Towards mass-customizing up/down sound cues for listeners: Issues concerning inter-subject variability

by

Au Tsun Lam John

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in Industrial Engineering and Logistics Management

October 2007, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Au Tsun Lam John

Towards mass-customizing up/down sound cues for listeners: Issues concerning inter-subject variability

by

Au Tsun Lam John

This is to certify that I have examined the above MPhil thesis and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the thesis examination committee have been made.

DR. Richard Hau Yue So Supervisor

PROF. Chung Yee LEE Head

DEPARTMENT OF INDUSTRIAL ENGINEERING AND LOGISTICS MANAGEMENT

14 January 2008

Acknowledgement

I would like to express my sincere gratitude to my thesis supervisor Dr. Richard H. Y. So. He provided guidance, advice and encouragement with much patience throughout the years of my postgraduate study. He enabled me to conduct the research continuously while I faced critical problems. My vision has been considerably broadened through in depth discussion with him. I am indebted to him for not only his continual support but for this entire experience.

Special thinks must also be given to Dr. Ravindra S. Gonnetilleke and Dr. Fugee Tsung as they served as my thesis committee members and I very much appreciated their advice for my thesis and defense examination.

I would like to take this opportunity to thank Dr. Richard H. Y. So, Dr. Ravindra S. Gonnetilleke and all of the members in the advanced seminar group who gave valuable advice during our discussions in any form.

I would like to express my appreciation to K. K. Yung, Denil Chan, Stanley Sze and K. C. Tin for their technical support in the experimental setup and for assistance with using many different types of equipment in the laboratories.

Moreover, I would like to thank Jennifer Ji, Brian Ngan, Jacky Chow, Leo Lau, Fan Sai Kit, Eric Chow, K. M. Cheung, Billy Chan for their encouragement and support during the discussion of different topics.

Last but not least my thanks go to Ki Ki for her love and support to overcome many difficult times.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Equations

Towards mass-customizing up/down sound cues for listeners:
Issues concerning inter-subject variability

by

Au Tsun Lam John

Department of Industrial Engineering and Logistics Management

The Hong Kong University of Science and Technology

Abstract

Sound cues filtered from individualized head-related transfer functions (HRTFs) can provide accurate up / down directional cues to that particular listener. When the same cues are presented to other listeners, errors occur in the perceived up / down directions and these errors vary greatly among listeners. This thesis presents a study to relate individual's localization errors of non-individualized HRTF-filtered up / down sound cues with a proposed index (referred to as the 'matching score'). These scores are calculated from individual's ear dimensions as well as the spectra of the HRTFs used in filtering the sound cues. It is hypothesized that for a particular listener and a particular up / down HRTF-filtered sound cue, the higher the matching score, the lower the localization errors (H1). The matching score, based upon the 'delay-and-add' theory proposed by Hebrank and Wright (1974), is new and original and forms part of the academic contribution of the thesis. If H1 is

proven, this thesis will be the first study to provide empirical evidence to support the proposed ‘delay-and-add’ theory. Three dimensional moulds of outer ears from thirty-three participants have been collected. Using the ear dimensions, matching scores have been calculated between each of the 33 participants and 192 open-copyrighted non-individualized HRTFs (from the LISTEN database: IRCAM and AKG Acoustics, 2004 and CIPIC database: Algazi *et al.,* 2001). These calculations have been repeated for each of the four selected elevation angles (30 and 15 degrees below ear level and 30 and 60 degrees above ear level) to give 25344 matching scores. Using these matching scores, five non-individualized HRTFs having the 0th, 25th, 50th, 75th, and 100th percentile average matching scores have been selected for each of the four elevation angles. These twenty HRTFs are then used to produce a total of twenty sound cues (4 angles x 5 HRTFs). Seventeen participants (randomly selected from the 33 participants of the survey) were invited back to take part in a within-subject design experiment in which each listener needed to localize 120 sound cues presented in random order. These 120 sound cues represent the combination of 20 sound cues and 6 repetitions. The objective of the experiment is to evaluate the relationship between the matching score and the localization errors. In this experiment, listeners obtained an average localization error of about 36 degrees to an elevation cue, which is comparable to the results from other studies (Zotkin et al.,2003, Seeber and Fastl, 2003). Our proposed matching scores have been found to significantly correlate with the localization errors and regression analyses confirm that as the scores increase, the errors reduce significantly. In other words, H1 has been supported. Further analyses using the ‘school-effect’ model indicates that the matching score alone can explain 27% of between listener variations in the localization errors. Discussions on inter- and intra-listener variability in perceiving up/down direction of a sound cue are included. The study is the first to provide empirical support to the 'delay-and-add' theory to explain human perception of up / down direction of a sound cue. Future work to refine the 'matching score' calculation is desirable.

1. Introduction

1.1 Up/Down sound localization by human listeners

1.1.1 Coordination system to describe sound directions

Human beings use two ears to localize the direction of a sound cue. To describe the sound source's direction, two angles are used: 'azimuth angle' and 'elevation angle' (see Figure 1). The azimuth angle describes the amount of angular deviation from the direct front direction (defined as zero degree) of a listener along the horizontal plane at ear level. The elevation angle describes the amount of angular deviation from the direct front direction of a listener along the median vertical plane (or 'median saggittal plane' as used in psychoacoustic researches). An illustration of the coordination system is shown in Figure 1.

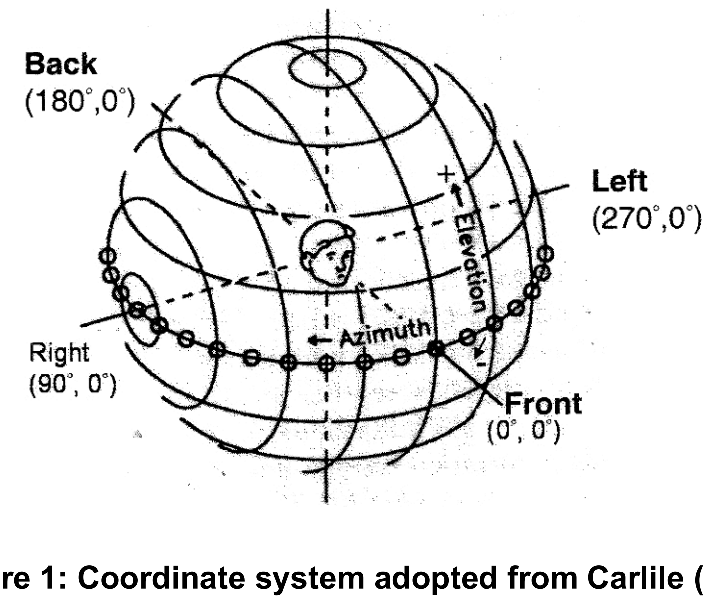


Figure 1: Coordinate system adopted from Carlile (1996)

1.1.2 Directional information (cues) used in sound localization

In this study, directional information embedded within sound stimulus is referred to as 'cue'. Cue can be classified into two categories in psychoacoustic studies: i) binaural cues and ii) monaural cues. Binaural cues are directional cues received at both ears. The differences of sound perceived at the two ears provide information on the leftward and rightward deviation of the incoming direction of the sound from the centre-front direction of the listener. In other words, these binaural cues help us to distinguish the azimuth angle of a sound source. The responsible sound cues for the identification of the azimuth angle of sound cues are inter-aural time differences (ITDs) and the inter-aural level differences (ILDs). ITDs originate from the arrival time difference at the two ears from the same sound source. This time delay generates a corresponding deviation in the perceived azimuth angle. The mechanism had been well explained in the literature (Jeffres, 1948). The other binaural cue, ILD, occurs as the amount of acoustic attenuation in the air differs by the different traveling path towards the left and right ears. Thus the perceived loudness at the two ears differs and thus the inter-aural level difference. Both ITDs and ILDs directional cues are easy to estimate from the geometric configuration between the sound and the listener (assuming a spherical head with known diameter). Therefore, the regeneration of one’s perception on sound cue’s azimuth angle (i.e. left and right) is relatively easy. (Blauert, 1997)

Localizing a sound cue’s elevation (i.e. up and down) requires additional directional cues. Mills (1972) described a condition, known as the ‘cone of confusion', that 2 distinct sounds coming from the same azimuth direction but different elevation directions to a listener. The binaural directional cues, ITDs and ILDs, of the 2 sound cues are identical (as they share the same azimuth source direction). However, listeners can still distinguish the different elevations of these 2 sounds. Therefore, additional directional cue is responsible for the listener's localization on a sound's elevation direction. It is later known as monaural cues.

This monaural cue, which can be generated by a single ear only, is actually the spectral variation embedded in a sound's spectrum (Blauert, 1997). Therefore it is also called the spectral cues or the spectral features in sound spectrum. These spectral features are originated from the acoustic interaction of sound waves and the ear concha, head and torso. These acoustic interactions include wave reflections, diffractions and resonances (Blauert, 1997) occur within the mentioned human organelles.

A number of peaks, troughs and the variations in a banded sound's spectrum had been identified as important features for listeners' localization for a sound cue's frontal and backward perception (Asano et al., 1990; Blauert. 1969/70). These monaural cues come from the acoustic interactions between physical ears and the sound sources and appear in the perceived sound spectrum, in the form of spectral peaks (local maximum in sound amplitude), troughs (local minimum in sound amplitude) and spectral variations.

To summarize, the perception of a sound's azimuth direction relies on the perceived binaural cues and the perceived elevation direction requires the information embedded in monaural cues.

1.1.3 Human ability in localization natural sounds

Combining the directional information from binaural cues and monaural cues a human listener can achieve localization accuracy with an error of 1 degree only for azimuth angle localization and 4 degrees for elevation angle localization. (Blauert, 1997, p.44)

1.2 Reproduction of directional cues with HRTFs (or HRIRs)

Since directional cues are caused by the acoustic interaction of a sound and the listener. For the sake of regenerating the same directional perception of a sound, measurements of these acoustic interactions have been developed. The interaction is recorded as a filter called Head-Related Impulse Response (HRIRs), which is in the time domain. They are then used for the regeneration of the same directional perception described in the recorded directional cues. HRIRs can be transformed into Head-Related Transfer Functions (HRTFs), which is in the frequency domain, by Fourier transformation for easier interpretation since monaural cues had been identified in different frequency bands in different psychoacoustic researches.

1.2.1 HRTFs – a record of all acoustic interactions

HRTFs are identified as the complete record of all acoustic interactions between the listener's body and a sound source placed at known relative position. Many studies (Bluaer, 1969/70; Hebrank and Wright, [1]; Asano et al. 1990; Langendijk and Bronkhorst 2002) had been conducted to understand the 'active ingredients' within the HRTF spectrum for sound localization. They found there are specific features (e.g. amplitude, spectral variations in HRTF spectrum) existed within certain frequency range of HRTF spectrum, which are readily interpreted by human for certain directional perception (e.g. frontal or backward; upward or downward perception of a incoming sound). However, the influence on localization error from the changes of these spectral features in a sound cue cannot be determined from their findings.

1.2.2 HRIRs: Measurements details and implementation on enhancing directional perception on mono sounds

Since HRIRs are records of the acoustic interactions between a sound source and a listener, the interactions are different when (i) for the same person listening to sounds at different position, or (ii) for different person listening to same sound source. In either case, multiple sets of HRIRs are required to be recorded for regenerating accurate directional perceptions for multiple directions of sound sources (as in (i)), or for multiple listeners on the same sound source (as in (ii)). In either case, background noise must be taken into account as well.

Therefore, listeners are required to seat in an acoustic-controlled room (e.g. anechoic chamber). A series of speakers are placed around the seated listener (often in the form of sphere, Figure 2). A microphone will placed at each of the listener’s ear opening to record the incoming sound energy from each of the speakers to each of the 2 ears (see Figure 3). Equalization of the speakers and microphones are done before the HRTF measurement.

Figure 2: Localization measurement: Speaker array

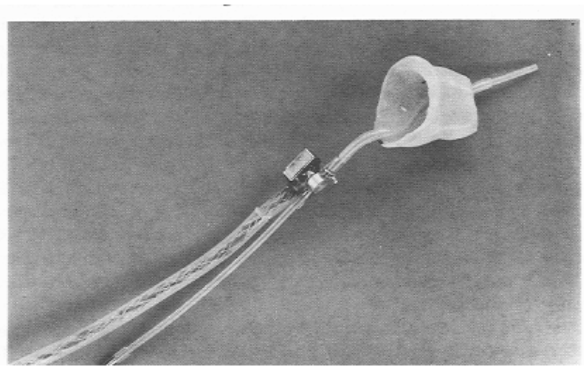


Figure 3: Localization measurement: in-ear microphone

The listener is required to sit still during the HRIR measurement. Therefore, with controlled audio signal from speaker and known geometric orientations between listener's ears and each of the speakers, the head-related Impulse Response (HRIR) for each of the channel, i.e. the right and left ear of the listener, can be calculated from the incoming sound impulses received at the 2 microphones in listener's ears. A Fourier transform is needed to transform the time-domain HRIRs into frequency-domain HRTFs for further studies. In later chapters, our main focus will be HRTFs. An example of calculated HRTF spectrum is illustrated in Figure 4. The same transformation can be captured for different positioned speakers. As a result, 2 arrays of transfer functions, or HRTFs, are captured for both ears. Each of the arrays consists of the HRTFs describing directional information for all speakers.

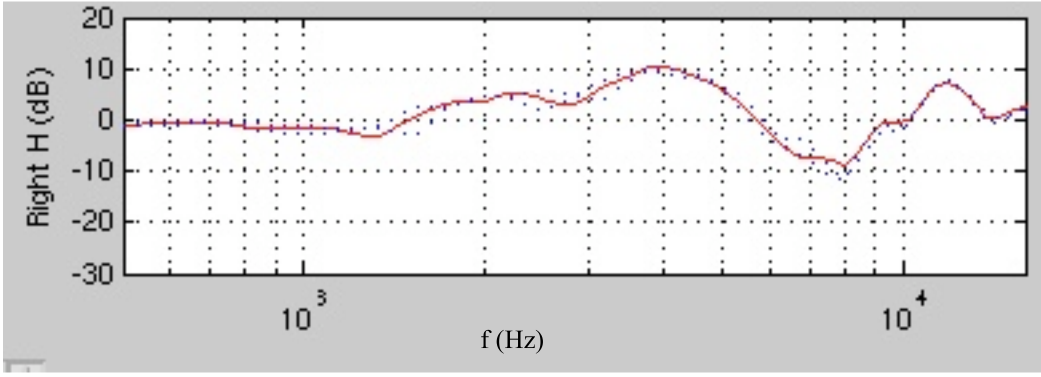


Figure 4: Frequency response of a single typical HRTF

Assuming there are N speakers used for the HRTFs measurement and the length of each HRTFs is L, the dimension for the 2 arrays of a particular person's HRTF is N by L. Its graphical representation is called a HRTF spectrogram and illustrated in Figure 5.

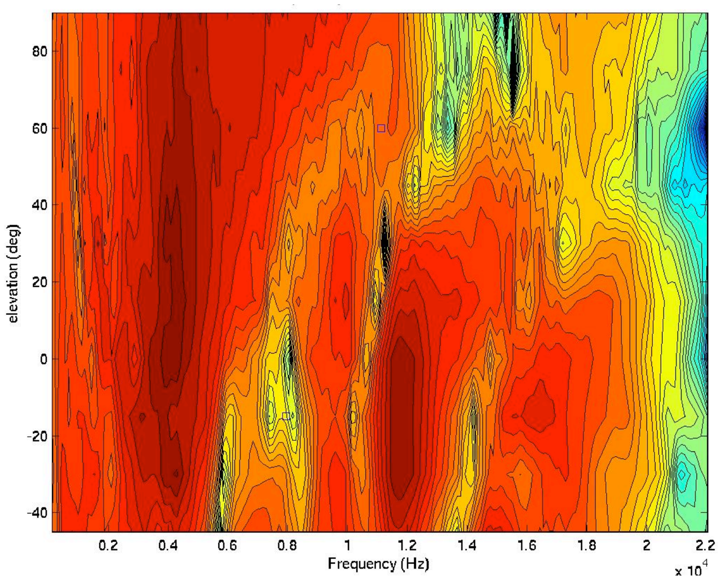


Figure 5: Multiple HRTF spectrums displayed with changing cue elevations (-40^o to +90^o, at an interval of 10^o)

The recorded HRIR sets can then be used for the regeneration of directional perception to listener. By definition, a mono sound lacks directional cues for listener's directional perception. By convolving the mono sound (X_A in Figure 6) with a pair of recorded HRIRs, one for right channel ($h_R(t)$ in Figure 6)and the other for the left channel ($h_L(t)$ in Figure 6), all the recorded directional cues can be inserted into the mono sound. A schematic diagram is shown below:

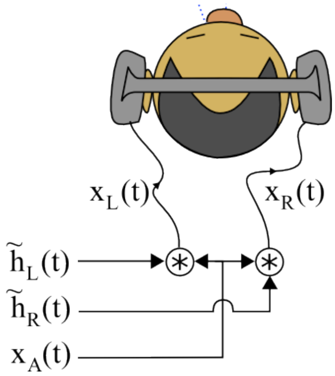


Figure 6: Schematic diagram of binaural sound regeneration with HRTFs (http://commons.wikimedia.org/wiki/Image:Hrir_binaural_synthesis.png)

1.2.3 Individualized HRTFs versus Non-individualized HRTFs: An issue of cost and performance

Theoretically a personally measured HRTF set (individualized HRTF set) can provide all the essential directional cues to the individual listener, provided that the audio delivery system (e.g. headphone) is ideal for the sound reproduction. Therefore individual listener can achieve a localization error in localization a simulated sound using personalized HRTF comparable to exposing to a natural sound. Specifically, this theoretical localization error for the determination of a sound cue's elevation is about 4 degrees (refer to section 1.1.3). The use of non-individualized HRTF sets, on the other hand, introduces an average localization error of about 15 to 20 degrees and with large variation (Seeber and Fastl, 2004). This variation is more severe when localization a up / down sound cue. (Zotkin et al., 2003).

Although the individualized HRTF set can provide higher localization accuracy for listeners (Blauert, 1997), this promising high performance also requires the listeners spending long HRTF measurement time in advance in a carefully acoustic-controlled room. Therefore using individualized HRTF sets for generating highly accurate binaural sound cues is not feasible for a large group of listeners.

Because we cannot reach the high localization accuracy with the use of non-individualized HRTF sets, and also the HRTF measurement on individual listener is very resource demanding (in terms of time and money). Therefore it is reasonable to explore different selection strategies on selecting non-individualized HRTFs and their associated localization accuracy.

1.3 Inter-subject variations on sound localization

1.3.1 Range of error in up/down localization with different HRTFs

Seeber and Fastl (2004) had reported that different listeners have different localization errors towards the same sound cue. The elevation localization error ranged from 10 degrees up to 50 degrees. However, there is still no method to predict this error for individual listeners. Therefore it is beneficial to investigate the possibility to predict the localization errors experienced by different listeners on sound cues generated from different non-individualized HRTFs. By doing so, we can see if there exist some non-individualized HRTF sets always produce low localization errors to different listeners, or the average localization performance achievable by a large population of listeners can be reduced by providing a few choices of sound cues for them to choose. In other words, we are customizing the usage of non-individualized HRTF sets to generate sound cues of similar localization performance as individualized HRTFs.

1.3.2 Need for customization on the use of generic HRTFs in localization applications

The concerns on implementation cost and localization performances of HRTFs-filtered sound cues are preventing the use of HRTF technology in current Hi-Fi audio systems. To reduce the development cost and increasing localization performance at the same time, different approaches are being developed to reduce the implementation cost. Using HRTF sets measured from mannequin is a way of reducing implementation cost. The KEMAR HRTF set (Gardner and Martin, 1995) from MIT has been a well-known HRTF set freely available on the Internet. This type of non-individualized HRTFs had been used in computer sound card (e.g. CMI-8738 from C-Media).

Another approach is selecting non-individualized HRTFs from public HRTF library (e.g. CIPIC HRTF Database, Algazi et al., [1]; LISTEN HRTF Database, IRCAM, 2004) by different algorithm: matching body dimensions (Zotkin et al., 2003), subjective selection by listeners (Seeber and Fastl, 2004) and data analysis on directional cues (Ngan et al., 2005).

1.4 Ear anthropometry - Origin of inter-subject variations in sound localization performances of HRTF-filtered sound cues

Because all spectral cues are originated from the acoustic interaction between listener's ear concha and the incoming sound cue, it is not strange to understand their resulting spectral cues generated by different listener's concha are very different. This is because every ear concha is unique in shape. Therefore each spectral cues generated in different listener's ear concha is unique as well.

However, the ear concha is generally of similar sizes with similar components (see Figure 1 for different part on a ear concha). The differences among different ear concha are the orientation of different components on the concha. In particular, the reflection surface on concha responsible for reflecting incoming sound waves within the ear cavity. Therefore there exists some kind of similarity of the 'pattern' of concha reflecting surface between different listeners ear. Zotkin et al. (2003) had found that by maximizing this 'pattern' similarity could reduce the listener's elevation localization error, comparing to sound localization without this similarity customization. However, the extent of improvement cannot be addressed for individual listener.

1.5 Benefits of understanding the cause of inter-subject variations in sound localization performance

The barrier in adopting HRTF technology in current audio systems is the lack of guaranteed fidelity in sound localization accuracy by non-individualized

HRTFs. In order to optimize the localization performance of listeners from the use of non-individual HRTFs, predicting listeners' perceived localization error on sound cues in advance could provide a good starting point in improving system fidelity in a cost effective way.

1.6 Purpose of this research

This study aims to investigate the relationship between individual's localization errors of HRTF-filtered up / down sound cues and his or her ear anthropometry. A proposed index (referred to as the 'matching score') calculated from individual's ear dimensions as well as the spectra of the HRTFs used in filtering the sound cues have beem developed to fulfill this aim. This study includes:

(i) Collection of spectral cues from individual's ear in a ear survey;

(ii) Collection of spectral cues from different non-individualized HRTFs;

(iii) Definition and calculation of matching scores between 33 individuals' ears and 192 sets of non-individualized HRTFs for foir pre-determined incident angles;

(iv) Study the predictive power of our proposed matching score on listeners' sound localization performances in a sound localization task through design and conducting an empirical experiment;

(v) Analyze the inter-subject and intra-subject variability on listeners' localization performances.

1.7 Organization of this thesis

The first chapter introduces the background and objectives of this research. It covers the relationships between HRTFs and how human localize a sound cue. Ear dimensions as the origins of inter-listener variability in localizing HRTF-filtered sound cues as well as impacts on localization accuracies when using personalized and non-individualized HRTFs to generate sound cues are reported. Chapter two contains a literature review on how listeners localize up / down directions of incoming sound cues. In particular, specific peaks and notches in HRTF spectrum (referred to as 'spectral cues') that have been identified to provide up / down directional information are summarized. The relationships among ear dimensions, spectral cues, and perceived sound localization error are also discussed. In Chapter three, our approach in building a proposed index (matching score) is explained in detail. Chapter four reports a survey of ear dimensions and the calculation of matching score between 33 listeners and 192 non-individual HRTFs.

Chapter five describes the design of an empirical sound localization experiment to evaluate the relationship between the matching scores and the localization errors. Results (e.g. regression analysis on matching score) are also included in the final section of this chapter. Chapter six presents an in-depth analysis on intre- and intra-listener variability on localization errors. A short review of the "school-effect" model is included and this method is used in this chapter to analyze the inter- and intra-listener variability.

Finally, the conclusions and recommendations are stated in Chapter seven.

2. Literature Review

2.1 Binaural cues and Monaural cues for sound localization

Directional cues are directional information embedded in sound cues that help listeners to determine the source direction of this sound cue. They are commonly classified into two categories: binaural cues and monaural cues.

Binaural cues require directional cues collected from both ears. It can be explained with one of the earliest theory of human sound localization formulated by Lord Raleigh (1907), the duplex theory.

The duplex theory assumed listener's head is spherical without ear pinnae. Since the two ears are separated on the left and right side of the head, unless the sound source is on the median plane, the distances traveled from the sound source tot he left and right ear are different. This phenomenon causes the sound waves to reach the ipsilateral ear before reaching the contralateral ear. This leads to the phase difference at both ears which is known as interaural time difference (ITD) and differences in sound level, which is known as the interaural level difference (ILD).

However, the incompleteness of the duplex theory has been illustrated with the cone of confusion by Carlile (1996). The interaural time and level differences are ambiguous by choosing any two points on the cone of confusion. However, front-back and up-down direction cannot be distinguished with only the information of ITD and ILD. Other cues in addition to ITD and ILD are needed to ITD and ILD are needed to explain the sound localization abilities of human beings.

The limitation of duplex theory leads to the study of spectral response of pinnae. The acoustical response of pinnae is the supplement to duplex theory. The folds of the pinnae, including the antihelical fold and antitragus and the hollow area in concha (see Figure 7 below), alter the spectral contents of incoming sound cues. These spectral differences are dependent on the sound direction. This leads to a series of studies on the role of pinnae and spectral cues in sound localization, i.e. Blauert (1969/70); Gardner and Gardner (1973); Hebrank and Wright [1]; Langendijk and Bronkhorst (2002).

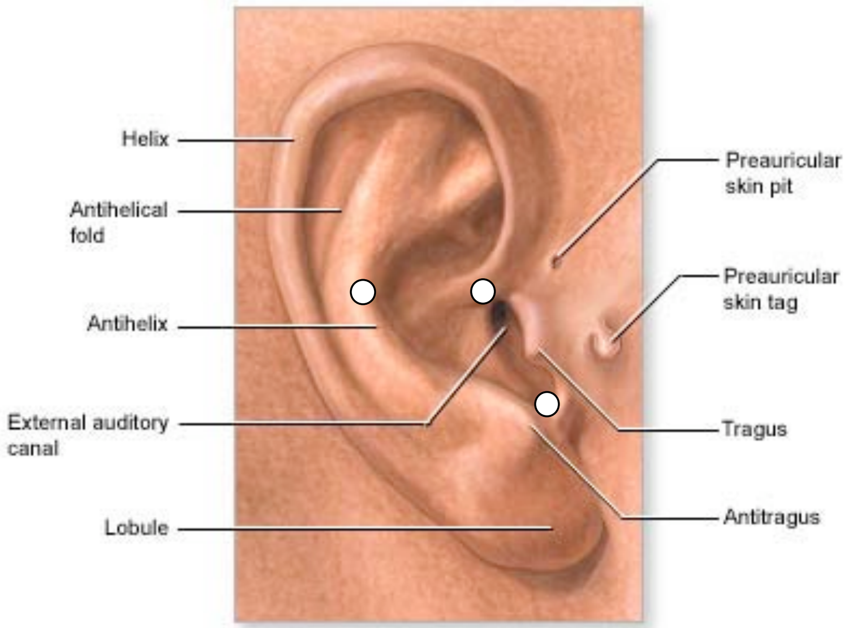


Figure 7: Ear Anatomy (with 3 landmarks)
(http://www.nlm.nih.gov/medlineplus/ency/imagepages/1126.htm)

This thesis focused on studying of the sound localization on the median plane, where the effects of interaural level difference (ILD) and interaural time difference (ITD) are minimized. ILD and ITD can be modeled using a physical model (Kuhn, 1987) that can be easily tuned for each individual. On the other hand, the inter-subject variability of the monaural spectral cues has not been explained, in terms of localization errors and ear anthropometry.

2.2 Spectral characteristics of monaural cues for localizing up/down direction of a sound source

The centre frequency of a spectral notch, or local minimum of amplitude, in sound frequency spectrum had been identified as a significant directional cue for localizing up / down direction of a sound source. (Hebrank and Wright, 1974; Langendijk and Bronkhors, 2002; Lopez-Poveda and Meddis 1996). Hebrank had found that sound cues with spectral notch moving from 5.7 kHz to 11.3 kHz can generate a upward movement perception on the sound cue's elevation direction. Langendijk and Bronkhorst found that spectral cues located between 6 kHz and 12 kHz are directional cues for up / down perception. Hebrank had hypothesized the generation of a spectral notch in the perceived sound cue is actually caused by a destructive interference occurs at the ear concha: the directly incoming sound wave superpose with the sound wave reflected on ear concha. Details are explained in section 2.4.

2.3 Detection of spectral characteristics of an up/down sound cue: a coding and decoding scenario

2.3.1 Interpretation of spectral cues into up / down directions

In order to allow listeners to determine the elevation direction of an incoming sound, our auditory system is build for the detection on mentioned spectral cues. Neuro-physiological studies had been conducted on various mammals to search for the mechanisms in detecting the up/down relevant spectral notches ('signal') within sound stimuli to give us elevation perceptions. Imig et al., (2000) had identified the dorsal cochlear nucleus (DCN) neurons in cats' mid-brain are responsible for the interpretation of sound cue elevation directions from spectral notches identified by the auditory nerves in cochlea. The neurons in DCN had shown an excitation pattern depending on the centre frequencies of these spectral notches. Also Vartanyan and Malinina (2004) had conducted another study on mice behavior in sound localization. They found the movement of centre frequencies of these spectral notches in sound stimuli can change the responses of these neurons, which are then interpreted as a moving sound source on the vertical plane (it coincide with the Hebrank's observation for human listeners in his study in1974). Further evidences had supported the neurons in DCN are detectors for spectral notches and signals are decoded as cues' elevations.

