Listener Training and Repeatability for Automobiles

4660 (P14-2)

Roger E. Shively
Harman Motive Inc. Martinsville USA

William N. House
Harman Motive Inc. Martinsville USA

Presented at the 104th Convention 1998 May 16-19 Amsterdam

AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

Listener Training and Repeatability for Automobiles

Roger E. Shively
William N. House

Harman-Motive, Inc.
1201 S. Ohio Street
Martinsville, IN 46151
USA

The benefits of trained listeners has been well established for listening in rooms and at computer workstations [1],[2],[3]. Recent studies have illustrated the use of a self-administered PC-based training program to improve listeners' ability to reliably identify and rate different types of spectral peaks and dips which have been added to a variety of programs (i.e., resonance detection)[4],[5]. Similar investigations have indicated that critical listening in automobiles might require a more detailed training regimen and additional repeats within a trial [6]. A modified training program was used to train and evaluate subjects for automobile listening tests. The listener's repeatibility was then studied when using a PC-based preference testing program and when listening in-situ. Criteria were established for determining listeners' accuracy in resonance detection, their proper application of that ability in a preference testing program, and for the number of listening repeats required for statistically significant ratings.

0 INTRODUCTION

The purpose of any listener training is to verify a listener's ability to make consistent and accurate evaluations of a sound system or to create a listener who has that ability. Such a trained listener should be able to provide results that are statistically meaningful and can be used to benchmark similar sound systems. The best approaches to listener training attempt the provide the listener with a method for identifying and quantifying the quality of sound they are hearing and provide meaningful feedback to a system designer or end user [1], [2], [3]. Some approaches fall short by merely giving the listener a descriptive method for evaluating sound and no decisive way to quantify that evaluation. These short-comings can lead to inconsistencies and high variability among similarly trained subjects, which in turn can lead to a lack of statistical significance and a lack of any useful way to inform the system designer or end user [7].

1 LISTENER TRAINING

1.1 Overview

The method of using a self-administered PC-based training program that requires the listener to identify spectral peaks and dips in a set of source material reinforces the listeners' ability to relate a perceived spectral aberration to a common frequency scale. However, experience has shown that because a subject consistently scores high on such a resonance detection test, it does not mean that they are well trained for listening evaluation purposes. Training sessions in the use of the preference ranking and timbral balance scales are also necessary [6]. The objective is to verify that the listener is applying the resonance detection training properly in the context of a listening evaluation or experiment.

1.2 Use of EarTrain

The PC-based resonance detection software that was used to begin the training is a homegrown program called EarTrain. The interface for this software is illustrated in Figure 1. During each round of training, there are 24 trials, in each of which the listener sees 4 different equalization curves on the screen. There are 6 buttons there, labeled A, B, C, D, Flat, and Done. Each of the first five buttons plays a different sound file. These sound files are the same source with different equalizations. The Flat button plays the source with no equalization. The general idea is for the listener to match up the equalization curves with the appropriate sound files, using the Flat curve as a reference. After the listener has completed the 4 matches for the trial, the Done button is pushed, and the listener is shown how many of the matches are correct and shown the correct match for any mismatches. At this point the listener can either go back and listen again to the current set of EQ's to help better understand any mistakes, or the listener can move on to the next trial.

The sources are *.wav files played back through a digital audio card on the PC, processed through an external D-to-A Converter, and amplified through a headphone amp where the output is maintained at a constant level. The analog conversion is done outside of the PC to reduce the noise that still exists on most PCs during *.wav file playback. The playback is listened to using Etymotic ER4s earphones, which, with a good seal in the ear canal, provide extremely linear full-band reproduction, at least 15dB of external sound isolation, and very high repeatability from one session to the next. There are three sources that are used during the training. The sources that are selected for use are characterized as having a broadband frequency response when averaged over the length of their use, which is approximately 20 to 30 seconds. The sources that were used for this training were (1) Pink Noise, (2) "Rubberband Man" by Yello, and (3) "Into the Night" by Little Feat. The broadband nature of these sources made the spectral aberrations that were added to them reasonably apparent to the listener.

The first level of the aberrations was either a 6dB peak or 6dB dip (2.5 octave) placed at 125Hz, 500Hz, 2kHz, or 8kHz. The criterion of 95% correct was set as a target for the subject. There were a total of 8 listeners that took part in the entire set of experiments. Their hearing was tested to be normal. Each listener was asked to complete one round (set of 3 sources repeated 4 times for a total of 24 trials) a day until they met the criterion and were able to repeat that performance at least once. Figure 3(a) shows the charted progress from one round to the next for most of the listeners. The majority of the listeners met the criterion of 95% correct within 3 to 5 rounds. The exception was Listener 4 who peaked at no better than 90% and rested on a plateau of ≈85% even after 13 rounds.

After the subjects achieved the target 95% correct for 6dB aberrations, the aberrations were replaced with 3dB (2.5 octave) peaks and dips to further refine the listeners' resonance detection ability. Listener 4 was also moved on, though not having met the criterion, in order to test the validity of the criterion. The results for the 3dB testing [Figure 3(b)] indicate that the starting points for %Correct were higher in general and that the majority of the listeners achieved 95% correct in 3 to 4 rounds. In this case Listener 4 achieved no better than an 85% correct. Also noteworthy is the performance of Listener 5. In both the 6dB and 3dB tests Listener 5, after having achieved 95% correct, had a round or more where the %Correct levels drop 5 - 10 points in what appears to be a lack of concentration on the part of the listener. With additional rounds, however, Listener 5 recovers and meets the criteria.

When a subject has reached a level of 95% accuracy for 3dB aberrations, their ability to rate overall preference and timbral balance has been assumed to be very high. But, as we will see, that is not always the case. All the listeners from the EarTrain sessions were moved on to the following sessions.

1.3 Use of PrefTest

For the next step of listener training another homegrown software program called PrefTest is used. The user interface for this self-administered PC program is illustrated in Figure 2. This is the same software that is used in collecting preference information from our listeners in listening experiments and automotive sound comparisons. In the role of a training tool, it is (1) used to determine agreement among the listeners for overall preference, and (2) used to evaluate the timbral balance results to determine if the listeners are correctly associating timbral changes with the appropriate frequency ranges. The same sound sources that were used in the resonance detection are used in a preference testing software. Four buttons are displayed to the user. These buttons play the same sources but with a different equalization. One of the 4 buttons randomly has a flat response assigned to it for each trial. This is used as a blind reference. There are again 24 trials total. For each equalization, the listener is forced to give a separate rating number for overall preference and for the timbral balance of treble, midrange, and bass.

The scale for the preference is 0 to 10, 1/10th pt increments. The listeners are instructed that 0.5 point is a slight preference, 1.0 is a moderate preference, and 2.0 is a strong preference. The scale for the timbral balances is +5 to –5, 1 pt increments, 0 being neutral.

The listeners repeated the PrefTest training until they demonstrated a consistent behavior. If they seemed to be having trouble with any of the concepts associated with PrefTest the administrator of the tests would attempt to help them understand better. In most cases proper use of the scales and ranking occurred after 2 rounds each of 6dB and 3dB aberrations. In some cases, consistent but less than ideal use of the scales and ranking occurred no matter how many rounds of training occurred. In either case, the quality of the listener was obtainable in quantitative terms.

The ANOVA (multivariate repeated-measures analysis of variance) results indicate that, in general, the listeners are using the scales consistently, preferring the Flat reference or the 125 Hz peak, and identifying all the peaks and dips correctly. When comparing the 6dB Overall Preference [Figure 4(a)] to the 3dB Overall Preference [Figure 4(b)], there is a definite narrowing of preference for most of the 3dB aberrations, yet the variances are small enough that the differences are significant. The ANOVA results for the timbre balances also show a similar behavior [Figures 5, 6,7].

