Characterization of Low Frequency Auditory Filters

Menachem Rafaelof
National Institute of Aerospace, Hampton, Virginia

Andrew W. Christian and Kevin P. Shepherd, Stephen A. Rizzi
Langley Research Center, Hampton, Virginia

James H. Stephenson
US Army Aviation and Missile Research,
Development and Engineering Center, Hampton, Virginia

January 2019

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compila-tions of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question to help@sti.nasa.gov

- Phone the NASA STI Information Desk at 757-864-9658

- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM–2019-220120

Characterization of Low Frequency Auditory Filters

Menachem Rafaelof
National Institute of Aerospace, Hampton, Virginia

Andrew W. Christian and Kevin P. Shepherd, Stephen A. Rizzi
Langley Research Center, Hampton, Virginia

James H. Stephenson
US Army Aviation and Missile Research,
Development and Engineering Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

January 2019

Abstract

The purpose of this study is to characterize auditory filters at low frequencies, defined as below about 100 Hz. Three experiments were designed and executed. They were conducted in the Exterior Effects Room at the NASA Langley Research Center, a psychoacoustic facility designed for presentation of aircraft flyover sounds to groups of test subjects. The first experiment measured 36 subjects' hearing threshold for pure tones (at 25, 31.5, 40, 50, 63 and 80 Hz) in "quiet" conditions. The subjects, male and female, had a wide age range. This experiment allowed the performance of the test facility to be assessed and also provided screened test subjects for participation in subsequent experiments. The second and third experiments used 20 and 10 test subjects, respectively, and measured psychophysical tuning curves (PTCs) that describe auditory filters with center frequencies of approximately 63 and 50 Hz. The latter is assumed to be the lowest (bottom) auditory filter; thus, sounds at frequencies below about 50 Hz are perceived via the lower skirt of this lowest filter. All experiments used an adaptive, three-alternative forced-choice test procedure using either variable level tones or variable level, narrowband noise maskers. Measured PTCs were found to be very similar to other recently published data, both in terms of mean values and intersubject variation, despite different experimental protocols, different test facilities, and a wide range in subjects' age.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AF	Auditory filter
nAFC	n-alternative forced-choice
ANOVA	Analysis of variance
BW	Bandwidth
CDF	Cumulative distribution function
CI	Confidence interval
nDmU	n-Down, m-Up
EER	Exterior Effects Room
ERB	Equivalent rectangular bandwidth
ISO	International Standards Organization
PMF	Psychophysical function
PoC	Point of convergence
PTC	Psychophysical tuning curve
SD	Standard deviation

1   Introduction

Noise associated with transportation sources, primarily aircraft and road traffic, has received much attention over the past several decades and has resulted in a range of standardized assessment methods and noise metrics, many of which rely on the integration, over time, of the A-weighted sound pressure level (e.g., refs. [1, 2]). This paper is concerned with the impact of low frequency noise, which typically arises either from sources with dominant low frequency content (e.g., helicopters and sonic booms) or sources at long distances for which high frequency content has been attenuated by the atmosphere and terrain (e.g., distant jet aircraft and road traffic). It is questionable whether A-weighted levels are suitable for assessing such sounds with dominant low frequency content. For the purpose of this study, "low frequency" is defined as being below about 100 Hz.

The human hearing system has been functionally characterized as a bank of bandpass filters and numerous experimenters have investigated the attributes of these presumed filters over the past eighty years (e.g., refs. [3-7]). Among other findings, it is generally recognized that when listening to a sound, humans are able, subconsciously, to place these bandpass filters at optimum frequencies. For example, if the signal of interest is a pure tone then a filter can be placed so that the tone is maximally resolved from any noise that is present at nearby frequencies. Most of this work has concentrated on the frequency range relevant to speech perception, with relatively little attention paid to the lowest frequencies. Furthermore, in contrast to higher frequencies, there are conflicting conclusions regarding auditory filter bandwidths at these lower frequencies [8, 9]. This lack of knowledge regarding human hearing is very important for the determination of the audibility and annoyance of low frequency sounds.

Recent work by Jurado [10, 11] indicates that the human hearing system has a "lowest" or "bottom" bandpass filter and thus sound below its tip frequency is perceived via the lower skirt of this lowest band. Jurado performed his tests using special-purpose headphones and a unique test chamber, a testament to the challenges of performing studies at these low frequencies. The differences in filter characteristics between test subjects were found to be quite large.

The main purpose of this research effort is to replicate Jurado's efforts using a different test facility and a somewhat different test methodology. In order that the results can be generalized and applied to environmental noise assessment and audibility, a larger number of test subjects, representative of the general population, is required to describe average filter shapes and bandwidths.

2   Experimental Approach and Test Procedures

2.1 Test Facility

The Exterior Effects Room (EER) [12] at the NASA Langley Research Center is an acoustically treated laboratory designed to produce high-fidelity sounds, primarily for evaluation by human test subjects. The EER is composed of 27 mid-to-high frequency satellite loudspeakers and 4 subwoofers, with an overall compensated frequency range of 16 Hz – 20 kHz. Although it has a seating capacity of 39, most psychoacoustic tests utilize far fewer seats in order to achieve uniform sound exposure of test subjects. For several reasons, the current test employed a single seat and a single loudspeaker. The low frequency nature of the test sounds requires only the subwoofers and not the satellite loudspeakers. Furthermore, in this frequency range the acoustic treatment on the walls and ceiling of the room is ineffective and results in unacceptable spatial variation in the sound field amongst the seats. A series of measurements within the facility identified the best arrangement as being one in which a single test subject was positioned close to a single subwoofer (Figure 1 and Figure 2). This arrangement has the advantage that higher sound levels can

be achieved at the test subject location due to the proximity of the subwoofer.   Figure 3 illustrates that this arrangement resulted in relatively small variations of the sound field in the vicinity of the test subject's head. The variation observed in the vicinity of 25 Hz was of some concern but subsequent tests in which subjects were presented with tones at that frequency showed no anomalous results. Measurements made at a location corresponding to the center of the head (in the absence of the test subject) were used to characterize the acoustic exposure. A number of acoustic wedges were placed in front of the subwoofer to attenuate any high frequency sound associated with movement of air through the port of the subwoofer.

Figure 1: Exterior Effects Room (EER) at the NASA Langley Research Center.
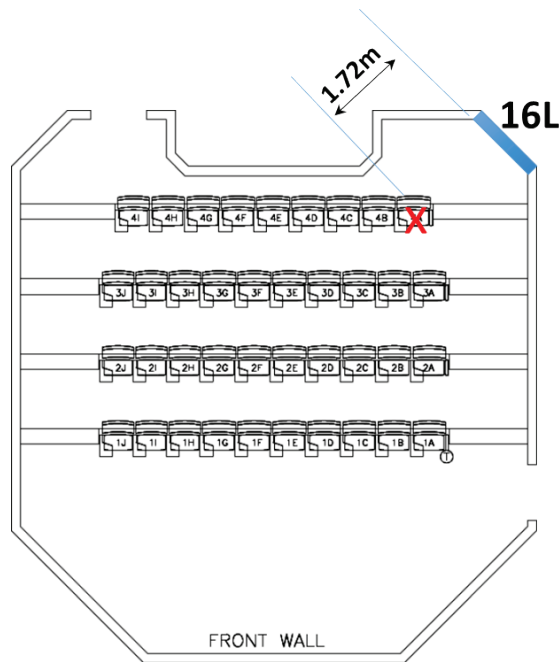
Figure 2: Single subject test setup: seat 4A in front of 16L subwoofer speaker.
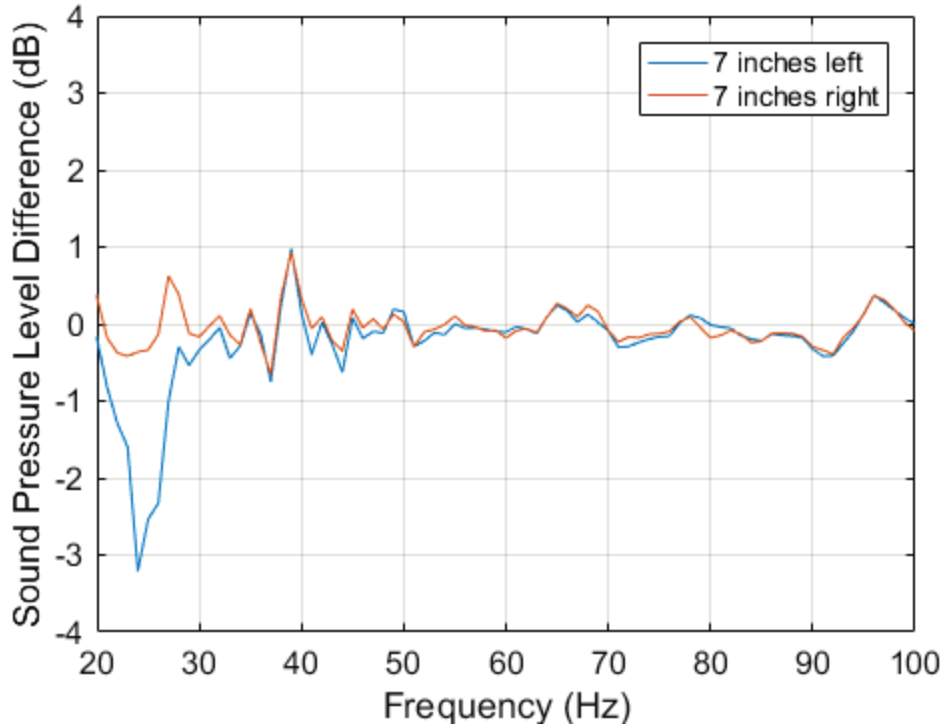
Figure 3: Variation of sound pressure level in vicinity of the test subject head center.

The operational range of the facility, in terms of frequency and sound level, is primarily determined by the performance of the loudspeakers. However, for current purposes in which test subjects were tasked with identifying signals in the presence of noise, there are two important additional factors. The first is the background noise in the laboratory due to the building's internal noise sources (primarily the heating and air conditioning system) and perhaps from other external noise sources. The second is the generation of rattle sounds by, for example, light fixtures vibrating in response to low frequency sound generated in the psychoacoustic testing. These rattle sounds could potentially provide auditory cues associated with the presence of inaudible signals, but fortunately they occur only at relatively high signal sound levels. The end result is that these two factors limit the operational range of the facility, as shown in Figure 4. The EER's heating and air conditioning system was switched off when test subjects were present.

2.2 Test Stimuli

The acoustic test stimuli utilized in the following experiments had common characteristics. All pure tones and broadband sounds were digitally generated. Sound pressure measurements made at a position corresponding to the center of a test subject's head were used to construct an equalization filter so that desired tone levels and uniform masker spectrum levels could be obtained. As will be described below, some of the test stimuli were random, band-limited sounds. Such sounds, when of short duration, have the potential to vary significantly from one realization to another because of the small bandwidth-time product. To overcome this problem, the bursts of noise used in the experiments were identical, ensuring that random differences did not impact test results. A three-interval, three-alternative forced-choice (3AFC) adaptive procedure was adopted for the experiments and further details are presented below. Each of the test stimuli, presented in groups of three, had a duration of 0.75 seconds with a 0.15 second interval of silence between stimuli. A Tukey window, with the first and last 4% of the samples equal to half of a cosine, was applied to all test stimuli to minimize transient effects.
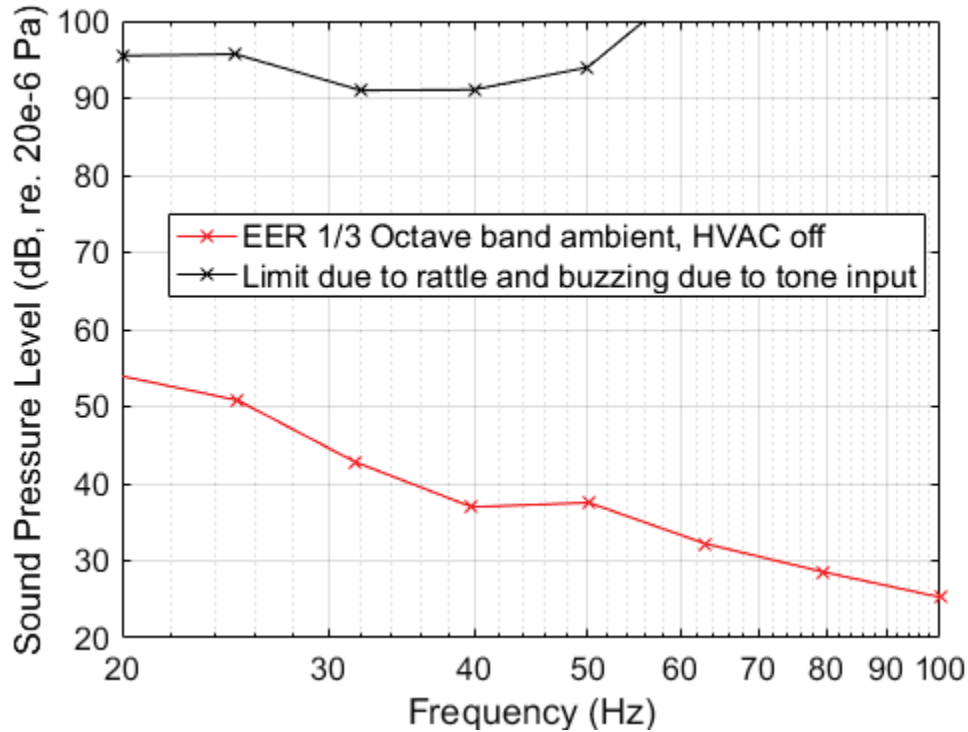
Figure 4: The operational range of the EER defined by ambient noise and upper excitation limits.