2.3.2 Detection of spectral cues in ears

Before the spectral-notch-stimulated neuronal signals are decoded in the DCN, the spectral notches are transformed into neuronal signals at the auditory nerves (ANs) in the cochlea. Instead of detecting a specific frequency in sound stimuli within a single neuron in AN, each neuron is more like an auditory filter showing excitation within a certain frequency bandwidth with centre at a specific frequency (called Best frequency, BF). This best frequency served as a characteristic for each AN.(see Figure 8 for their relationship)

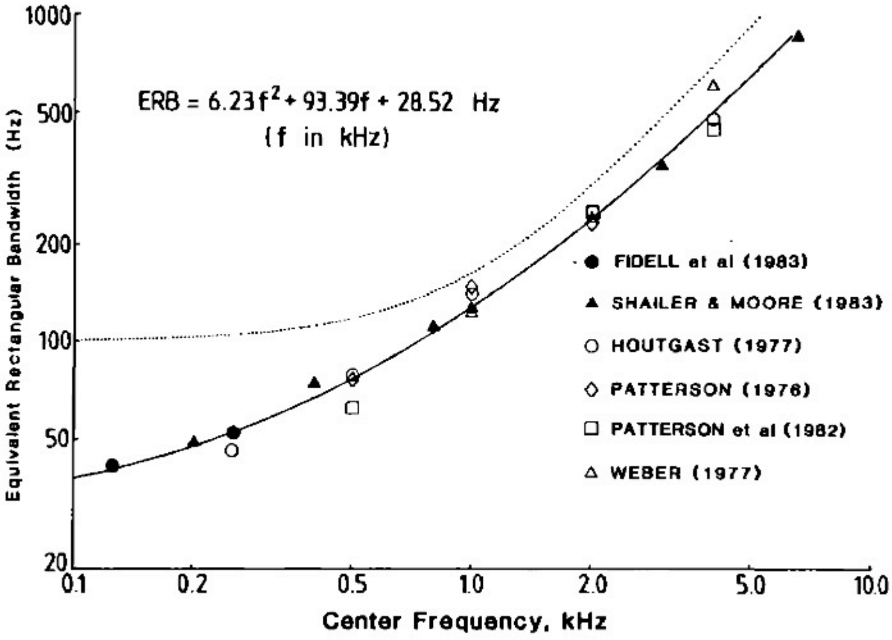


Figure 8: Estimation of ERB for various centre frequencies of spectral notches detected by auditory nerves (quoted: Moore & Glasberg, 1983)

The rectangular bandwidth is called an equivalent rectangular bandwidth (ERB) and its width is dependent on its centre frequency. .Figure 8 shows the relationship between ERB and BF from various studies. Moore summarized their result with a model equation as shown in the figure (Moore and Glasberg, 1983). He had also modeled the excitation pattern of these neurons at different centre frequencies. (See this pattern in Figure 9) The mathematical model of these filters was developed by Patterson et al in 1992, named as gammatone filters:

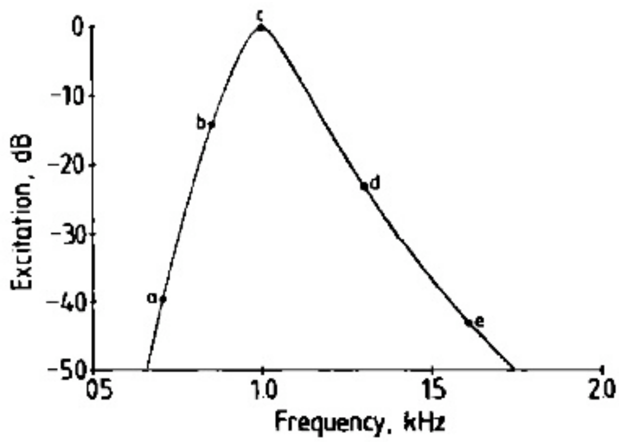


Figure 9: Excitation pattern from a single auditory nerve cell modeled by gammatone filter (quoted: Moore and Glasberg, 1983)

The cochlea contains large amount of neurons with different BFs, therefore, in the signal processing point of view, the cochlear is considered as a bank of gammatone filters, or a filterbank. Implementation of gammatone filters had been carried out on different computer platforms (Slaney, 1993). As a result, we can model the excitation pattern detected by the whole cochlea if the centre frequencies of spectral notches in sound stimuli are known.

2.4 Relationship between one’s ability to detect cue directions and his/her ear anthropometry

In addition to the connection between spectral cues (in particular, spectral notches) and listener's perception on sound source direction, spectral notches has also been hypothesized by Hebrank and Wright (1974) for the creation of a spectral notch is originated from the superposition of a direct incident sound wave and a reflected sound wave from ear reflective surface (e.g. antihelical fold on concha, see Figure 7) at the ear canal opening. His hypothesis is known as the 'single delay-and-add theory'. It stated the centre frequency of a spectral notch could be transformed from the relative distance between the reflection surface on concha the ear canal opening. The transformation equation is shown below:

$$Freq_{notch}(Hz) = \frac{34300}{4 \times RelativeDistance(cm)}$$

Equation 1: Single delay-and-add theory:
Relationship of spectral notch and physical dimension on concha

Raykar et al. (2005) had built an algorithm for extracting spectral notches from individualized HRTFs and then reverse-transformed the extracted spectral notches into physical dimensions. A mapping of these physical dimensions on the listeners' photo had shown a large resemblance to the actual ear shape of these listeners. (Figure 10)

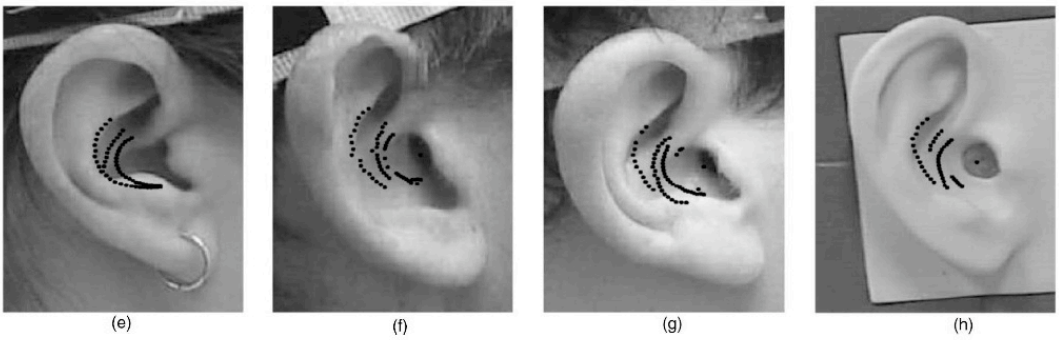


Figure 10: A picture of the physical dimensions identified with Raykar's spectral notch extraction algorithm, overlay on the listener's ears (quoted: Raykar et al., 2005).

2.5 Past studies on reducing sound localization error using ear shapes

Sound cues filtered with non-individualized HRTFs give listeners unpredictable localization accuracies (Ranged from 10 degrees up to 50 degrees or errors; Seeber and Fastl, 2004). Therefore, lots of researches had been conducted on seeking solution to reduce localization errors while using non-individualized HRTFs. Zotkin et al. (2003) had studied the reduction of localization errors when non-individualized HRTFs are selected by minimizing disparity in ear anthropometry between listener and the source of HRTFs. He found that this minimization of anthropometric disparity could reduce the listener's localization error for about 15 degrees.

3. A proposed Matching Score method: Definition and Rationale

3.1 The proposed Matching Score (MS): definition

The matching score is intended for predicting the perceived localization error of a particular listener when determining the up / down incident angle of a HRTF-filtered sound cue. This score matches the ear dimensions of a listener with the corresponding spectral features of a HRTF. The score ranges continuously from 0 to 1, with "0" is defined as 'unmatched' and meaning the perceived localization error will be very large (equal or larger than 50 degrees) and "1" is defined as 'perfectly matched' meaning the error is smaller or close to 10 degrees. These values are based upon the reference performance, reported in Seeber and Fastl (2003). .

3.2 Rationale of the proposed Matching Score

The proposed Matching Score (MS) is used to differentiate non-individualized HRTFs for their compatibility of spectral cues in sound cues generated, which would be reflected as different localization errors perceived by listeners. In order words, MS can be used to rank different non-individualized HRTFs. A perfectly matched HRTF to a listener means the centre frequency of spectral notches are same as the listener's estimated spectral cue from his/her ear shape. Therefore an HRTF-listener pair with MS equals to "1" means the spectral notches' centre frequencies are matched to the listener's and the listener's localization error will be close to 4 degree (comparable to human ability in localization).

3.3 Steps to define the Matching Score

Since MS is used to describe the compatibility of spectral cues to be provided to listeners, this compatibility calculation requires the spectral cues (i.e. spectral notches) from the sound cues ("signal", which originate from the non-individualized HRTF being used) and the spectral cues ('decoder') estimated from listener's ear anthropometry (transforming ear shape into corresponding spectral notches). An ear survey can be performed to collect spectral cues carried by our listeners. These spectral cues can be transformed from image capture of the listeners' ear shapes. This transformation process is based on the equation hypothesized by Hebrank and Wright (1974) (see equation 1, delay-and-added theory). For M participants (2*M ears in total) in our ear survey, there will be 2M set of spectral notches' centre frequencies for each definite cues' elevation.

Another set of spectral cues ("signal") from the sound stimuli (or HRTFs) will be extracted with Raykar's spectral notch extraction algorithm (Raykar et al., 2005). In our HRTF database within this study, there are 90 HRTF sets employed from the CIPIC HRTF database (Algazi et al., 2001) and 102 HRTF sets from LISTEN HRTF database (IRCAM, 2004), therefore there will be 192 set of spectral notches' centre frequencies for each definite cues' elevation.

As a result, we will calculate all the possible pair of spectral compatibility from both set of spectral cues, or the calculation will result 2M * 192 matching scores for 1 selected cue elevation. If we pick 5 choices of elevation angle for our experimental setup, we need to calculate 2*M*192*5, or 1920*M matching scores.

Malcolm Slaney (1993) had developed a toolbox to create the gammatone filters. Since the gammatone filters, which models the responses of type III Auditory Nerve cells, are used for modeling the frequency response of electrical stimuli in our auditory neurons, we can generated a filterbank of gammatone filters specific for our ear survey participants with their spectral features collected in a ear survey. Malcolm's toolbox create gammatone filters in two steps with 2 functions: 'MakeERBFilters' and 'ERBFilterBank ' to create a single gammatone filter with input centre frequency. The gammatone filters corresponding to the same listener at the same elevation are overlapped in the frequency domain and normalized to represent the final gammetone filter for that individual listener's AN excitation pattern for that particular elevation angle. (See Figure 9)

3.4 Using excitation level in auditory nerve cells to estimate compatibility of spectral cues from stimuli on listeners

To quantify the amount of excitation in the nerve cells (using spectral cues transformed from ear photos and then modeled by the gammatone filterbanks), each of the gammatone filter in the filterbank can be considered as a single target for the stimuli to 'hit' or 'match'.

For a filterbank having 4 gammatone filters ('decoders'), it is characterized by 4 different center frequencies. Its frequency response can be calculated by Malcolm's toolbox (see section 3.3) and will be normalized to be between 0 and 1. Assuming there are three spectral notches identified in the sound stimuli with 3 different centre frequencies, therefore 3 different values of excitation level in the filterbank's frequency response is resulted.

The matching score between this sound cue and the listener will be the average value of the above 3 of excitation level. If there are N spectral notches identified in a sound stimulus, the matching score will be the average of N values of excitation levels in the filterbank's response.

3.5 A final note on the Matching score calculation

In some cases, due to the limitation of our image-processing algorithm on identifying ear concha curves, no spectral notches can be identified at some cue elevations for some ear photos. Therefore the matching score at this elevation would be denoted as "0" although the true score maybe larger than zero. More work is needed to improve the image-processing algorithm.

3.6 Evaluation of matching score

After all the matching scores are calculated, the HRTF sets representing the 0th percentile, 25th percentile, 50th percentile, 75th percentile and 100th percentile matching scores for each selected incident angles of sound cues are selected for further experimental testing. In this study, 4 cue elevations have been selected (see Chapter 5) and there are 20 (5 x 4) testing sound cues for each participant in the experiment on evaluating the performance of our MS. Details will be discussed in Chapter 5.

4. Ear Survey to determine a set of Matching Score (MS) among 33 listeners and 192 non-individualized HRTFs

4.1 Survey of outer ear dimensions and theoretical estimation of the corresponding frequency notches to be imposed on incident evaluated sound cues due to echoes in the outer-ears

4.1.1 Introduction

First, a survey of outer-ear dimensions are conducted for 33 listeners. Secondly, Hebrank and Wright (1974)'s single delay-and-add theory (see Equation 1) is used to estimate the frequency notches that are corresponding to the acoustic interactions between an incident up / down sound cue and an outer-ear with its particular dimensions. This estimation procedure is repeated for 4 incident up / down angles of sound cues and outer-ears of the 33 listeners. Hebrank and Wright (1974)'s delay-and-add theory explains that when an elevated incident sound is reflected from the surfaces of the concha area inside one's outer ear, the superposition of the reflected wave and the original incident wave will generate a perceivable frequency notch in the resultant sound cues. The estimation procedure for the frequency notches has been demonstrated in Raykar and Duraiswami (2005), In this study, the procedure reported in Raykar and Duraiswami (2005) is adopted. From another perspective, our estimation procedure is determining the frequency notches of one's HRTF without actually measuring one's HRTF.

The spectral features to be collected are spectral notches within the HRTF spectrum between 5.7kHz and 11.3kHz as suggested by Hebrank and Wright (1974) and Langendijk and Bronkhorst (2002). By converting into physical dimensions by Equation 1, we will search for sound reflection paths ranged between 0.759c, and 1.504 cm.

The physical dimensions of these 33 listeners' ears will be collected from digital images of their 66 ears and the mold casting on the same 66 ears. Different procedures will be taken to collect physical dimensions from these 2 sources.

4.1.2 Procedure (hardware and software)

There are 33 participants participated in our ear survey, 23 chinese males and 10 chinese females recruited within university campus. The ear survey is split into 3 parts:

i) Anthropometric measurements;

ii) Outer ear shapes captured by digital camera; (a sample photo is shown in Figure 11)

iii) Casting of negative molds for the construction of a plaster mold on each ears.

The ear survey lasts for about 30 minutes and each participant is given HK$30 for remuneration after the survey.

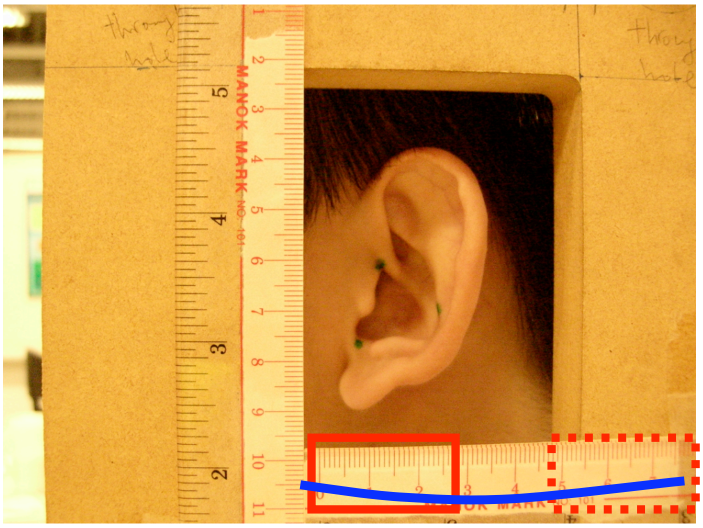


Figure 11: Sample of ear photo captured in ear survey

After anthropometric measurements on head and ears is completed, the participants would be invited to stand besides a wooden frame with a digital camera (Model: Nikon Coolpix 4800) mounted at about 1.6 meter above the ground. The wooden frame contains a window for the participant to place their ear within it, allowing the digital camera to capture its shape at side view.

The participant is asked to align their head at upright without turning with respect to the photo plane. In case of a distortion in the digital image captured from a head-turned participant, an error sensitivity analysis was performed.

There are a few possible sources of errors: (i) Mis-alignment of head orientation introduced by subject; (iI) Radial distortion in digital image introduce by camera lens

Since the ear shape capture equipments were unfamiliar to participants, they may pose with a nonzero pitch angle head orientation. This would lead to a shift of incoming sound elevation angle predicted from digital. To prevent it from happening, we asked the participant to aim with their eyes at a far target 8 meters away. This will lead to the formation of a natural head orientation (in pitch direction) as in natural standing posture. Therefore the incoming elevation angle predicted from digital images would coincide with listeners' daily experience.

To prevent the participants posed with a non-zero yaw and/or non-zero roll angle in head orientation, it is accomplished by a visual inspection by experimenter and instructs the participants to re-orient their heads if a mis-orientation is observed.

Since our final goal is to seek for a quick and simple measurement method (compare with actual HRTF measurement in an acoustic-controlled environment), we have chosen a consumer-grade digital camera for image capture. However, spherical lens are popularly-used in many consumer-grade cameras. Images will be radial-distorted with the use of spherical lens, the distortion is more severe when capturing in macro mode (as in our application). Therefore we will evaluate the distortion introduced in our current instruments and image correction will be made if the severe distortion is found.

As seen in the 2 square brackets in Figure 11, the straight ruler was more distorted in the dotted-line bracket. To be more specific, the photo is distorted into a 'radial-out' curved shape (see the solid curve in Figure 11). Therefore the horizontal interval markers on the ruler are denser near the photo edges than those markers near the centre of photo. The same observation is assumed to be the same as spherical lens are used in the camera. It is observed that the compression of markers on photo is constant. For example, if 50 pixels are shifted for a marker 10mm away from photo centre, the 20 mm far marker would be 100 pixels shifted to the same direction as well.

Therefore we can calculate the actual distance of the ear dimensions based on the distorted distance on the same photo. A detail calculation of the compensation is described in Appendix A.

4.1.3 Transformation of ears' physical dimensions into spectral features

The physical dimensions of ear shapes are extracted and transformed into spectral features in MATLAB environment. It was done in 4 steps:

4.1.4.1 Step 1: Defining ear canal opening and radial distortion ratio on imported ear photo.

All physical dimensions are defined with ear canal opening. Our MATLAB program required a program user to import a ear photo first. Then the user is required to locate the ear opening visually and click on it with a single left button mouse click. This procedure is repeated 3 times and the average value of the coordinates is taken as the final coordinates of the ear opening.

After the ear canal opening is defined, the user will be asked to click on the 0 cm, 1 cm, 2 cm, 3 cm and 4 cm marker on the horizontal ruler in the photo. The average rate of change of the pixel differences for each 1 cm gap is taken as distortion ratio, Kx, in the horizontal direction. The same estimation is also performed in the vertical direction by left mouse-clicking on the 3 cm, 4 cm, 5 cm, 6 cm and 7 cm markers on the vertical ruler and the distortion ratio, Ky, in the vertical direction is also obtained. The values of Kx and Ky are used for compensating the radial distortion during the search for reflection surface in the next step.

4.1.4.2 Step 2: Extracting edges of ear shapes in MATLAB

Once we have completed acquiring the basic parameters as described in 4.1.4.1. The color photo will be transformed into a grayscale photo, followed by the search for the reflection surfaces on ear concha with an image processing function 'edge' available in MATLAB image processing toolbox. The 'edge' function requires two parameters to run: type of edge-searching method and a numerical threshold value for edge tolerance. In our program, the two parameters are 'canny' and '0.034'. The numerical value is chosen by user's visual inspection on the quality of output ear shape in picture. A large tolerance value would give a picture with lots of noise points, while a small tolerance value would lose lots of details on the ear shape.

4.1.4.3 Step 3: Extracting points of reflection surface in MATLAB

After the surface extraction step by MATLAB's 'edge' function, we will search for reflection points along 4 paths, which are describing 4 distinct elevation angles (-30 degree, -15 degree, +30 degree and +60 degree; +: above ear level; -: below ear level) (Figure 12) starting from the ear canal opening.

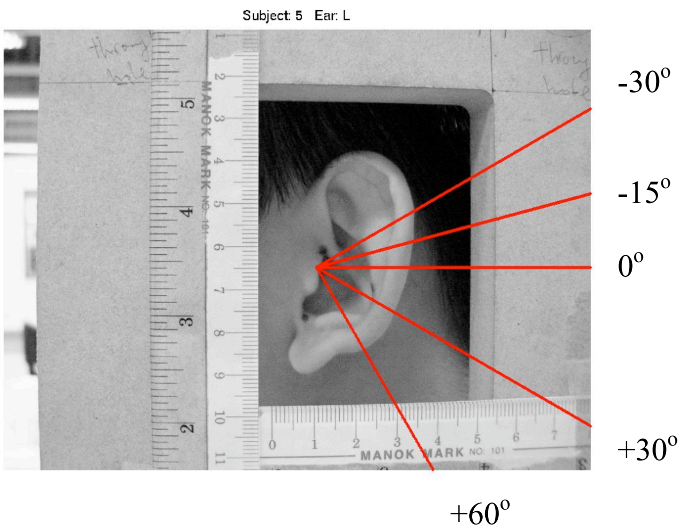


Figure 12: An illustration of the elevations on a ear photo (top to bottom lines:-30, -15, 0, +30, +60 elevations)

Since the reflection points on the photo will be transformed into centre frequencies of spectral notches by Hebrank's single delay-and-add theory with Equation 1 repeated below:

$$Freq_{notch}(Hz) = \frac{34300}{4 \times RelativeDistance(cm)}$$

Based on the relevant spectral notches identified by Hebrank and Wright (1974) and Langendijk and Bronkhorst (2002) is between 5700 Hz and 11300 Hz, therefore we only allow reflection points located within the range between 0.759cm and 1.504cm from the ear canal opening to be recorded and used for calculating matching score in later steps. Because different people have different ear shapes, thus the number of valid reflection points may be different for different listeners. As a result, the total number of reflection points collected for each ear can be different.

4.1.4.4 Step 4: Transformation of physical dimensions into spectral features

After all the relevant reflection points (as observed on ear foldings on concha in photo) are extracted, they are transformed into values of centre frequencies of 'simulated spectral notches' by Equation 1.

In general, the whole spectral feature collection computation procedure on one ear is less than 3 minutes on a Pentium III computer. Defining the ear canal opening and the distortion parameters is most time consuming, as it required human intervention. Improvements can be made for auto-detecting the ear canal opening in ear photo.

4.1.4 Another spectral characteristics collection method: Casting 3D plaster Ear Molds on participants' ear shapes

In addition to collecting spectral features from ear photos (an reverse process by Raykar, 2005), we also made ear moulds on the 66 ears as well. Silicone gels were filled in the concha cavity to make negative molds. They are then put into plaster to make into plaster molds. They are put carefully into plaster to maintain its orientation as if the participant is having their ear openings facing upward. A sample plaster moulds is shown in Figure 13 below:

Figure 13: A sample plaster mold made in ear survey

After the plaster moulds are set, Markers are marked on the plaster moulds to show the reflection points found along the 4 elevations (same as the 4 elevations used in ear photos). The whole plaster mold will be put on computer measuring machine (CMM) to measure the 3D coordinates of these markers, as well as the coordinates of ear canal opening. After the coordinate measurement on CMM is completed, relative distances of these markers to the ear canal opening is calculated from the 3D coordinates. These distances are then transformed into respective simulated spectral notches with Equation 1. Since every marker is at a known elevation angle, therefore each of the resultant spectral notches will be associated with a specific elevation angle. This second set of simulated spectral notches will be collected for calculating the 'ear mould version' of the final matching score, while the spectral notches from ear photos would be used for calculating the 'ear photo version' matching score.

4.2 Collecting spectral characteristics from 192 open-copyrighted non-individualized HRTFs

4.2.1 Introduction

After collecting spectral characteristics (to be specific, spectral notches) from ear shapes of 33 participants, equivalent spectral features are also collected from 192 open-copyrighted non-individualized HRTFs. To fulfill this goal, we adopt an HRTF's spectral notch extraction algorithm developed by Raykar et al. (2005).

4.2.2 Procedure

To maintain compatibility of spectral notches collected from ear survey (photos and moulds), we use the HRTFs describing the same 4 elevation angles on median plane (i.e. -30, -15, +30 and +60 degrees). The 192 non-individualized HRTF sets employed into this extraction procedure are from the CIPIC HRTF database (Algazi et al., 2001) and the LISTEN HRTF database (IRCAM, 2004). The 768 HRTFs (192 x 4) are open-copyrighted, publicly available on the internet and are equalized for their measuring equipments.

Basically, we rerun Raykar's extraction algorithm on our 768 HRTFs. Although we cannot have the same programming code as Raykar et al. used in their study. Our program follow the steps listed in their work and extracted same set of spectral notches for the illustrated HRTF (CIPICE #10, Right ear).

Our MATLAB program can extract spectral notches from a single HRTF within 1 minute on Macintosh platform with PowerPC G4 at 1.33 GHz processing speed. Therefore the whole extraction procedure on the 768 HRTFs takes less than 90 minutes. Our extraction algorithm is documented in Appendix B.

4.2.3 Calculating Matching scores based on the spectral characteristics collected from 33 listeners and 192 non-individualized HRTFs

The calculation of matching score between ear X and HRTF Y comprises of 3 steps:

1. (Refer to section 3.3) we use Malcolm’s toolbox to create gammatone filters on “simulated spectral notches” from ear survey (i.e. spectral notches from 33 listeners. If there are 3 “simulated spectral notches” for ear X along particular elevation angle on ear concha, the summation of the 3 gammatone filters will be the resultant filterbank to represent the spectral characteristics of this ear at this cue elevation.

2. Check the excitation level (i.e. amplitude of gammatone filters normalized between 0 and 1) of the new filterbank at the frequencies of the spectral notches collected from HRTF Y (associated with the same cue elevation as in step 1). If there are 4 spectral notches identified in HRTF Y, there will be 4 excitation level values.
3. The final matching score between ear X and HRTF Y will be the average value of all the excitation levels obtained in step 2.

A detail example is illustrated in Appendix C.

5. Experimental evaluation of the benefits of matching scores

5.1 Objective and hypothesis

A sound localization task is designed to evaluate the predictive power of our matching score in predicting the localization errors perceived by different listeners when listening to sound cues carrying different spectral features.

To fulfill this goal, some non-individualized HRTFs are selected from the 768 (192 HRTF sets x 4 target elevations) HRTFs to generate testing sound stimuli for our participants to localize. In particular, we are interested the extent of deviation of their response elevation deviated from the target elevation described by each of the sound stimuli.

We hypothesized that the perceived localization errors in listening these sound stimuli is significantly correlated to the matching scores which describe the compatibility between listeners and sound stimuli.

5.2 Experimental Designs

5.2.1 Independent variables and dependent variables

Three independent variables were used in this experiment: (i) the 4 directions of cues ((0^{o}, -30^{o}), (0^{o}, -15^{o}), (0^{o}, 30^{o}), (0^{o}, 60^{o})); (ii) 5 HRTF sets for each direction (details of selection described in next section); and (iii) repetitions.

The dependent variable is the perceived absolute localization errors along elevation direction. In this experiment, there are 340 conditions, which are the exhaustive combination of 17 listeners and 20 standard HRTF sets. Each of the 340 conditions was repeated 6 times.

5.2.2 Design of experiment

The experiment is not a simple full factorial design, because each HRTF is unique to the cue elevation described. In other words, the factor “HRTF” is nested with another factor “cue elevation”. Another parameter for regression analysis is the matching score between a specific ear-HRTF pair. This matching score is so unique that its value can only apply to a particular ear paired with a particular HRTF at specific cue elevation. Therefore another method of analyzing the significance, the hierarchical analysis, will be introduced and applied in chapter 6.

5.2.3 Apparatus

A sound localization measurement system was designed to collect listeners' perception of sound cue's elevation angle. A block diagram (see Figure 14) showed the overview of the system.

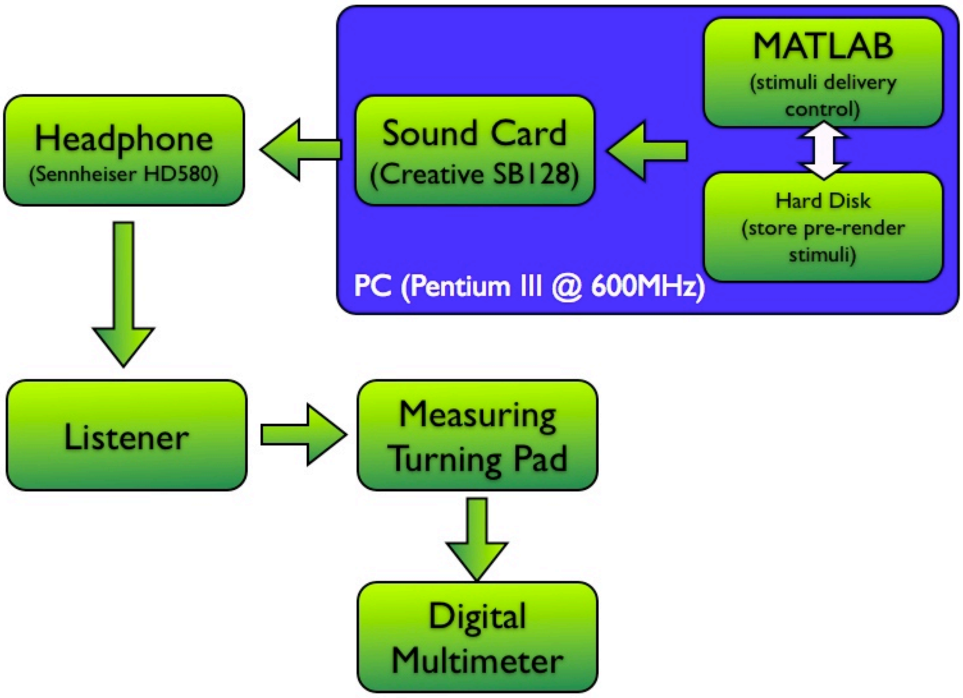


Figure 14: System architecture of sound stimuli delivery and response measurement

5.3.1 Hardware

A Pentium III 600MHz personal computer was used to control the sound delivery to participants in the listening task. The computer was equipped with a Creative Live sound card. A pair of Sennheiser HD 580 headphones (Hannover, Germany) was connected to the speaker output of the sound card, which was used to deliver sound stimuli to listeners. To capture the listener's response on the perceived elevation of the sound, a home-made measuring pad was designed. It is a printed circuit board of about A4 paper size with a metal stick attached to a rotating variable-resistor which is mounted on the PCB board. The working principle of this measuring pad is like a clock. The variable resistor would be at a specific resistance value when the metal stick is turned to rotate to a specific amount, which is done to reflect the perceived elevation of sound stimuli.