However looking at the Agreement Among Listeners for Program for the first two rounds of the experiment [Figure 8], it is seen that there is a decidedly different use of the preference scale for Listener 2 and Listener 5. As somewhat of a saving grace, Listener 5 does move closer to the norm in the second round. Yet, again, looking at the Agreement Among Listeners for EQ, it is seen in round 1 [Figure9] that Listener 2 and Listener 5 are not in agreement with the others in the use of the scale and the ranking of their preference. In the round 2 data of Agreement Among Listeners for EQ [Figure 10], there is a general reduction of variance for the listeners, showing a trend toward group consensus (especially on the Flat reference) except for Listener 2 and Listener 5. They do have lower variance levels for low- to mid- bass aberrations, but they still don't agree with the norm on the Flat reference and show larger variances for the mid to high frequency aberrations. If we were to look at their individual ANOVA results for each of the timbre categories of bass, midrange, and treble, we would see larger variance values there than for the other listeners. This type of listener behavior could pose a problem in the presence of more complex automotive sound system aberrations. The size of this listener variance and its effect will be pursued further in the last section of this paper [Section 3] where a sample of data from an experiment using some of the above listeners is discussed.

1.4 Discussion and Criteria

At the beginning of the training, the subjects completed a questionnaire, one question of which asked the subjects to rate their level of audio experience. The results of the resonance detection training indicate a correlation between the number of sessions of resonance detection training and the level of audio experience. Listener 2 and Listener 4 were subjects who rated their audio experience as low. For Listener 4, less experience meant more sessions were required. In the case of Listener 4, lower experience also meant the inability to obtain a 95% correct score. Listener 4, who was moved onto the preference training program without reaching 95% correct, but had at least 85%, did demonstrate an ability to accurately identify resonances. However, both "inexperienced" Listener 2 and "experienced" Listener 5 achieved the 95% correct score and later displayed signs of having difficulty in providing stable answers or difficulty in joining the consensus of the group. Listener 2 did begin to stabilize, but Listener 5 never did. After the training and experiments were over, the questionnaire was completed again. Listener 2 and Listener 4 had better opinions of themselves but did not think they were expert listeners. Listener 5 never did change its opinion of itself.

The end result from the training would be that lowering the criterion from 95% correct to 85% correct should be allowed if required. However, the drawback of using the lower criterion could be that it will take the subjects longer to master the preference testing. The ultimate watch point in this training method is the application of the resonance detection to actual preference and aberration ranking. A group average variance of 0.5 point on a 10 point scale seems to be a valid acceptance criterion in the ANOVA of PrefTest results. But it is equally important to monitor the individual variance performance and how they relate to the group and to the blind Flat reference.

2 REPEATABILITY

2.1 Overview

Listeners trained by the above method were also used in a repeatability study. The purpose of the experiment was to determine the number of repeated listening evaluations that were necessary to obtain a variance in each of the sound system rating categories that was 0.5 point, or less, on a 10 point scale both on an individual and group mean basis. This was investigated for both binaural and in-situ vehicle listening evaluations.

Four separate cars with different types of sounds systems were evaluated in this study. The PrefTest software was used for both the binaural and in-situ listening tests. For the binaural listening, binaural recordings of each of the cars were made and edited into *.wav files and presented to the listener blindly as either A, B, C, or D. There was no reference for the binaural listening because there would be no reference for the in-situ listening. For the in-situ listening, the program was used on a laptop computer. The program randomly ordered the sources to be played, and the listeners entered their answers the same as they did for the binaural listening.

2.2 Results

The data was collected after each round of listening for both binaural and in-situ and analyzed in terms of variance from the previous round. The variance was analyzed on an individual and a group mean basis. The point past which no significant improvement in either the subject's or the group's variance was determined, and a criterion for the number of necessary repeats was established for each.

The data for the variance among listeners is shown in Figure 11. In Figure 11(a), listener variance for the binaural listening, it is seen that the criterion of 0.5 variance is obtainable with stability for some after 5 rounds, but others can only stabilize to 1.0 variance. There is one exception here and that is Listener 5, who tends to hover around 2.0 variance. In Figure 11(b), listener variance for the in-situ listening, it is seen that the criterion of 0.5 variance is readily obtainable with stability for most of the listeners in 5 rounds, with the exception again of Listener 5, who is still hovering around 2.0 variance.

In Figure12 we see the group mean variance of binaural and in-situ. Figure 12(a) is the average of all the listeners involved, Figure 12(b) is the average of all the listeners except Listener 5. From what we have learned about Listener 5, we are better off not including Listener 5 in the average. In Figure12(b) we can see that for in-situ listening a stable group mean variance of 0.5 can be achieved in 5 rounds of listening and that for binaural listening the same stable results require 7 rounds. If the binaural listening was to be run for only 5 rounds its group mean variance would be approximately 0.75.

Based on these results, the following can be expected:

For:

In-situ listening:	5 Rounds	Individual Var. = 0.5	Group Var. = 0.5
Binaural listening:	5 Rounds	Individual Var. = 1.0	Group Var. = 0.75
Binaural listening:	7 Rounds	Individual Var. = 0.75	Group Var. = 0.5

Remember that the listeners were instructed that 0.5 point was to be considered a *slight* preference. Given that: We should be able to use listeners that have been trained under the criteria associated with EarTrain and PrefTest to significantly determine anything from a slight preference to a strong preference of one vehicle over another with the minimum of 5 to 7 rounds of data, using either the binaural or in-situ listening method.

3 SPATIAL EXPERIMENT: Binaural vs. Live

3.1 Overview

As an illustration of the effectiveness of using properly trained listeners for a listening evaluation, some results of an experiment that compared the use of binaural recordings and in-situ ("live") evaluations of various sound fields in a car are presented here. This Binaural vs. Live experiment investigated the ability of both methods to discern two types of aberrations. The first type were spectral aberrations, wherein the listener listened to the same car and sound system with three different equalizations applied to it. The results of that experiment support earlier work that illustrates the benefit of using binaural recordings for evaluating spectral differences in a sound field [8]. The second type of aberrations that were used for the investigation were spatial aberrations, wherein the listener listened to the same car with speakers in four different locations but with identical sound system frequency responses. The variability and the spectral characteristics of binaural versus live data for the spectral experiment are identical to those in the spatial experiment data. Therefore, for the sake of brevity, only the results from the spatial experiment will be presented here

PrefTest was used again for this experiment. In addition to the selections used in training (Preference, Treble, Midrange, and Bass), selections for Clarity/Definition, Distortion, Sound Stage Accuracy, and Definition of Images were used [Figure 2.]. The listeners were instructed as follows on each of the additional selections:

> Clarity/Definition [-5 less clear to +5 more clear]: Refers to the ability to hear and distinguish different instruments and voices within complex orchestrations. The individual notes should also be distinguishable, with well-defined attacks, not diffuse or muddled.
>
> Distortion [0 Bad to 10 Excellent]: Is the sound clean and natural sounding, or do you hear artifacts like buzzes, rattles, etc.?
>
> Sound Stage Accuracy [0 Bad to 10 Excellent]: Refers to the extent that different sounds are accurately positioned in space in terms of sound stage width (left and right) as well as the impressions of distance and depth (forward/backward) of the instruments. Are the images located in logical groupings and continuous or are their large gaps in between? Are the various distances between the instruments accurate or not?
>
> Definition of Images [0 Bad to 10 Excellent]: Refers to the extent that different sources of sound are spatially separated and positionally defined. Images should not move as the pitch of the music falls and rises. The size of the image should be appropriate for the source of the sound.

3.2 Results

The binaural and in-situ ANOVA results for Preference, Treble, Clarity, and Distortion (with respect to each of the five program sources used in the experiment, shown on the x-axis) are shown in Figures 13 and 14. And binaural and in-situ ANOVA results for

Midrange, Bass, Stage, and Image (with respect to each of the five program sources used in the experiment, shown on the x-axis) are shown in Figures 15 and 16.