2.3 Test Subjects

A large number of test subjects with a diversity in age and gender was desired in order to provide test results that are representative of the general public. Thirty-six test subjects were selected from a pool of local residents with a wide range of ages and socioeconomic backgrounds and were compensated for their participation time in the experiments. All subjects participated in Experiment A and subsets participated in Experiments B and C. Many of the subjects had previously participated in other noise-related experiments, but none were experienced in an *n*-alternative forced-choice procedure. All test subjects' hearing was tested prior to the experiments to verify normal hearing within 20 dB over the frequency range of 125 Hz to 4000 Hz [13, 14]. A wide range of subjects' age was desired, as was gender balance. The age and gender distributions of the selected subjects are described in Table 1 – Table 3. The gender imbalance is largely due to poorer hearing exhibited by males, particularly the older ones.

Table 1: Distribution of subjects' gender and age for Experiment A.

Age	No. of Subjects	Male/Female	Replications (No. of subjects)	Male/Female
18-28	10	3/7	2	1/1
29-38	10	4/6	2	1/1
39-48	8	5/3	0	0/0
49-60	8	0/8	0	0/0

Table 2: Distribution of subjects' gender and age for Experiment B.

Age	No. of Subjects	Male/Female
18-28	4	1/3
29-38	1	1/0
39-48	1	1/0
49-60	4	0/4

Table 3: Distribution of subjects' gender and age for Experiment C.

Age	No. of Subjects	Male/Female
18-28	7	2/5
29-38	3	1/2
39-48	4	3/1
49-60	6	0/6

2.4 Test Method

A test method was needed that would enable a large number of tests to be completed in a reasonable amount of time and with acceptable experimental error. Discussions and analyses described in Appendix B explore a range of experimental methods and test procedures. In particular, given the requirement that a relatively large number of subjects be tested, the importance of intersubject and intrasubject variances and their effect on overall test efficiency was a key consideration in the selection of the chosen method.

Threshold measurements of signals in the presence of noise (the EER ambient noise with and without masking noise) were made using an adaptive, three-alternative forced choice procedure. As shown on the left side of Figure 5, the signal was randomly assigned to one of the three intervals. As illustrated on the right-hand side of Figure 5, a tablet computer touchscreen indicated the sound being presented within a trial, followed by another screen for the subject to indicate which of the three intervals contained the signal. This was followed by the third screen that provided feedback for correct/incorrect responses. An adaptive, 3-down 1-up staircase procedure was employed [15] in which three correct responses result in a reduction in sound level and a single incorrect response results in an increase in sound level. The starting amplitude of each sound was well above the expected threshold. In order to rapidly approach the threshold level, the initial step size was 6 dB and a 2-down 1-up procedure was employed until the second reversal. The step size was then reduced to 4 dB and a 3-down 1-up procedure was followed for a further two reversals after which the step size was reduced to 2 dB. Each staircase was limited to 45-55 trials, depending on the test condition. An explanation of the limit placed on the number of trials is provided in Appendix B. The threshold was estimated as the average of the levels of the turnpoints past the third reversal. This corresponds to 79.4% correct and a d' value of 1.61 [16]. Examples of staircase data spanning the range of subjects' performance are illustrated in Figure 6.

Staircases were interleaved within a test session so that subjects were not able to discern the pattern of the adaptive procedure. The sequence of trials within a session was based on a random selection from the two or three test conditions being interleaved. This selection was different for each test subject in all of the experiments. Further testing details are provided for each experiment below.

3AFC Trial Sequence.
Blue: narrowband noise masker
Green: signal

Please indicate which interval contained the sound

Subject Interface:

tablet screen 1, stimuli (top)
tablet screen 2, response (middle)
tablet screen 3, feedback (bottom)
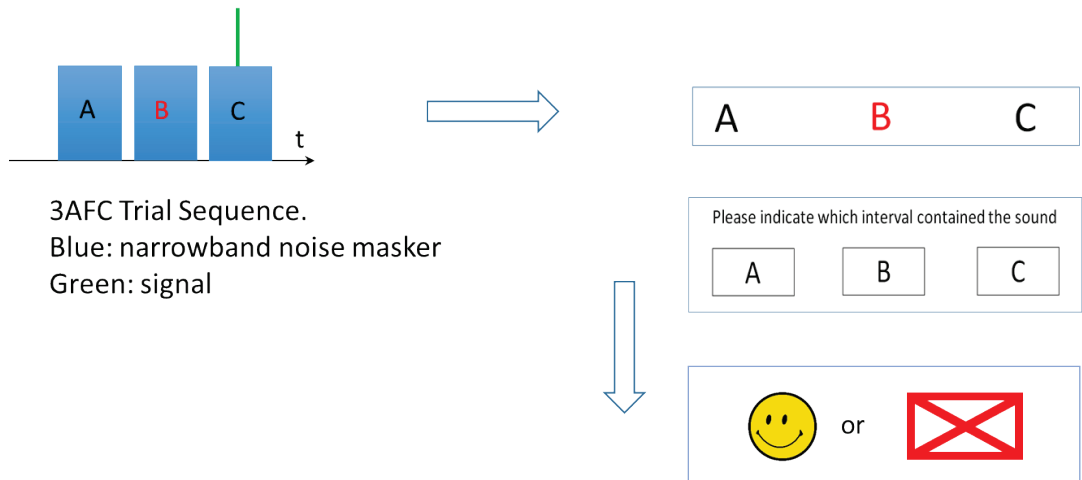
Figure 5: 3AFC trial sequence vs. time (left) and its representation on subject's tablet (right) followed by response and feedback tablet screenshots.

2.5 Laboratory Procedure

Upon arrival at the laboratory the subjects were given written instructions detailing the test. After reading the instructions, the subjects completed consent forms. Copies of the test instructions and informed consent forms are given in Appendix A. The subjects were then given a verbal explanation of the task that was to be performed using the tablet computer, followed by familiarization and training sessions. The familiarization session consisted of listening to some representative sounds and the training session required the subjects to provide their responses using the tablet computer. Test sessions were approximately 8 or 12 minutes in length for two and three interleaved staircases, respectively, and separated by short rest breaks during which the subjects were free to leave the EER. Further details are given for each experiment below.
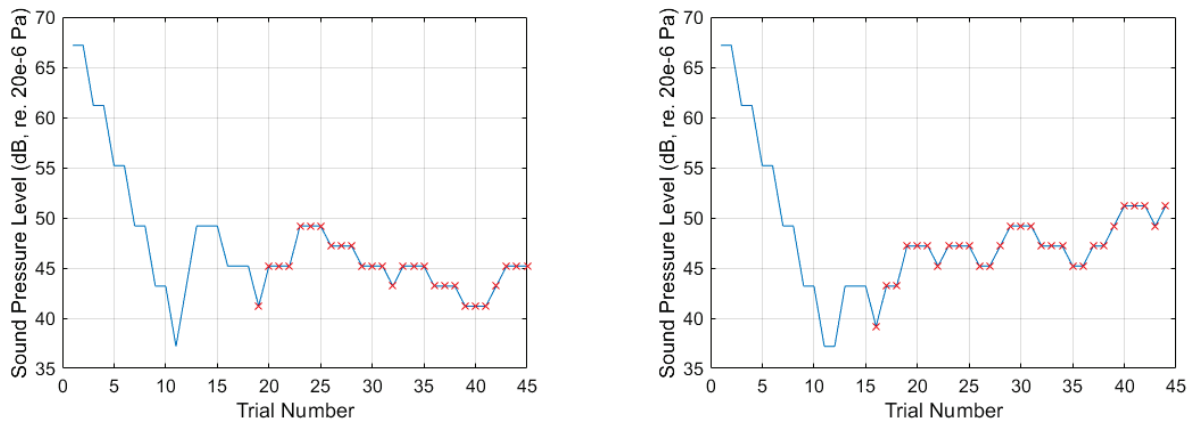
Figure 6: Sample adaptive staircase responses.

6

3   Experiment A – Determination of Low Frequency Pure Tone Thresholds

3.1 Experiment A – Objective

The primary objective of this experiment was to screen a number of test subjects for their potential for participation in the subsequent experiments to characterize auditory filters. Test subjects were needed with acceptable low frequency hearing and the ability to reliably perform the adaptive 3AFC procedure. Furthermore, in order to evaluate the performance of the EER test facility, it was important that an experiment be conducted for which the results could be compared with well-accepted findings from previous studies. It was determined that measurements of pure tone thresholds in a "quiet" background would satisfy these objectives.

3.2 Experiment A – Design

Thirty-six test subjects participated, one at a time, in the determination of pure tone thresholds in the EER using the 3AFC adaptive procedure described above. As shown in Table 1, the test subjects had a wide range of ages. The selected tone frequencies were 25, 31.5, 40, 50, 63, and 80 Hz. A practice session of about 3 minutes in duration was composed of two interleaved staircases and was followed by three test sessions of about 8 minutes duration separated by short rest breaks. Each test session was composed of two interleaved staircases: 25 and 50 Hz, 31.5 and 63 Hz, 40 and 80 Hz. Four of the 36 subjects repeated the test on a different day in order to assess within-subject variability.

3.3 Experiment A – Results and Discussion

Figure 7 presents measured thresholds for all subjects and tone frequencies. The plotted points have been jittered along the abscissa to aid the reader. Several statistics have been derived from the plotted points: the solid red line in the center of each red box indicates the mean threshold at each frequency; the vertical extent of each box represents the standard error about the mean; the vertical blue line indicates plus and minus one standard deviation from the mean. The solid red line presents the one-third octave band sound pressure levels of the natural ambient noise (i.e., "quiet") present in the EER. The three dashed lines are the 10th, 50th, and 90th percentiles of the hearing threshold distribution for young, healthy ears under binaural listening conditions and frontal incidence [17].
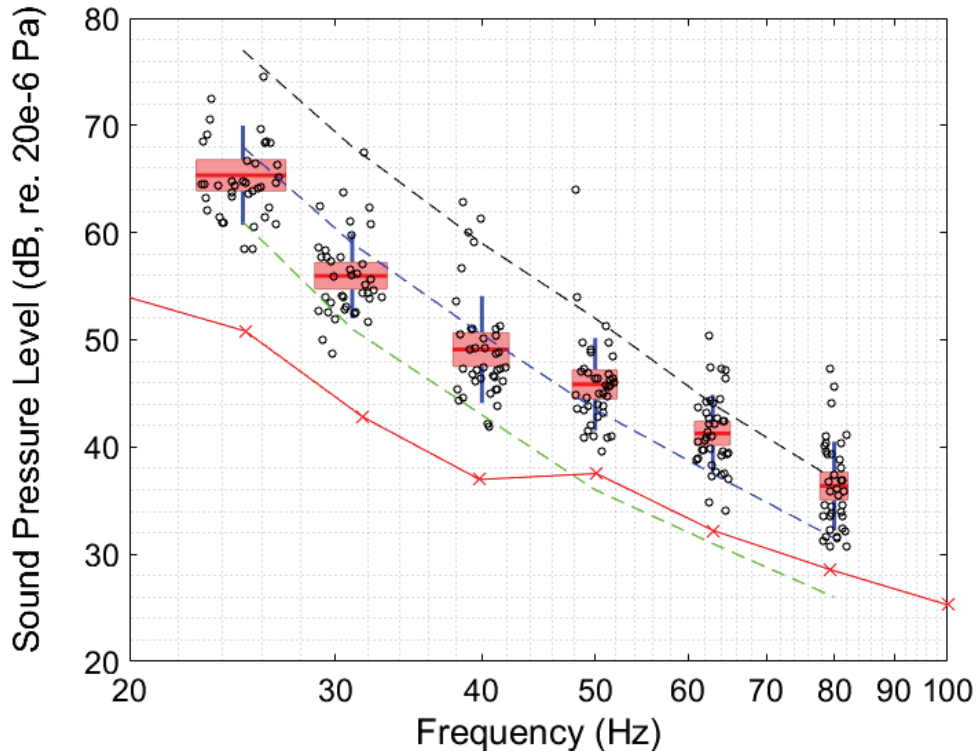
Figure 7: Pure tone threshold for all subjects (circles). Mean values (horizontal red lines); standard deviation (vertical blue lines); standard error of the mean (red box). Dashed traces: Pure tone threshold 10/50/90 percentiles (green/blue/gray) according to ISO 28961. Solid trace (red) is EER ambient noise one-third octave band spectrum.

The measured thresholds in Figure 7 are reasonably consistent with the ISO standard, which represents young, healthy ears. The measured mean thresholds are slightly below the ISO standard values at the lowest frequencies, with a clear trend toward exceeding the standard values at higher frequencies. Differences are expected to occur for several reasons. Unlike the ISO standard, the measured thresholds were acquired, by design, from subjects with a large range in age, although they were screened for normal hearing. It is clear that ambient noise above about 40 Hz has affected the measured thresholds to a significant degree. A thorough analysis to address this question requires knowledge of auditory filter bandwidths and shape, which is the subject of the experiments described below, and is considered in Appendix C. Another source of differences between the measured data and the standard could well be due to differences in experimental methods that were employed. Data in the standard typically come from studies using a yes/no methodology whereas the EER measured data used an adaptive 3AFC method. The latter is expected to yield lower threshold values since the influence of the decision criteria adopted by the test subjects is absent for the AFC method.