In addition to the presentation and measurement hardware, the listeners are required to fixate their head orientation during the experiment. Markers are marked on a wall in front of them for reference.

MATLAB program was developed for the experimenter to control the system. This program was used to present sound clips in random order. The graphical user interface is shown on next page. Also the sound clips are generated in MATLAB environment by convolving a sample mono sound. The MATLAB code for generating testing sound stimuli is included in Appendix D.

5.2.4 Stimuli

Recall that our objective for this experiment is to evaluate the predictive power of our matching score in predicting individual listener's localization errors. Therefore it is preferred to select HRTF sets covering a larger range of matching score over our 17 participants. Therefore, for each elevation angle, we selected the 0th, 25th, 50th, 75th and 100th percentile of the averaged matching score among the 33 participants in our ear survey. A total of 20 HRTFs were selected out of the 768 non-individualized HRTFs. There 20 corresponding HRIRs (the HRTF in time domain) will be used to generate a sound stimuli of 3 white noise burst with 0.3 second of duration separated by 0.3 second of silent gap (Figure 15):

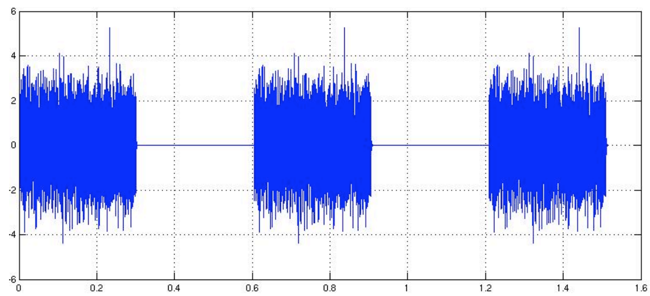


Figure 15: Sample of sound stimulus deliver in a single condition

5.2.5 Experimental Procedure

Participants were asked to sit comfortably and take rest for 5 minutes. A turning pad for listeners to give responses on perceived elevation was given to the participant. Instructions on using the turning pad as responding equipment were given by the experimenter. For the instruction sheets for participants, please refer to Appendix F.

Before the experiment was started, calibration of the turning table was done to acquire the relationship to transform resistance measured from multimeter into perceived elevation.

Stimuli were presented through Sennheiser HD580 headphones. After each stimulus, each participant had to put the turning pad as shown in Figure 16. Then turn the metal stick to represent their perceived elevation angle. Then the multimeter will give a resistance value of the built-in variable resistor. Then the experimenter input the resistance value into computer. This value was transformed into corresponding elevation angle by the relationship derived during the calibration step. 20 Trial measurements were conducted to allow listeners to get familiarized with the experimental procedure. The data acquired in the 20 trials are discarded.

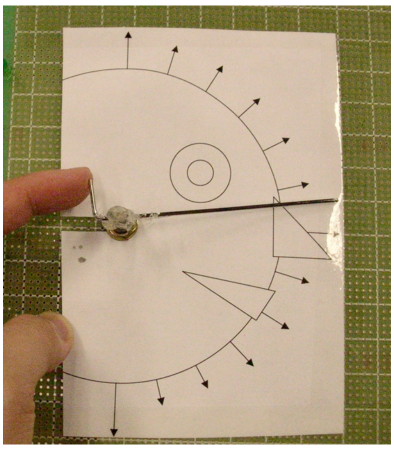


Figure 16: Turning pad for indicating perceived elevation

To avoid participants getting tired, they were given a 5-minutes break for every 20 minutes of experiment time. Each participant completed the whole experiment within 2 hours.

5.3 Result and Discussion

In our experimental design, the dependent variable is the localization error experienced by listeners in each condition. A normality test (Minitab® 14, MiniTab Inc.) reveals a Box-Cox transformation with an alpha value of 0.5 is appropriate:

$$transformed\_error = raw\_error^{0.5}$$

Equation 2: Box-Cox Transformation of raw data for normality

For analysis in later sections, we denote the transformed localization error as "transformed error" and it will be our dependent variable under investigation.

5.3.1 Analysis on Variance

In our study, the independent variables are: i) cue elevation ("tar_ele"); ii) choice of non-individualized HRTF for the addition of directional cues into sound stimuli ("HRTF") and repetition ("repeat"). There are 4 levels for the factor "tar_ele", namely they are set at -30o, -15o, +30o and +60o. One of the intrinsic properties of HRTFs is each of them contains directional cues describing a single direction only. Therefore, the factor "HRTF" is nested with the factor "tar_ele". And we have selected 5 choices of non-individualized HRTFs for each cue directions (Please find the selection criteria of the HRTFs in section 3.6). Therefore each participant will be given 20 testing conditions and these conditions will be repeated in 6 repetitions (i.e. "repeat").

Since the factor "HRTF" is nested with "tar_ele", we then perform ANOVA test (SAS® 8, SASon data separated by different cue direction, which will give us 4 ANOVA tables:

Effect	df	Sum of Squares	Mean Squares	F	p
HRTF (H)	4	83.5805	20.8951	2.7934	0.0334
Participant (P)	16	1802.5452	67.6591	15.39	<0.0001
P * H	64	478.7307	7.4802	1.7	0.0012
Error	425	1868.1055	4.3955		
Corrected total	509	3512.9618			

Table 1: ANOVA table for data subset in cue direction=-30^{o}

Effect	df	Sum of Squares	Mean Squares	F	p
HRTF (H)	4	46.0018	11.5005	3.47	0.0084
Participant (P)	16	821.295	51.3309	15.49	<0.0001
P * H	64	307.4255	4.8035	1.45	0.0183
Error	425	1408.7614	3.3147		
Corrected total	509	2583.4837			

Table 2: ANOVA table for data subset in cue direction=-15^{o}

Effect	df	Sum of Squares	Mean Squares	F	p
HRTF (H)	4	56.0562	14.0141	4.00	0.0034
Participant (P)	16	696.9822	43.5614	12.43	<0.0001
P * H	64	385.2913	6.0202	1.72	<0.0001
Error	425	1489.1627	3.5039		
Corrected total	509	2627.4925			

Table 3: ANOVA table for data subset in cue direction=+30^{o}

Effect	df	Sum of Squares	Mean Squares	F	p
HRTF (H)	4	49.6632	12.4158	2.13	0.0767
Participant (P)	16	1281.1814	80.0738	13.71	<0.0001
P * H	64	587.9492	9.1867	1.57	0.0051
Error	425	2481.954	5.8399		
Corrected total	509	4400.7478			

Table 4: ANOVA table for data subset in cue direction=+60^{o}

The significance for the factor "HRTF" on 'transformed error' for each cue direction is:

Cue direction	Mean Squares of "HRTF" (dof=4)	Mean Squares of "HRTFxParticipant" (dof = 64)	$F_{4,64}$	p
-30^o	20.90	7.48	2.794	0.0334
-15^o	11.5	4.8	2.396	0.0594
$+30^o$	14.01	6.02	2.327	0.0656
$+60^o$	12.42	9.18	1.353	0.2602

Table 5: Summary of significance level of factor "HRTF" in each cue direction conditions

It is observed that the significance level for the 4 data subsets above ranged from 0.03 to 0.26. Therefore the effect of the HRTF choices is marginally significant on participants overall localization error. Also this significance is dependent on the cue direction as well. Although we acknowledged the choices of non-individualized HRTFs are different at different cue directions, however we expect the factor "HRTF" should show significant effect on the participants' localization error in all four cue-directions. This unexpected result can due to many causes. But all leads to the selection criteria of the HRTF choices, which is based on the matching score between sound stimuli and listeners. More details on this question will be discussed in section 5.4.

In later analysis, we will separated the whole data set into 4 data subset in the same manner (i.e. each data subset refers to a specific cue direction).

5.3.1 Regression Analysis: Using Matching Score to predict localization errors

Our interest in this study is to use a simple index ("Matching score" in our study) to predict listener's localization error when listen to a particular sound stimulus. Therefore a regression analysis is performed to evaluate the significance of our predictor, Matching Score (MS), in the following model:

$$Transformed\_Error = A \times (MS) + B \times (tar\_ele) + C \times (repeat) + D$$

Equation 3: Model equation for regression analysis

In addition to doing a regression analysis on the whole dataset, similar to ANOVA test, we will separate the regression analysis into 4 regressions, each on data subset with specific cue elevation:

$$Transformed\_Error = A \times (MS) + B \times (repeat) + C$$

Equation 4: Model equation for regression analysis (single cue direction)

Results can be found in the 5 tables below:

Predictor	Parameter Estimate	p
Intercept	7.579	<0.0001
tar_ele	0.007	<0.0001
Repeat	-0.005	>0.9
Matching Score (MS)	-1.865	<0.0001
	R^2=2.5%	

Table 6: Regression analysis on data from all 4 directions condtions

Predictor	Parameter Estimate	p
Intercept	8.1375	<0.0001
Repeat	0.0205	0.8
Matching Score (MS)	-2.529	0.004
	R^2=2.4%	

Table 7: Regression analysis on data from cue direction=-30^o

Predictor	Parameter Estimate	p
Intercept	6.778	<0.0001
Repeat	-0.05	0.4041
Matching Score (MS)	-1.1045	0.0687
	R^2=0.9%	

Table 8: Regression analysis on data from cue direction=-15°

Predictor	Parameter Estimate	p
Intercept	5.57	<0.0001
Repeat	0.013	0.8711
Matching Score (MS)	0.317	0.7669
	R^2=0.04%	

Table 9: Regression analysis on data from cue direction=+30°

Predictor	Parameter Estimate	p
Intercept	8.86	<0.0001
Repeat	-0.018	>0.8
Matching Score (MS)	-2.8	<0.001
	R^2=3.4%	

Table 10: Regression analysis on data from cue direction=+60°

Among the four regression analyses (Table 7 to Table 10), there are 3 test results (cue direction = -30°, -15° and +60°) showing the matching score is significant (cue direction = -30° and +60°) or marginally significant (cue direction = -15°) in predicting the localization error. Also their relationship is negatively correlated (see "parameter estimate"): the higher is matching score, the smaller will be the perceived localization error by the listener on the sound stimulus. In a way, we have successfully creating a new index for ranking sound stimuli for smaller perceivable localization error. However, we also observe that the model fitting, denoted by R^2, is extremely small for the 3 cases (cue direction = -30°, -15° and +60°). Therefore we suspect the low fitting of model may be caused by within-subject variability (since the between-subject variability should be predicted by matching score).

To compare intra- and inter-subject variability, Hierarchical Analysis (for detail, see Gelman and Hill, 2007), another class of regression analysis will be performed in the next chapter.

6. Multilevel analysis on inter-listener variability and its relationship with matching scores

One of the basic assumptions in classical regression analysis (as used in chapter 5) is that all predictors in the regression model (see Equation 3) are required to be independent to each other. However, the matching score in our model, by definition, is dependent on HRTF, hence cue direction ("tar_ele"), and participant. Therefore the data structure would be in a hierarchy. Therefore we will use another class of regression analysis, which relax the assumption of independent variables in the model equation, for our study. This analysis method is called hierarchical analysis (Gelman and Hill, 2007), or multilevel analysis (Hox, 2002). It had been applied widely in human behavior studies in social science studies (Hox, 2002; Singer, 1998). A classical example to illustrate the details of this analysis is using the "school effect model".

6.1 School Effects model: An introduction (Literature Review)

The school effect model describes a behavior study using questionnaires to collected data from different students, who belong to different classes in a school and also from many other schools. Normally the scale of this type of study is quite large with sample population over 1000. And the most important characteristic of the data collected is the nested data structure of variables being studied.

For a particular student, he or she must belong to only one class, which in turns belonging to only one school. If the scale of study is really large, we may need to consider the geographical location of the school as well. In other words, all the variables in the model are nested together.

In hierarchical analysis, we need to clearly define the 'level' of predictor (i.e. variable) in our regression model. In the 'school effect model' example, students are within class and classes are within school. If we are considering a two-level hierarchy and students' individual behavior are assumed to be a repeat measure within the class, then 'class' will be known as level-1 (lower level) predictor and 'school' will be known as level-2 (higher level) predictor. Lower level predictors are grouped together by higher-level predictors. Also more levels can occurs in a hierarchical model.

Since the assumption of independent variables is relaxed, therefore variables are expected to be affected by other variables. Thus the concepts of intra-group and inter-group variances are introduced.

6.2 School Effects model analyses to determine inter-listener variability

Singer (1998) had illustrated a very detail example in using PROC MIXED in statistical software SAS® (SAS Inc.) to perform hierarchical analysis. In general, a two-level school effect model requires a 4-step analysis. The 4 steps require building a different regression model and start from a model without any predictor, called "unconditional model" (Singer, 1998) or

"Intercept-Only model" (Hox, 2002). Then one more predictor will be added into the model and the final model will include both level-1 and level-2 predictors. Apart from predictors at different levels, another type of variables also exists as an indicating variable ("SchoolType" in Figure 17). A simple illustration of the data hierarchy in Singer's example:

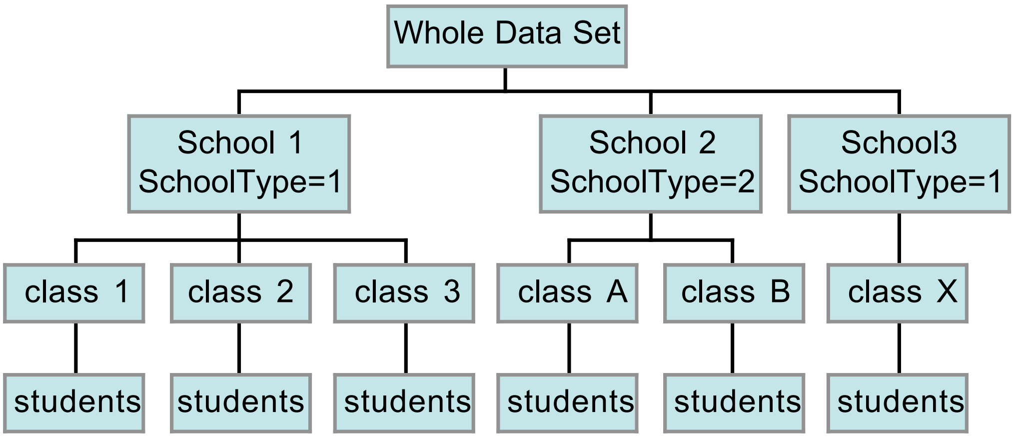


Figure 17: Outline of Singer's school effect model

For details of creating SAS code for analysis and the result interpretation method, readers can refer to Singer's work (Singer, 1998). Our data structure is also similar to this example:

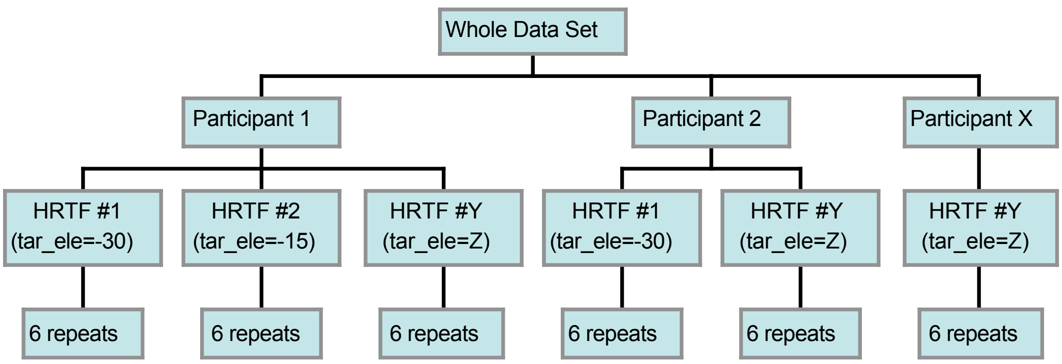


Figure 18: Outline of Our data as in "school effect model" strucutre

In Singer's example, the grouping variable is "school", while "participant" is the grouping variable in our study. Also "HRTF" carries an indicating variable "tar_ele" denoting the directional cues carried by the associated HRTF. Therefore we can rerun Singer's analysis code with following conversions:

Variable names in Singer's example	Our variables
mathach	Transformed error
school	Participant
meanses	Average MS for individual listener
sector	tar_ele

Table 11: Variable names comparison between our data and Singer's data (1998)

6.2.1 Results

By following analysis procedure suggested by Singer, which involves comparing within- and between-group (in our study, "subject" is the group) variances between the 4 regression models (mentioned in section 6.2). We can conclude how much variance is predicted by our matching score predictor. And compare the proportion of total variances divided by the grouping variable. Results of the 4 hierarchical analyses are shown below:

	Model 1: Intercept-only model		
	Parameter estimate	Standard error	p
Covariance estimate			
between-subject	**1.1361**	**0.4298**	**0.0041**
within-subject	**5.4413**	**0.2027**	**<.0001**
Fix Effect			
Intercept	**6.0764**	**0.2661**	**<.0001**

Table 12: Hierarchical analysis result on model 1: Unconditional model

	Model 2: level-2 predictor model		
	Parameter estimate	**Standard error**	**p**
Covariance estimate			
Between-subject	**0.8255**	**0.3292**	**0.0061**
Within-subject	**5.4412**	**0.2027**	**<.0001**
Fix Effect			
Intercept	**6.0379**	**0.2297**	**<.0001**
subject average MS (SMS)	**-12.6847**	**4.9395**	**0.0214**

Table 13: Hierarchical analysis result on model 2: level-2 predictor model

	Model 3: level-1 predictor model		
	Parameter estimate	**Standard error**	**p**
Covariance estimate			
Between-subject	**1.1909**	**0.4531**	**0.0043**
Within-subject	**5.3353**	**0.1991**	**<.0001**
Fix Effect			
Intercept	**6.0323**	**0.2721**	**<.0001**
HRTF_average_MS (HMS)	**-0.7679**	**0.3872**	**0.0475**
tar_ele	**0.008223**	**0.001849**	**<.0001**
HMS*tar_ele	**-0.02039**	**0.01035**	**0.049**

Table 14: Hierarchical analysis result on model 3: level-1 predictor model

	Model 4: All predictors model		
	Parameter estimate	**Standard error**	**p**
Covariance estimate			
Between-subject	**0.8915**	**0.3691**	**0.0079**
Within-subject	**5.332**	**0.1989**	**<.0001**
Fix Effect			
Intercept	**5.9937**	**0.2381**	**<.0001**
HRTF_average_MS (HMS)	-0.5583	0.4516	0.2166
tar_ele	0.008253	0.001848	<.0001
subject average MS (SMS)	**-12.5072**	**5.1169**	**0.0273**
SMS*HMS	16.4224	11.4523	0.1518
HMS*tar_ele	**-0.02039**	**0.01035**	**0.049**

Table 15: Hierarchical analysis result on model 4: All predictors model

Model	Between-subject variance	Total variance	Proportion of total variance as inter-subject variance
Intercept-only	1.1361	5.4413	17.3%
level-2 predictor	0.8255	5.4412	13.2%
level-1 predictor	1.1909	5.3353	18.2%
all predictors	0.8915	5.332	14.3%

Table 16: A summary of variance components for the 4 regression models

From Table 16, the proportion of variances explained by the inter-subject variation equals 1.1361 / (1.1361+5.4413) = 17.3% .

By Snijders' equation (Snijders and Bosker, 1999, p. 102-103) for calculating explained variance by our predictors, we find that the proportion of between-subject variances explained by our matching score predictor equals (1.1361-0.8255)/1.1361 = 27%.

6.2.3 Discussion

The hierarchical analysis reveals that the between-subject variation is much smaller than the within-subject variation. Further analysis provides evidence that our matching score can reduce 27% of unexplained variance when it is included in our regression model. More questions are opened: why our listeners are having so large variation in their responses? How to predict the other 73% of between-subject variation?

However these questions cannot be answered at the moment as it is outside our scope of study.

7. Conclusion

7.1 General Discussion

7.1.1 Relating sound localization errors in up / down directions and our proposed matching scores and its implications

This study proposes and evaluates the possibility in predicting listeners' sound localization errors in up / down direction with a new matching score proposed and developed by the author. This score is calculated based on the similarities of the acoustics responses of listeners' audio organelles (namely the acoustical structure of the pinna) and the frequency content of the sound cue. This score has been based upon the 'delay-and-add' theory proposed by Hebrank and Wright (1974) (see Chapter 3). Details of how the scores are calculated are documented in Chapter 4 and Appendix C. An empirical experiment has been conducted (see Chapter 5) and the results of regression analyses show that this score is significantly correlated with the measured localization errors in four up / down directions perceived by seventeen listeners (see Chapter 6). This finding provides empirical support to the 'delay-and-add' theory and implies that individual's sound localization errors in the up / down directions are statistically related to the dimensions of their outer ears. However, further analyses reveal that the levels of significance in correlation depend on the cue directions. This dependence is not predicted in our theory, as the 'delay-and-add' theory should apply to all up / down incident sound angles. Further analyses show that the significant relationships are maintained only when the incident sound angles are from 60 degrees up and

30 degrees down. Another unexpected observation is the low R^2 value in the regression analysis. This low R^2 value is due to the unexpectedly large intra-listener variability in perceived sound localization errors. In this study, the intra-listener variability is larger than the inter-listener variability. More discussion on this can be found in the next section (Section 7.1.2). Regardless of the unexpected observations listed above, the matching score index is statistical correlated with the perceived localization errors. Given that this matching score is the first of its kind and the lack of empirical verification on the 'delay-and-add' theory, the significant correlation results are of important implications to the theory. Furthermore, if the matching scores have significant relationships with the perceived sound localization errors, the scores can be used as selection criteria for individual listeners to choose their most appropriate sound cues. In due course, this score can be an important tool for mass-customizing 3D sound cues based upon low-cost non-individualized HRTF-filters that are freely available in the Internet. This study is just the first step and future work to improve the matching score is desirable.

7.1.2 Intra-listener and inter-listener variability in perceived sound localization errors in up / down directions

Since the R^2 value is very low (~2% of total variance) in the regression analyses, further analyses have been conducted to examine the causes of variance in the data. Hierarchical analysis (namely the 'School Effects Model') has been used (see Chapter 6). By isolating the intra-listener variability and the variance due to different incident angles of sound, we found that our matching score can explain 27% of total variance in localization errors. In other words, the matching score can successfully explain 27% of the inter-listener variability in sound localization errors. This indicates that about two-third of total variances have not been explained by our matching scores. As explained in Chapter 6, this suggests that the 'delay-and-add' theory can only partially explain the human perception of up / down direction of a sound cue. Other possible reasons for the non-explained inter-listener variance include (i) inaccuracies in measured ear dimensions; (ii) inaccuracies in aligning the incident sound path with the 3D ear moulds; (iii) arbitration in deciding the locations of the reflecting points inside one's outer-ear in response to a particular incident sound; and (iv) finally, the lost of information when only the notch frequencies are extracted from the spectra of the sound cues (i.e., the magnitude of the notch is disregarded). Readers can refer to Chapter 6 for details. Given that this study is the first such research effort in substantiating the 'delay-and-add' theory into a matching score index to predict up / down sound localization error, the relatively low percentage in explained variance can be a target for improvement. The important finding is that the matching score has a statistically reliable relationship with the measured up / down sound localization error.

7.2 Conclusion and Limitations

7.2.1 Conclusion

A matching score based upon the 'delay-and-add' theory (Hebrank and Wright, 1974) has been developed and presented.

A survey of ear dimensions of 33 participants has been conducted and 25,344 scores have been calculated between these 33 participants and up / down sound cues filtered by 192 open-copyrighted non-individualized head-related transfer functions selected from the LISTEN database (Ircam and AKG Acoustics, 2004) and the CIPIC database (Algazi *et al.,* 2001).

An empirical sound localization experiment has been conducted and the perceived up / down sound localization errors have been shown to significantly correlate with the matching scores. This provides the first empirical support for the 'delay-and-add' theory using measured sound localization data.

An unexpected large intra-listener variability in sound localization errors are found although the average sound localization errors (36 degrees) have been shown to be consistent with other studies (Zotkin et al.,2003, Seeber and Fastl, 2003). Using hierarchical analyses (namely the 'School Effects Model'), the contribution of the intra-listener variability and the contribution of the different incident sound angles have been successfully isolated. Results indicate that the matching score can explain up to 27% of the inter-listener variability in the sound localization performance.

7.2.2 Limitations

There are some limitations in this study, regarding the generation of matching score (which is the core of this study). The first one is related to the image edge extraction algorithm and the ear dimension extraction procedure. In our current ear dimension extraction procedure, human intervention is needed to determine the position of ear opening in the ear photos (see Chapter 3). Also parameter tuning is needed in processing the ear photos. Since human ear shape is very complex, it is reasonable to expect some inaccuracies associated with the selection of the above mentioned parameters. To facilitate future researchers to duplicate the current studies, all ear photos have been included in this thesis as Appendix G.

Another limitation of our study is the number of choices available for listeners. Among our 17 participants, the range of matching score between the stimuli and the participants is between 0.3 and 1.0. Therefore more non-individualized HRTFs, or more participants, are needed to increase the range of matching score being studied. This limited range of scores may also have caused the low value of R^2 in the regression analyses (see Chapter 5).

7.3 Future Work

There are two-third of the total variance in the measured sound localization errors cannot be explained by our matching score. Therefore additional predictors (maybe from other directional cues) are needed to improve the significance of matching score. One suggested approach is to include the magnitude of the frequency notches in the sound spectra rather than ignoring them.

Another improvement can be made is related to subject stability. The insufficient stability of participants had been observed to be problematic. In this study, subject-selection has already been done before experiment using a sound localization performance test (3 of 20 subjects have been excluded because they failed to detect up / down direction of incident HRTF-filtered sound cues). The author recommends that in future study, listeners can be grouped according to their stability in localize sound. A review of literature indicates that intra-listener variability in sound localization performance is rarely reported and studied, future work on this topic is desirable.

8. Reference

[1] V.R. Algazi, R.O. Duda, D.M. Thompson, C. Avendano (2001). The CIPIC HRTF database. In 2001 IEEE ASSP Workshop on Applications of Signal Processing to Acoustics, pp. 1111-123, New Paltz, NY, USA, October 2001.

[2] F. Asano, Y. Suzuki, T. Sone (1990). Role of spectral cues in median plane, J. Acoust. Soc. Am., vol. 88(1), pp. 159-168.

[3] J. Blauert (1969/70). Sound localization in the median plane, Acustica, vol. 22, no. 4, pp. 205-13

[4] J. Blauert (1997). Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press, London, England.

[5] S. Carlile (1996). Virtual auditory space: generation and applications, R.G. Landes Company and Chapman & Hall, Texas.

[6] B.U. Seeber and Hugo Fastl (2003). Subjective Selection of Non-individual Head-Relate Transfer Functions. Proceegins of the 2003 International Conference on Auditory Display, ICAD03-1 – 4, Boston, MA, USA, July 2003.

[7] M.B. Gardner, R.S. Gardner (1973). Problem of localization in the median plane: effect of pinnae cavity occlusion, J. Acoust. Soc. Am., vol. 53(2), pp. 400-408

[8] W.G. Gardner, K.D. Martin (1995). HRTF measurements of a KEMAR. J. Acoust. Soc. Am., 97(6), pp. 3907- 3908, 1995.

[9] Andrew Gelman, Jennifer Hill (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press, New York.

[10] J. Hebrank, D. Wright (1974). Spectral cues used in the localization of sound sources on the median plane, J. Acoust. Soc. Am., vol. 56(6), pp. 1829-34.

[11] Joop Hox (2002). Multilevel Analysis: Techniques and Applications. Lawrence Erlbarum Associates, Publishers, Mahwah, New Jersey

[12] T.J. Imig, N.G. Bibikov, P. Poirier, and F.K. Samson (2000). Directionality derived from pinna-cue spectral notches in cat dorsal cochlear nucleus. J. Neurophysiol. 83: 907 – 925, 2000.

[13] Ircam and AKG Acoustics (2004). LISTEN HRTF Database. http://www.ircam.fr/equips/salles/listener/index.html

[14] Jeffres, L.A. A place theory of sound localization. J. comp. Physiol. Psychol. 61: 468 – 486, 1948

[15] G.F. Kuhn (1987), Physical acoustics and measurements pertaining to directional hearing, Directional hearing (ed. by Yost, W.A. and Gourevitch, G.), Springer-Verlag, New York, pp. 3-25)

[16] E.H.A. Langendijk, A.W. Bronkhorst (2002). Contribution of spectral cues to human sound localization, J. Acoust. Soc. Am., vol. 112(4), pp.1583 - 96.

[17] E.A. Lopez-Poveda and R. Meddis (1996), A physical model of sound diffraction and reflections in the human concha, J. Acost. Soc. Am., 100 (5), pp. 3248 – 3259, 2000.

[18] Mills A.W. (1972).”Auditory Localization in Foundations of Modern Auditory Theory, edited by J. V. Tobias (Academic, New York), Vol. 2.

[19] B.C.J. Moore and B.R. Glasberg (1983). Suggested formulae for calculating the auditory-filter bandwidths and excitation patterns.

[20] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang and M.H. Allerhand. (1992). Complex sounds and auditory imges. In Auditory Physiology and Perception, (Eds.) Y. Cazals, L. Demany, K. Horner, Pergamon, Exford, 1992.