There is very good agreement between the binaural and in-situ results for the Preference and timbre rankings. Bass, in Figures 15(b) and 16(b), is the exception. Illustrated there is the effect of a lack of perceived bass for the binaural tests due to the lack of bone conduction in binaural recordings and playbacks. The rankings have meaning and agree in relative terms with the in-situ rankings in that the same preferences are shown, just on a different scale. There are also some very low variances for the preference and timbre rankings. Values are in the range of 0.25 to 0.5 (sometimes less). We were not expecting large differences in the frequency domain from one speaker setup to the next, and the differences are slight in the results. So, the binaural method is working very well in terms of the spectral aspects of a sound field. And our listeners are doing a very good job of keeping the variances low so that "slight preferences" are statistically significant.

Surprisingly, there is very good agreement between the binaural and in-situ results for the spatial rankings of Stage and Image. [Figures 15(c,d) and Figures 16(c,d)]. Because the binaural recordings and playbacks do not allow for the spatial cues that come with head movement, normally the results can be less than encouraging when asking questions regarding the spatial nature of a sound field. But in this case we were not asking the listeners to provide us with a detailed map of the soundfield (azimuth angle degree-by-degree, etc.). We asked them simply to rate the general spread and depth and placement of images. And the two methods agreed well. The rankings also made sense in terms of what speaker setup we expected to be preferred. They both also had variances of 0.5 or better. So, once again our listeners were doing a good job of keeping the results in the realm of statistical significance whether they are listening binaurally or in-situ.

4 SUMMARY

1. The acceptance criterion for listeners completing the resonance detection training should be 95% correct for experienced listeners or 85% if training less experienced listeners (more time for the completion of the training should be expected).

2. The additional preference training, using PrefTest to verify the proper use of the preference and timbral rankings, was very effective in producing listeners that provided useful evaluation results.

3. A group mean of 0.5 point variance for the preference training should be expected, and individual listeners should approach that same level of variance as well or not be considered for subsequent experiments or listening evaluations.

4. For experimental or comparative listening evaluations, to obtain a group mean of 0.5 point for in-situ listening, 5 rounds should be used, and listeners should be expected to maintain the same 0.5 point variance. For binaural playback listening evaluations, to obtain a group variance of 0.5 point, 7 rounds should be used, and listeners should be expected to maintain 0.5 – 0.75 point variances.

5. The training method described here is a self-administered method and is highly automated. It requires little administrative involvement beyond monitoring the results and correcting any unclear issues for listeners who are having problems. These administrative functions are also easily automated in a PC-based environment.

6. The software used for training also lends itself well to listening evaluations, either binaurally or in-situ. This too can lead to higher levels of productivity for those involved in automotive sound system design.

7. This training method in general has proven capable of producing valuable listeners from many backgrounds and providing a meaningful, quantitative description of a perceived soundfield for experimental and comparative analysis of automotive sound systems.

5 ACKNOWLEGEMENT

We would like to acknowledge the invaluable assistance of Brian Lennon in collecting and compiling the data for these experiments.

6 REFERENCES

[1] Floyd E. Toole, Sean E. Olive, "Listening Test Methods for Computer Workstation Audio Systems", presented at the 99th Convention of the Audio Engineering Society, New York (1995 Oct.)

[2] Soren Bech, "Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment," *J. Audio Eng. Soc.*, Vol 40, (1992 July/August) pp. 590-610.

[3] F.E. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance", *J. Audio Eng. Soc.*, vol. 33, pp. 2-32 (1985 Jan./Feb.)

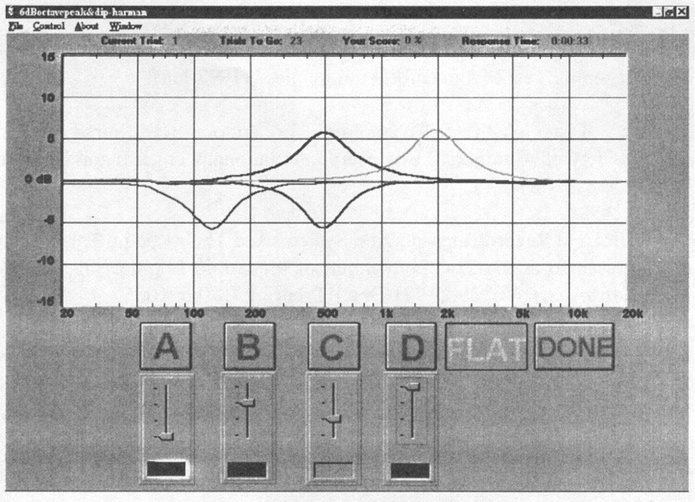
[4] Sean Olive, "A Method for Training Listeners and Selecting Program Material for Listening Tests," Presented at the 97th Conv. of the AES , San Francisco, (1994 November), Preprint 3893.

[5] Sean Olive, "A Method for Training Listeners: Part II", presented at the 101st Convention of the Audio Engineering Society, Los Angeles (1996 Sept.)

[6] R.E. Shively, W.N. House, S.E. Olive [unpublished report] "Pilot Results for: An Examination of Listening Test Methods for Automobiles" (1997 Jan)

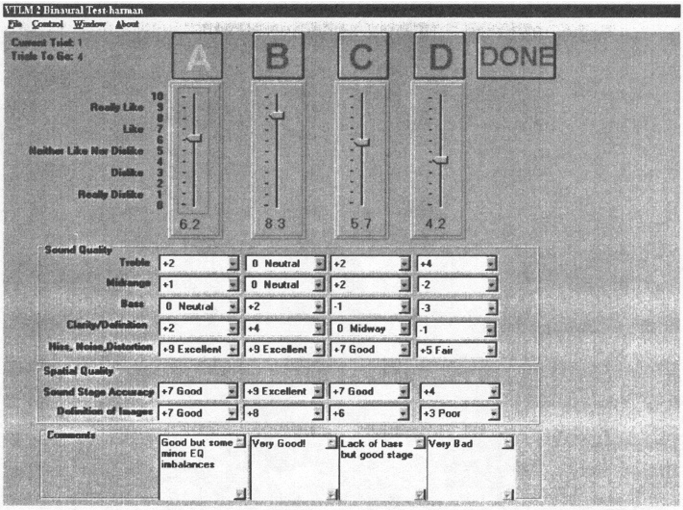
[7] David Clark, "Listening Test Technology for Automotive Sound Systems", presented at the Society of Automotive Engineers International Congress and Exposition, Detroit (1987 Feb.), paper 870145.

[8] F.E. Toole, "Binaural Record/Reproduction Systems and Their Use in Psychoacoustic Investigations," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, Vol 39, (1991 Dec.), Preprint 3179 p. 1005.



Current Trial: 1 Trials To Go: 23 Your Score: 0 % Response Time: 0:00:33

Figure 1. EarTrain Software User Interface

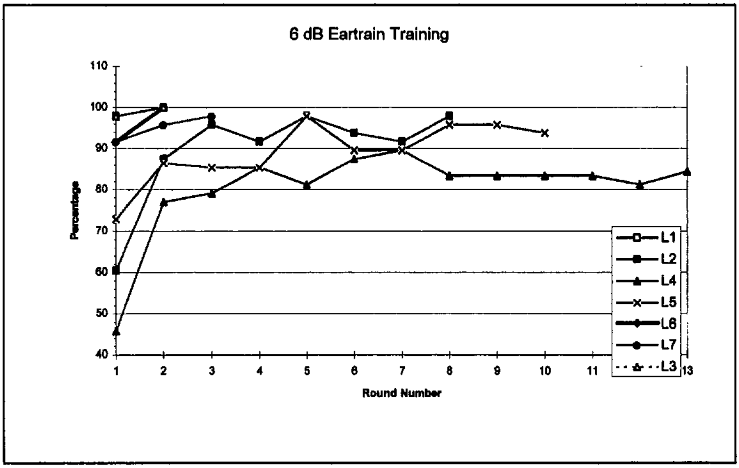


VTLM 2 Binaural Test-harman
File Control Window About
A
B
C
D
DONE
10
Really Like 9
8
Like 7
6
Neither Like Nor Dislike 5
4
Dislike 3
2
Really Dislike 1
0