Selection of test subjects for Experiments B and C was based on the measured auditory thresholds described above. Sixteen test subjects were identified based on the sum, across the six frequencies, of positive deviations from the ISO standard median values. This is illustrated in Figure 8, which identifies the 16 rejected subjects in red. Of the remaining 20 subjects, all participated in Experiment C and 10 were randomly selected for inclusion in Experiment B. It is noteworthy that the rejection of subjects with the poorest or most inconsistent hearing did not result in the elimination of the older subjects. This is shown by a comparison of the original 36 subjects (Experiment A, Table 1) and the 20 subjects that were retained (Experiment C, Table 3). Figure 9 presents average pure tone thresholds for each age group and each tone frequency. Analysis of variance indicates no significant effect of age group on tone thresholds. This might

be viewed as surprising, but it should be remembered that all the test subjects were screened for reasonably good hearing. Also, at these low frequencies (below 100 Hz), there is little expected hearing loss due to presbycusis or excessive noise exposure.
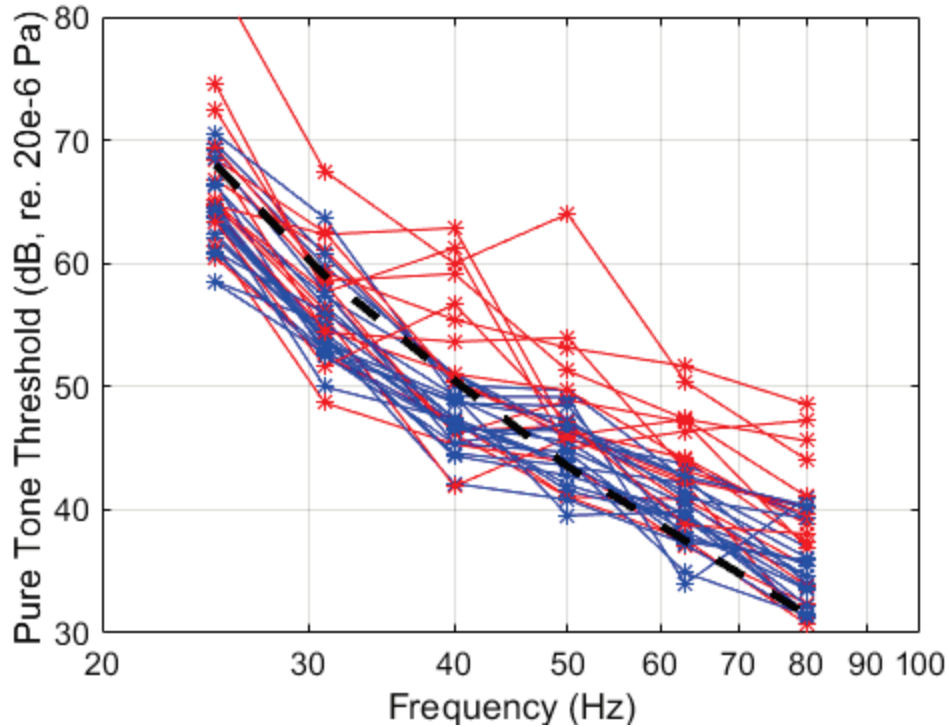
Figure 8: Pure tone thresholds for 20 test subjects. Dashed black line is 50th percentile according to ISO 28961 [17]. Shown in blue are the 20 subjects selected for subsequent tests. Shown in red are those not selected.

As indicated in Table 1, four of the test subjects repeated the experiment in order to get a measure of within-subject variability. The average change in threshold (across subjects and test frequencies) was found to be 1.4 dB, which is an insignificant change based on an analysis of variance of the data for the four replicate test subjects. The standard deviation of the change in threshold was found to be 2.6 dB. Further discussion of this topic is provided in Appendix B.4, in which analyses designed to inform the experimental design are assessed following execution of the experiment.

In summary, a group of test subjects having good low frequency hearing and the ability to perform the 3AFC method was identified for participation in the subsequent experiments. Furthermore, a comparison between the measured thresholds and the ISO standard enabled the performance of the test facility to be well understood.

9
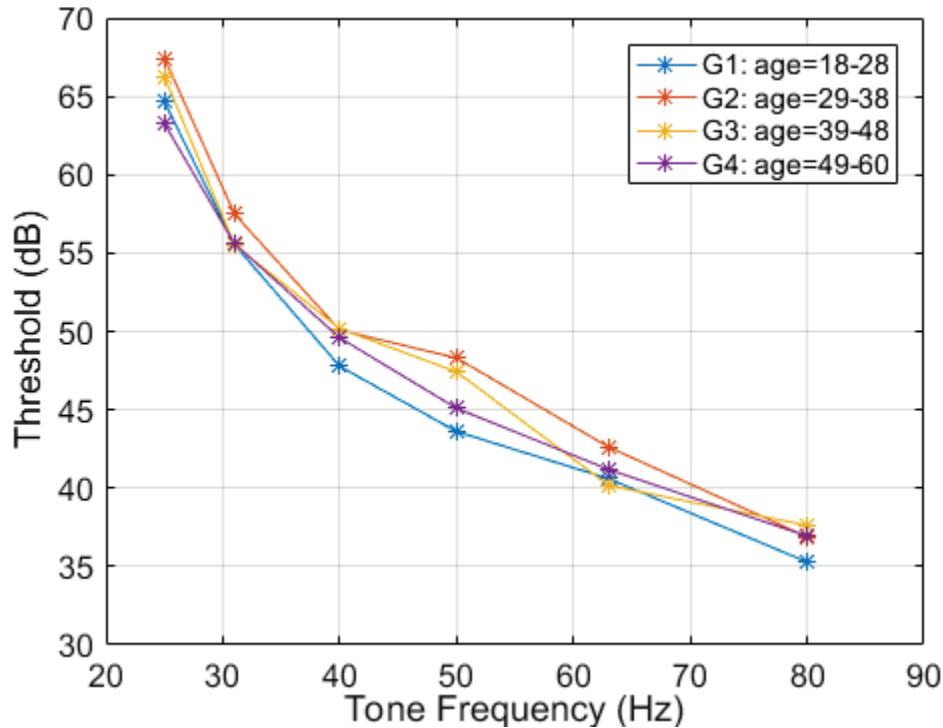
Figure 9: Average pure tone thresholds for each age group.

4 Experiment B – Determination of Auditory Filter Shape at 63 Hz Center Frequency

4.1 Experiment B – Objective

As described in the introduction, there has been relatively little attention paid to auditory filters below about 100 Hz. The work of Jurado [10, 11] has demonstrated that there is a "lowest" auditory filter, estimated to have a center frequency near 50 Hz. Before attempting to examine this lowest filter, the method employed by Jurado [11] was replicated to examine an auditory filter centered at 63 Hz. The objective was to supplement Jurado's data (acquired with 8 subjects) with an additional 10 subjects, and potentially identify any weaknesses in test methods and facility performance. It should be noted that unless stated otherwise, the use of the term "auditory filter" is used in the context of the auditory system as a whole, and includes outer and middle ear transfer functions, the shunt effect of the helicotrema, and the filtering effects associated with the inner ear, the cochlea.

4.2 Experiment B – Design

Psychophysical tuning curves were obtained with ten test subjects using the 3AFC adaptive procedure described above. The signal was a pure tone at 63 Hz fixed at a level of 57.5 dB, equivalent to a sensation level of 20 dB (relative to the median hearing threshold [13, 14]). This level is well above the ambient noise present in the EER (Figure 7). Narrowband noise maskers with a bandwidth of 31 Hz (approximately equal to the 63 Hz signal frequency divided by two) were centered at 31, 37, 50, 63, 75, 88, and 113 Hz. The masker bandwidth of 31 Hz was chosen as a compromise; wide enough to reduce the influence of beats as a cue, while being narrow enough to have only a small influence on the measured frequency selectivity. The masker center frequencies have a slightly narrower range (31 – 113 Hz vs. 25 – 138 Hz) than used by Jurado [11] due to previously described facility limitations in which high amplitude tones excite audible rattling/buzzing in the EER. A practice session of five minutes consisted of two interleaved staircases and was followed by four test sessions of 12 minutes in duration separated by short rest breaks. Each test session was composed of three interleaved staircases. The selection of the maskers in each session and the

presentation sequence of sessions was varied for each subject by using a Latin square design that sought to minimize any ordering effects such as those due to subjects' learning and fatigue. The seven narrowband maskers were supplemented by five additional conditions, all of which used the same 63 Hz signal tone, to make up the four sessions of three staircases each. The supplemental sounds included replications that enabled estimates of within- and between-subject variability to be made. This is discussed in Appendix B.4 and includes an analysis that identifies and quantifies a source of variance which is a direct result of the test method employed in the experiment.

4.3 Experiment B – Results and Discussion

Figure 10 presents psychophysical tuning curves (PTCs) obtained for the 10 test subjects. Also included in the figure are the mean values and the upper and lower 95% confidence intervals obtained by Jurado [11] using a very similar test method. The confidence intervals were calculated from the standard deviations and the reported number of degrees of freedom. There is considerable variation amongst the ten test subjects, but clearly these results obtained in the EER are quite consistent with those of Jurado. This conclusion is further illustrated in Figure 11, which shows consistency between the mean values and also the standard deviations of the two data sets. The shapes of the two average PTCs are very similar, even at the lowest frequencies where the slope is steepest. The valleys of both data sets coincide with the signal frequency of 63 Hz. An auditory filter shape is simply the inverse of the PTC, so that the valley of the PTC becomes the tip frequency of the filter. By setting the filter gain to be 0 dB at the tip frequency, the equivalent rectangular bandwidth of the filter can easily be calculated. The equivalent rectangular bandwidth for the EER data was calculated to be 34.6 Hz, which should be compared to 34.3 Hz and 28.4 Hz reported by Jurado (Table 4). This good agreement implies that small changes in signal amplitude (signal sensation levels were 15 dB and 20 dB for Jurado and the EER, respectively) do not appreciably affect filter shapes. In summary, it is clear that the results of this experiment are very consistent with those previously reported by Jurado.
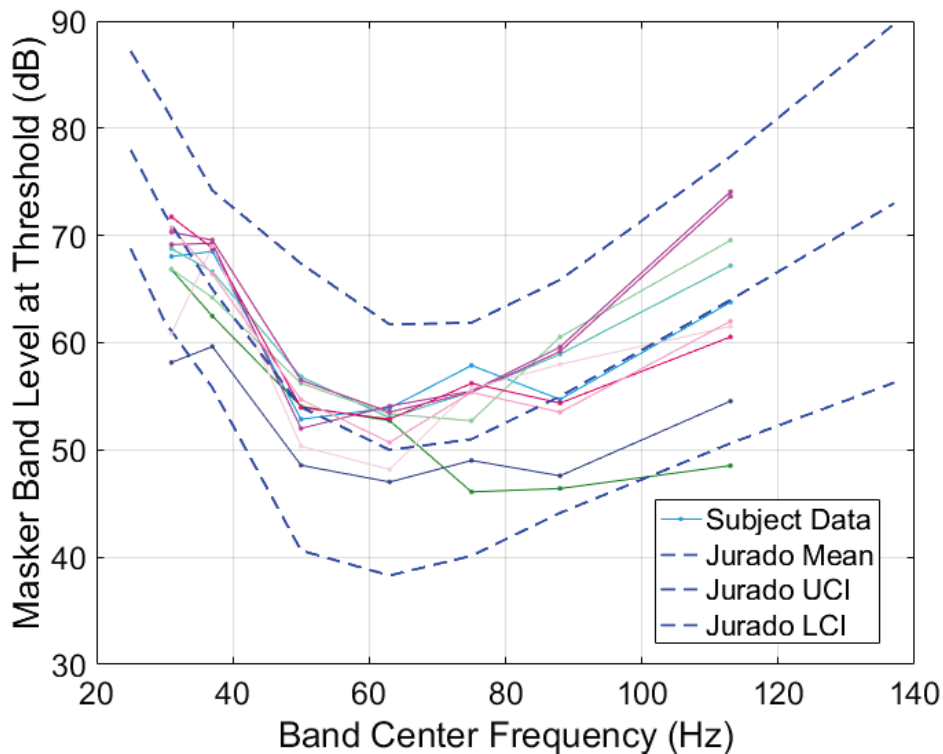
Figure 10: Experiment B PTCs for 10 subjects and similar data
(mean and its 95% lower and upper confidence intervals) reported by Jurado [11].
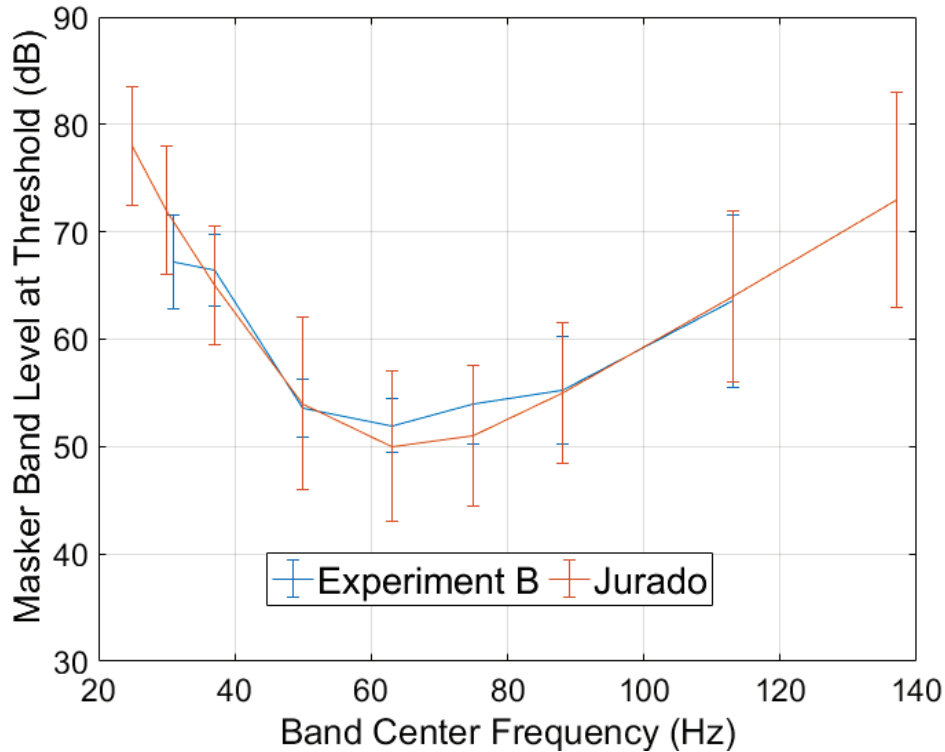
Figure 11: Mean PTC obtained with 63 Hz signal and similar data by Jurado [11].
Error bars for both curves represent ± 1 standard deviation.

