27] H. S. Leff. Acceleration for circular motion. *American Journal of Physics*, 2002.

[28] L. Li and Q. Huang. HRTF personalization modeling based on RBF neural network. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3707–3710. IEEE, 2013.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[30] K. J. Lohmann and C. M. Lohmann. Detection of magnetic field intensity by sea turtles. *Nature*, 380(6569):59, 1996.

[31] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3317–3326, 2017.

[32] W. Mao, J. He, and L. Qiu. Cat: high-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM.

[33] S. Prakash and P. Gupta. An efficient ear localization technique. *Image and Vision Computing*, 30(1):38–50, 2012.

[34] N. Roy, H. Wang, and R. Roy Choudhury. I am a smartphone and I can tell my user's walking direction. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 329–342. ACM, 2014.

[35] C. Schissler, A. Nicholls, and R. Mehra. Efficient hrtf-based spatial audio for area and volumetric sources. *IEEE transactions on visualization and computer graphics*, 22(4):1356–1366, 2016.

[36] I. Tashev. HRTF phase synthesis via sparse representation of anthropometric features. In *2014 Information Theory and Applications Workshop (ITA)*, pages 1–5. IEEE, 2014.

[37] E. W. Weisstein. Least squares fitting. 2002.

[38] X. Zhong and B. Xie. Head-related transfer functions and virtual auditory display. *Soundscape Semiotics-Localization and Categorization*, 1, 2014.