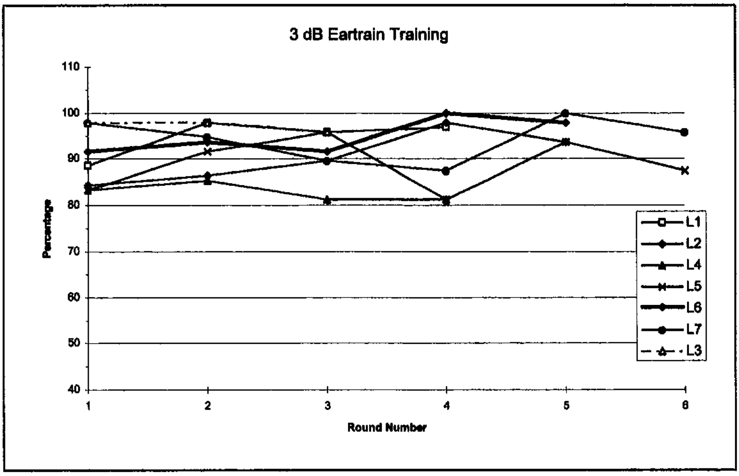
Current Trial: 1
Trials To Go: 4

	A	B	C	D
	6.2	8.3	5.7	4.2
Sound Quality				
Treble	+2	0 Neutral	+2	+4
Midrange	+1	0 Neutral	+2	-2
Bass	0 Neutral	+2	-1	-3
Clarity/Definition	+2	+4	0 Midway	-1
Hiss, Noise,Distortion	+9 Excellent	+9 Excellent	+7 Good	+5 Fair
Spatial Quality				
Sound Stage Accuracy	+7 Good	+9 Excellent	+7 Good	+4
Definition of Images	+7 Good	+8	+6	+3 Poor
Comments	Good but some minor EQ imbalances	Very Good!	Lack of bass but good stage	Very Bad

Figure 2. PrefTest Software User Interface.

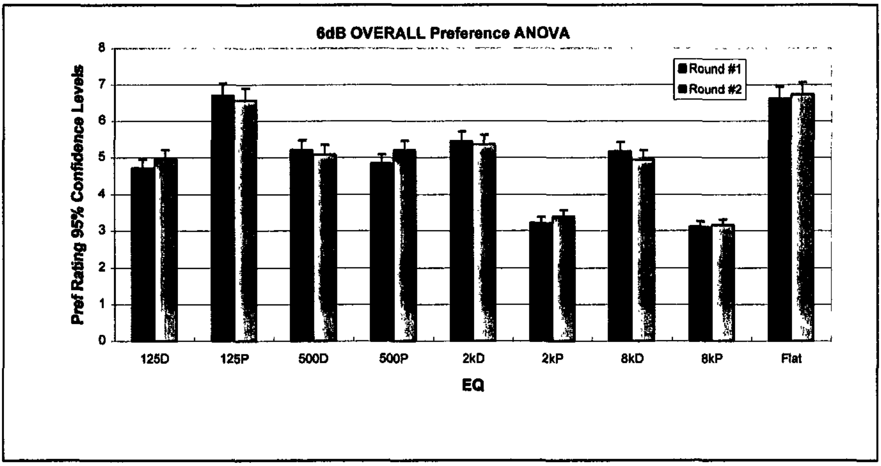


(a)

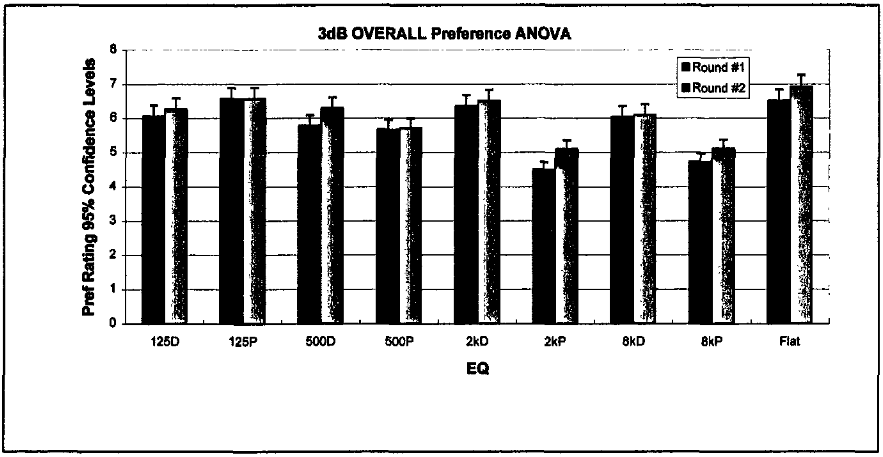


(b)

Figure 3. Results from EarTrain (a) 6dB, (b) 3dB

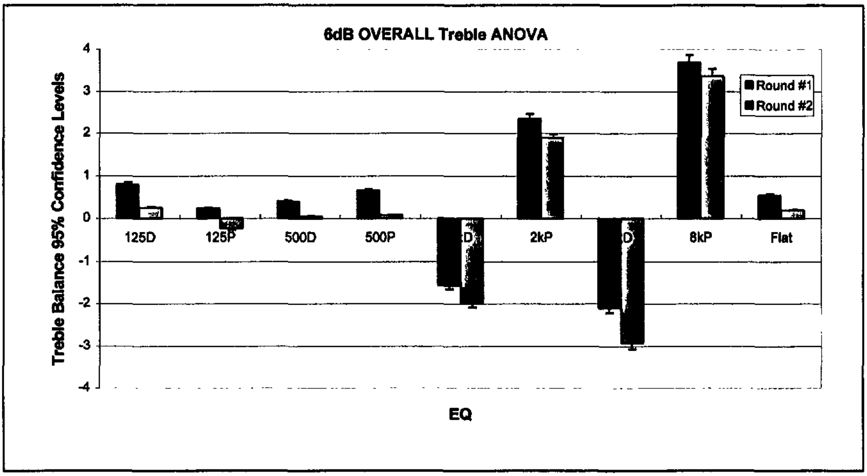


(a)

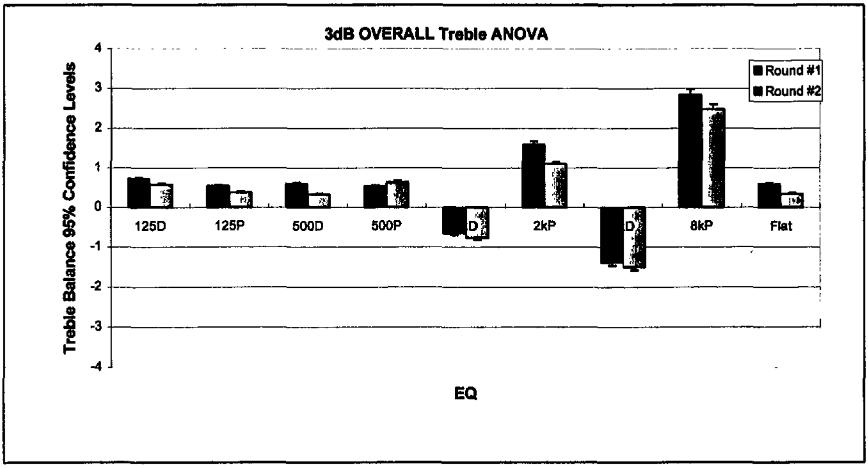


(b)