[21] Lord Raleigh, J.W. Strutt (1907). Our perception of sound direction. Philosophical Magazine, 12:214-232

[22] V.C. Raykar and R. Duraiswami (2005). Extracting the frequencies of the pinna spectral notches in measured head related impulse responses, J. Acoust. Soc. Am., vol 118(1), pp. 364 – 374

[23] M. Slaney (1993). An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank, Apple Computer Technical Report #35. (http://web.interval.com/~malcolm/pubs.html#Gammatone)

[24] Judith D. Singer (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics, Winter 1998, vol. 24(4), pp. 323-355

[25] Tom Snijders and Roel Bosker (1999). Multilevel Analysis: An introduction to basic and advanced multilevel modeling. SAGE Publications, London

[26] Richard H.Y. So and B. Ngan (2006). Optimizing the accuracies of generic binaural directional cues simulating frontward and backward directions: cluster analyses and an experimental study. In Press.

[27] I.A. Vartanyan and E.S. Malinina (2004). The selectivity of Neurons in the Auditory zone of the Mouse Midbrain to the Direction of Movement of a Spectral Notch in Wide-Band Noise. Neuroscience and Behavioral Physioilogy, vol. 34(2), 2004.

[28] D.N. Zotkin, J. Hwang, R. Duraiswami, L.S. Davis (2003). HRTF Personalization using Anthropometric Measurements. In 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 156 – 60, New Paltz, NY, USA, October 2003

9. Appendix

A. Compensation algorithm in photo taking

Let X1, X2, X3, X4 and X5 are x-coordinates of the 5 points pin-pointed by experimenter on the horizontal ruler in ear photo.
Let Y1, Y2, Y3, Y4 and Y5 are y-coordinates of the same 5 points.

```
dE=mean(   abs(   X5-X4,  X4-X3, X3-X2, X2-X1  )  )
dx=abs(X2-X1);
dy=abs(Y2-Y1);

/* the proportion of distortion for each cm on ruler: horizontal (Kx) and vetrtical (Ky)
*/
Kx=dE/dx;
Ky=dE/dy;
```

Let (H/2, V/2) be the centre point of the ear photo.

```
/* Correction to points coordinates on search path */
x=H/2+(x-H/2)/Kx;
y=V/2+(y-V/2)/Ky;
```

B. Spectral notch extraction algorithm (Raykar et al., 2005)

2 Functions are developed to follow Raykar et al. (2005) notch extraction algorithm.

First Function is: mass_notch_extract.m

```
%Date: 14-Nov-2005 Author: John Au
%Function use for extracting notch-related parameters from extracted notches
%Syntax:
%[freq_l, freq_r, notch_num_l, notch_num_r, angle_median_db_out_l,
angle_median_db_out_r angle_var_num_notch_l
angle_var_num_notch_r]=mass_notch_extract(f_low,f_high)
%Input: f_low = lowest allowable notch frequency. f_high = highest allowable notch
frequency
%Output: freq_x=[sub_idx elev_idx elev_angle freq hrtf_x(db)
angle_var_num_notch_l angle_var_num_notch_r]
function [all_feature_vector]=mass_notch_extract(f_low,f_high,total_subject,angles)

elev=-45+15*(0:9);
wd=cd;
%f_low=5700;f_high=11300;
notch_dir=wd;
%Matrix for storing result (we assume for each direction there are 15 notch
frequencies at most)
res=zeros(15,size(elev,2),total_subject*2);


total_hrtf=size(res,3);
%OUTPUTs
all_feature_vector=[];
all_notch_freq=[];   %Feature 1: ALL NOTCH FREQUENCIES
all_notch_region=[]; %Feature 3: location of notches

HRTF and the corresponding notch frequencies

for i=1:total_hrtf

    %Extracting Raw HRIR (Right and Left) from each subject folder
    wd=cd;
    hrir_path=['../../Combined_HRTF/' num2str(i)];
    cd(hrir_path);
    load hrir;
    cd(wd);


    if (i<=97)
```

```
        IR=L;
    else
        IR=R;
    end
    %Now extract notches from each elevation angles (j:1 to 10)
    for j=angles
        cd(notch_dir);
        %Finding Left ear notches

        p=10;f_out=[];
        while (size(f_out,1)==0 & p<13)
            [db_out, f_out]=anatomical_notch_extract(IR(:,j),p);
            f_out=f_out(find(f_out<=f_high & f_out>=f_low));
            db_out=db_out(find(f_out<=f_high & f_out>=f_low));
            p=p+1;
        end
        if (numel(f_out)>0)
            for k=1:max(size(f_out))
                %all_notch_freq=[all_notch_freq;i j f_out(k)];
                %all_notch_region=[all_notch_region;i j bin_find(f_out(k))];
                %all_feature_vector=[all_feature_vector;i j -45+(j-1)*15 f_out(k)
bin_find(f_out(k))];
                all_feature_vector=[all_feature_vector;i j -45+(j-1)*15 f_out(k) db_out(k)];
            end
        end

    end

    assignin('base','all_feature_vector',all_feature_vector);
    ['Ear ' num2str(i) ' is done at time: ' num2str(clock)]
    save all_feature_vector_23may07.mat all_feature_vector;
end
cd(wd);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Second function used: Anatomical_notch_extract.m:

```
%Notch frequencies extraction from HRIR
%Usage: [d_out, f_out]=anatomical_notch_extract(X,p)
%X is the input HRIR and p is the order of linear prediction filter (optimal: 10-12)
function [db_out,f_out]=anatomical_notch_extract(HRIR,p)

%p=Order of Linear prediction (LP) filter

%Frequencies below 2kHz are excluded, so start of pt in frequency spectrum is set to
8
start=8;
%Sampling frequency
FS=44100;

%Delay Cutoff
% i=find(HRIR==max(HRIR));
i=find(HRIR==max(HRIR));
[pp,tt]=extremes(HRIR);

cutoff=find(abs(tt-i)==min(abs(tt-i)));
cutoff=tt(cutoff);
%Calculate the original HRTF with a length of 200
HRTF=20*log10(abs(fft(HRIR,400)));
HRTF=flipud(HRTF(numel(HRIR)/2+1:end));

HRIR=HRIR(cutoff:size(HRIR,1));

%Paremter estimate from LP residual
A=lpc(HRIR,p);

%Calculation of predicted HRIR(n) and the residual err
Xp=zeros(size(HRIR));
err=zeros(size(HRIR));
for n=p+1:size(HRIR,1)
   for k=1:p      %(correction on 26-nov-2006: k=2:p --> k=1:p)
      Xp(n)=Xp(n)-A(k)*HRIR(n-k);
   end
   err(n)=HRIR(n)-Xp(n);
end


%Hann Window of around 1.0ms creation
w=hann(FS*0.002);
%Auto-correlation function of the

hw0=zeros(size(HRIR));
%Hann Window creation
if (size(HRIR)>=size(w))
   hw0(1:size(w,1)/2)=w(size(w,1)/2+1:size(w,1));
```

```
else
   hw0=w(1:size(hw0,1));
end

%window the LP residual and perform auto-correlation
residual=err.*hw0;
autocorr_residual=xcorr(residual);    %residual is in time series

%Window the autocorrelated LP residual with Half Hann Window
hw1=zeros(size(autocorr_residual));
%Hann Window creation
if (size(autocorr_residual)>=size(w))
   hw1(1:size(w,1)/2)=w(size(w,1)/2+1:size(w,1));
else
   hw1=w(1:size(hw1,1));
end
win_auto_resid=autocorr_residual.*hw1;

%Group delay function
[b,a]=invfreqz(fft(win_auto_resid,400),linspace(0,pi,400),400,400);  %Lengths are set
to 200 as HRIR
[Gd,W]=grpdelay(b,a,400);
[p,t]=extremes(Gd(1:200));
W=W(1:200);
%%%%%%[23-Nov-2006 Correction]%%%%%
%only troughs with value <1 are count as spectral notches%
t=t(find(Gd(t)<-1));

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
f_out=W(t)*44100/pi;
f_out=f_out';
d_out=343./(4.*f_out);  %unit is metre (m)

%Added on 3-Dec-2006
%Filtering out the redundant notch frequencies according ot the equivalent
rectangular bandwidths (ERBs)
%of the corresponding Auditory Nervers (ANs)
%f_interval=linspace(1,22050,numel(HRTF));
%f_out=erb_select(f_out,f_interval(2),HRTF);

f_step=22000/numel(HRTF);
db_out=HRTF(floor(f_out/f_step));
```

C. A detail example of Matching Score calculation

Assuming 3 spectral notches are identified from an ear photo (ID=100) along the elevation =-15 degrees: 6.9kHz, 8.8kHz and 10.7kHz.

Build 3 gammatone filters based on the 3 spectral notches' centre frequencies with Slaney's toolbox. (Slaney, 1993):

```
fs=44100;
  len=1024;
  wd=cd;
  cd('./AuditoryToolbox');
  if (numel(ear_notch_vector)==1)
     ear_notch_vector=[ear_notch_vector_ori ear_notch_vector_ori];
  end
  [fcoefs]=MakeERBFilters(fs,ear_notch_vector);
  y = ERBFilterBank([1 zeros(1,len-1)], fcoefs);
  cd(wd);

  resp = 20*log10(abs(fft(y')));
  resp=resp(1:len/2,:);
  freqScale = (0:(len/2)-1)/(len)*fs;

if (numel(ear_notch_vector_ori)==1)
     resp=resp(:,1);
  end

  %Find maximum excitation for each model filter by each spectral
  %notch from HRTF
  idx1=round(hrtf_notch_vector./freqScale(2));


  %normalize the gammatone filter values in all ear notches filters
  for i=1:size(resp,2)
     resp(:,i)=(resp(:,i)-min(resp(:,i)))./(max(resp(:,i))-min(resp(:,i)));
  end

  plot(freqScale,max(resp'))
```

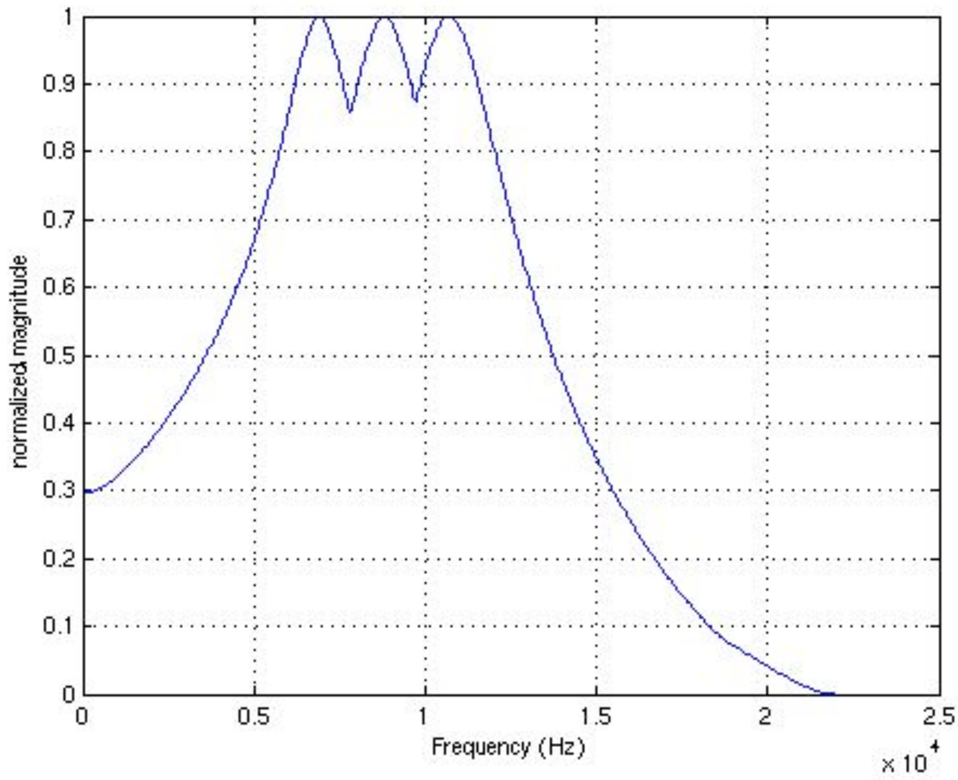


Figure 19: Graphical output of final gammatone filterbank

On the other hand, 2 spectral notches are extracted from a non-individualized HRTF describing a cue at elevation= -15 degrees: 7.7kHz and 9.1kHz

Find the magnitude at 7.7kHz and 9.1kHz on the above gammatone filterbank: 7.7kHz (~0.88) and 9.1kHz (~0.94)

The final matching score between this ear-HRTF pair will be the average value of all the magnitude values identified. In this case, the matching score = (0.88 + 0.94)/2 = 0.91

D. MATLAB code for sound clip convolution (i.e., generation of a HRTF-filtered sound cue).

```
wd=cd;
hrir_folder=['/Users/johnau/Documents/Research/Combined_HRTF/'];
noise_folder=['/Users/johnau/Documents/Research/Elevation/notch_detection/'];
snd_folder=['/Users/johnau/Desktop/expt/expt_snd/'];
cd(noise_folder);
load noise.mat;