Table 4: Equivalent rectangular bandwidths for 63 Hz center frequency filter
from Experiment B and from Jurado [10, 11].

Equivalent Rectangular Bandwidth (Hz)		
Experiment B	Jurado [11]	Jurado [10]
34.6	34.3	28.4

5   Experiment C – Determination of the Shape of the Lowest Auditory Filter

5.1 Experiment C – Objective

The work of Jurado [10, 11] has indicated that there is a "lowest" auditory filter, estimated to have a center frequency near 50 Hz. Precise determination of the shape of this filter is challenging for a number of reasons. One such difficulty is a result of using narrowband maskers with a bandwidth equal to the signal frequency divided by two. As mentioned above (and by Jurado), this bandwidth is a compromise; wide enough to reduce the influence of beats as a cue, while being narrow enough to have only a small influence on the measured frequency selectivity. This compromise becomes particularly problematic when determining the shape of the lower filter skirt below about 50 Hz where the slope of the filter is very steep due to the middle ear transfer function and the dominance of the shunt effect of the helicotrema. The components of the masker spectrum below the masker center frequency will be highly attenuated relative to those above the center frequency. This results in higher masker levels being required to mask the signal. A similar issue concerns the characterization of the narrowband masker according to its center frequency. In the region where the filtering is very steep, the "effective" center frequency of the masker will be above

its actual center frequency. The wider the bandwidth of the masker and the steeper the filter, the greater will be the divergence of the "effective" frequency from the center frequency.

The objective of Experiment C is to investigate the shape of the lowest auditory filter using a method that is a combination of that employed in Experiment B, along with an approach aimed at overcoming the difficulties regarding the lower filter skirt.

5.2 Experiment C – Design

Psychophysical tuning curves were obtained with 20 test subjects using the 3AFC adaptive procedure. Different methods were employed to determine the upper and lower skirts of the lowest auditory filter.

Determination of the upper skirt used the same method as that employed in Experiment B and is illustrated in Figure 12. The pure tone signal at 40 Hz was presented at a fixed level of 66.5 dB (re. 2 x 10 $^{-5}$ Pa), equivalent to a sensation level of 16 dB, relative to the median hearing threshold [13, 14]. This level is well above the ambient noise present in the EER (Figure 7). Narrowband maskers with a bandwidth of 20 Hz (equal to one-half of the 40 Hz signal frequency) were centered at 40, 48, 56, 72, 88, 104, and 128 Hz. The amplitudes of the narrowband maskers were varied through the adaptive 3AFC staircase procedure in order to determine psychophysical tuning curves at 40 Hz and above.

Part 1: Upper skirt
Masker: Narrowband variable level noise with constant BW = F_s / 2
       F_m = 40, 48, 56, 72, 88, 104, & 128Hz

Signal: pure tone Fs = 40Hz
       fixed level at 16dB Sensation Level
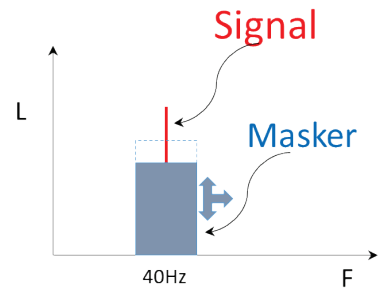
Part 2: Lower skirt
Masker: fixed level noise
       100Hz BW (50 to 150Hz)
       Spectrum Level = 51dB/Hz

Signal: variable level pure tones at 20, 25, 30, 35, & 40Hz
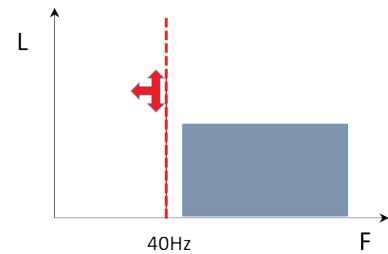
Figure 12: The two methods used to determine the upper (40 Hz and above)
and the lower (40 Hz and below) skirts of the PTC for Experiment C.

The method used to determine the shape of the lower filter skirt is also shown in Figure 12. The masker was broadband, extending from 50 to 150 Hz with a spectrum level of 51 dB per Hz. Using the same 3AFC adaptive staircase procedure, the sound levels of pure tones at 20, 25, 30, 35, and 40 Hz were varied and masked thresholds were determined. The broadband masker ensured that the test subjects utilized the lower skirt of the lowest filter to listen for the pure tones. The amplitude of the masker was chosen in an attempt to be consistent with the amplitude that was chosen for determination of the upper skirt (16 dB sensation level at 40 Hz). For example, if the broadband masker was presented at a very low level then the masked thresholds for the low frequency tones would differ little from their thresholds in quiet. Conversely, if the broadband masker level was high resulting in masked tone thresholds with sensation levels much greater than 16 dB, then it is possible that the measured upper and lower skirts of the PTC are not compatible. This is because there is evidence that auditory filtering is a nonlinear process, although this has been shown at

frequencies much higher than those used in the present experiments [18]. The degree of nonlinearity is likely small for small differences in amplitude. The results of Experiment B support this notion; very good agreement was observed with Jurado's results despite a 5 dB difference in signal presentation levels.

The two methods employed for the upper and lower skirts each had a measurement at 40 Hz, thus enabling the upper and lower filter shapes to be joined together to form a composite filter shape. In other words, the masked threshold for the 40 Hz tone derived for the lower skirt was set to the numerical value of the 40 Hz masked threshold derived for the upper skirt. Masked thresholds for the other frequencies on the lower skirt were adjusted accordingly. This overall approach has an implicit assumption that there is a lowest auditory filter and that its center frequency is in the vicinity of 50 Hz and not at a much lower frequency.

A practice session of about 5 minutes in duration was composed of two interleaved staircases and was followed by four test sessions of 12 minutes duration. Each test session was composed of three interleaved staircases with a short rest break between sessions. The selection of the sounds in each session and the presentation sequence of sessions was varied for each subject by using a Latin square design.

5.3 Experiment C – Results and Discussion

The psychophysical tuning curves (PTCs) obtained for the 20 test subjects are presented in Figure 13. Recall that the data presented for frequencies below 40 Hz were derived using broadband maskers and are normalized to the 40 Hz data point obtained using the narrowband masker. Thus, the ordinate values in Figure 13 only apply to frequencies of 40 Hz and above; those below 40 Hz are relative values only. Also included in the figure are the mean values and the upper and lower confidence intervals obtained by Jurado [11] for a tone at 40 Hz and narrowband maskers. There is considerable variation amongst the 20 test subjects, but clearly, the results obtained in the EER are quite consistent with those reported by Jurado. This conclusion is further illustrated in Figure 14, which shows consistency between the mean values and also the standard deviations of the two data sets. The shapes of the two average PTCs are quite similar, perhaps with an indication of a steeper slope for Jurado's at the lowest frequencies. As noted by Jurado [11], the shape of the PTC below about 40 Hz is very similar to that of the 40 phon equal loudness contour. If the latter is assumed to represent equal cochlea excitation, this suggests that the measured PTC is indeed the "bottom" or "lowest" auditory filter. In other words, the auditory system is unable to center PTCs on frequencies below about 50 Hz. The valleys of both data sets are at 48 Hz. The equivalent rectangular bandwidth for the EER data was calculated to be 34.9 Hz, which is compared to Jurado's results in Table 5. As described earlier, Jurado's test design differs from that used in Experiment C. As a result, Table 5 presents Jurado's results for signal frequencies of both 40 and 50 Hz since both are at or below the expected tip frequency of the lowest-frequency filter. There is reasonable agreement between the estimates made by Jurado and those derived from the current experiment. The shape of the lowest filter derived in the EER and by Jurado is also noteworthy because of its large asymmetry with the lower skirt being much steeper than the upper one. This is in marked contrast to the generally accepted filter shapes for higher center frequencies [6, 7]. Furthermore, the equivalent rectangular bandwidth of the lowest filter is, relative to its center frequency, much larger than bandwidths for higher frequency filters.
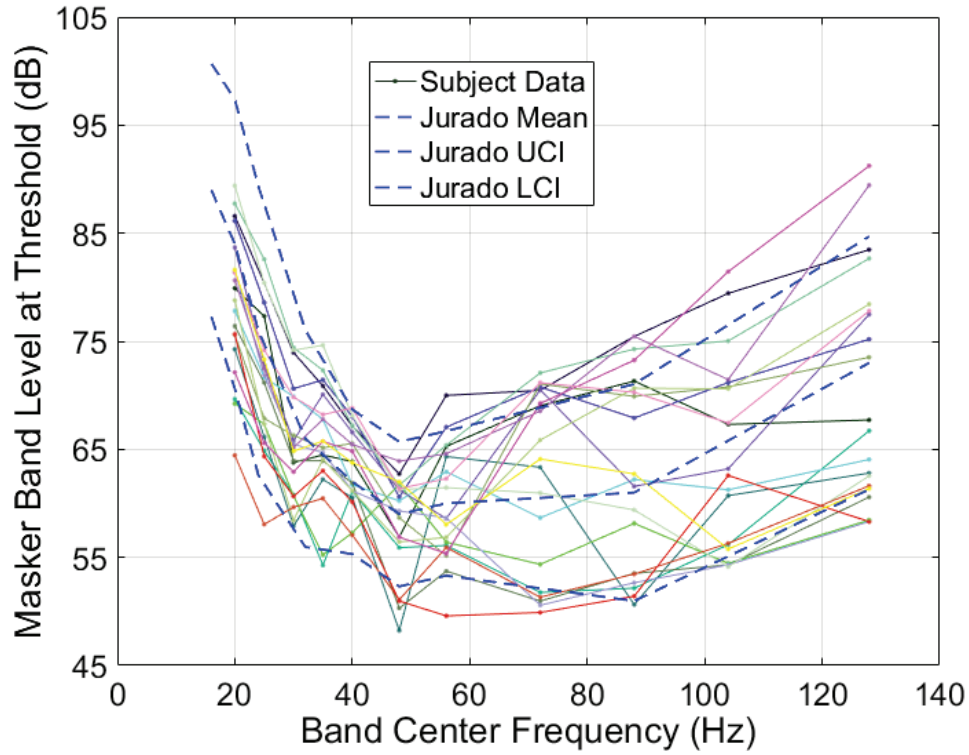
Figure 13: PTCs for 20 subjects (Experiment C) and similar data (mean and its 95% lower and upper confidence intervals) reported by Jurado [11].
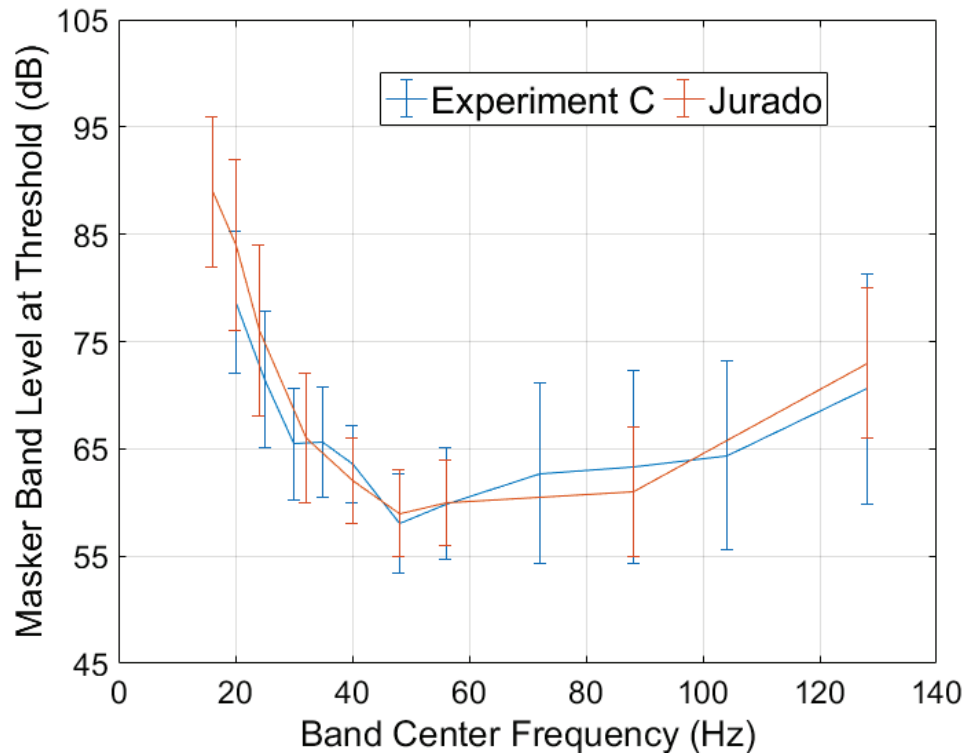
Figure 14: Mean PTC for Experiment C and similar data by Jurado [11].
Error bars for both curves represent ± 1 standard deviation.

Table 5: Equivalent rectangular bandwidths for "lowest" auditory filter
from Experiment C and from Jurado [10, 11]. Signal frequencies (Fs) from Jurado.