Figure 4. Overall Preference PrefTest (a) 6dB, (b) 3dB

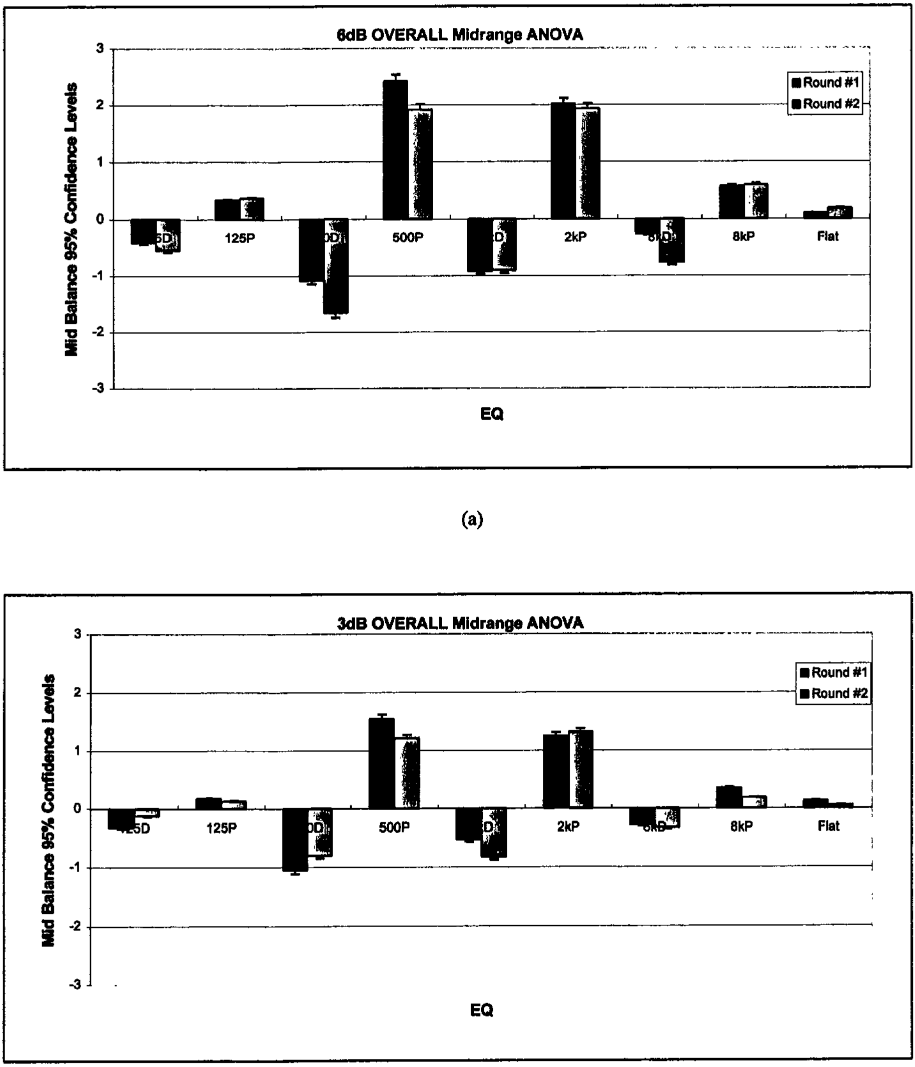


(a)



(b)

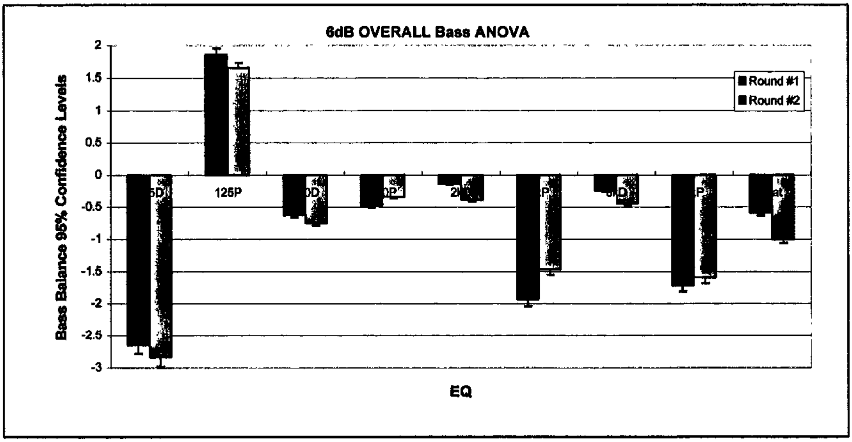
Figure 5. Overall Treble PrefTest (a) 6dB, (b) 3dB



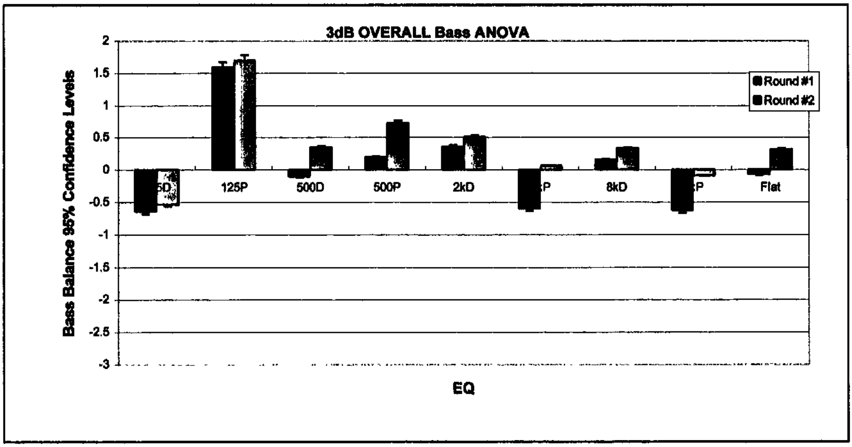
(a)

(b)

Figure 6. Overall Midrange PrefTest (a) 6dB, (b) 3dB

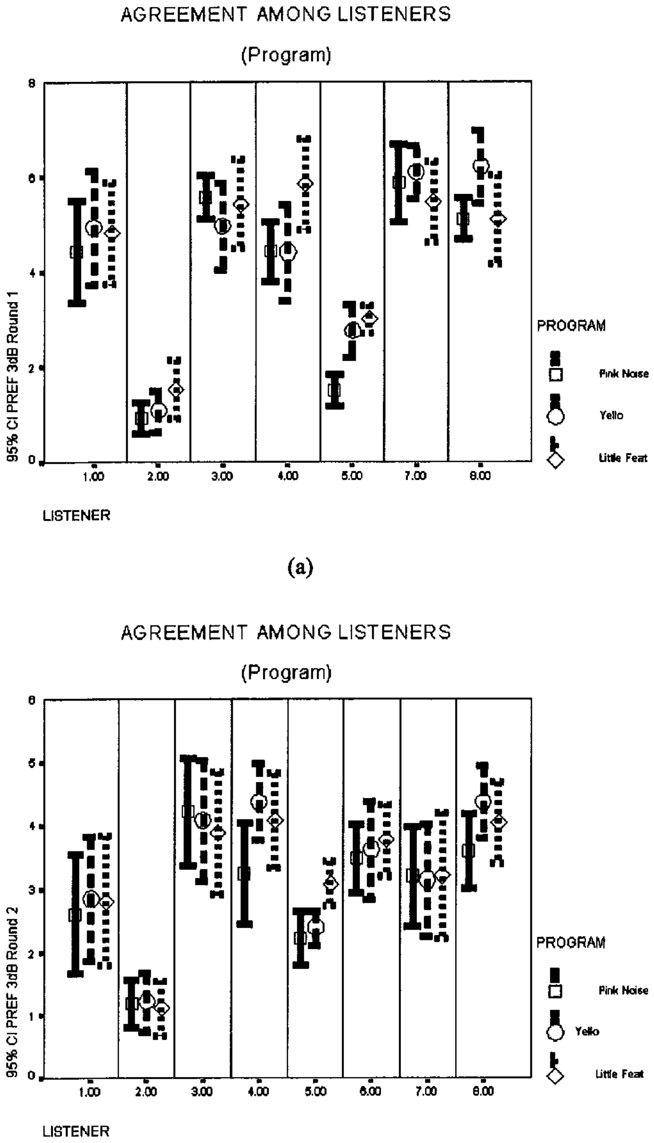


(a)