for i=1:194
   cd([hrir_folder num2str(i)]);
   load hrir.mat;
   if (i>97)
      IR=R;
   else
      IR=L;
   end
   for elev=[2 3 4 6 8]
      Y=conv(noise,IR(:,elev));
      gap=zeros(0.3*44100,1);
      Y=[Y; gap; Y; gap; Y];
      if elev<4
         ele_val=['en' num2str(abs(elev*15-60))];
      else
         ele_val=['e' num2str(abs(elev*15-60))];
      end
      snd_fn=[num2str(i) '_1_a0' ele_val '.wav'];
      cd(snd_folder);
      wavwrite([Y Y],44100,32,snd_fn);
   end
end
cd(wd);
```

E. Sample of listening task consent form

Consent Form
for
Experiment of Human sound localization task
Venue: Acoustic Room

Date: ______________
Time: ______________
Subject ID: ______________
SID: ______________

1. Name __________________________________ (Mr/Mrs/Miss/)

2. Are you feeling ill in any way? Yes / No

3. Do you suffer from any form of hearing problems? Yes / No

4. Have you suffered any serious illness? Yes / No

5. Are you under medical treatment? Yes / No

If you answer is 'Yes' to question (2), (3), (4), or (5), please give details to the Experimenter.

Declaration
I consent to take part in an experiment. My replies to the above questions are correct to the best of my belief, and I understand that they will be treated as confidential by the experimenter.

I understand that I may at any time withdraw from the experiment and that I am under no obligation to give reasons for withdrawal or to attend again for experimentation.

I undertake to obey the regulations of the experiment and instructions of the experimenter regarding safety only to my right to withdraw declared above.
The purpose and methods of the research have been explained to me and I have had the opportunity to ask questions.

Signature of Subject ___________________________ Date ________________

This experiment conforms to the requirement of the University Research Ethic Committee.

Signature of Experimenter ______________________ Date ________________

(When completed this form should be filed in the Exposure Archive).

Subject Consent Form / June 2007

F. Sample instruction set used for localization task

<u>**Instruction for Sound Localization Experiment**</u>

Welcome to this experiment! Please read through this instruction first. If you have any question about the procedure, please ask the experimenter at any time.

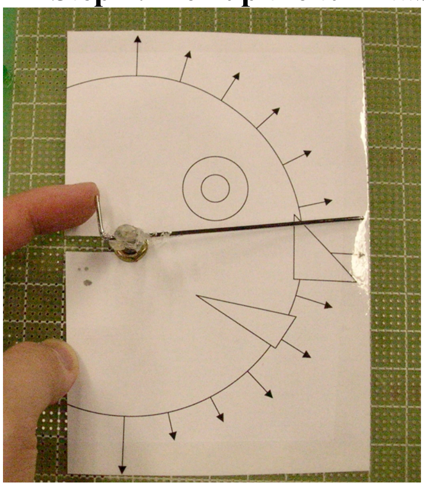
This experiment investigates the perceived direction of sound cues. During this experiment, about 200 sound cues will be presented to you and you will be asked to identify their incoming directions. The whole experiment will last for about 80 minutes. A rest will be given every 20 minutes.