Equivalent Rectangular Bandwidth (Hz)		
Experiment C	Jurado [11]	Jurado [10]
34.9	39.7 (Fs = 50 Hz)	34.9 (Fs = 50 Hz)
	51.9 (Fs = 40 Hz)	

As discussed in the experimental design, the use of two different methods for the determination of the upper and lower skirts of the PTC was of some concern due to potential nonlinear auditory filtering effects. Pure tone thresholds derived in a "quiet" background in Experiment A can be compared to masked thresholds from Experiment C in order to estimate the sensation levels that were used to determine the lower skirt of the PTC. Figure 15 presents the mean and standard deviation of the measured thresholds for the 20 test subjects that participated in both experiments. It is clear that the broadband masker used in Experiment C resulted in elevated thresholds relative to those measured in "quiet." The sensation levels of the masked thresholds vary from about 13 dB at the lowest frequency (25 Hz) to 25 dB at 40 Hz. Recall that the upper skirt of the PTC was derived using a 40 Hz tone at a fixed sensation level of 16 dB. It is thus possible that the small differences in the shapes of the PTC reported herein and by Jurado are due to nonlinear filtering effects. It is also quite probable that the small differences are simply experimental error.
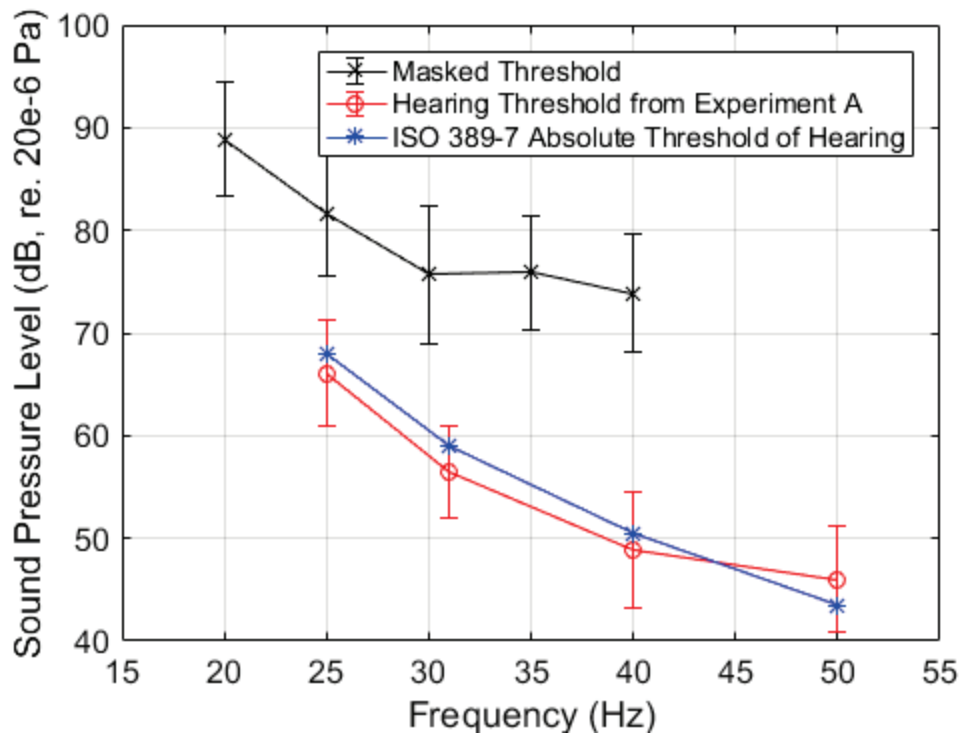
Figure 15: Mean masked thresholds for points on lower skirt of the lowest PTC (experiment C) and mean thresholds in 'quiet' (Experiment A). Error bars represent ± 1 standard deviation. The median absolute threshold of hearing according to ISO 389-7 [13] is also shown.

Having characterized low frequency auditory filters, it is interesting to return to the results of Experiment A in which pure tone thresholds were determined in the EER under "quiet" conditions (Figure 7). Comparisons with the ISO standard led to the observation that ambient noise in the EER was affecting

16

measured values at frequencies above about 40 Hz. Also, recall that the thresholds at the lower frequencies, unaffected by the ambient noise, were consistently ~2 dB below the ISO values. A detailed analysis (Appendix C) shows that the ambient noise in the EER, when passed through the 63 Hz filter derived in Experiment B, was consistent with the observed deviation of the measured threshold of a 63 Hz tone from the expected value of ~ 2 dB below the ISO standard at that frequency. A similar analysis for the 50 Hz tone threshold and the "lowest" filter had a similar result; the level of the ambient noise was consistent with the observed deviation of the tone threshold from its expected value. This finding also means that the ambient noise in the EER is expected to have no significant effect on the results of Experiments B and C because the signal levels were well above the ambient noise levels within the auditory filters.

In summary, it is clear that the results from this experiment are in substantial agreement with those reported by Jurado, despite the different method used to determine the lower skirt of the filter.

6   Conclusions

A series of experiments were performed in the Exterior Effects Room to characterize auditory filters at low frequencies, defined as below about 100 Hz. One test subject at a time was seated close to a subwoofer loudspeaker and psychoacoustic measurements were made using an adaptive three-interval, three-alternative forced choice test procedure. The first experiment determined auditory thresholds, measured under "quiet" conditions, for 36 test subjects. The measured thresholds were consistent with the international standard and enabled the influence of the background noise in the test facility to be assessed. This experiment served as a "shakedown" test to explore the performance of the test facility and the implementation of the testing methods and protocols. It also provided a screening of test subjects' performance prior to their participation in subsequent experiments.

The second and third experiments were designed to determine auditory filter shapes. The shape of the filter with a center frequency of 63 Hz was measured for 10 test subjects and found to be consistent with published data acquired using a very similar test design. It is noteworthy that the average filter shape and the variation amongst test subjects, expressed in terms of standard deviation, were both very similar for the two data sets.

Twenty subjects participated in the third experiment to characterize the "lowest" auditory filter. This experiment utilized a hybrid design in which two different test methods were employed; one to determine the upper filter skirt shape and another for the lower skirt. A test frequency common to both allowed the two sets of data to be joined together to yield a composite filter shape. The resultant filter was found to be very similar to published data acquired using a different test method, both in terms of the average filter shape and the observed variation amongst test subjects. Based on the two data sets, it can be concluded that the lowest auditory filter has a peak response (tip frequency) at about 48 Hz. The auditory system is unable to center a filter at a lower frequency and thus sound below 48 Hz is perceived via the lower skirt of this lowest filter. This has important consequences, particularly for the prediction of the audibility of a low frequency signal in the presence of ambient noise. This lowest filter is highly asymmetrical and its bandwidth is far wider relative to its tip frequency than filters at higher frequencies.

These experiments were conducted in a facility that was designed for other purposes but which proved adequate for the task. A relatively large number of subjects was tested. This prompted the design of a test protocol that could measure a single PTC for an individual subject in a short period of time (approximately one hour). Also, unlike the typical selection of young, healthy test subjects, those employed in these experiments represented the broader community, and had no discernable differences in low frequency hearing characteristics.

7   Acknowledgments

8   References

[1]   "Information on levels of environmental noise requisite to protect public health and welfare with an adequate margin of safety," U.S. Environmental Protection Agency, 550/9-74-004, March 1974.

[2]   "Acoustics – Description, measurement and assessment of environmental noise – Part 1: Basic quantities and assessment procedures," International Organization for Standardization, ISO 1996-1, 2016.

[3]   Fletcher, H., "Auditory patterns," *Reviews of Modern Physics*, Vol. 12, No. 47, pp. 47-65, 1940.

[4]   Hawkins, J.E.J. and Stevens, S.S., "The masking of pure tones and of speech by white noise," *The Journal of the Acoustical Society of America*, Vol. 22, No. 6, pp. 6-13, 1950.

[5]   Zwicker, E., "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America*, Vol. 33, No. 2, pp. 248, 1961.

[6]   Patterson, R.D., "Auditory filter shapes derived with noise stimuli," *The Journal of the Acoustical Society of America*, Vol. 59, No. 3, pp. 640-654, 1976.

[7]   Moore, B.C.J., Glasberg, B.R., and Baer, T., "A model for prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, Vol. 45, No. 4, pp. 224-240, 1997.

[8]   Greenwood, D.D., "Auditory masking and the critical band," *The Journal of the Acoustical Society of America*, Vol. 33, No. 4, pp. 484-502, 1961.

[9]   Zwicker, E., Flottorp, G., and Stevens, S.S., "Critical band width in loudness summation," *The Journal of the Acoustical Society of America*, Vol. 29, No. 5, pp. 548-557, 1957.

[10]   Jurado, C. and Moore, B.C.J., "Frequency selectivity for frequencies below 100 Hz: Comparisons with mid-frequencies," *The Journal of the Acoustical Society of America*, Vol. 128, No. 6, pp. 3585-3596, 2010.

[11]   Jurado, C., Pedersen, C.S., and Moore, B.C.J., "Psychophysical tuning curves for frequencies below 100 Hz," *The Journal of the Acoustical Society of America*, Vol. 129, No. 5, pp. 3166-3180, 2011.

[12]   Faller II, K.J., Rizzi, S.A., and Aumann, A.R., "Acoustic performance of a real-time three-dimensional sound-reproduction system," NASA TM-2013-218004, June 2013.

[13]   "Acoustics – Reference zero for the calibration of audiometric equipment – Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions," International Organization for Standardization, ISO 389-7:2005, 2005.

[14]   "Reference threshold of hearing at 20 Hz and 18 000 Hz under free-field listening conditions and at 20 Hz under diffuse-field listening conditions," International Organization for Standardization, ISO 389-7:2005/Amd 1:2016, 2016.

[15]   Levitt, H., "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical Society of America*, Vol. 49, No. 2B, pp. 467-477, 1971.

[16]   Hacker, M.J. and Ratcliff, R., "A revised table of d' for M-alternative forced choice," *Perception & Psychophysics*, Vol. 26, No. 2, pp. 168-170, 1979.

[17]   "Acoustics - Statistical distribution of hearing thresholds of otologically normal persons in the age range from 18 years to 25 years under free-field listening conditions," International Organization for Standardization, ISO 28961:2012, 2012.

[18]   Baker, R.J. and Rosen, S., "Auditory filter nonlinearity across frequency using simultaneous notched-noise masking," *The Journal of the Acoustical Society of America*, Vol. 119, No. 1, pp. 454-462, 2006.

[19]   Gescheider, G.A., *Psychophysics: The fundamentals (3rd ed.)*, Lawrence Erlbaum Associates, London, 1997.

[20]   Shmueli, G., "To explain or to predict?," *Statistical Science*, Vol. 25, No. 3, pp. 289-310, 2010.

[21]   Glasberg, B.R. and Moore, B.C.J., "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds," *Journal of the Audio Engineering Society*, Vol. 53, No. 10, pp. 906-918, 2005.

[22]   Macmillan, N.A. and Creelman, C.D., *Detection theory: A users guide (2nd ed.)*, Lawrence Erlbaum Associates, London, 2005.

[23]   Swets, J.A., *Signal detection and recognition by human observers (Revised)*, John Wiley & Sons, New York, 1988.

[24]   Green, D.M. and Swets, J.A., *Signal detection theory and psychophysics*, John Wiley & Sons, New York, 1966.

[25]   Schlauch, R.S. and Rose, R.M., "Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency," *The Journal of the Acoustical Society of America*, Vol. 88, No. 2, pp. 732-740, 1990.

[26]   Green, D.M., "Stimulus selection in adaptive psychophysical procedures," *The Journal of the Acoustical Society of America*, Vol. 87, No. 6, pp. 2662-2674, 1990.

[27]   Box, G.E.P., Hunter, W.G., and Hunter, J.S., *Statistics for experimenters*, John Wiley & Sons, New York, 1978.

[28]   Kuehler, R., Fedtke, T., and Hensel, J., "Infrasonic and low-frequency insert earphone hearing threshold," *The Journal of the Acoustical Society of America*, Vol. 137, No. 4, pp. EL347-EL353, 2015.

[29]   Kurakata, K., Mizunami, T., and Matsushita, K., "Percentiles of normal hearing-threshold distribution under free-field listening conditions in numerical form," *Acoustical Science and Technology*, Vol. 26, No. 5, pp. 447-449, 2005.

[30]   Hartl, J. and Christian, A., "A real-time method for computing confidence intervals for trial-based psychoacoustic tests," *The Journal of the Acoustical Society of America*, Vol. 143, No. 3, pp. 1812, 2018.

[31]   Montgomery, D.C., *Design and analysis of experiments (3rd ed.)*, John Wiley & Sons, New York, 1991.

[32]   Madigan, R. and Williams, D., "Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations," *Perception & Psychophysics*, Vol. 42, No. 3, pp. 240-249, 1987.

[33]   Fastl, H. and Zwicker, E., *Psychoacoustics: facts and models (3rd ed.)*, Springer, Berlin, 2006.

[34]   Moore, B.C.J., "Frequency analysis and masking," In *Hearing (2nd ed.)*, Academic Press, Inc., San Diego, 1995.

[35]   Buus, S., "Auditory masking," In *Handbook of acoustics*, Crocker, M.J., Ed., John Wiley & Sons, Inc., Hoboken, NJ, pp. 1147-1165, 1995.

[36]   Wright, B.A. and Dai, H., "Detection of unexpected tones with short and long durations," *The Journal of the Acoustical Society of America*, Vol. 95, No. 2, pp. 931-938, 1994.

[37]   Stillman, J.A., "A comparison of three adaptive psychophysical procedures using inexperienced listeners," *Perception & Psychophysics*, Vol. 46, No. 4, pp. 345-350, 1989.