(b)

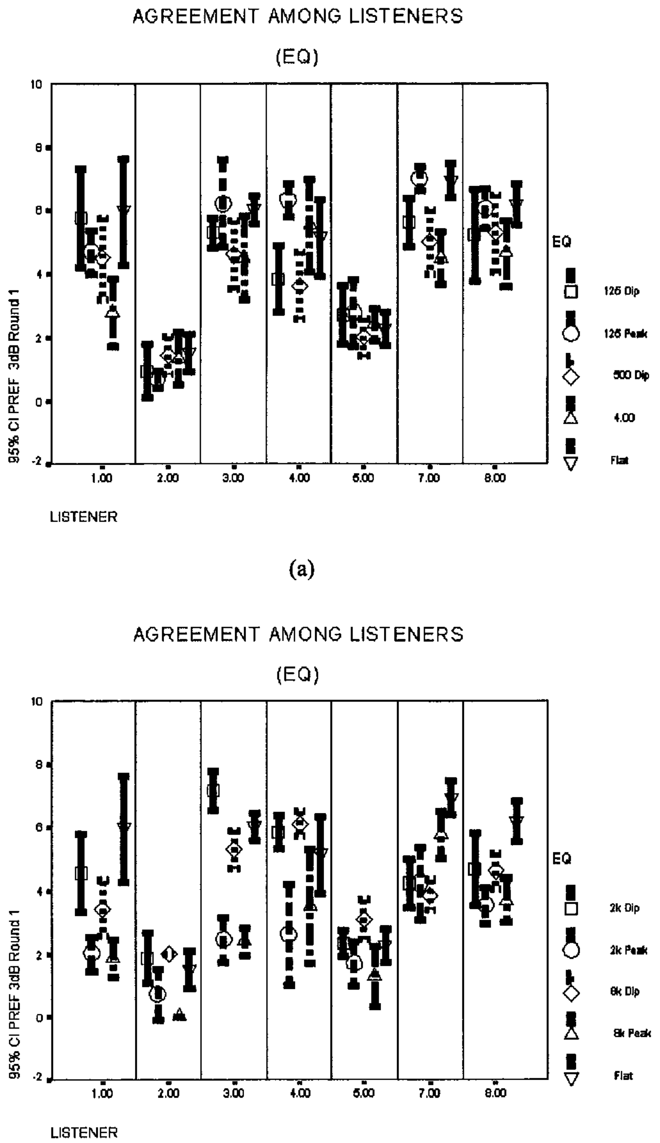
Figure 7. Overall Bass PrefTest (a) 6dB, (b) 3dB



(a)

(b)

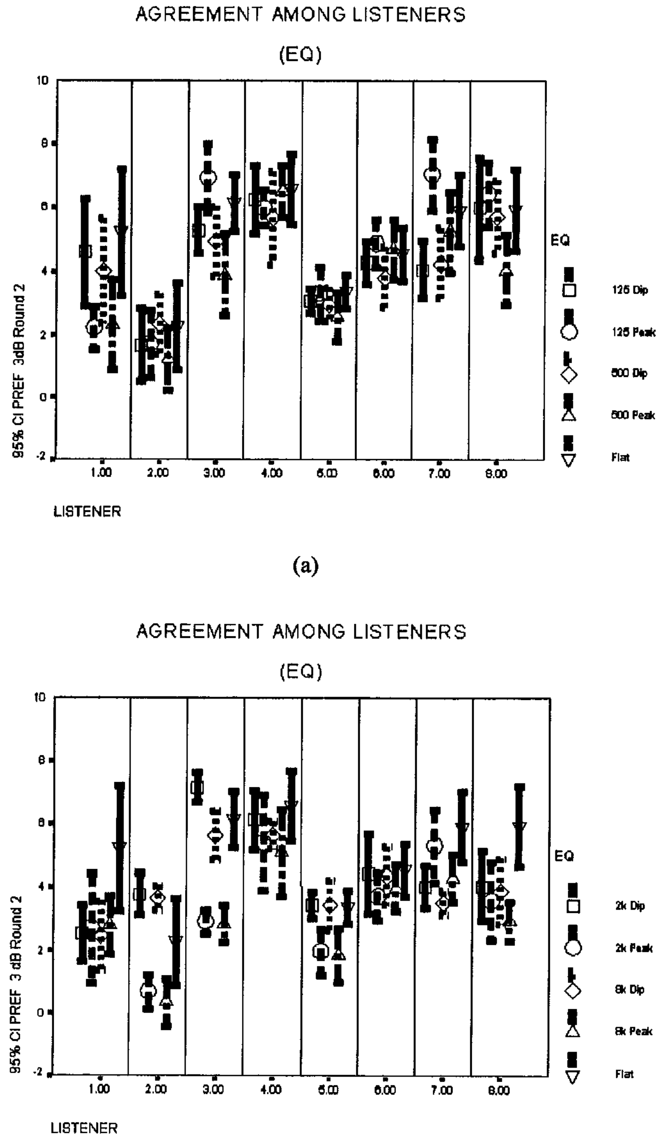
Figure 8. Agreement Among Listeners for Program: PrefTest 3dB (a) 1st Round, (b) 2nd Round.



(a)

(b)

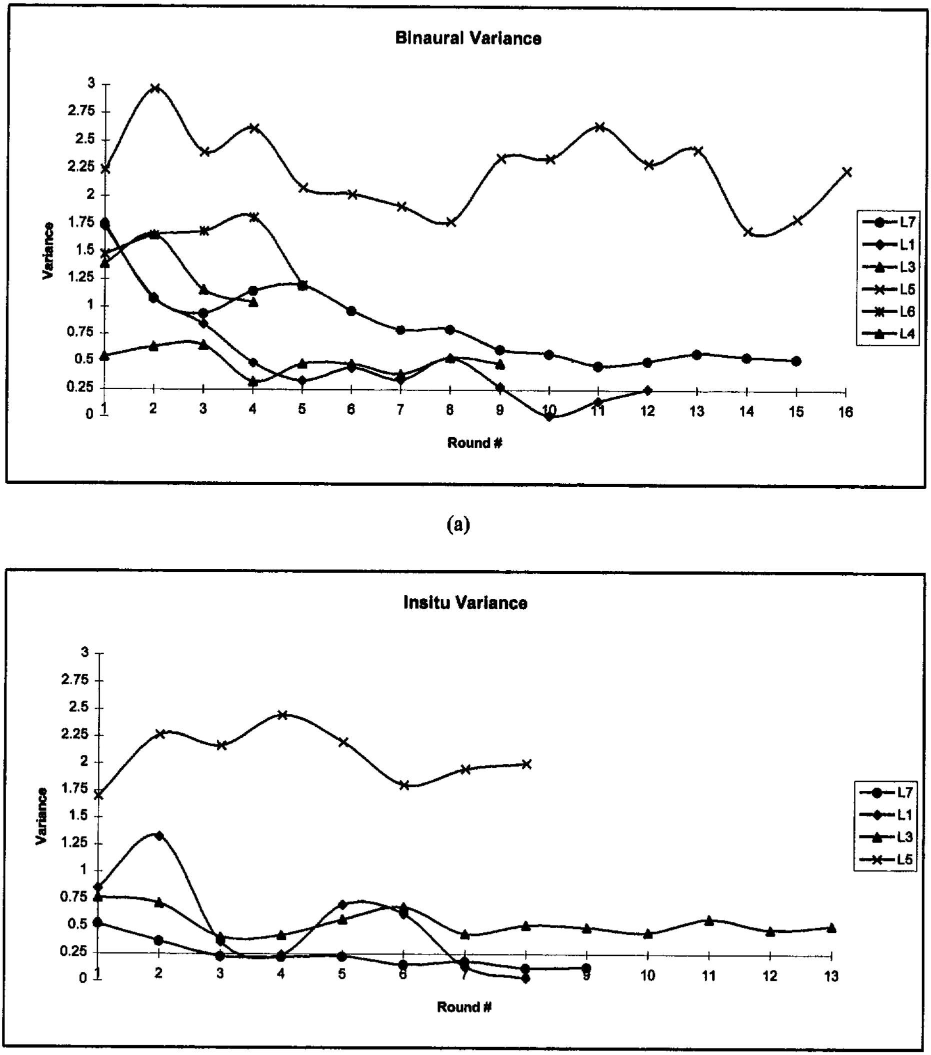
Figure 9. Agreement Among Listeners for EQ PrefTest 3dB
1st Round: (a) 125Hz, 500Hz, Flat (b) 2kHz, 8kHz, Flat