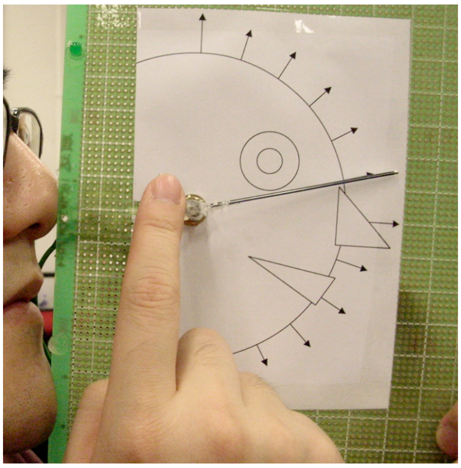
In this experiment, you are asked to:

1. Read and understand the procedure of this experiment.
2. Complete the consent form.
3. Enter the inside chamber to prepare for the listening test (training session).
4. Taking the listening test (breaks will be given in-between)

Tips on preparing to listen to a sound cue:
Step 1: <u>**Focus your eye view**</u> (i.e, line-of-sight) on the door marker.
Step 2: **Maintain your head orientation and upright sitting posture**
Step 3: **Close your eyes** and get ready to listen to the next sound.

Tips on responding the perceived direction of the sound cue:
Step 1: Pick up the turn-table:





Step 2: Point the cartoon's nose (on turn-table) to your front direction and hold the turn-table to a level so that the nose is at about same height as your nose.

Step 3: Rotate the needle to an orientation showing your perceived elevation.
If there is any problem during the experiment, please ask the experimenter. **Thanks for your help!**

Caution
Please beware of the equipment cables on the lab floor to avoid any accident. Please report to experimenter at once if you experience any illness.

When you have finished reading this page, please continue on the next page.

Details of listening test:

1. Please **switch off your mobile phone** and put all your belongings on the cupboard at the chamber corner.
2. Sit on the blue chair.
3. Put on the headphone (red cable lead on your right side)
4. Before the training session, we need to find out your natural upright sitting posture

Hints on sitting with natural sitting posture:

   a. Sit comfortably on the chair
   b. Imagine you are looking at very distant object in the direction of the door.
   c. Keep this eye position and report your eye position on the paper ruler on the door to the experimenter
   d. You may relieve your eyesight after you are told to do so.

5. The training session will start after **resting on chair for 30 seconds**.
6. In each run in the training/test, you need to do three things:
   a. Before you hear the sound:
      i. **Focus your eye view** (i.e. line-of-sight) on the marker (mark on door during the test preparation) in front of you.
      ii. **Maintain your head orientation and sitting posture** during the whole listening test.
      iii. **Close your eyes** and get ready to listen to the next sound.
   b. While hearing to the sound:
      i. Each sound clip contains several 'buzz' sounds
      ii. **Listening carefully and determine your perceived incoming direction** of this sound clip.