Appendix A   Test Instructions and Informed Consent Form

Test Instructions

This study is designed to investigate human capability to detect sounds. This test consists of three sessions, each of which lasts between 15 and 20 minutes. During each test session, you will be presented with many test trials. Each trial will involve 3 time intervals (A, B and C) during which you may or may not hear a sound. The screen on your tablet will display the letters A, B, and C in black. These letters will turn red, in order, to indicate each time interval as it occurs. The figure below shows as an example of how the screen will appear during time interval "B." Among these three sounds <u>one</u> will include an additional sound.

Following the three time intervals, you will be shown a screen on your table that looks like:

Please indicate which interval contained the sound

| A | B | C |
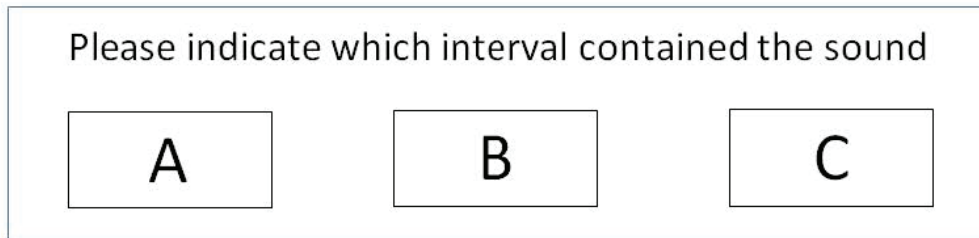
Please touch the box that contains the letter of the interval that contained the sound. After you have made your selection, the tablet will show another screen indicating whether or not your selection was correct. This process will be repeated throughout the test session many times.

Before starting the actual test, we will play some sample sounds so you can hear the types of sounds that we are asking you to listen for. You will then complete a practice session in which you answer these questions while the Test Director is with you in the room. This will let you become familiar with using the tablet computer to indicate your responses and allow you to ask any questions you may have about the test.

Thank you for your help with this investigation.

Human Subjects Informed Consent Statement

Principal Investigators/Phone:

Federal regulations require researchers to obtain signed consent for participation in research involving human subjects. After reading the information and the Statement of Consent below, if you wish to consent, please indicate so by signing this form. Also initial each of the previous pages (pages 1-3) of the form in the spaced provided at the bottom of the pages.

I.   **Statement of Procedure:**

Thank you for your interest in this research. By this time you have had the experimental rationale and procedures discussed with you in detail. You will find a summary of the major aspects below, including the risks and benefits of participating. Please feel free to ask any questions about the procedures at any time. Carefully read the information provided below. IF YOU WISH TO PARTICIPATE in this study, please sign your name and date the form in the space provided. Also please initial each page. Any information you provide will be maintained in strict confidence to protect your privacy.

II. **I understand that:**

- This is a research experiment.
- I will be participating in an experiment designed to evaluate the characteristics of noise.
- This study will be performed in the Exterior Effects Room, Building 1208, at NASA Langley Research Center.
- Audiograms will be taken before and after the test, and I will be informed whether I meet the program requirements.
- Noise exposure monitoring at the test facility will confirm that the noise exposure to test subjects is limited to levels that comply with NASA Noise Exposure Requirements (NPD 1800.1 c, 4.8.3.3, Table 1) and the allowable limits defined by the Occupational Safety and Health Administration (OSHA) regulations associated with this research testing.
- I will receive instructions and training on the task I am to perform. I will be allowed time to familiarize myself with the task prior to starting the experiment.
- During the course of the experiment, I will provide my assessment of the sounds I hear using a tablet computer.
- I have been briefed on the reasonable risks involved with the testing procedures. I will adhere to the safety requirements as specified by the Principal Investigator, and I will obey any and all regulations.
- The duration of my participation will require approximately 3 hours of my time including hearing tests at the start and the end of the test. There may be up to 6 sessions in the test, each lasting about 20 minutes. I will be given a break between sessions. In each session I will be presented with sets of sounds and asked to evaluate their differences. Each set will last about 5 seconds.
- Federal representatives are eligible to review research records for the purposes of protecting human subjects.
- I may contact the Principal Investigator listed at the beginning of this document if I have any questions regarding this experiment before or after my participation.

Page 1 of 4

III. Compensation

There are no costs associated with participating in this study. If I am not a U.S. Federal Government civil servant, and if the results of the initial hearing test I take confirms my eligibility to participate as a test subject, I will be compensated by receiving $75 for each day of my participation. I will not receive a separate reimbursement for travel.

I understand that:

- If I am a NASA civil servant, I am participating in the research in my official capacity. I understand that "official capacity" means that my supervisor is aware of my participation as a test subject in this research, and has approved of my participation in my official capacity. If I am injured during the course of this research, I am eligible for compensation through the Federal Workers Compensation System. For additional information, I may contact the NASA Shared Services Center at (XXX) XXX-XXXX.
- If I am a non-civil servant volunteer injured as a result of participating in the research, I am eligible to file a claim under the Federal Tort Claims Act by filing Standard Form 95. For additional information, I may contact the LaRC Office of the Chief Counsel at (XXX) XXX-XXXX.
- If I am a NASA contractor employee participating in the research as part of official duties as approved by my appropriate contractor manager, in the event of injury, I may be eligible for workers compensation. I may contact my company's human resources office for additional information.
- If I am not a NASA civil servant or NASA contractor participating in this research in my official employment capacity, I am participating as a subject through the NASA XXXX contract with YYYY Corp, and, in the event of injury, insurance coverage is provided to me as a research subject volunteer under the NASA XXXX contract. For additional information, I may contact X.YYYY at (XXX) XXX-XXXX.

IV. Potential Risks

- There is minimal risk associated with participation in this study other than that associated with or encountered in everyday life, however, I may request to stop the experiment at any time. A pre-test and post-test audiogram is performed to verify that no change in hearing ability has resulted from the test.
- In the unlikely event that I am injured or otherwise experience discomfort while at NASA Langley, I may visit the on-site Occupational Health Clinic. The Clinic has hours of operation from 7:00 a.m. to 3:30 p.m. The clinic number is (XXX) XXX-XXXX. Emergency medical personnel and ambulance service is also available within a mile of the test facility to transport me to nearby health care providers.
- If I have questions about the research and my rights should I experience any injury, I may contact the Principal Investigator listed at the beginning of this document.
- I will hear noises no greater than those experienced on a daily basis from heavy traffic or other sources. I will experience a maximum A-weighted sound level of no more than 95 decibels (dBA). A noisy home vacuum cleaner or lawn mower will normally equal the 95 dBA level. A automatic volume limiter system will automatically stop playback in the unlikely event that signals exceed a safe level.

V. Potential Benefits

I understand that:

- Except for being able to assist NASA and receiving information about my hearing ability, I will derive no direct benefit from participation in this study. I may request copies of the screening and test day audiograms, however these will be made available only after review by the audiologist and may take up to 30 days to process.
- The results of my participation may advance the development of a new simulation tool that will allow the assessment of human response to aircraft sounds.

VI. Confidentiality

- The test will be monitored on closed circuit television and audio intercom, and this visual and audio monitoring data will not be recorded. Some personal data may be collected, but my name will not be associated with these collected data.
- Any information obtained about me from the research will be kept strictly confidential. Data from the research could be used in reports, presentations, or publications, but I will not be identified.
- Records of my participation will be kept confidential by encoding them with subject identification numbers. There will be approximately 40 individuals participating in the test.
- Any research data stored electronically will not be confidential, but I will not be personally identified in any event. The results of the hearing tests will be kept confidential, maintained in the files associated with this research, and stored in a secured location.
- I understand that any public release of data obtained from my participation in this study will be done in a manner that does not associate me with the data.

VII. Voluntary Participation

I understand that taking part in this study is voluntary. I may withdraw from participating or be asked to withdraw from participating at any time. Such a decision will not result in any penalty or loss of benefits to which I may otherwise be entitled.

VIII. Safety

As a voluntary test subject participating in this research, I understand that:

- NASA is committed to ensuring my safety, health, and welfare plus the safety and health of all others involved with this research.
- I should report any accident, injury, illness, and changes in my health condition, hazards, safety concerns, or health concerns to the Principal Investigator listed at the beginning of this document. If I am unable to reach the above named individuals or am not satisfied with the response I receive, I should contact the LaRC Safety Office at (XXX) XXX-XXXX or the Chairperson of the LaRC Institutional Review Board, Mr. Y.XXXX, at (XXX) XXX-XXXX.
- If I detect any unsafe condition that presents an imminent danger to me, or others, I have the right and authority to stop the activity or test. In such cases the Principal Investigator and associated research personnel will comply with my direction, stop the activity, and take action to address the imminent danger.

IX. Statement of Consent:

- I certify that I have read and fully understand the explanation of procedures, benefits, and risks associated with the research herein, and I agree to participate in the research described herein. My participation is given voluntarily and without coercion or undue influence, and I also voluntarily consent to sharing the data collected during my test session, as long as my identity is not disclosed. I understand that I may discontinue participation at any time. I have been provided a copy of this consent statement. If I have any questions or modifications to this consent statement, they are written below.

_____	_____
Participant Signature	Witness Signature
_____	_____
Participant Name	Witness Name
	NASA Langley Research Center
_____	_____
Participant Address	Witness Address
	Hampton, VA 23681
_____	_____
Participant City, State, & ZIP	Witness City, State, & ZIP
_____	
Participant Age	
_____	
Participant Phone Number	
_____	_____
Date	Date

Appendix B   A Treatise on Test Design Considerations for Experiments A, B, and C

[Contributed by Andrew W. Christian]

This appendix contains information regarding a multitude of considerations that were brought into account during the design of Experiments A, B, and C. Generally, the Exterior Effects Room (EER) is a significantly different environment than the typical facility used for fundamental psychoacoustic experiments. This puts an impetus on the researchers using that facility to review and challenge, if necessary, aspects of experimental design that appear to be canonical approaches to psychoacoustic experimentation. The aspect principally discussed here is the tradeoff made between the number of subjects that participated in the test, the precision to which each subject was tested, and the overall precision expected in the final result, which can be cast in terms of the efficiency of the facility (the time taken to achieve a given precision).

The first section of this appendix reviews general information about the facility and provides motivation for the following discussion. The second, third and fourth sections regard the above-mentioned tradeoff between the number of subjects and the precision desired overall: The second deals with determining a prediction of the irreducible variance generated by the transformed up-down staircase procedure through simulations of staircase runs. This prediction allows comparison of two important parameter settings within the staircase method in terms of the efficiency of a single staircase. This prediction is used as an input to the third section which, combined with information from other sources regarding the variance between subjects, produces estimates of the overall test efficiency aggregated across subjects. The fourth section reviews the data collected during the test in the light of these predictions, and revisits the assumptions that were made. The fifth and final section reviews other aspects of experimental design germane to the proposed experimental method.

B.1 The Facility and the Goal

It is important to first discuss attributes of the test facility that will be used,[1] especially those that differ from the typical experimental situations encountered in psychoacoustics.

First, given safety considerations at NASA, it is necessary to occupy at least 3 full-time employees during any test: two employees to administer the test who are also charged with the safety/privacy concerns of the subject, and a certified occupational hearing conservationist to administer pre- and post-test audiograms. Further, the subjects, who are members of the surrounding community, are compensated for the time they spend travelling to, and participating in the test. This time commitment is nontrivial – even getting on and off of the Langley Research Center can be a chore. This is in contrast to many academic situations, where those conducting the test are unpaid (e.g., graduate students), those participating in the test are unpaid (e.g., undergraduates (commonly music majors) receiving course credit), and all are in the proximity of, for instance, a university campus.

This situation puts an impetus in at least three facets of test design that may be absent in other studies:

1. To test an individual to their maximal single-visit usefulness during each of their visits.
2. Not to invite subjects back for more visits than necessary.
3. To have a comprehensive test plan in place that takes into consideration the time-efficiency of the facility before inviting the majority of paid subjects.

---

[1] Again, Exterior Effects Room (EER), see Section 2.1.

It is also important to recall that the final result of the testing protocol that is being developed here will be a single set of curves that are a best fit across the entire population tested. This is again contrasted to the typical goal of most psychophysical experiments, which is to produce data to gain insight into the transductive mechanisms that are being employed in perception [19]. In this way, the goal of this overall effort is predictive rather than explanatory [20]. Therefore, in contrast to most psychoacoustic experiments, this effort will be expected to make use of a larger number of subjects, more test conditions, and a smaller amount of time devoted to each condition than is typical. This produces an impetus to have *a priori* confidence in the test design that appears to be absent in the literature.

It is noted that most expositions on psychoacoustic experiments – primarily those appearing in peer reviewed journals – do not provide much rationale for the selection of the particular testing procedures used for the given test. The pertinent example being Jurado's work [10, 11], but also consider the well-respected work of Glasberg and Moore [21] in which they discuss 4 different psychoacoustic tests, employing 3 different methods, and give little to no rationale for the selection of their procedures. One hypothesis for this effect is that, at most laboratories that deal in such testing, there is a great deal of 'institutional knowledge' that guides the selection of such procedures and the designs of tests in general. The only systematic exception to this rule is typically found in literature that is focused on the development, comparison, and adaptation of general psychophysical methods to psychoacoustic applications, but even references in this arena do not provide direct guidance for the design of experiments.