(a)

(b)

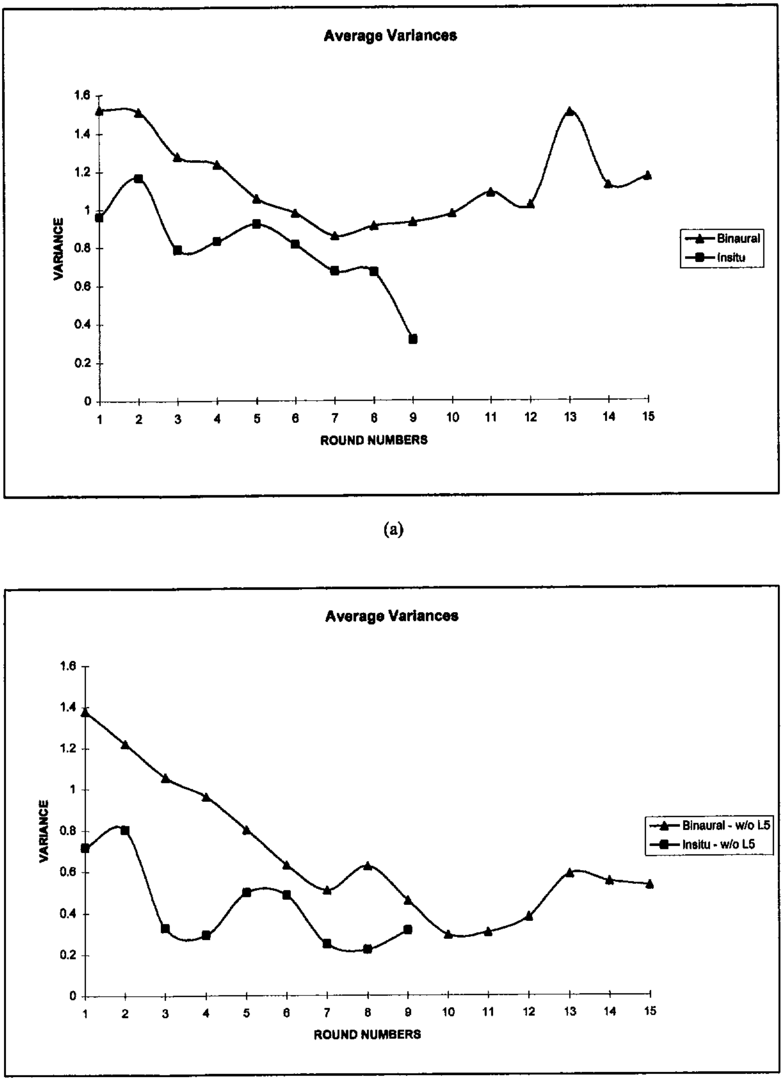
Figure 10. Agreement Among Listeners for EQ PrefTest 3dB
2nd Round: (a) 125Hz, 500Hz, Flat (b) 2kHz, 8kHz, Flat



(a)

(b)

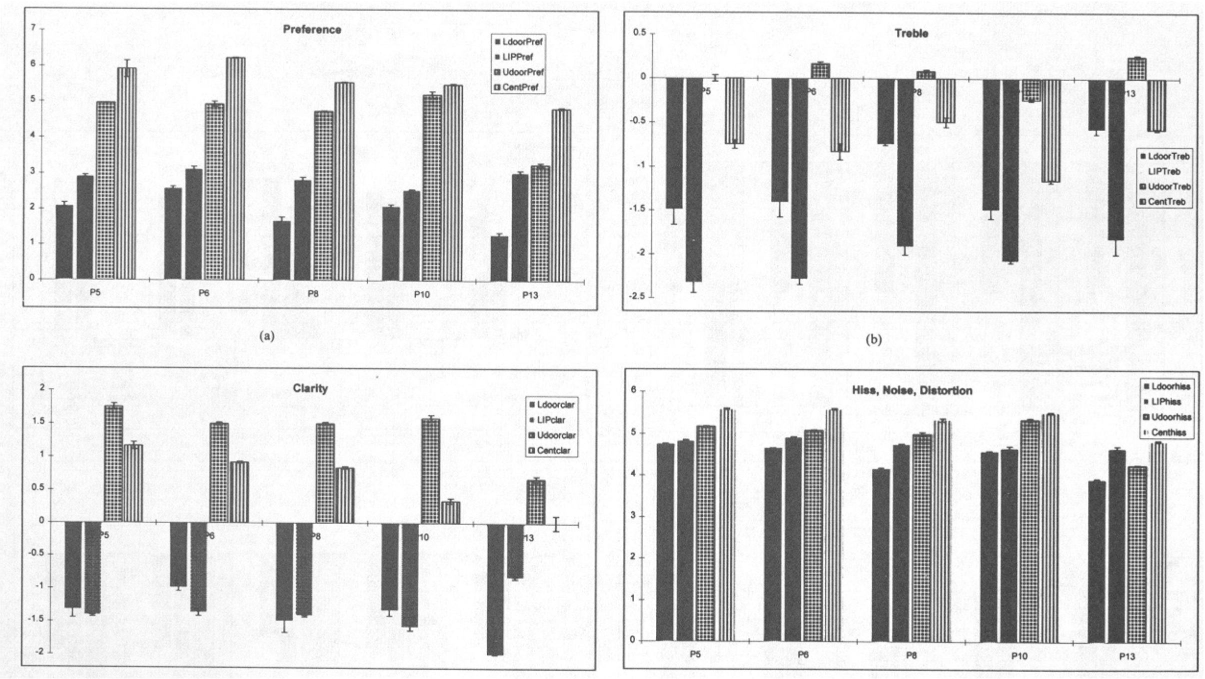
Figure 11. Variance Among Listeners: (a) Binaural, (b) In-situ



(a)

(b)

Figure 12. Binaural and In-situ (a) Average Variances, (b) Average Variances w/o Listener 5



(a)

(b)

(c)

(d)

Figure 13. Spatial Binaural ANOVA: (a) Preference, (b) Treble, (c) Clarity, (d) Distortion