   c. After hearing the sound:
      i. **Show your perceived incoming elevation direction** by rotating the arrow on the turn-table:

      (pic: whole turn table) (picture: turning button on turn table)

      Tell the experimenter when you have complete the arrow re-orientation.

      ii. Return to your previous head orientation (back to (a))
7. For about every 20 minutes, a 2-3 minutes break will be given to you. You can leave the room during the break.

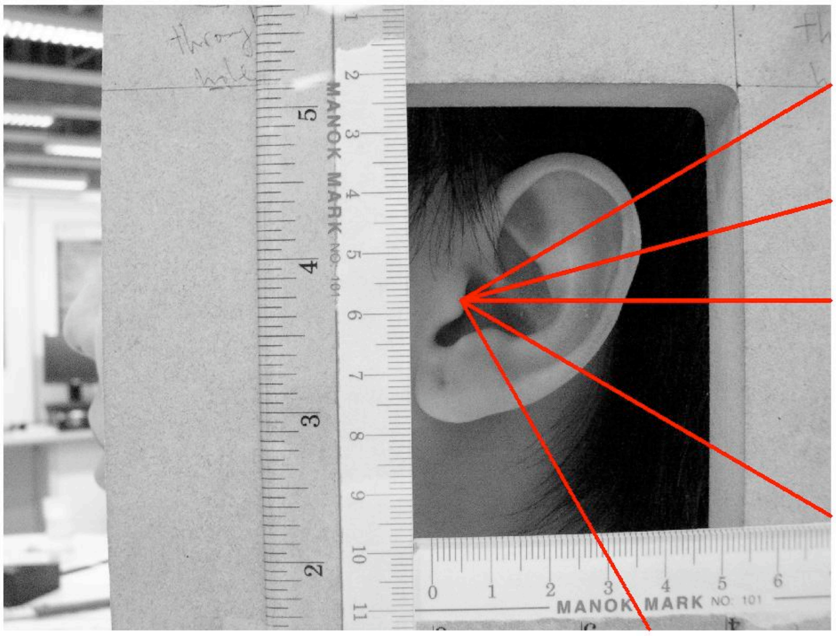
(When you have finished reading this page, please continue on the next page.)

G. Photos of Paired Ears from 36 participants

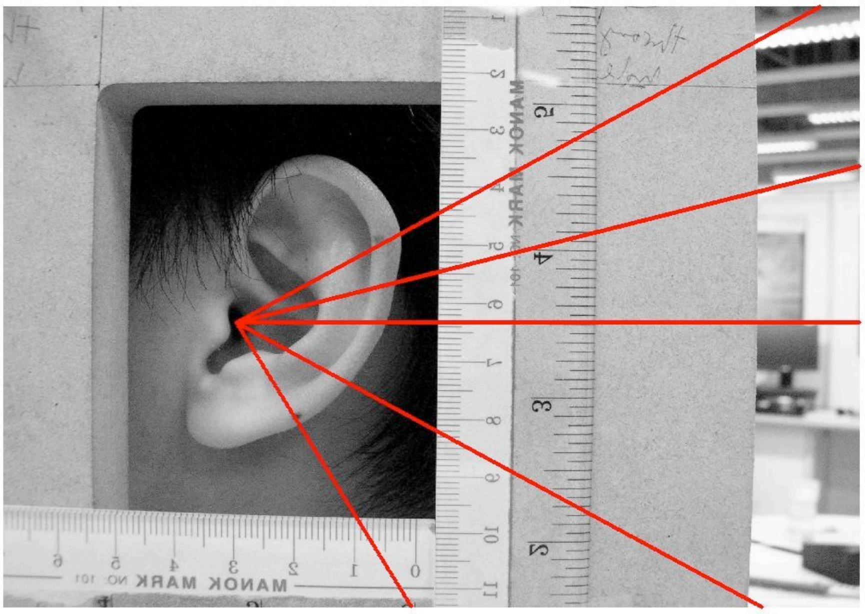
Subject #1 to #3 (p.87 to 89) are 3 test participants

Subject #4 to #36 (p.90 onwards) are 33 participants in ear survey.

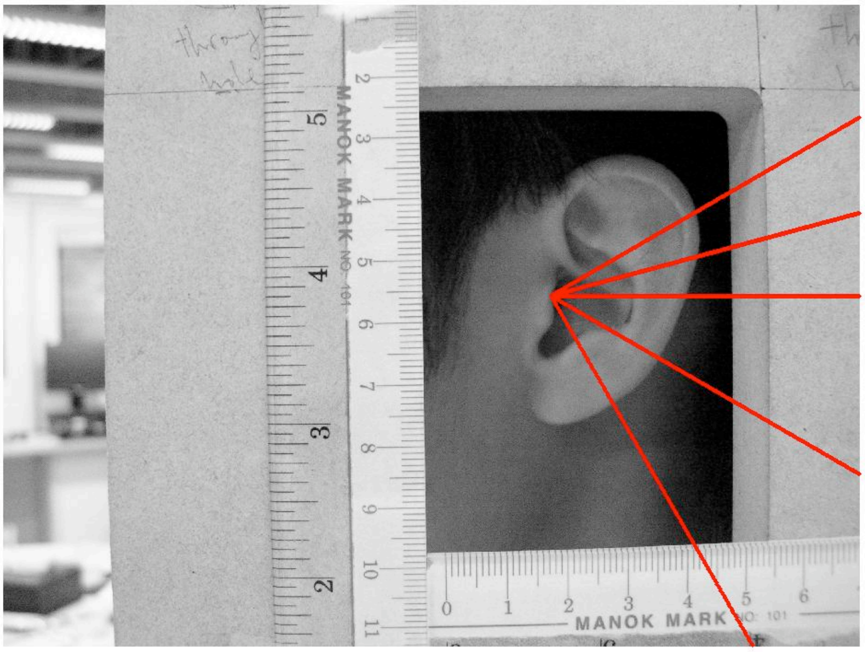
Subject: 1   Ear: L



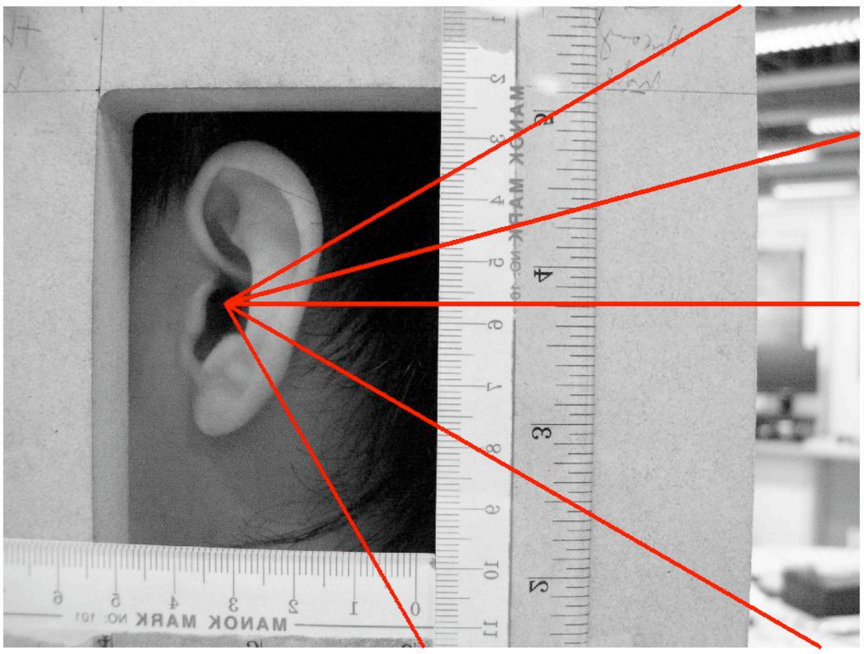
Subject: 1   Ear: R



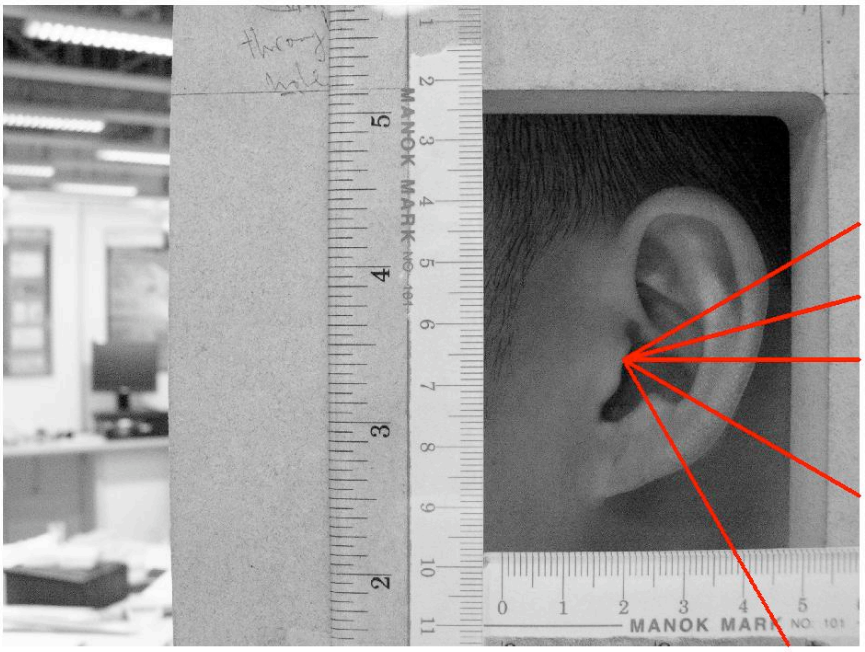
Subject: 2 Ear: L



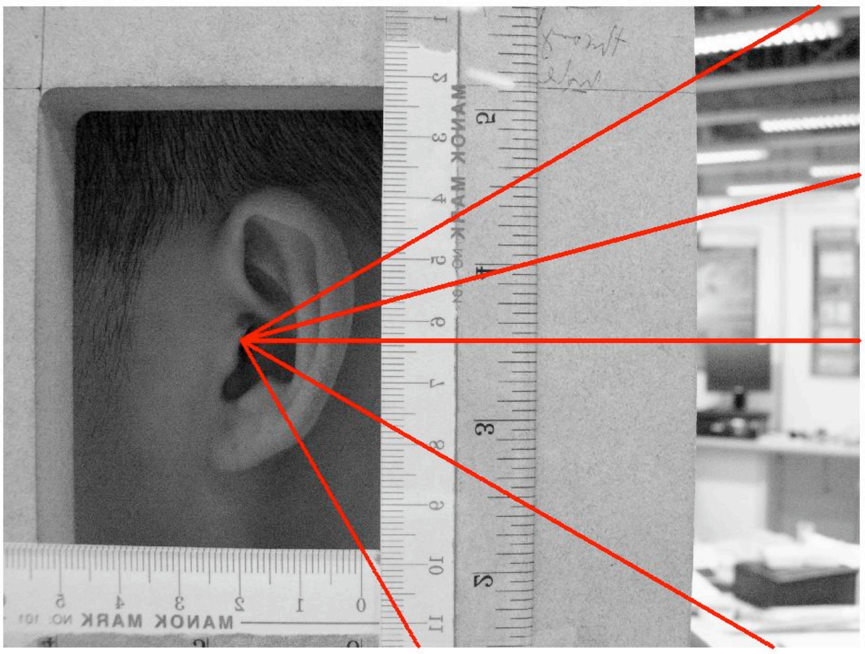
Subject: 2 Ear: R



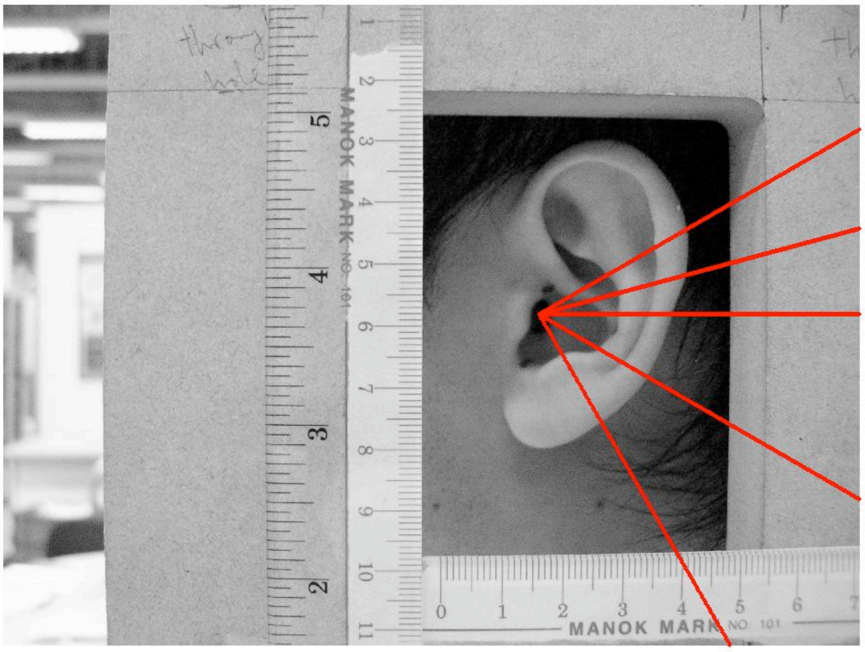
Subject: 3   Ear: L



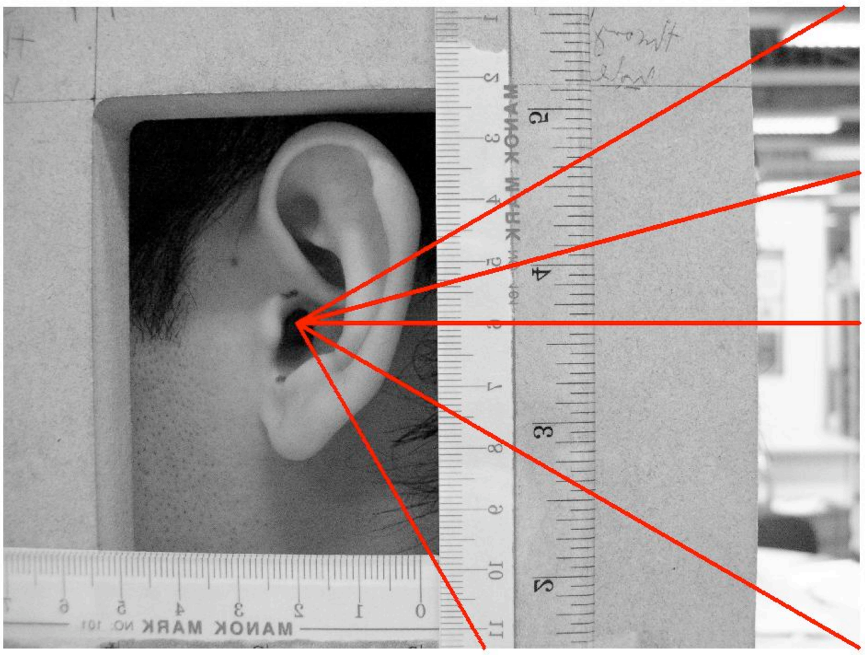
Subject: 3   Ear: R



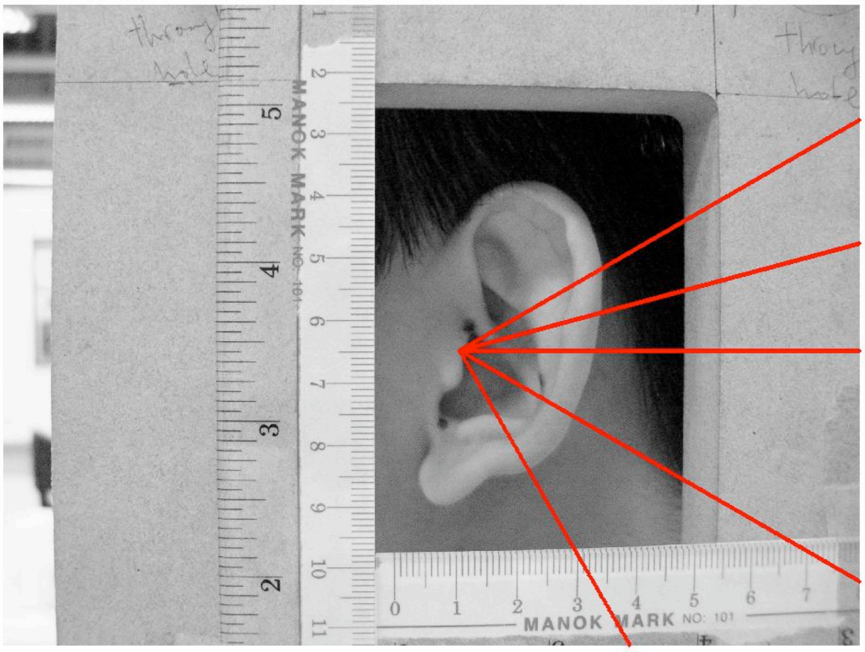
Subject: 4   Ear: L



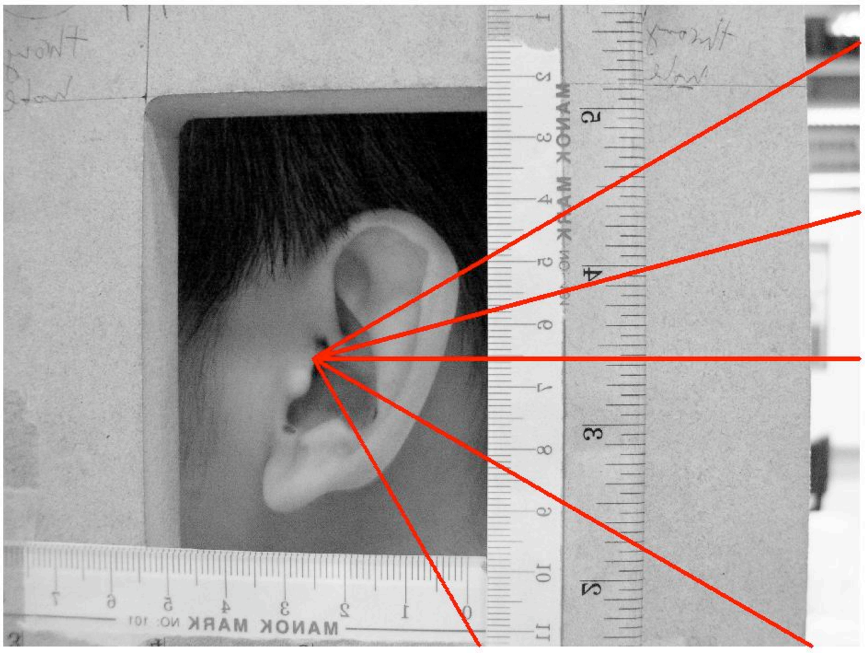
Subject: 4   Ear: R



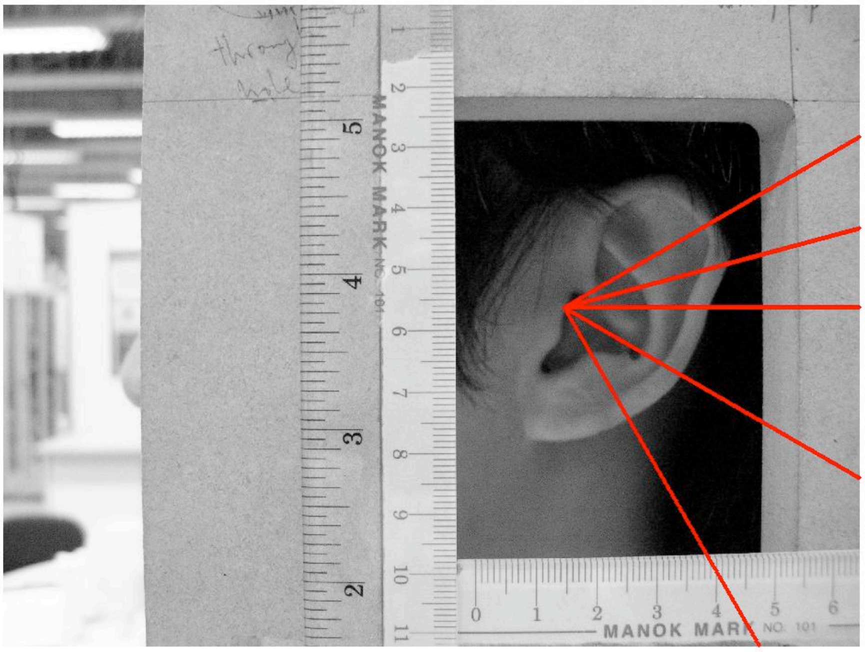
Subject: 5 Ear: L



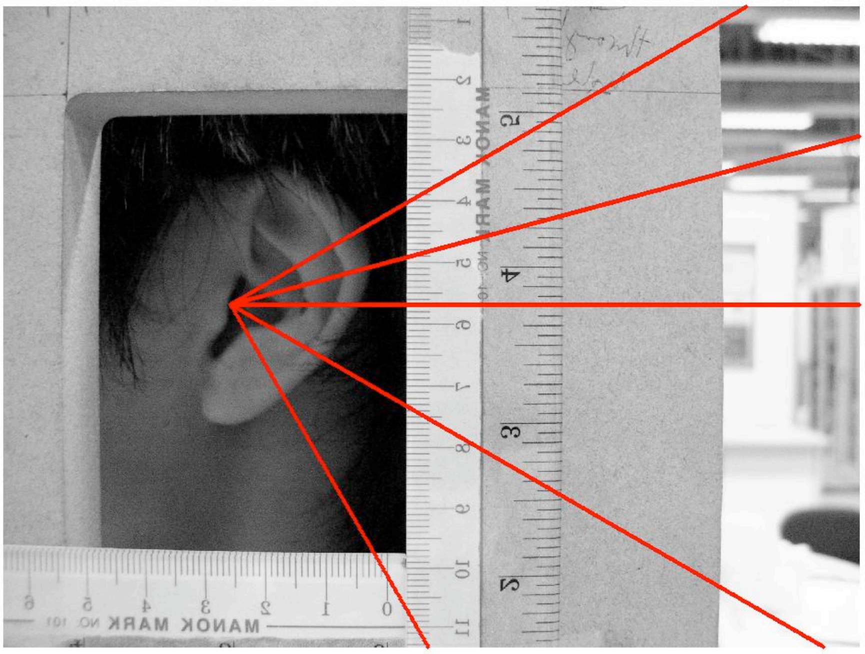
Subject: 5 Ear: R



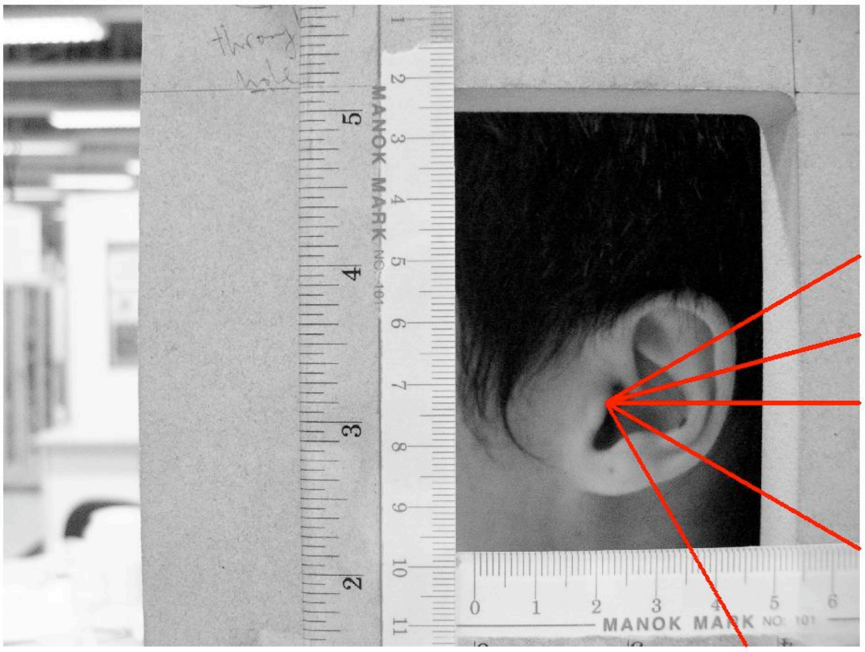
Subject: 6   Ear: L



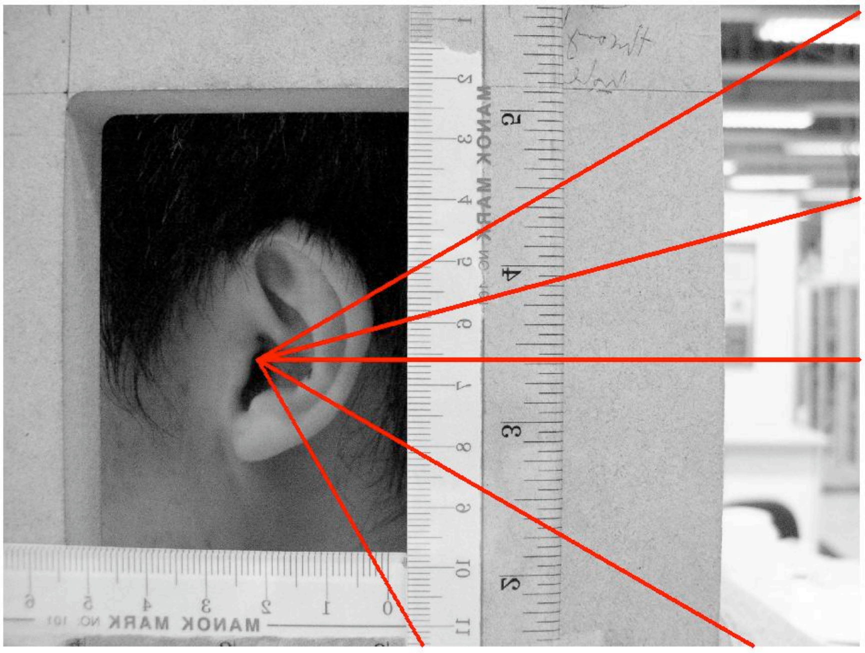
Subject: 6   Ear: R



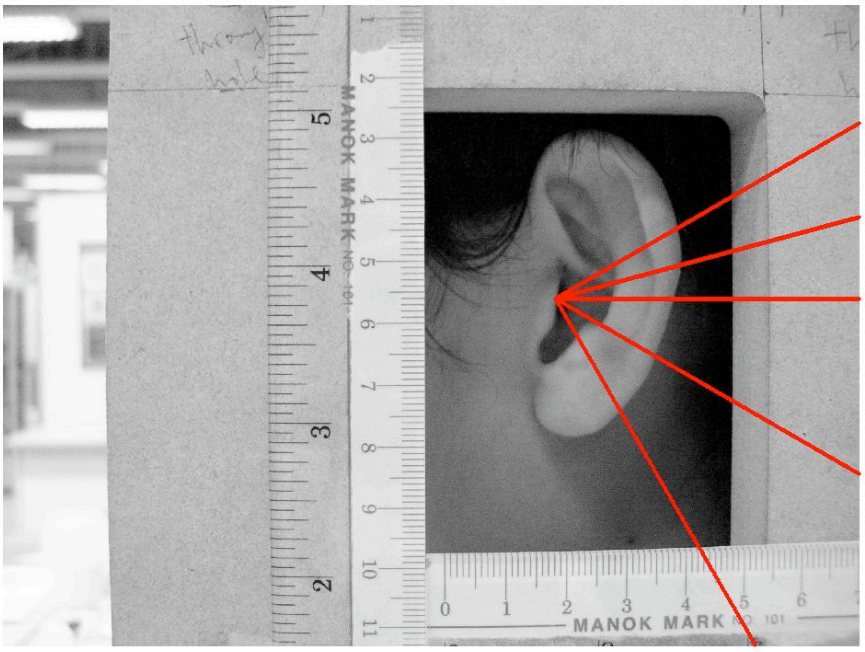
Subject: 7   Ear: L



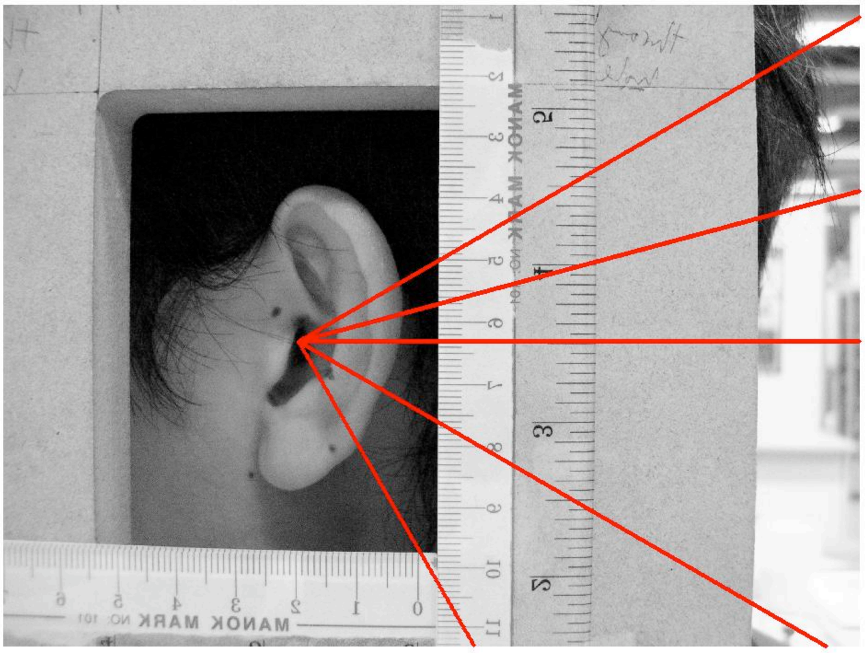
Subject: 7   Ear: R



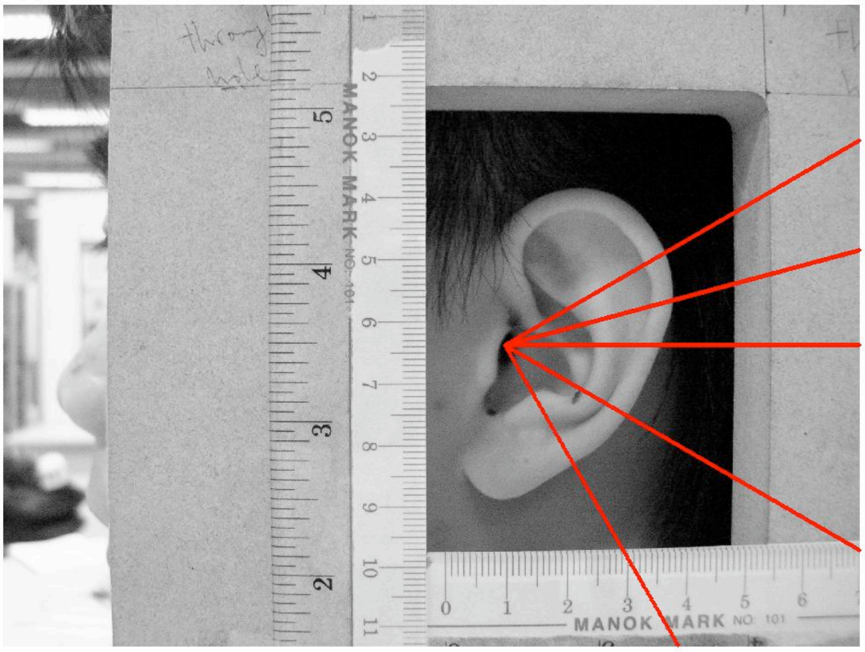
Subject: 8   Ear: L



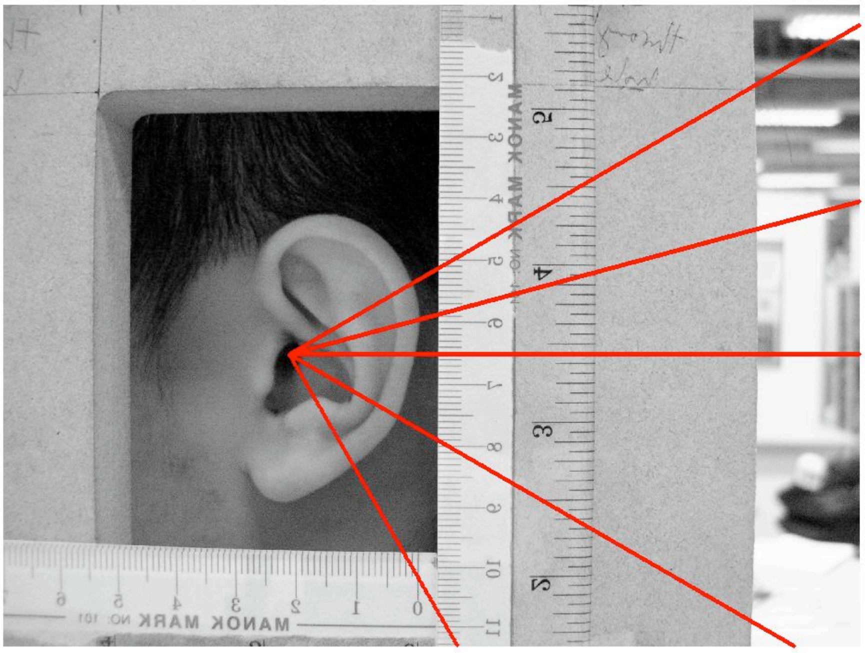
Subject: 8   Ear: R



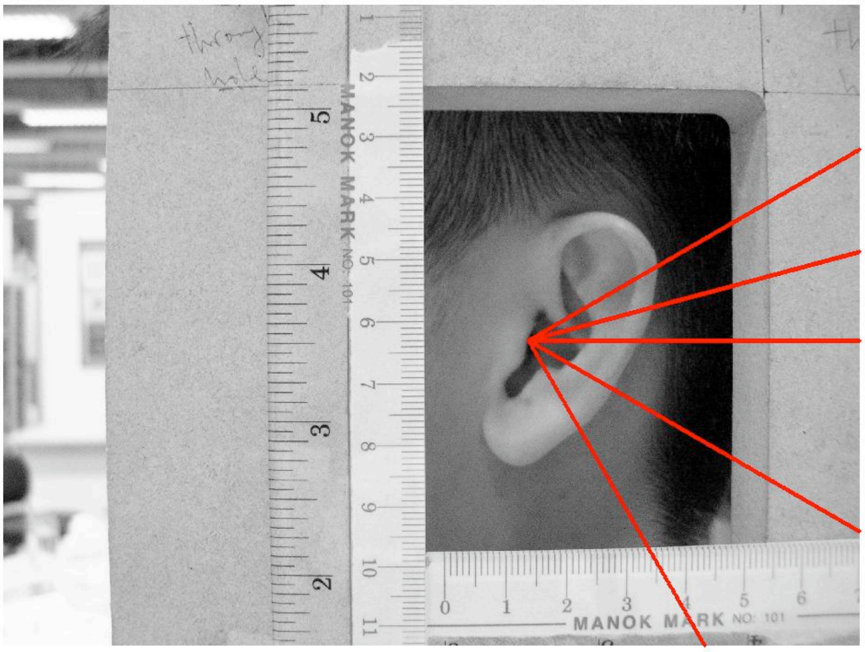
Subject: 9 Ear: L



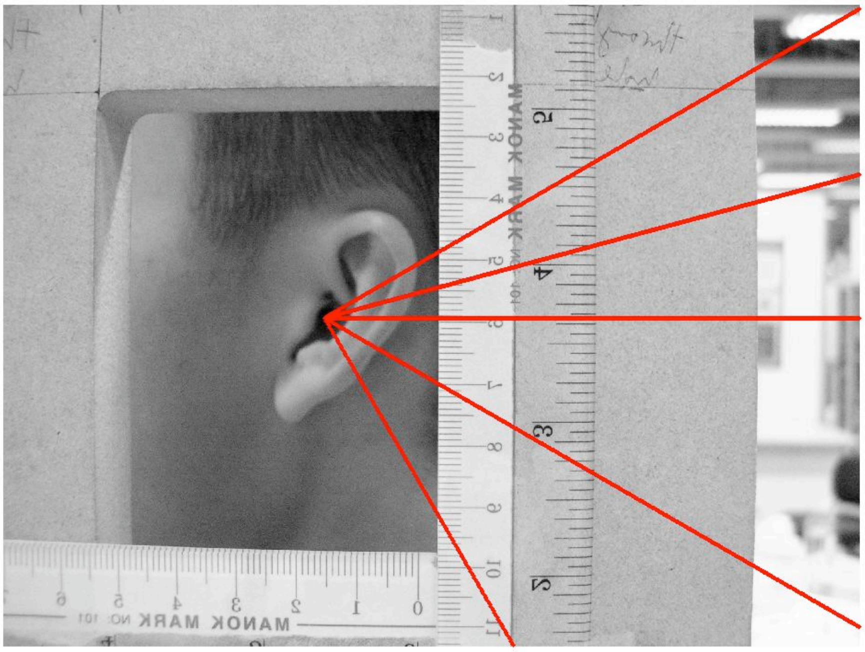
Subject: 9 Ear: R



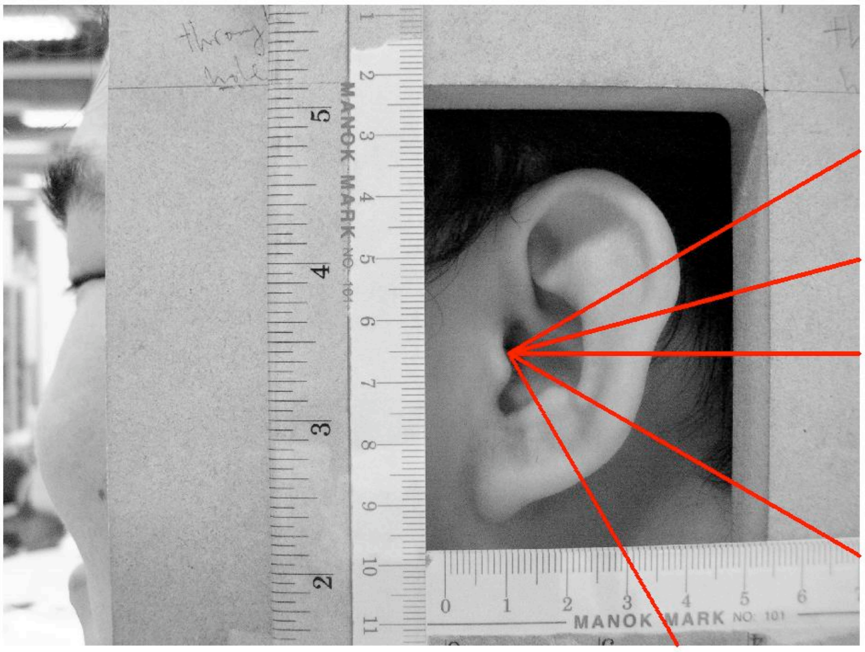
Subject: 10 Ear: L



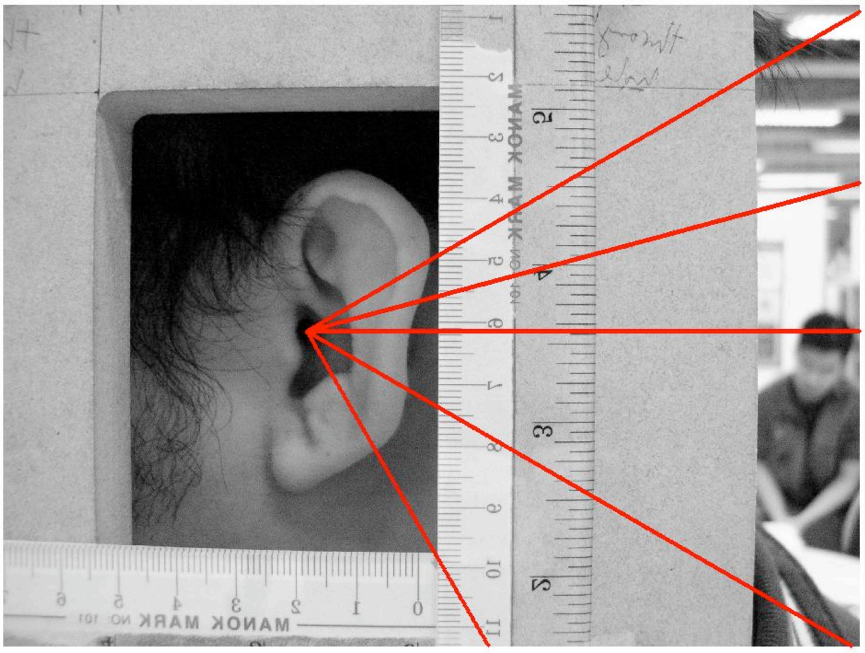
Subject: 10 Ear: R



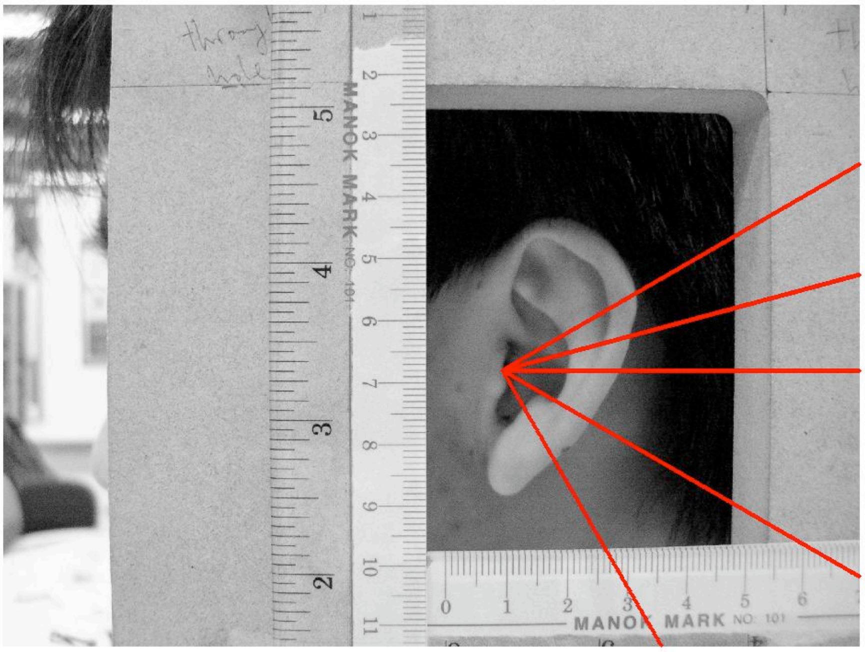
Subject: 11 Ear: L



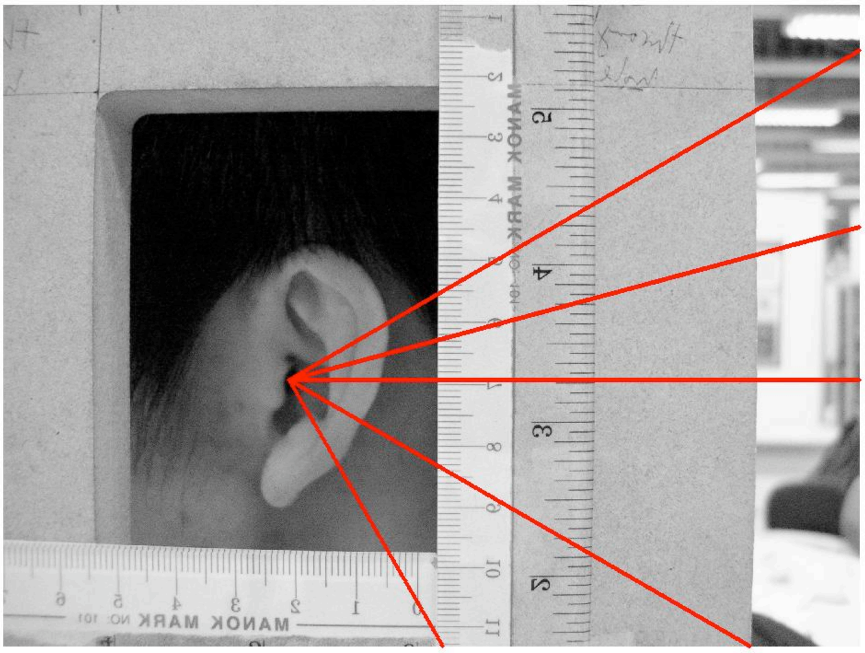
Subject: 11 Ear: R



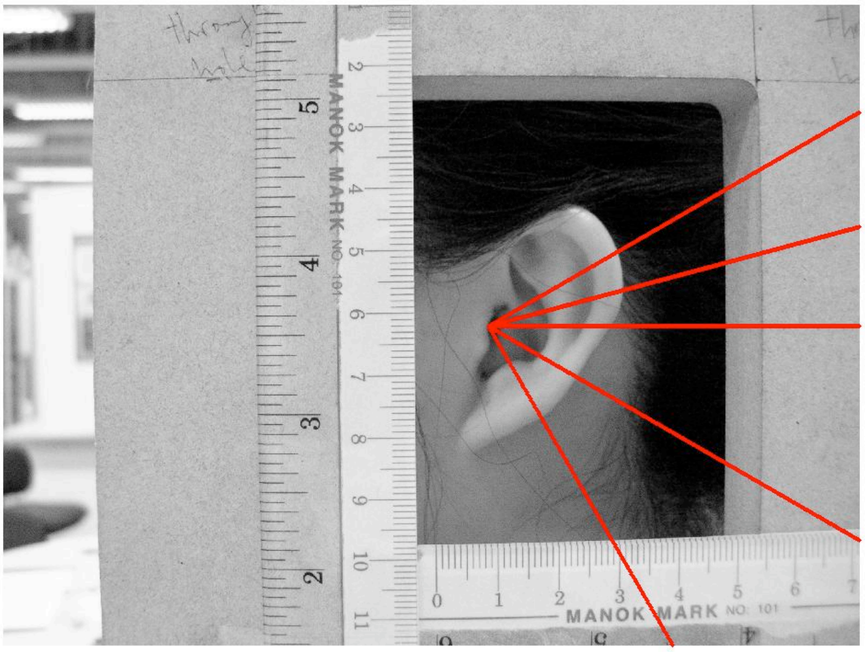
Subject: 12   Ear: L



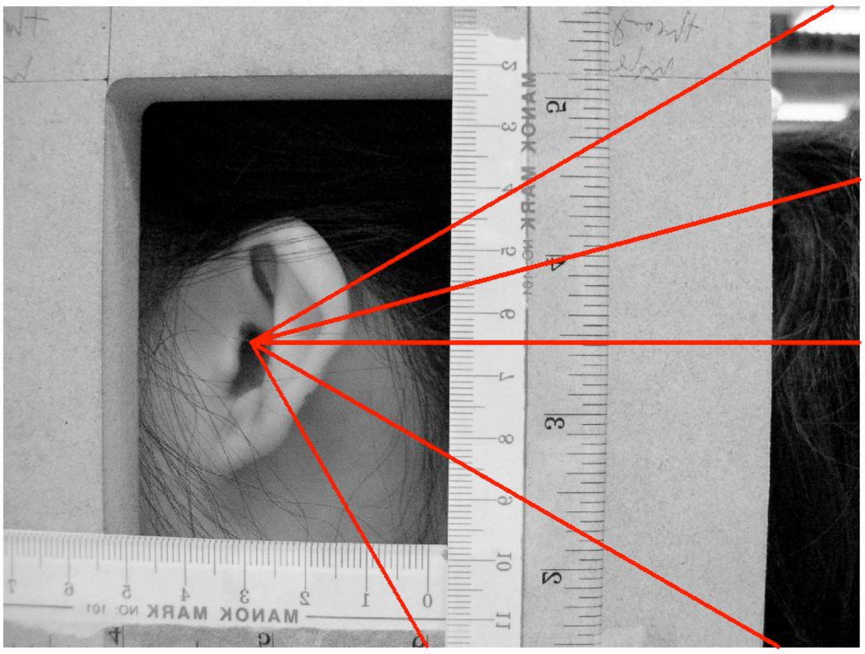
Subject: 12   Ear: R



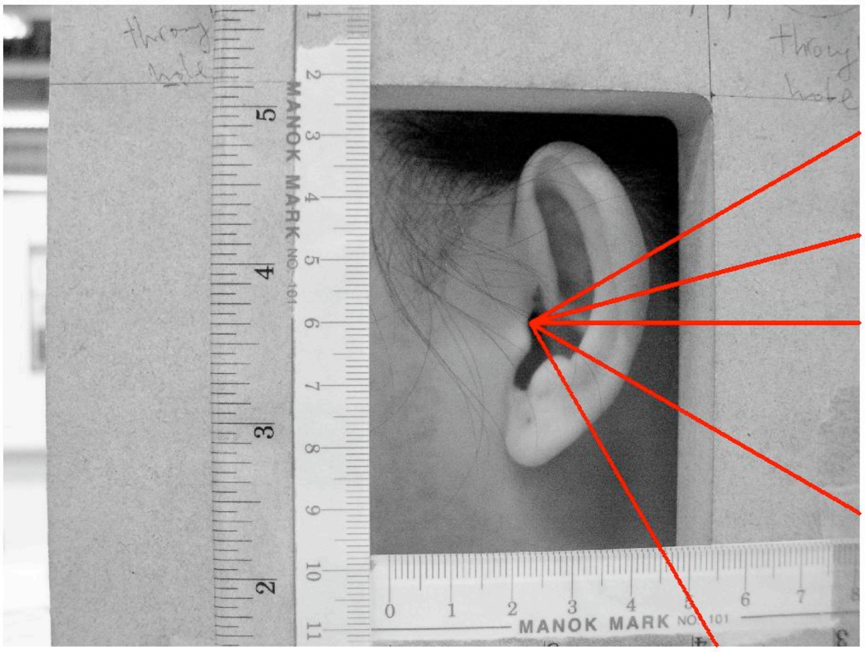
Subject: 13   Ear: L



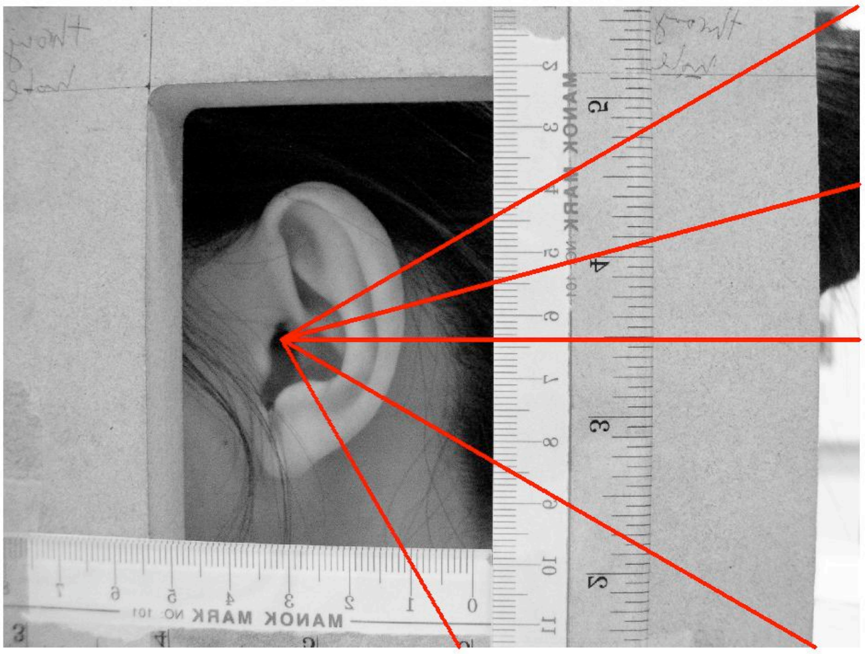
Subject: 13   Ear: R



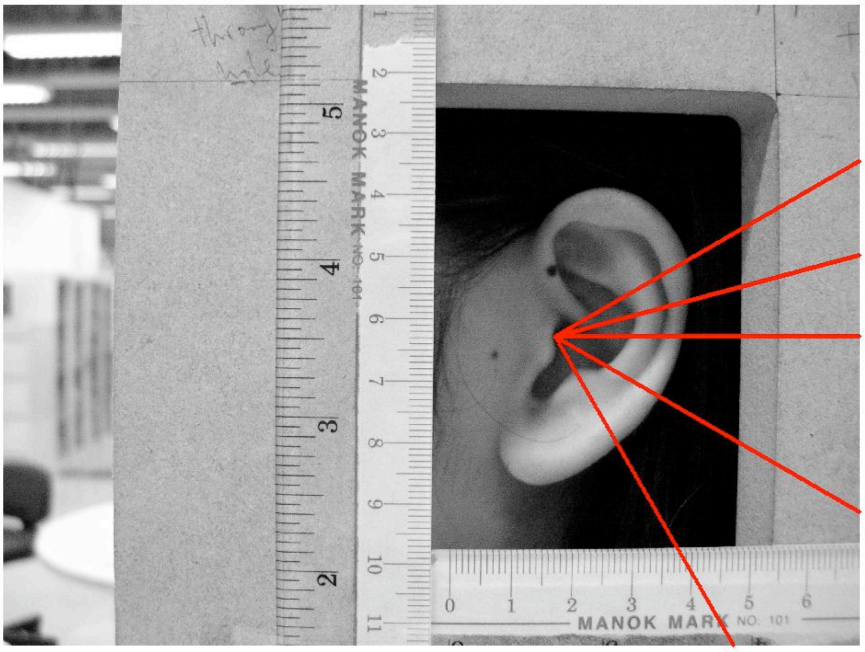
Subject: 14 Ear: L



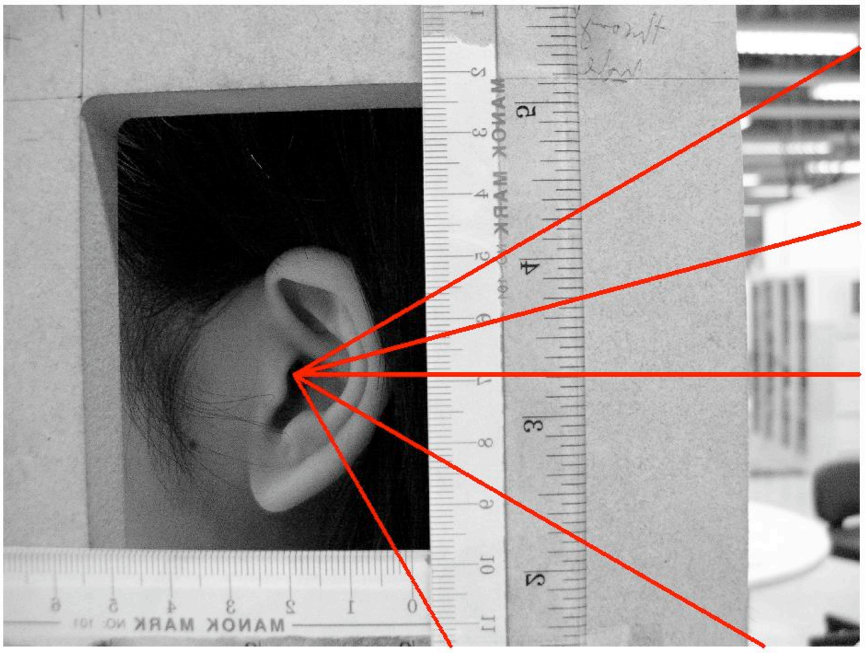
Subject: 14 Ear: R



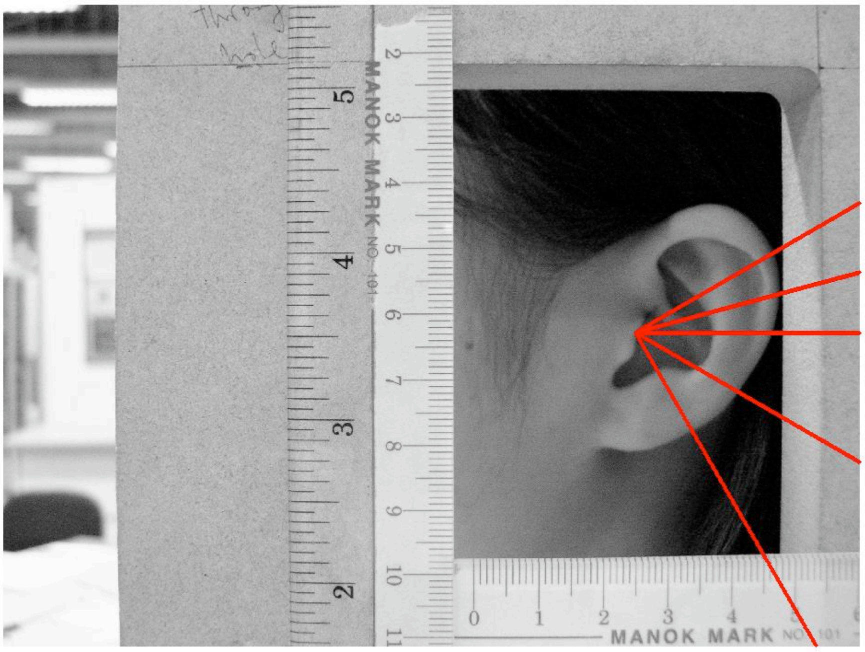
Subject: 15 Ear: L