B.2 The Relationship Between Staircase Attributes and Precision

Given the desire to gain an understanding of the design of experiments employing, in this case, transformed up-down staircases [15], and the noted paucity of guidance available in the literature, an effort was undertaken to perform a simulation experiment to determine the relationship between the attributes of the staircases and the resultant precision offered by the method. The experiments employed a staircase testing procedure using an x-down 1-up (xD1U) algorithm for adjusting the intensity of the target signal presented to the subject. The staircase will be made up of a series of n-alternative forced-choice trials (nAFC).[2] The goal of this section is to determine what values for x and n will create the most efficient test. A review of some general details of staircases and detection is appropriate before explaining the specifics of the simulations that were run.[3]

Some Psychophysics

First, assume that an individual has a monotonically increasing relationship between the intensity of a stimulus, for example, the gain of some tone in noise, and the probability that they have of accurately detecting that tone. This relationship is known as the underlying psychometric function (PMF), and is shown as the black curve in Figure 16. In this case, it is a cumulative normal distribution with some mean and standard deviation.[4] Here, the mean is set to 0 dB (re. level for 50% detection), and the standard

---

[2] These methods were discussed above in Section 2.4, and the reader is turned to the text by Gescheider [19] for a general overview.

[3] This approach of simulating staircases is not new. The results of this effort will be discussed in relation to published data at the end of Appendix B.3.

[4] A normal CDF corresponds to many classical and contemporary psychophysical models [19]. For example, signal detection theory predicts this as the function when the variances of the underlying signal and noise distributions are both normal and have equal variance. In practice, it is very difficult to discriminate between PMF forms using real data, especially in the absence of a microscopic theory of the detection process that would predict a particular function. Normal CDFs are often used in practice as archetypal sigmoids, which are then appropriately scaled, see [22], Chapters 11 and 12.

deviation is set to 2 dB.[5] This curve spans from 0% chance of detection for very low-level stimuli (left y-axis), to 100% chance for high level ones.

A main effect of using an *n*AFC procedure is that the PMF is compressed upward by a factor of $1/n$. The reason for this is simple: even if the subject cannot hear the tone they still have a finite chance of guessing the correct alternative. This effect is shown in the two red curves in Figure 16, which correspond to 2- and 3AFC, and are compressed upward especially on the left hand side of the plot.[6] The right-hand y-axis indicates the probability of a correct *n*AFC response for these two curves, which is distinct from the probability that the subject actually detected the tone.
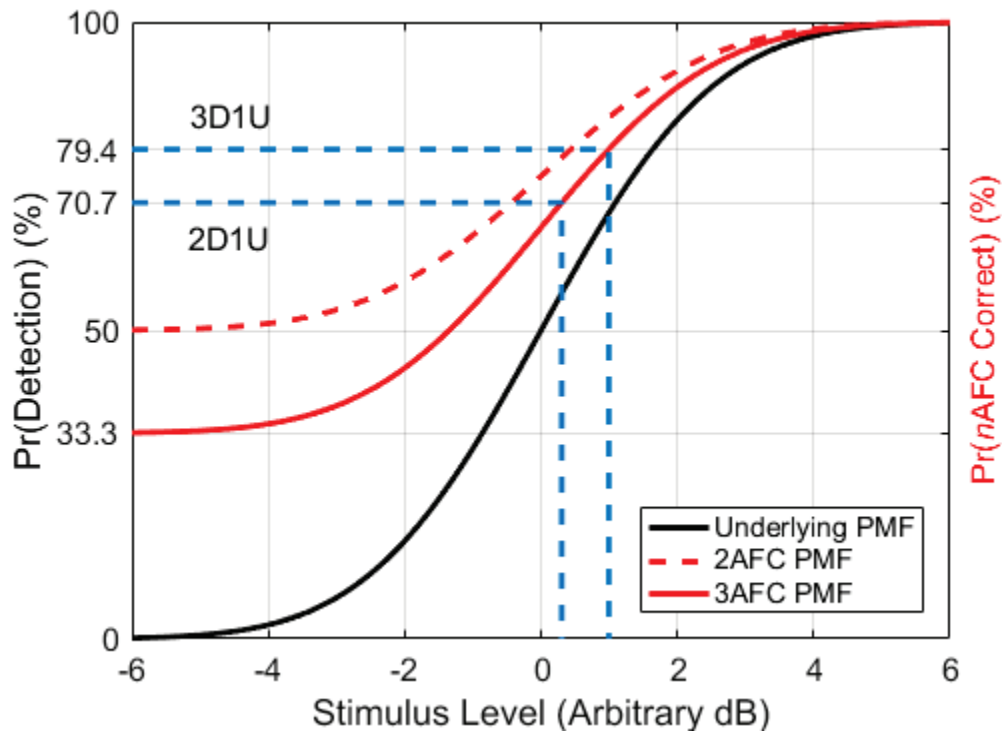
Figure 16: The relationships between the underlying PMF and the corresponding *n*AFC responses.

The response to a single *n*AFC trial is not particularly enlightening to the researcher, and the method must be combined with an overall testing paradigm that distributes many *n*AFC trials around the PMF. The transformed up/down staircase method used here has another parameter, which sets how many correct responses in a row a subject has to produce in order to reduce the level of the stimulus. Over time, the staircase will converge to a point on the PMF that corresponds to the probability of a correct response being the *x*-th root of 1/2. The staircase can be thought of as 'stepping' around the PMF to the left and right of the point of convergence (PoC). For example, a 3-down 1-up (3D1U) method will converge on the point

---

[5] This value is largely empirical. Psychometric functions for detection tasks derived from energy-based signal detection theory models tend to predict equivalent standard deviations near 4 dB. Experiments routinely produce results with functions that are approximately twice as steep, c.f., the foundational work on the subject: [23], especially chapter 27. N.B. that psychometric functions for other psychoacoustic tasks can have radically different slopes (e.g., increment detection, see [24], page 195).

[6] In the case of *n*AFC, this compression is a linear factor – a benefit of *n*AFC, and not always the case. For more complicated psychophysical tasks the compression can significantly distort the form of the PMF (see [22], Chapter 10).

where the subject's chance of detecting the stimulus is $\sqrt[3]{1/2}$, or 79.4%, while a 2D1U staircase will converge to 70.7%. Both of these PoCs are indicated in Figure 16, and their corresponding levels in dB are the levels that the experimenter is interested in determining (shown for a 3AFC task).

Clearly, setting x and n will impact the length of time a staircase takes to converge, with larger xs and ns producing longer times. However, the optimal efficiency, in terms of accuracy vs. overall time, is not clear. Other factors, such as the steepness of the PMF near the PoC, and the chance the subject has of guessing correctly can introduce variation that will result in slower convergence.[7]

In order to understand this variation, it is necessary to understand the fact that staircase methods will induce a significant amount of variance into the measurement of the PoC due to the granularity of the response. In fact, this is the case for any psychophysical method that is based on a choice between two or more alternatives (i.e., Yes/No or nAFC, as opposed to a rating on a continuous scale). As an illustration of this, consider the performance of a 'perfect' subject - one whose psychometric function does not wander over time with effects of attention/memory/fatigue/etc. This subject, when presented with multiple trials at the same stimulus level, would always give a response that was drawn from the same percentage correct, but the response itself would still be stochastic, and each trial would either result in a correct or incorrect response. Further, if presented with an inaudible stimulus, the subject would still have a 1-in-n chance of getting the correct answer, regardless of their ability to report to the experimenter that they could not in fact hear the sound. The subject would also be prone to suffer *apparent* 'lapses' in attention during times when they, by chance, produced a series of incorrect responses (perhaps after having 'gotten lucky' for a while).

If this perfect subject were to participate in a staircase experiment, the evolution of the staircase would be different every time it was run. Accordingly, the answer recovered by the experimenter would change for each staircase in accordance with the varying results, the standard deviation of that wandering would be an irreducible result of the staircase procedure, and the magnitude of that variation would be a function of the attributes of the staircase protocol. This is the variation investigated in this section. It will be discussed in terms of the staircase-induced standard deviation – that is, the square root of the variance, in units of dB – and will be termed σ_{SC}.

The underlying assumption then, is that the variance represented by σ_{SC} is much larger than the variance that will be induced by changes in the observer performance over the course of a single staircase. This assumption will be tested using the experimental data in Appendix B.4. Further, the long-term stability of psychophysical performance has been demonstrated for human observers for tasks that are considerably more arduous than the ones investigated here (see, e.g., [24], Appendix III, which demonstrates stability over a period of thousands of trials), though this is not a measure of fluctuations that may happen and be significant for tests that run on smaller time scales (of 50 or so trials).

Modeling σ_{SC}

A computer simulation experiment was performed in order to determine the relationship between x, n, and σ_{SC} for various lengths of staircases. A simulated subject was implemented in MATLAB using the underlying PMF shown in Figure 16. Individual trial probabilities were simulated using a binomial process given the probability determined by the PMF for the signal magnitude of that trial.

---

[7] Other realistic factors will further constrain what is possible for x, n, and the length of a staircase (e.g., a subject becoming fatigued). These issues are disregarded for now, and will be discussed in Sections B.3 and B.5.

The staircases were formulated similarly to those used in the EER psychoacoustic tests, with both x and n ranging from 2 to 4.[8] Each staircase was allowed to run to 150 trials – near the maximum that might be expected for a single session (cf. [24] Appendix III). A large number of staircases were run for each (x, n) condition. After the staircases were run for each condition, the value of σ_{SC} was computed for each trial as the sample standard deviation of the point estimates given by the midpoint averaging procedure discussed in Section 2.4.[9]

Figure 17 shows the results of the simulation experiment. 100,000 staircases were run for each condition in order to produce the smooth curves shown. The results in the figure are shown in terms of the estimated wall-clock time it would take to execute such a staircase. For this time estimate, the components of the nAFC paradigm were modeled as follows: 1 second for each alternative, 0.1 seconds for the pauses between alternatives, and 2 seconds total for the subject to decide, respond, and receive instruction as to whether they answered correctly or not. This timing scheme was corroborated by data from pilot testing.

The Results

Figure 17 shows that, for the most part, the combinations of x and n studied produce very similar results, except for the case of 2AFC/2D1U. All of the 3AFC procedures, and two of the 4AFC ones congregate in a corridor that is never more than 0.1 dB wide.[10] This indicates that, from a statistical point of view (again, regardless of other 'human' factors that may impact this range of procedures), the choice of which combination to use is open.

As discussed in Section 2.4, the combination settled upon for all phases of the human subject testing done here was 3AFC/3D1U. The reasons for this are manifold:

- There is no clear advantage in testing efficiency (in terms of σ_{SC}) provided by another transformed up-down staircase procedure.
- 4AFC methods are deemed to be suboptimal, as the effects of a subject's memory and attention across the 4 alternatives may play a detrimental role (see [25]). Further, higher-n procedures may be more susceptible to response bias – the tendency of human subjects to respond more accurately to stimuli presented in, for instance, the first alternative (see discussions in [22]).
- The 4D1U methods are discarded as they produce a smaller number of reversals for given wall-clock time. While this may not be an impediment for maximum likelihood data analysis techniques, as well as for simulations that employ 100,000 repetitions, it may contribute to scatter in the data for midpoint averaging analyses on realistic amounts of human-subject data.

---

[8] The initial up/down cycle of all staircases used a 2D1U rule as described in Section 2.4. It was determined through pilot testing that 3D1U was too slow for this first 'rough convergence' cycle, and that 1D1U left too much chance open that the subject would get lucky several times in a row at the beginning of the test. The latter would cause the magnitude of the target signal to drop considerably below audibility, thereby negating any effect that this 'lead-in time' would have to get the subject accustomed to the signal. This is generally good practice ([24], Appendix III). Clearly, in the simulations, this is not a consideration, though it was important to model this first cycle in order to get accurate estimations of σ_{SC}.

[9] Note that this estimate of σ_{SC} is independent of possible bias in the estimation of the PoCs. That being said, such a bias was not observed.

[10] It is expected that higher-n and -x procedures (i.e., 5AFC and 5D1U, etc.) would produce curves that started to trend back up toward the 2AFC/2D1U and inefficiency. This may be evidenced by the sub-optimal performance of the 4AFC/4D1U trace.

- The 3AFC/2D1U combination is discarded because it has been demonstrated that staircases produce more consistent results when allowed to converge to higher-intensity signals (see, e.g., [26]).
- Lastly, and perhaps only a point of comfort, this is the procedure that is used by Jurado in his experiments, against which the results of this experiment will be compared.
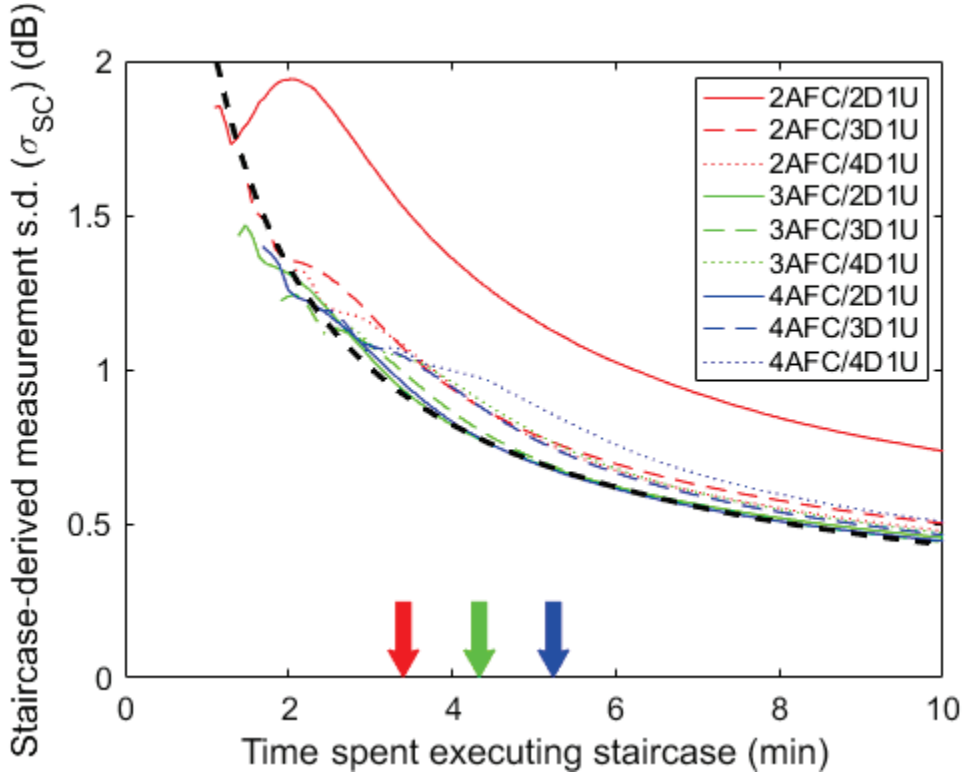
Figure 17: Results of the staircase simulation effort. The colored arrows indicate the time it takes to perform 50 trials for 3 values of n. The black dashed trace approximates the best performing procedures.