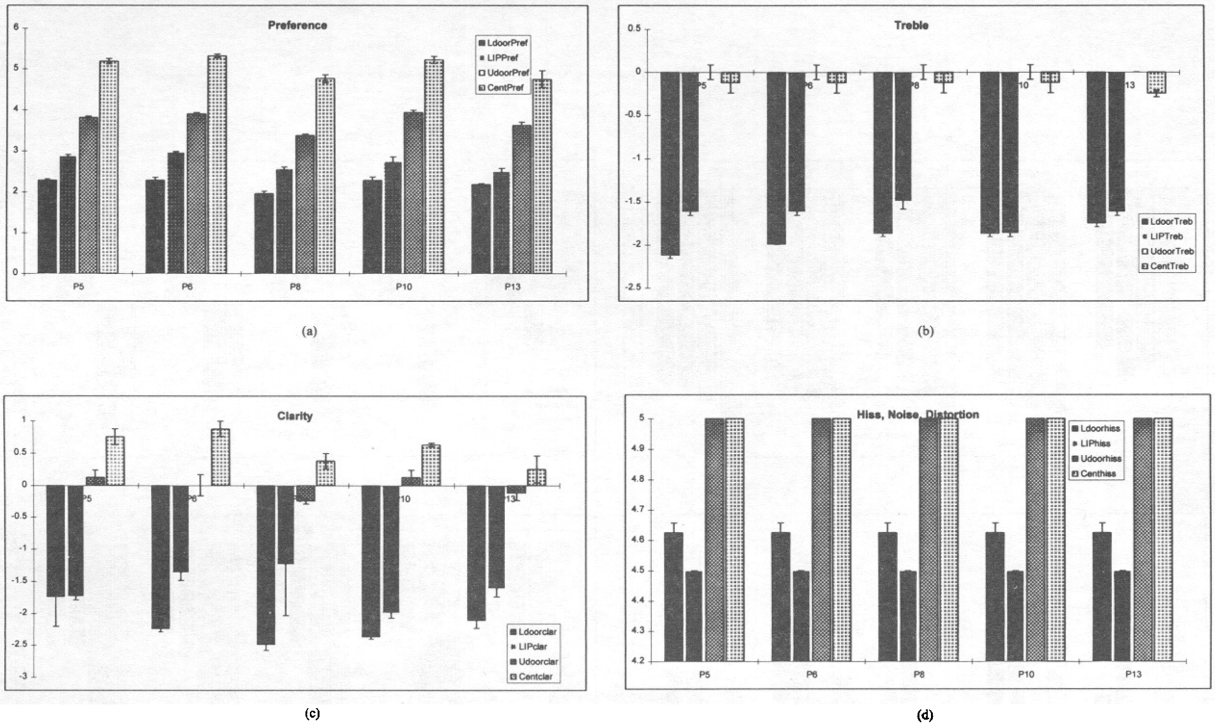
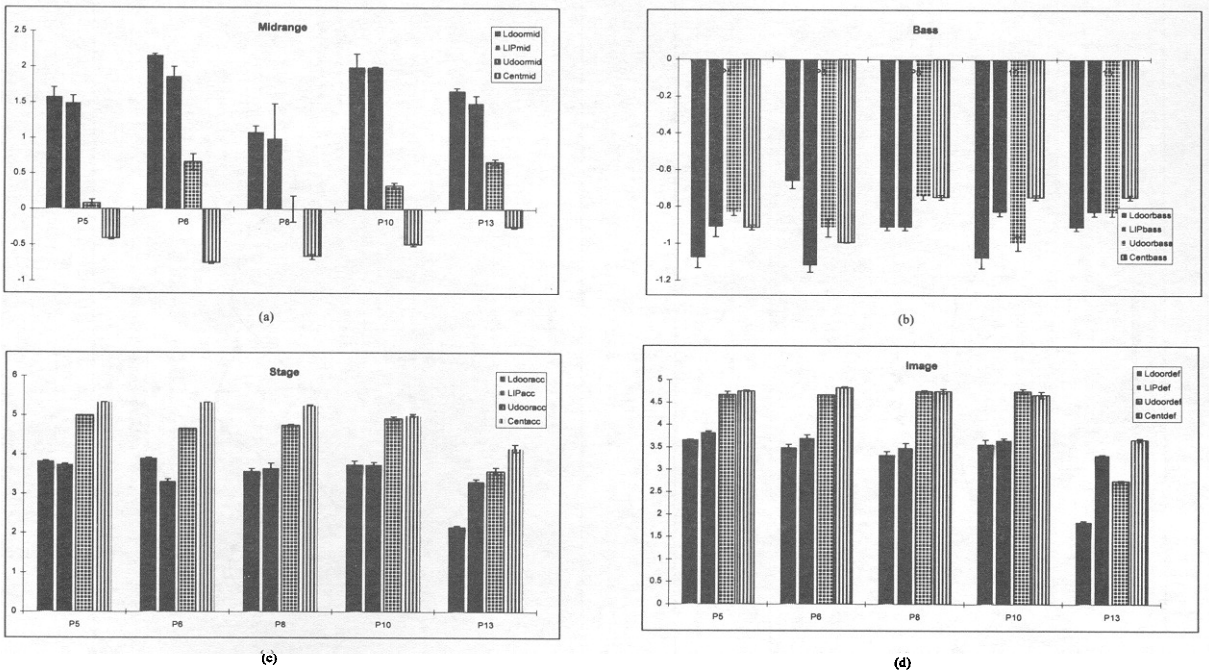


Figure 14. Spatial In-situ ANOVA: (a) Preference, (b) Treble, (c) Clarity, (d) Distortion



(a)

(b)

(c)

(d)

Figure 15. Spatial Binaural ANOVA: (a) Midrange, (b) Bass, (c) Stage, (d) Image

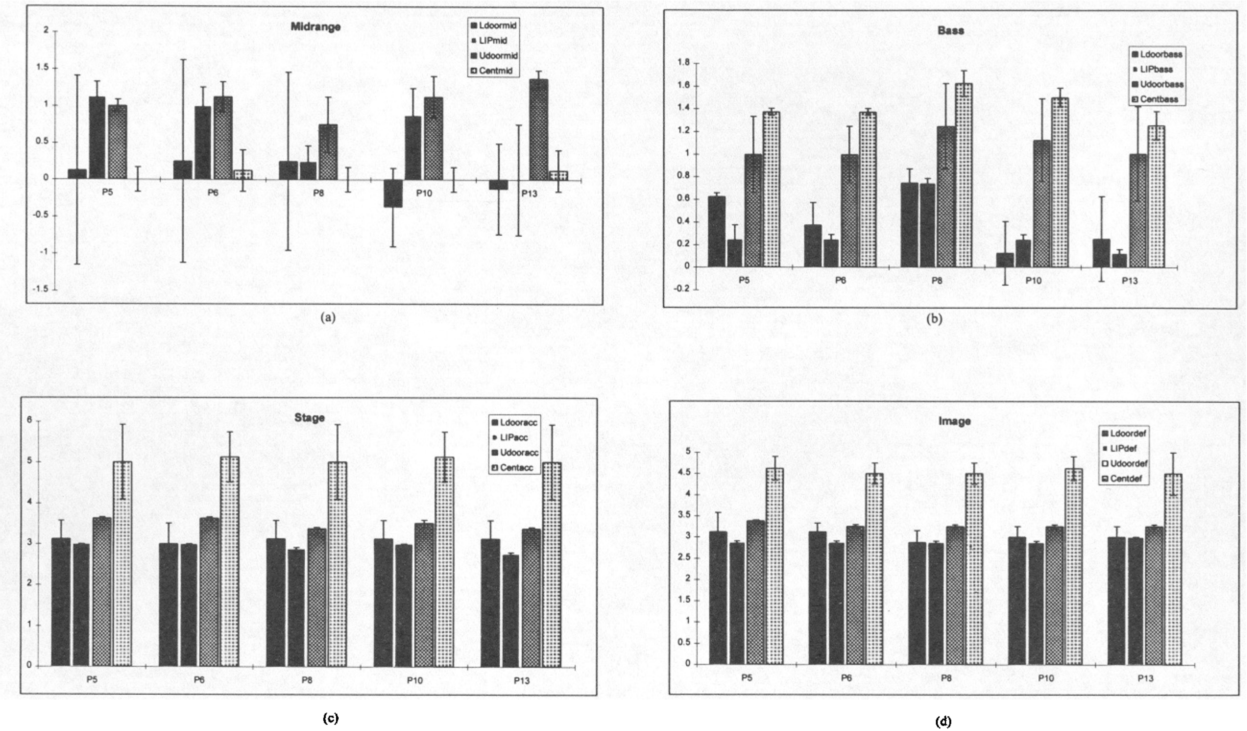


Figure 16. Spatial In-situ ANOVA: (a) Midrange, (b) Bass, (c) Stage, (d) Image