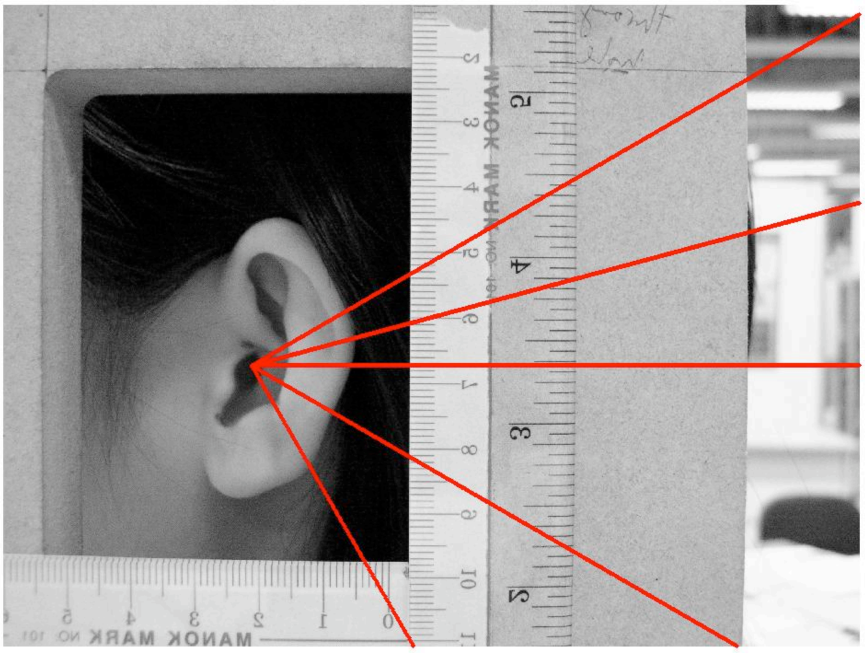
Subject: 15 Ear: R



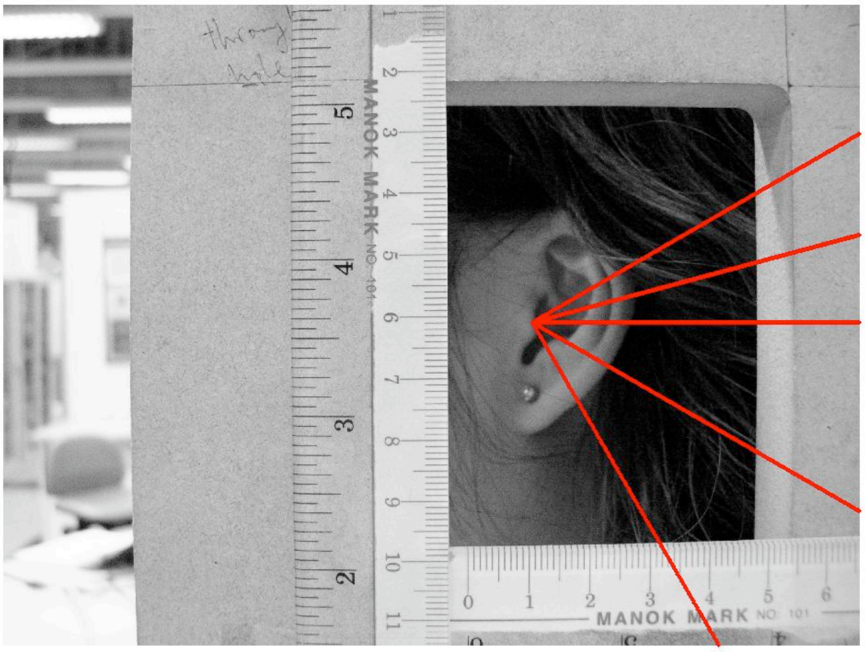
Subject: 16   Ear: L



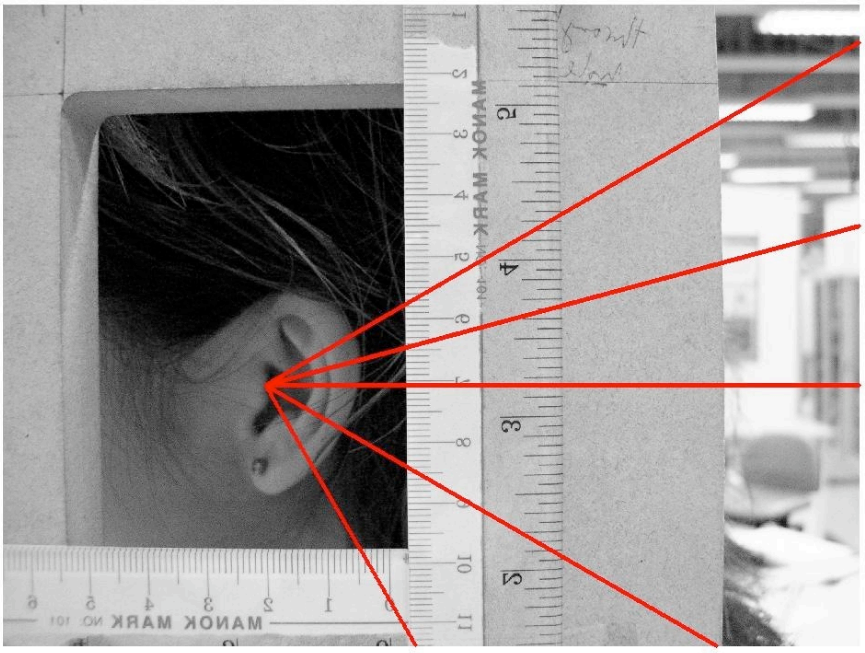
Subject: 16   Ear: R



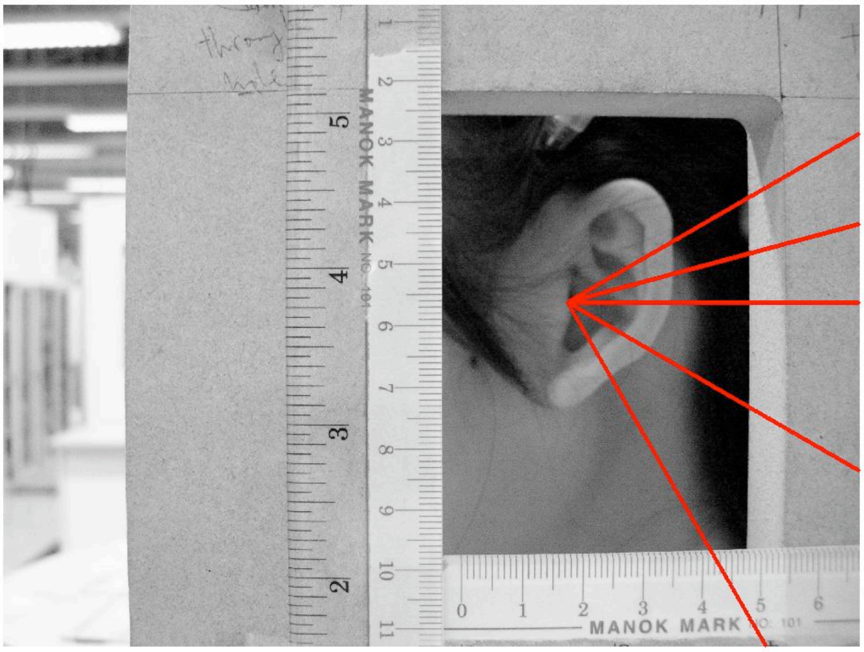
Subject: 17   Ear: L



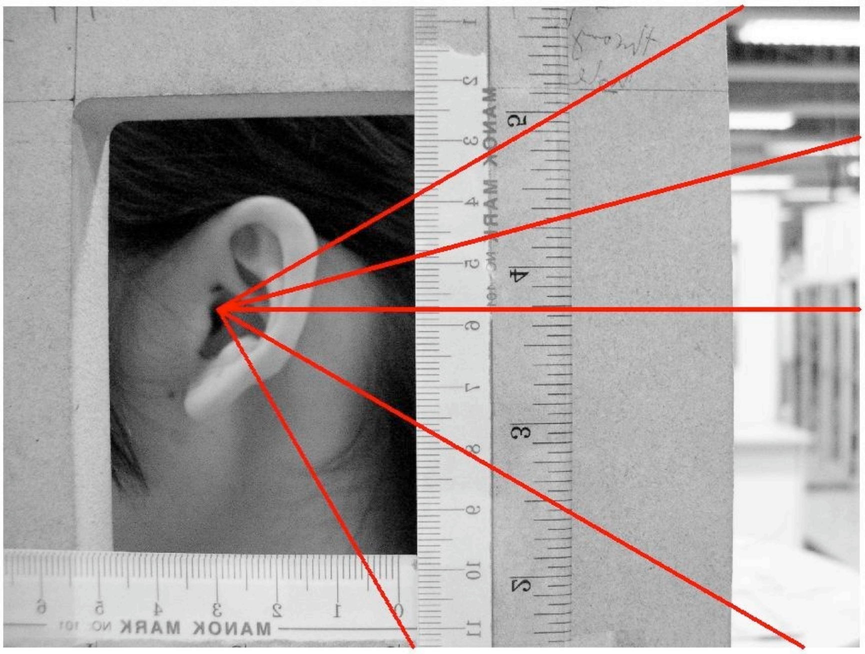
Subject: 17   Ear: R



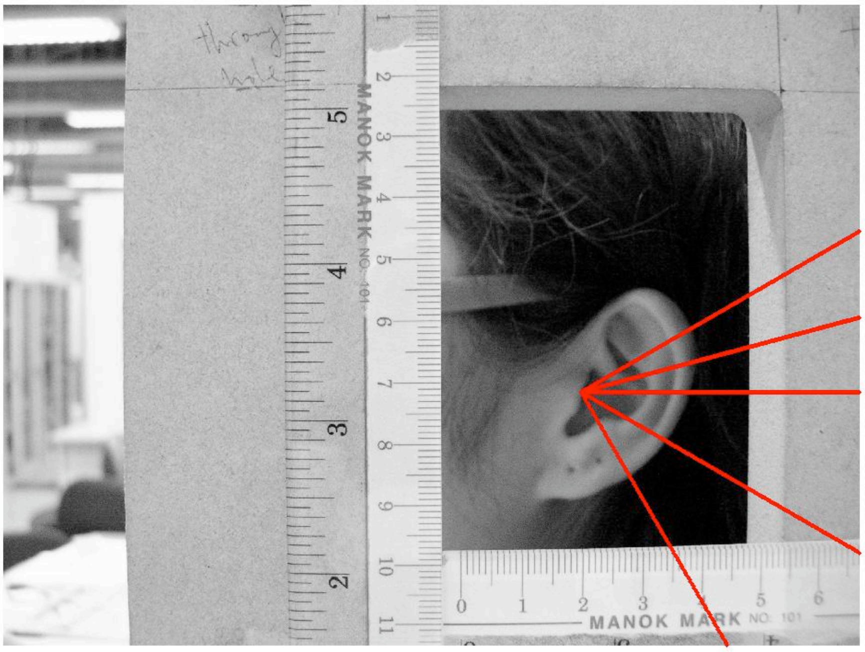
Subject: 18 Ear: L



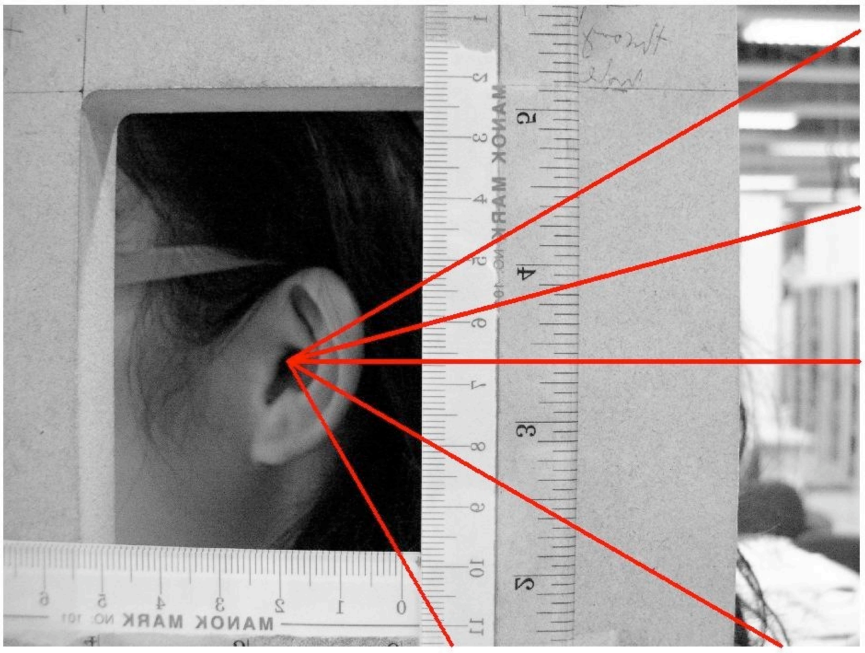
Subject: 18 Ear: R



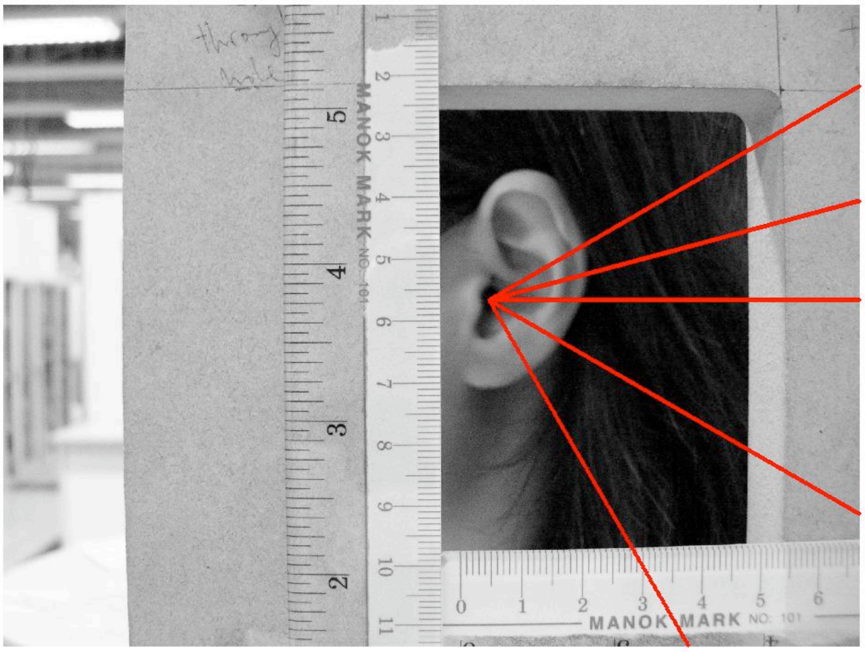
Subject: 19 Ear: L



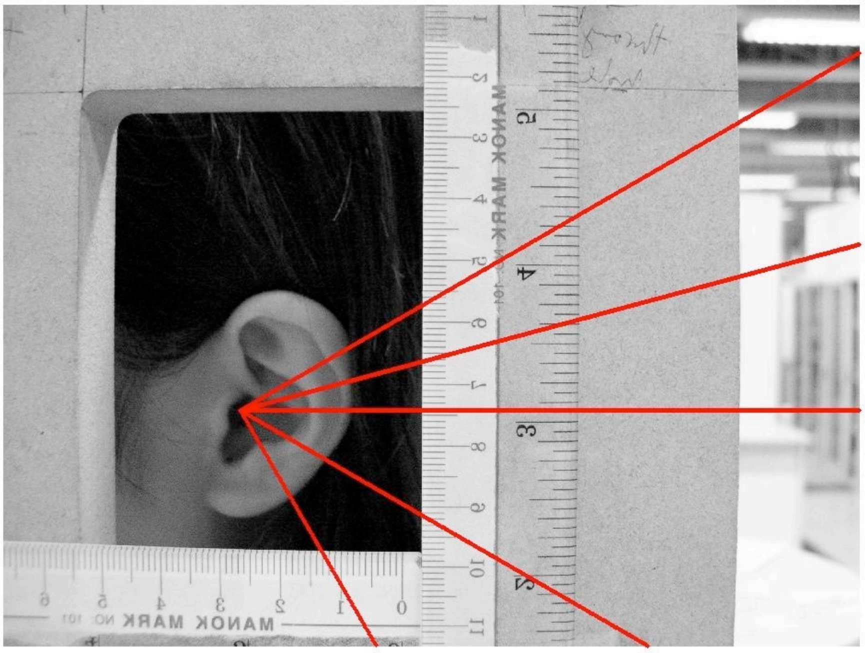
Subject: 19 Ear: R



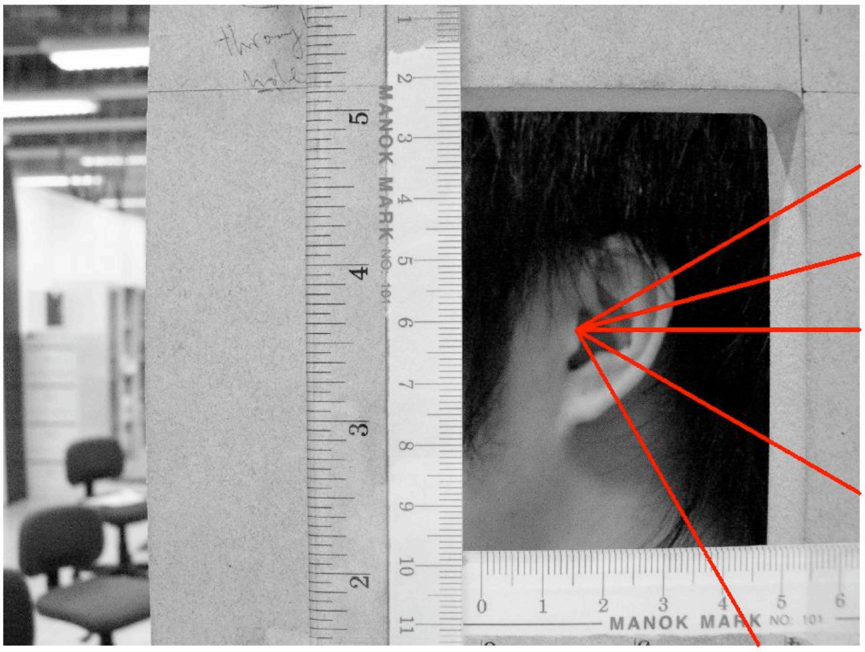
Subject: 20   Ear: L



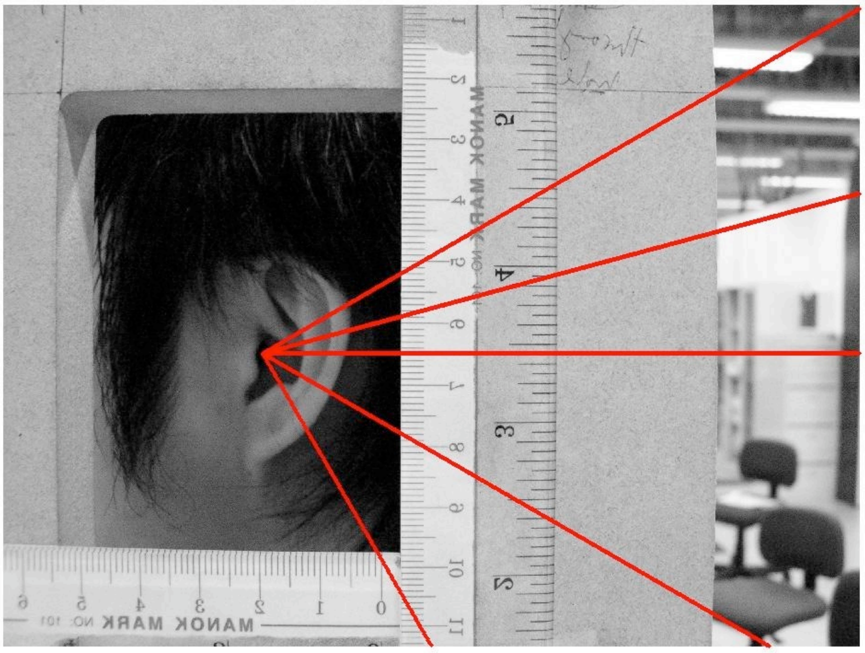
Subject: 20   Ear: R



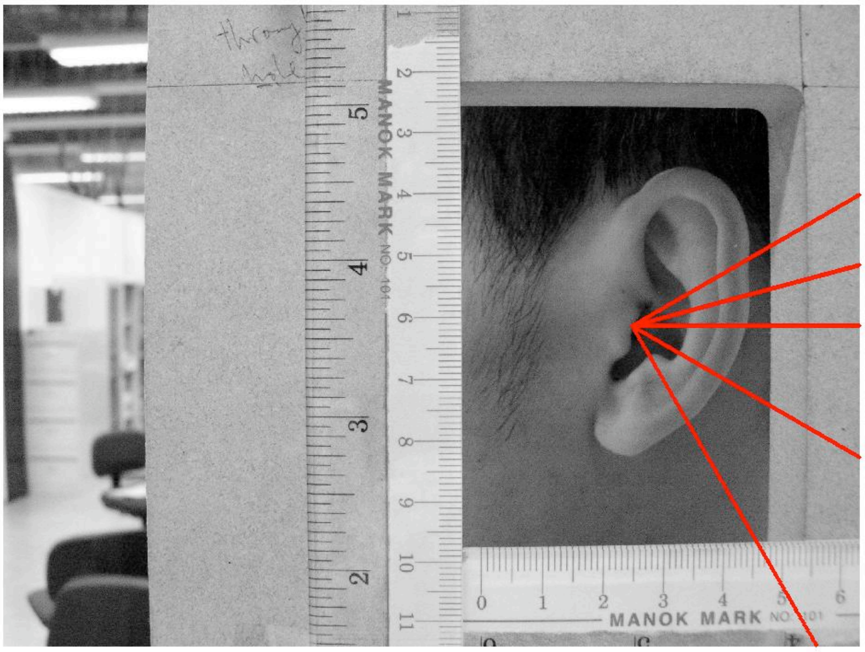
Subject: 21 Ear: L



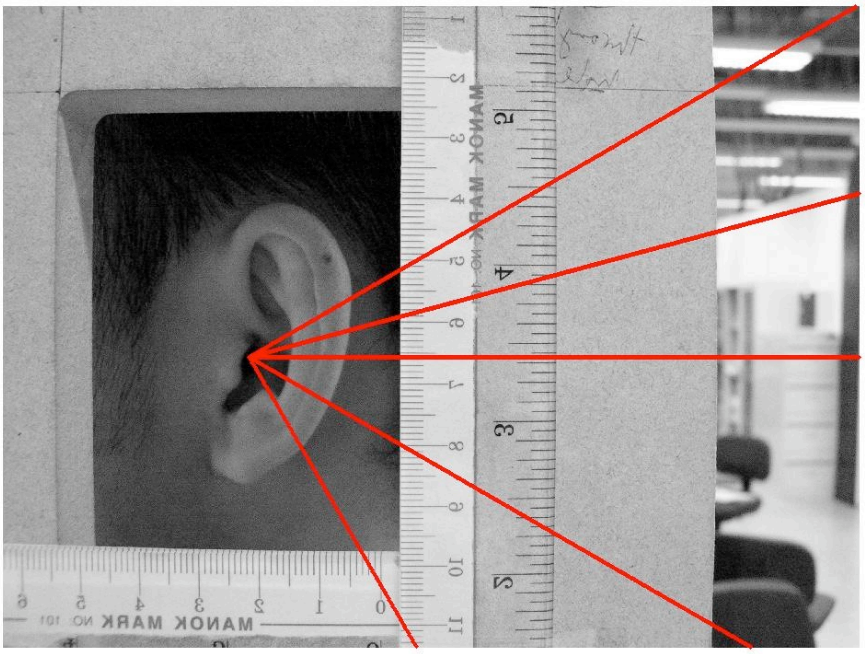
Subject: 21 Ear: R



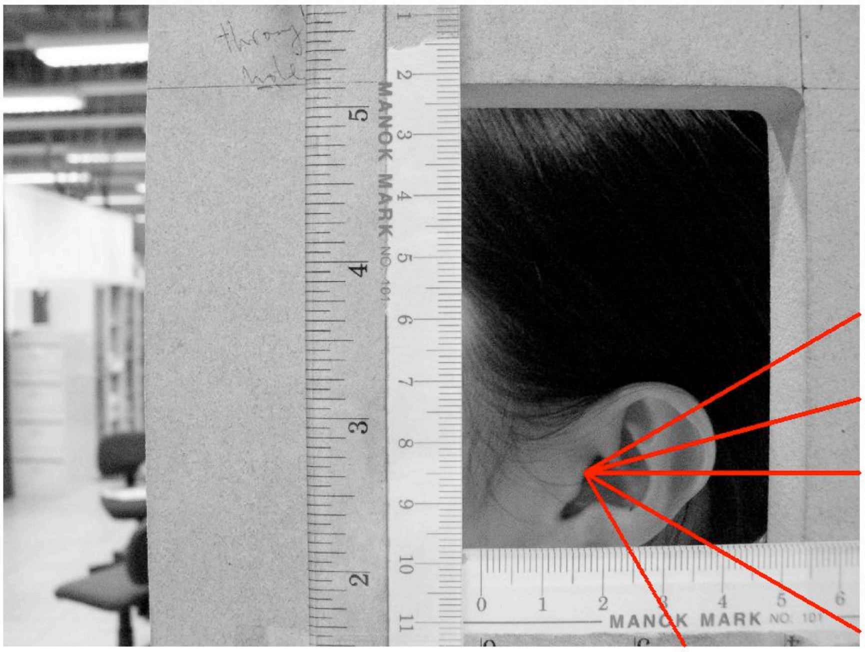
Subject: 22 Ear: L



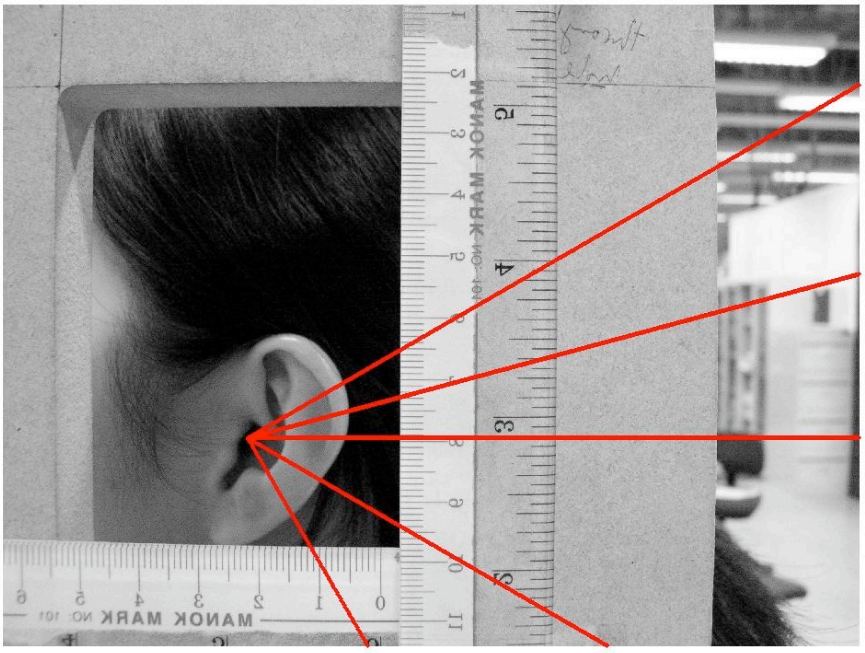
Subject: 22 Ear: R



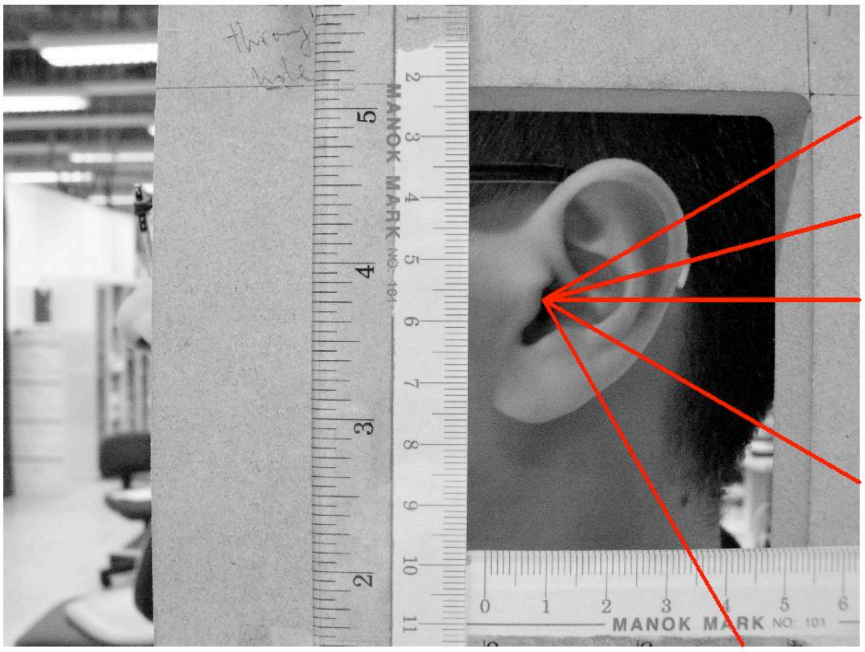
Subject: 23 Ear: L



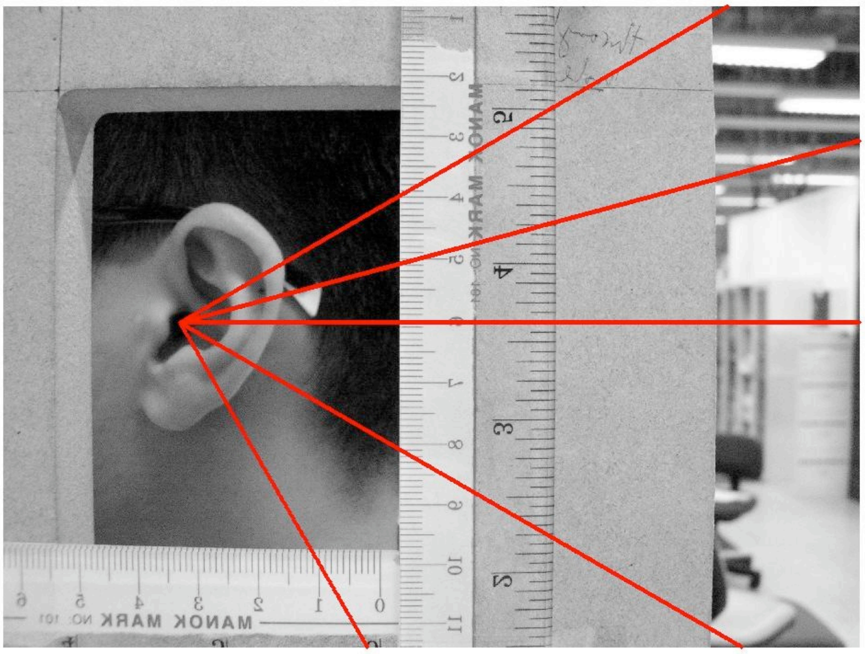
Subject: 23 Ear: R



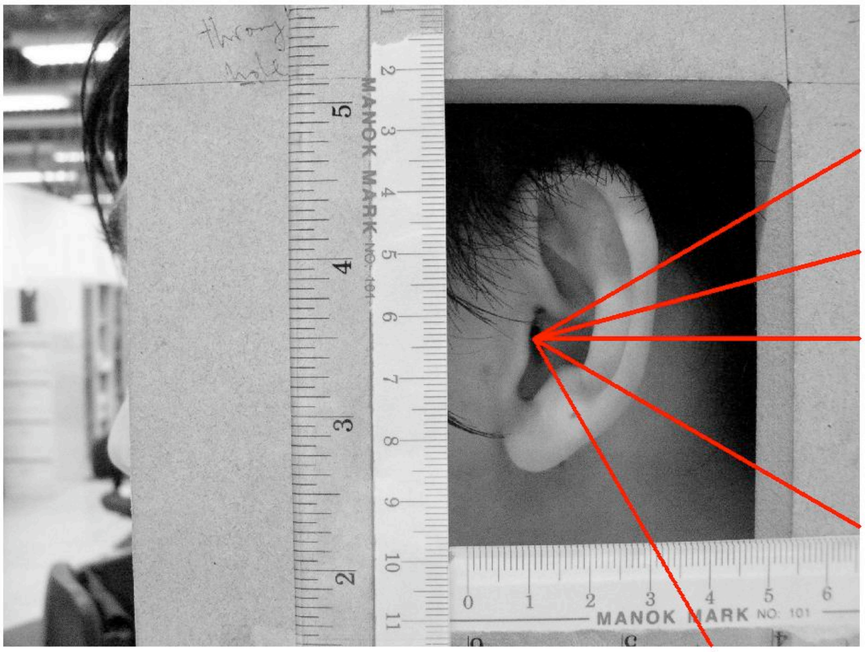
Subject: 24 Ear: L



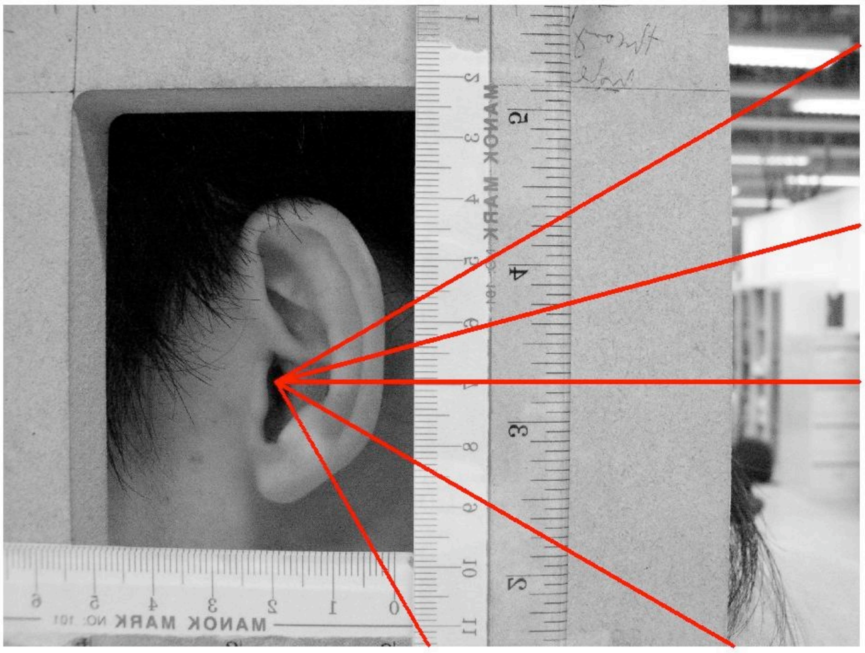
Subject: 24 Ear: R



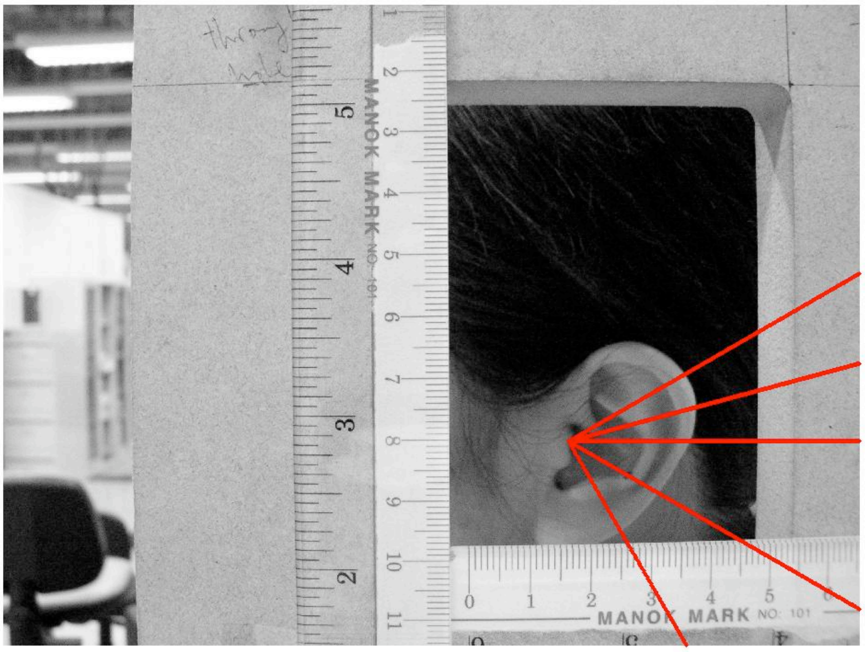
Subject: 25   Ear: L



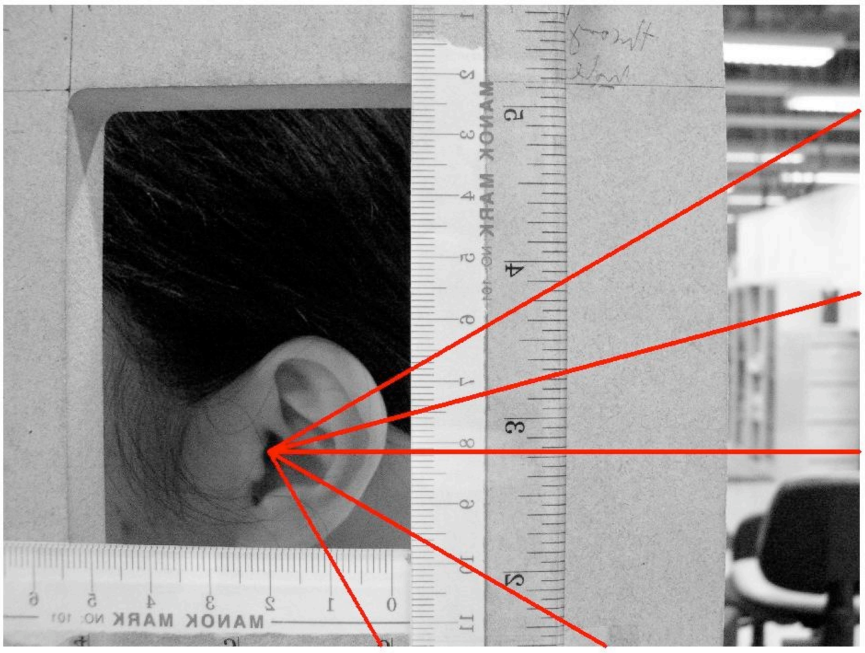
Subject: 25   Ear: R



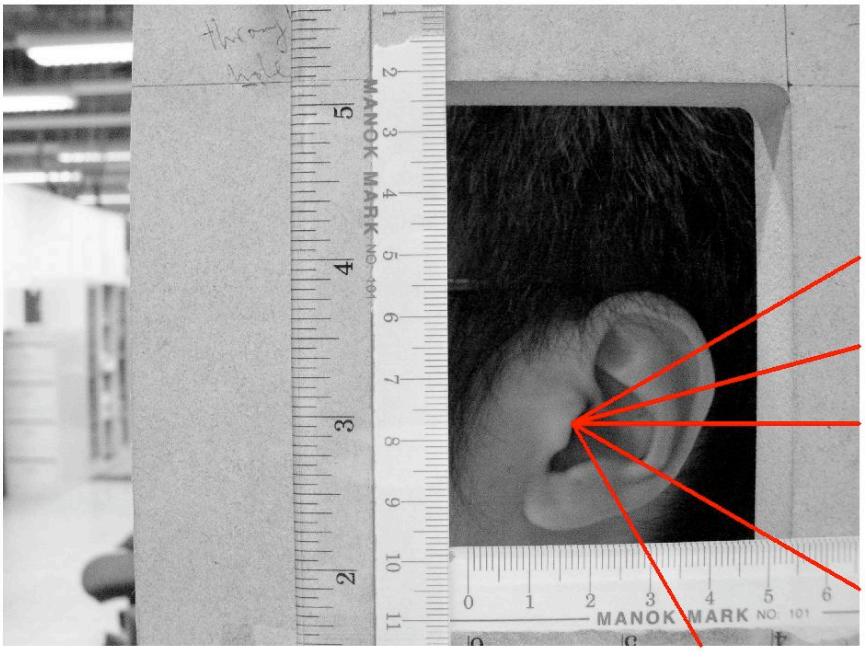
Subject: 26 Ear: L



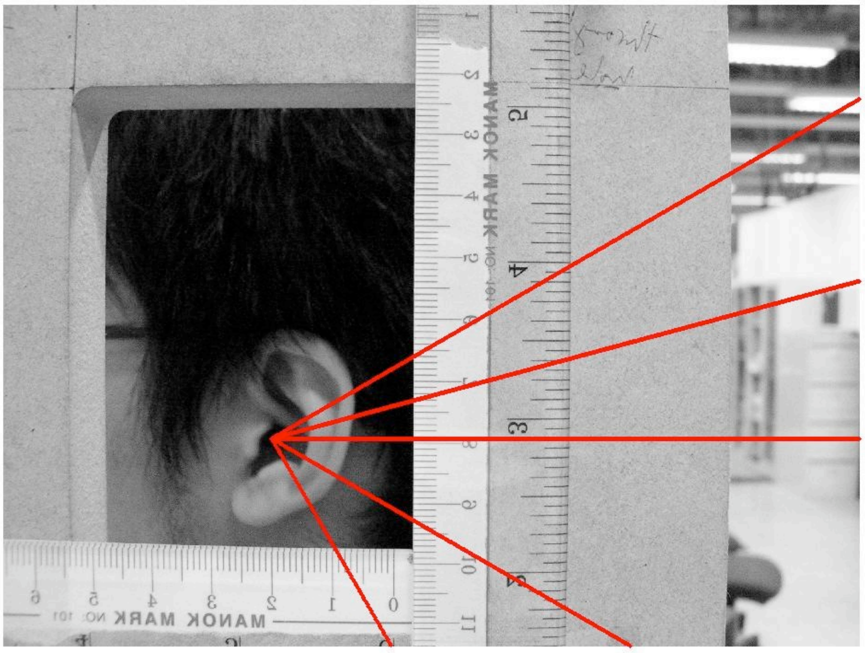
Subject: 26 Ear: R



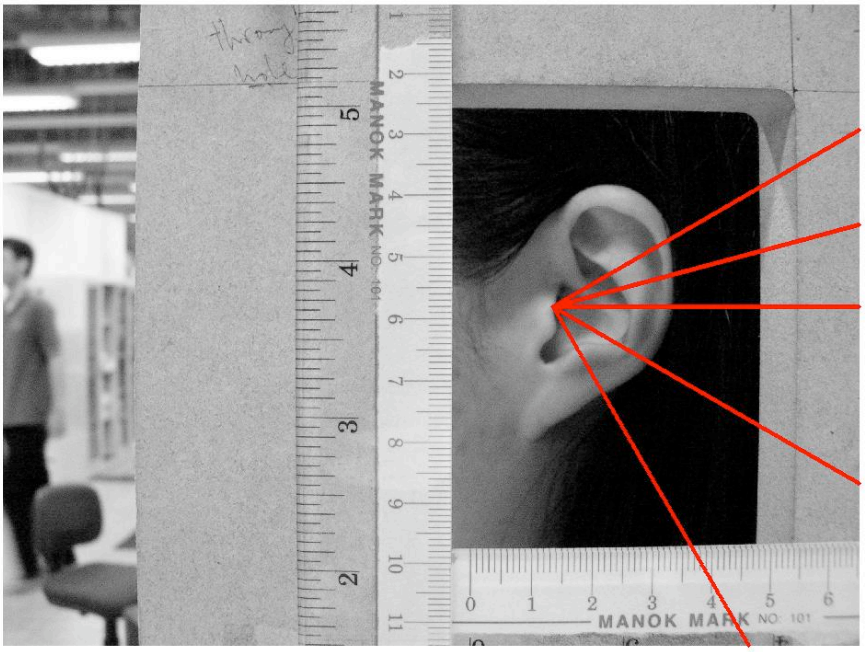
Subject: 27   Ear: L



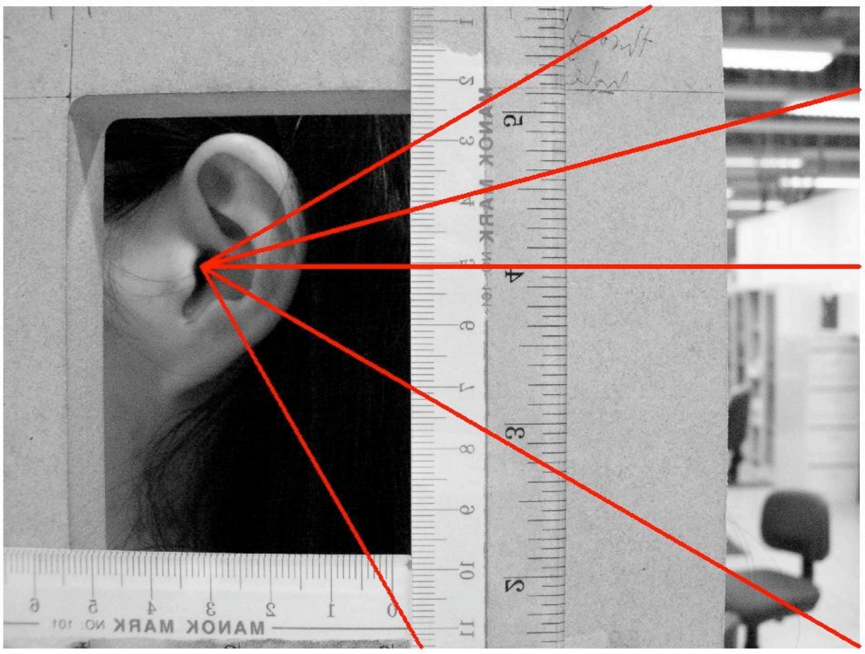
Subject: 27   Ear: R



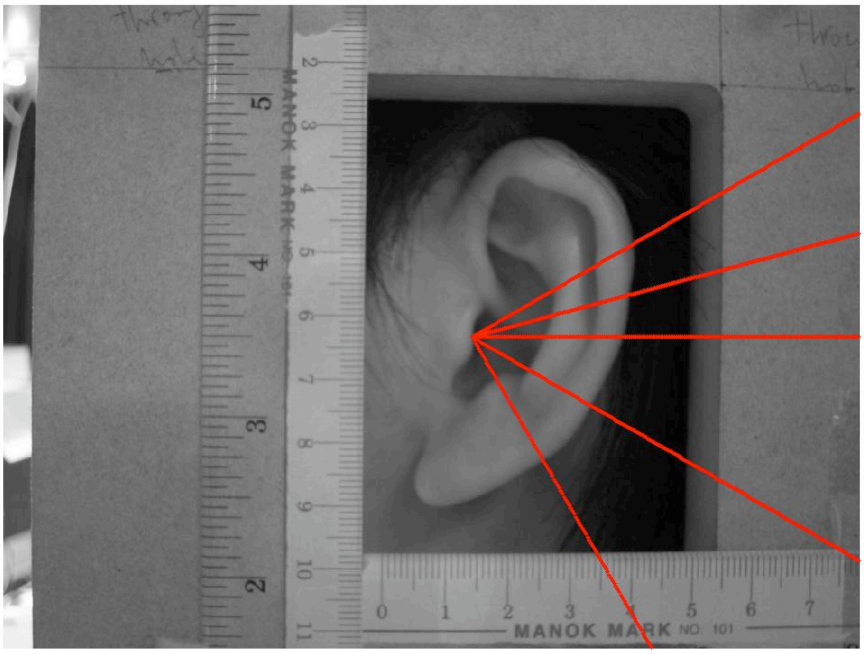
Subject: 28 Ear: L



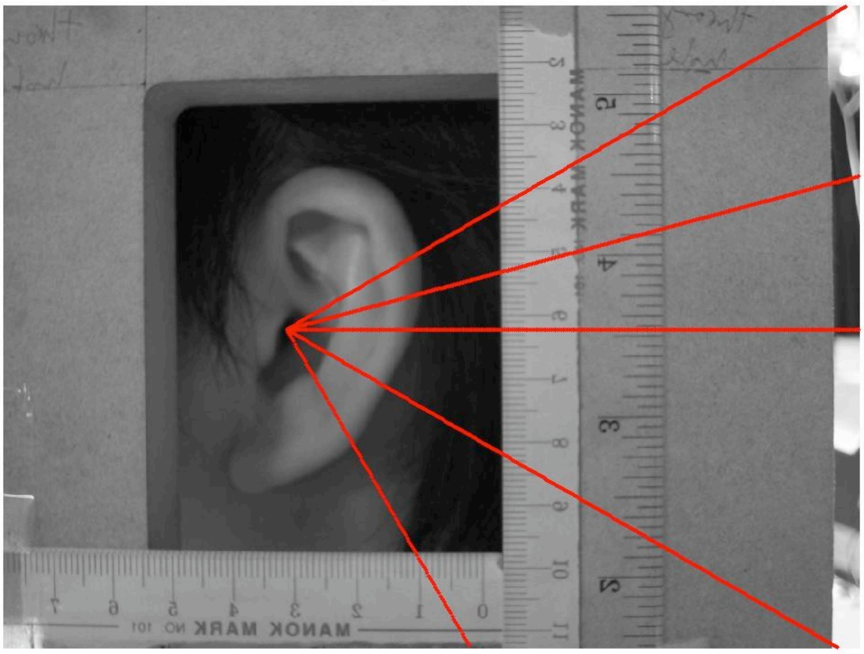
Subject: 28 Ear: R



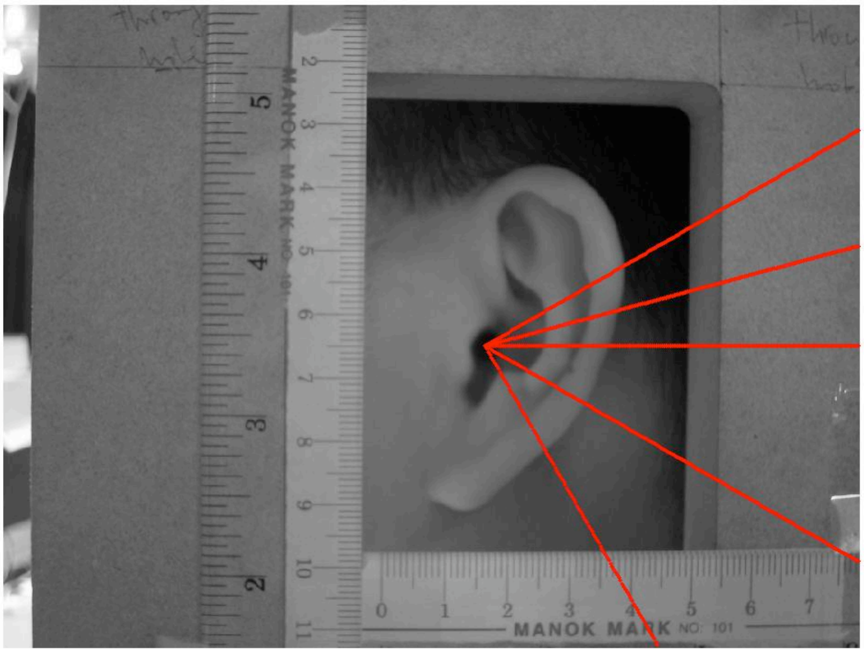
Subject: 29   Ear: L



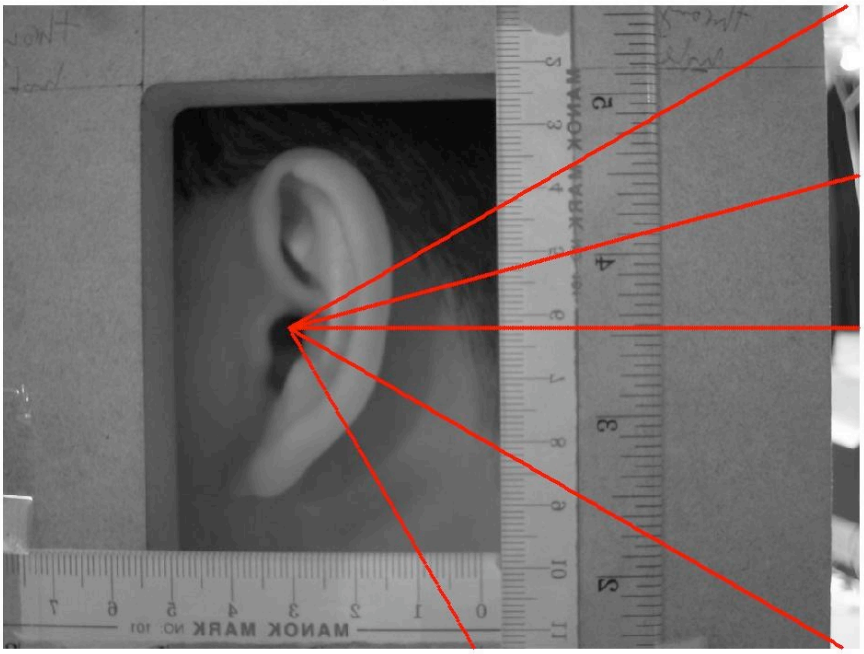
Subject: 29   Ear: R



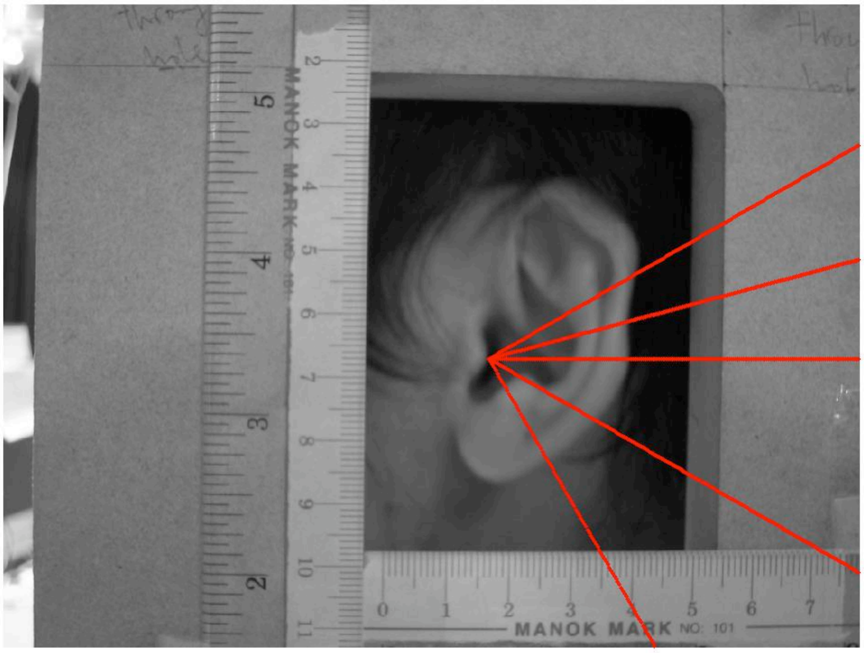
Subject: 30   Ear: L



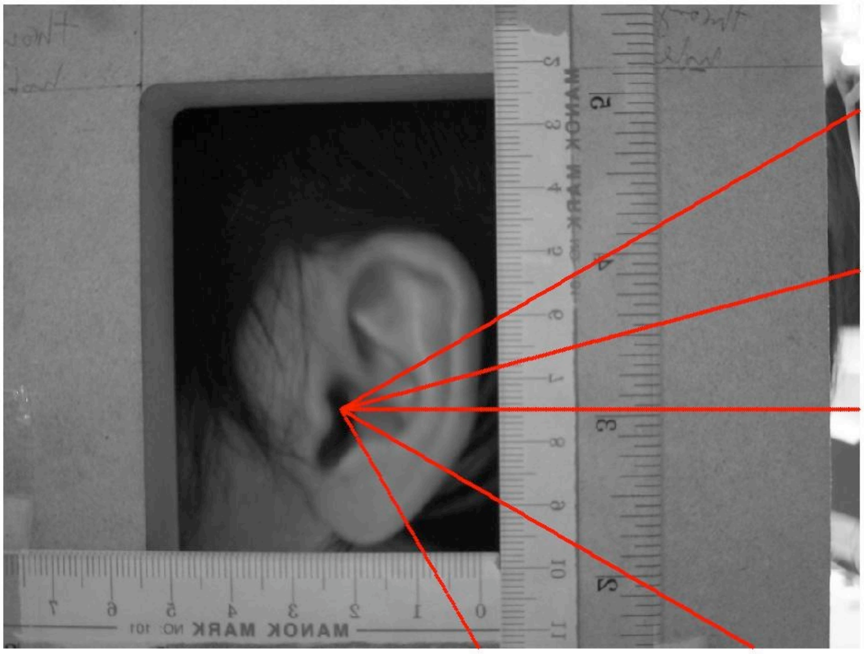
Subject: 30   Ear: R



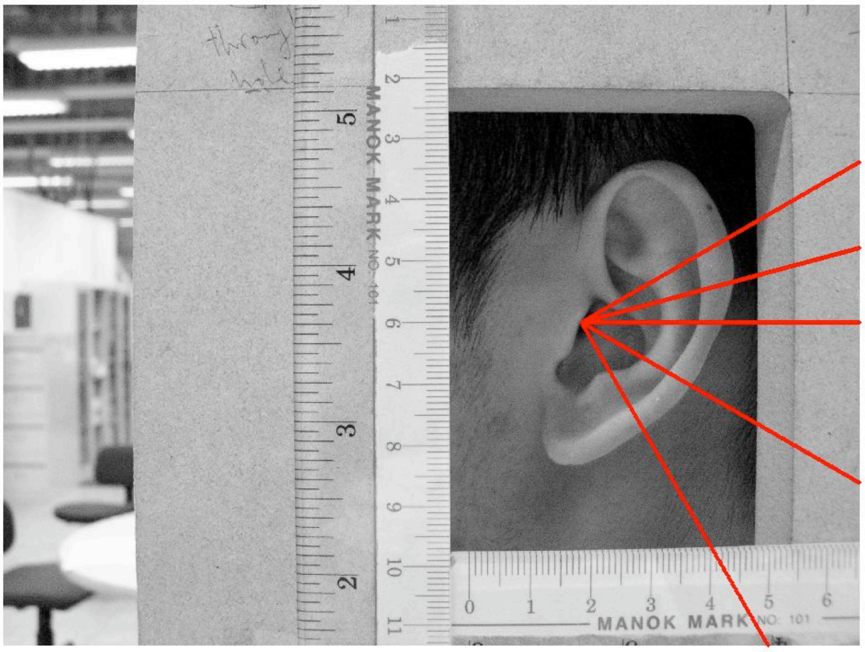
Subject: 31   Ear: L



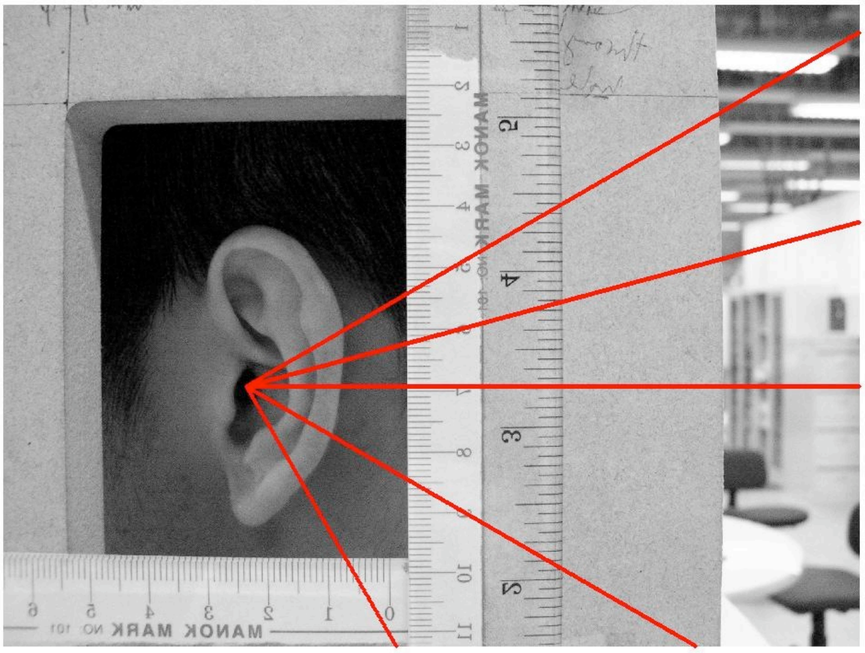
Subject: 31   Ear: R



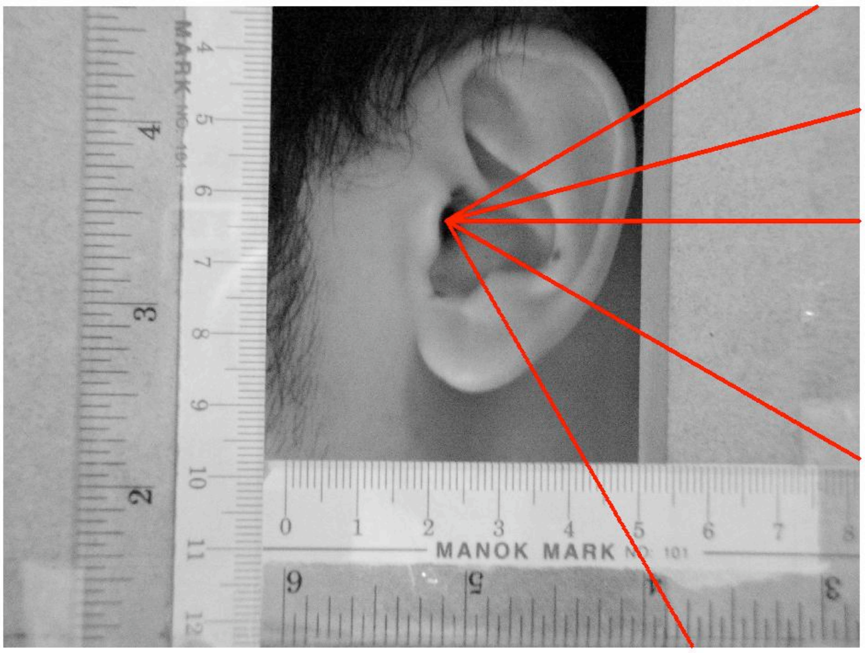
Subject: 32 Ear: L



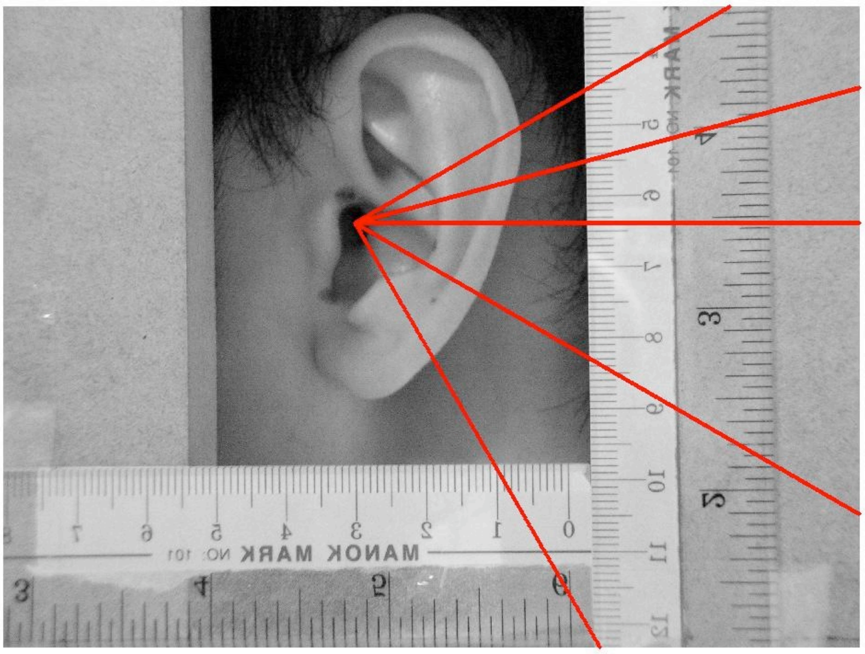
Subject: 32 Ear: R



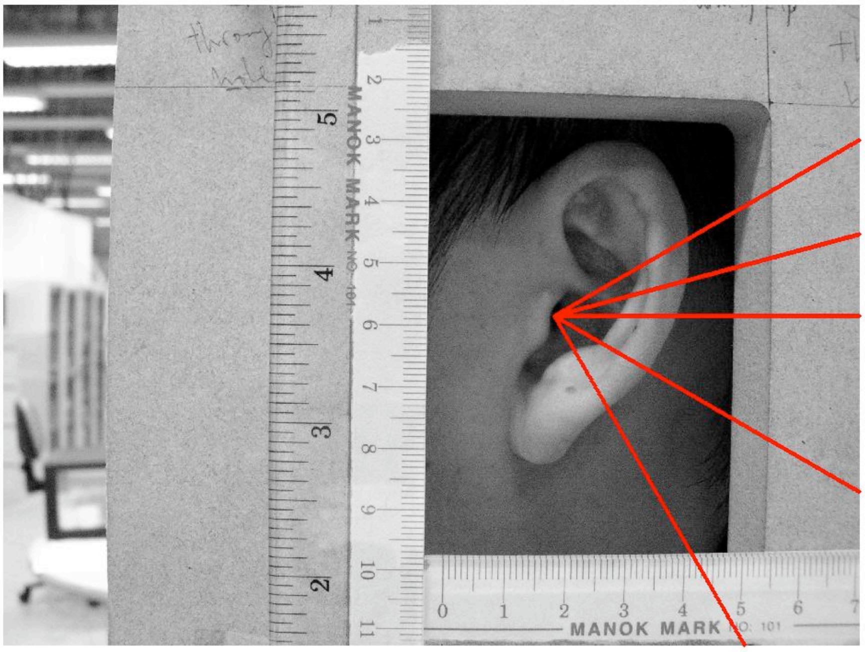
Subject: 33   Ear: L



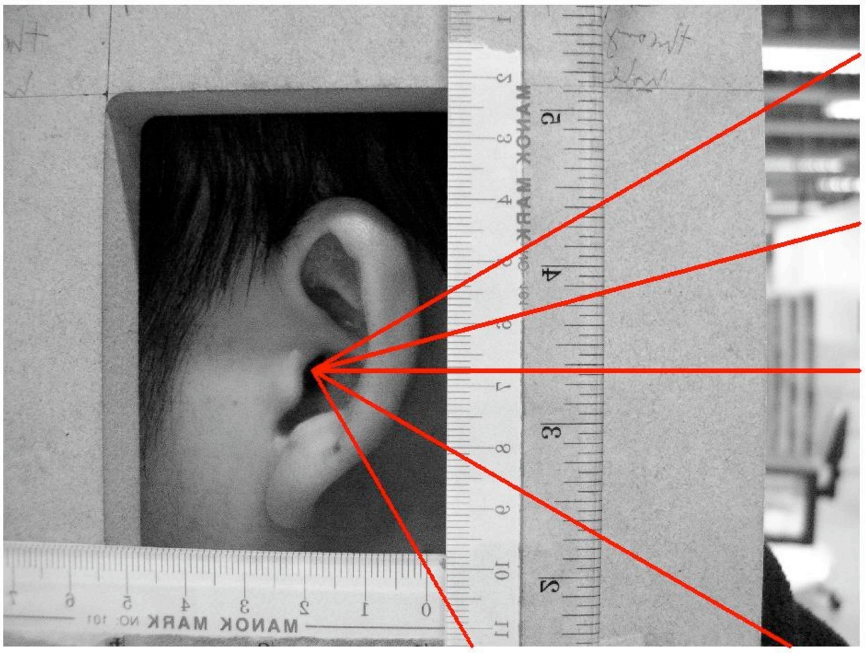
Subject: 33   Ear: R



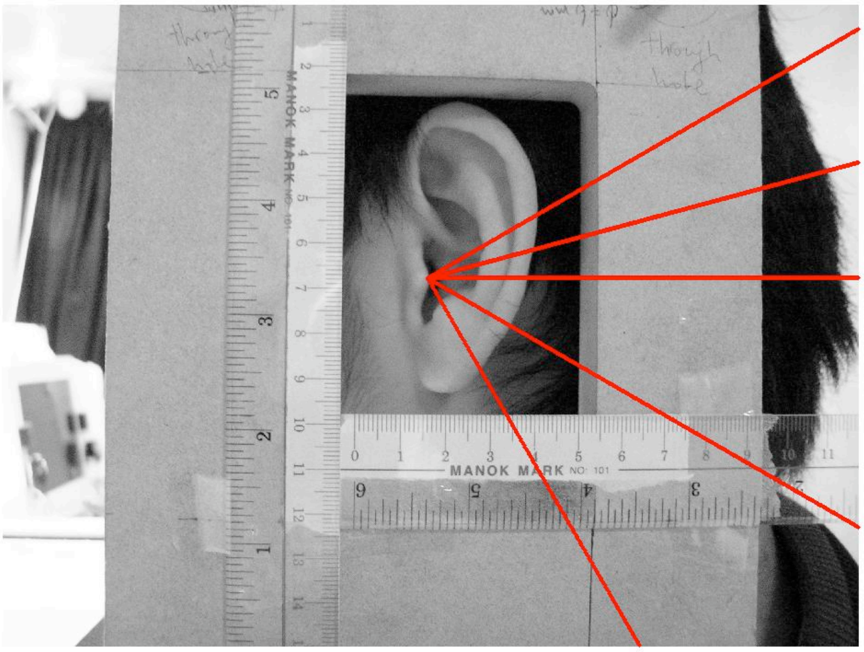
Subject: 34 Ear: L



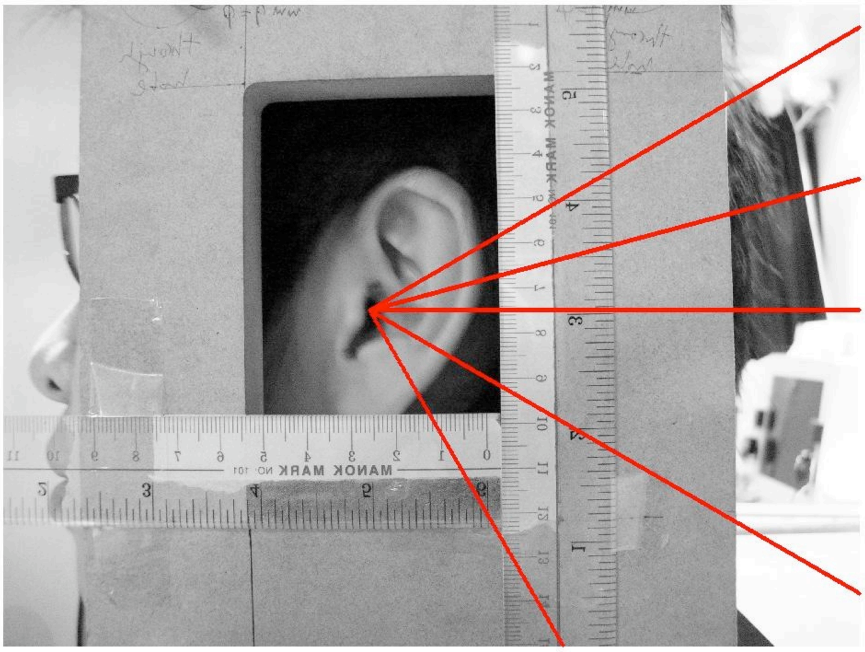
Subject: 34 Ear: R



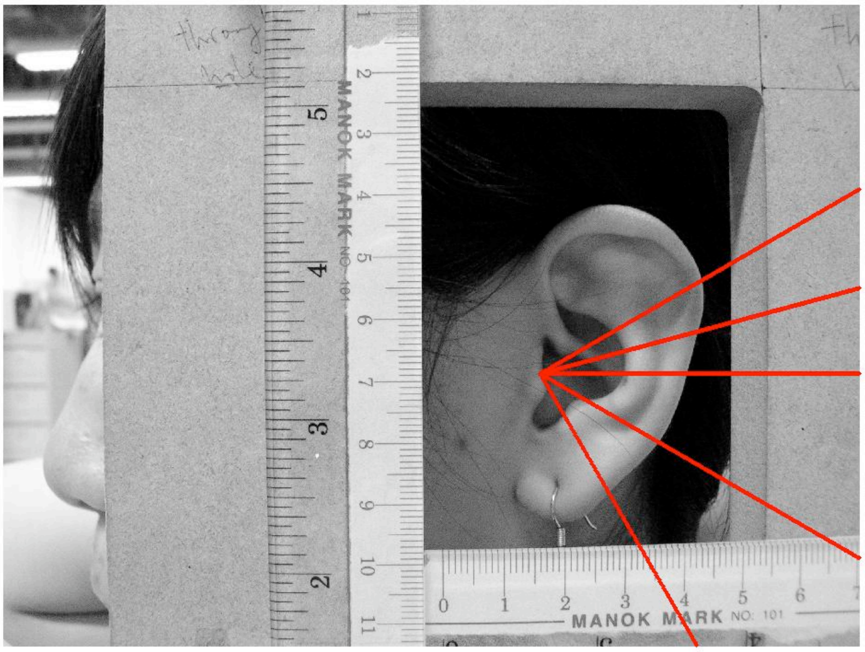
Subject: 35 Ear: L



Subject: 35 Ear: R



Subject: 36 Ear: L



Subject: 36 Ear: R