The corridor, which contains most of the traces in Figure 17, including 3AFC/3D1U, can be approximated by:

$$\sigma_{SC} = 12 / N_{\{T,3\}}^{0.7} \tag{B.1}$$

in which $N_{\{T,3\}}$ is the number of trials for a 3AFC staircase.

By plotting this equation vs. time using the estimated time for 3AFC trials from above, the black trace in Figure 17 is produced. This expression will aid in the next step of the test design phase below.

It is noted that the exponent in Eq. (B.1) is 0.7, and not 0.5 as is usually encountered (for instance, in standard error or confidence interval calculations). The likely reason for this is that Eq. (B.1) is a function of trial number, and not of the number of reversals that the staircase has achieved. The beginning of a staircase is somewhat 'wasted' with trials that are clear to the subject, so that they can orient themselves to the task. For midpoint-averaging analyses, those trials will bear no effect on the result. Therefore, there will seem to be faster convergence, but from a higher coefficient (in this case 12), for a given trial number.

This trend is expected to asymptote to $\sqrt{N}$ over the order of hundreds of trials (cf., [26], esp. Fig. 5).

The arrows in Figure 17 indicate how much time the various tasks require to generate 50 trials. Around this location is the 'knee' in the traces. It is also the region in which σ_{SC} (for the best performers) passes below 1 dB. Beyond this knee, adding more trials becomes a matter of diminishing returns.

It should be stated that these results are not readily generalizable – any change in the test protocol (for instance, changing the first reversal to 1D1U) would necessitate a resimulation in order to produce a new empirical relationship such as Eq. (B.1).

B.3 The Overall Test Design

This section concerns itself with the application of the prediction of σ_{SC} to a test that is designed with both the practical limitations of the EER facility and the desired outcome of the experiment in mind. A basic analysis regarding the size of the expected confidence intervals on the final result informs the tradeoff made in this study between the number of subjects used and the time spent testing each subject.

Confidence Intervals

An appropriate measure of the accuracy of a "grand mean" result is the size of a confidence interval (CI) upon the mean. Such a CI, formed on the experimental data, will implicitly include all of the sources of variation that will appear in the data. The typical expression for calculating a CI on the mean of normally distributed data is [27]:

$$CI = \mu \pm \frac{\sigma_{tot}t}{\sqrt{N_s}} \tag{B.2}$$

in which

- μ is the observed mean (the average of the observations)
- σ_{tot} is the known population standard deviation (SD)
- t is the quantity derived from the Student's-t distribution that encodes both the number of observations as well as the desired confidence interval [27].[11]
- N_s is the number of observations (subjects, in this case).

For this analysis, σ_{tot} is made up of two components: the σ_{SC} derived above, and an intersubject variance component that describes how subjects differ from one another (σ_{IS}). These two components add as a pythagorean sum [27]:

$$\sigma_{tot} = \sqrt{\sigma_{SC}^2 + \sigma_{IS}^2} . \tag{B.3}$$

This expected intersubject SD will be modeled with a value of 5 dB. This expectation may seem large, but it is based on several pieces of literature that all seem to indicate a comparable spread for various detection tasks undertaken by audiologically 'young and healthy' ears. This includes an ISO standard [17], a recent peer-reviewed study of very-low frequency thresholds [28], and a review of percentiles of hearing thresholds [29].

The radius of Eq. (B.2) – the second term on the RHS – can be computed for various numbers of subjects, and for various lengths of staircases using Eq. (B.1) to generate σ_{SC}. The result of this is the contour plot of Figure 18, which shows lines of equal-CI radius vs. the two independent variables, which are the parameters of the test design. The numbers that follow the lines are the dB value of the CI radius.

---

[11] This analysis will only be concerned with 95% CIs, in which case t is about 2 for large n.
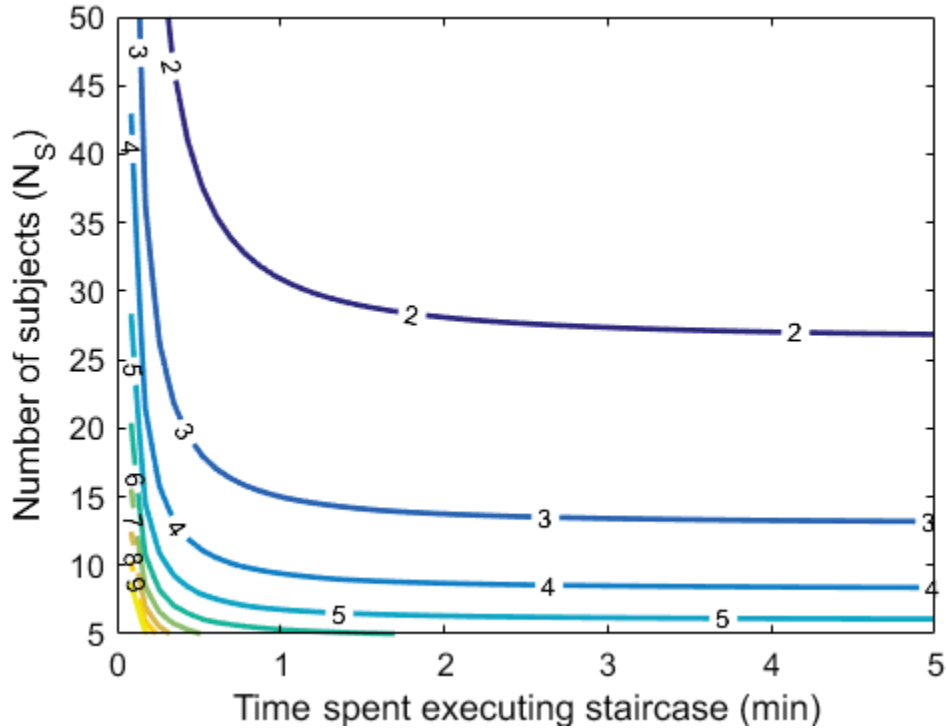
Figure 18: Contours of predicted confidence interval radius as a function of time spent executing a staircase and number of subjects.

Discussion

It can be seen in Figure 18 that there is little reason to test a single subject beyond the first several minutes for a given condition. This is evidenced by the lines of constant-CI becoming horizontal quickly as they move from left to right. This is due primarily to the fact that there is an imbalance between σ_{SC} and σ_{IS}. For instance, at 4 minutes – nominally the design point that will be settled upon below – the value of σ_{SC} is approximately 1 dB, which is dwarfed in the Euclidean sum by σ_{IS} at 5 dB. The overall implication is that producing a design that accommodates the most subjects should be the primary concern of the experimenter if the goal is the smallest possible CI on the grand mean, as long as the staircases are sufficiently long.

It is important to keep in mind that the left hand side of Figure 18 is heavily based on the approximation from Eq. (B.1), as time goes to 0. It is likely a good idea to not push the design to very short times – where the discrete nature of the staircases are liable to betray the simplicity of Eq. (B.1). Therefore, the sufficient staircase length is deemed to be 4 minutes, which corresponds roughly to 50 trials, or about 5 reversals of the staircase. As discussed in Section 2.4, some conditions are allowed to go slightly longer than this number of trials in an attempt to ensure good convergence of the staircase This was done at critical points such as the 40/40 Hz conditions in Experiment C at which the response for the upper and lower skirts are merged (as discussed in the Section 5.2).

A fixed number of trials is used, instead of an implied length criterion (i.e., either number of reversals as used by, for instance, [11], or more elaborate statistical criteria such as real-time computed confidence intervals [30]). This approach is generally supported by results in the literature, which report various problems with more exotic criteria. Using a variable stopping criterion also creates a dataset that is heterogeneous in the distribution of variance between staircases (vis-à-vis the results in Figure 17, for instance), which might complicate repeated measures and diagnostics.

Pilot testing and much of the literature reviewed here indicates that blocks of 100 to 175 trials are appropriate for testing with real human subjects due to constraints related to fatigue. In between these blocks, subjects should be allowed to take a break from the test. Using 4 minutes as a design target for individual staircases, indicating about 50 trials per-staircase, it is seen that 3 staircases will fit nicely within one session. These staircases are interleaved between test conditions (masker/signal combinations) as described in Section 2.4. Implications for this interleaving are discussed in Appendix B.5.

Given the other constraints of the EER testing protocol, a single subject's participation is approximately 2 hours per visit. About 1 hour is spent with actual testing. The first 12 or so minutes are dedicated to familiarization/practice, as discussed in Section 2.5. The remaining useful time is divided among 4 sessions of 3 staircases each, each of which are nominally 4 minutes. This gives a figure of 12 useful test conditions per visit. This combination of effort on the part of the subject is in line with recommendations given by [24], Appendix III, as well as with observations of other test protocols, where given, throughout the literature.

One major shortcoming of this approach that deserves mention is that, given that there will be a large scatter in the data for an individual subject, it will be disconcerting for an experimenter to try to judge the quality of data coming from an individual subject. The likelihood of finding data points that appear to be outliers will be high, and the auditory filter shapes for individual subjects may seem to be noisy (cf., Figure 10 and reference [13]). The pattern of the auditory filter will likely only reveal itself from the noise when averaged over many subjects, or replications by the same subject. Therefore, the application of this test protocol will require somewhat of a leap of faith as the experiment progresses, especially in its early stages. Perhaps methods can be developed in the future that allow experimenters to gain confidence that small samples of data are in fact conforming to their expectations, even when that sample appears to be of poor quality.

Relationship to Previous Results

In this subsection, the results of several sources are compared with the above conclusions and test design. This is not meant to be an exhaustive review, but a list of other sources that were influential on (and generally corroborative with) the decisions made for the test design.

- Schlauch and Rose [25] corroborate the finding that 2AFC is generally suboptimal, and 3- and 4AFC are relatively interchangeable when viewed on a time-basis in staircase procedures. This work used both computer simulations similar to the ones here, as well as a limited number of human subjects. They note similar reasons for favoring low-x and -n procedures due to response bias and memory capacity in human subjects. It is noted that they indicate (and provide other references that indicate) that probit analysis (curve fitting the PMF) reduces bias and variance in the result (relative to midpoint averaging).

- Green [26] provides a statistical argument for the use of adaptive procedures that target higher Pr(Correct) locations on the subject's psychometric function. The discussion is given in the context of more exotic adaptive methods and data analysis techniques than are used here (again, curve fitting, etc.), but are noted to be portable to staircase methodologies. He presents a considerable body of experimental data (using both naïve and experienced listeners) to corroborate numerical simulation results. This gives credence to the preference of 3D1U over 2D1U.

- In their textbook on signal detection theory, Macmillan and Creelman [22] provide hypotheses for why 2AFC had been favored in many past experiments. They echo the sentiments of Schlauch and Rose: that $n > 2$ is generally preferable statistically, and low-n is preferable due to other factors, such as response bias among the alternatives, found in human observers. They also agree with Green's statements regarding the optimal PoC, and report that other experimenters have found 3D1U to be superior to 2D1U.

The last point of comparison is with the procedures used by Jurado [10, 11], as they turn out to be strikingly similar; though the developments detailed in this Appendix were not informed directly by their design. The typical design settled on for these experiments is to target a number of trials that approximates 5 staircase reversals per PoC estimate. Jurado and Moore use 6 reversals as their stopping criterion. Further, their test protocol is to retest subjects if they provide two responses on a single element of the test matrix that varies by more than 2 dB. The difference between two samples from a normal distribution with 1 dB standard deviation will itself have a standard deviation of 2 dB. This implies that subjects will not need to be retested roughly 70% of the time (if σ_{SC} is the dominant source of variation). The difference between the two test methods then is that Jurado treats this variation as pathological – something to be treated, perhaps – whereas this approach views that variation as an inevitable component of the testing procedure. Since, by our assumptions, the expected intersubject variation is going to be dominant, the extra time taken to double test/retest is not worth the overall effort. Jurado makes no mention of how he arrived at this combination of staircase length and implied confidence, but it is hard to imagine that the similarity in the two methods is complete coincidence.

As a final word on the subject, it seems appropriate to quote from Gundy, writing in the collected works of Swets [23], Chapter 8. This comes during his conclusion regarding an experiment with similar considerations to those expressed above:

> *"Finally, it should be noted that the requirements of this investigation demand an analysis of very small segments of the performance of experimentally naïve subjects – conditions somewhat alien to the tradition of psychophysical experimentation. Under these circumstances, the results were surprisingly stable: both the average level of detectability and the variability around this average remained roughly constant throughout the experiment, and no essential changes in these results appeared when the data was pooled for analysis in larger segments."*

B.4 Checking the Assumptions

This section presents results from repeated measures that were done during part of the Experiment B (63 Hz AF) testing. The 10 subjects that participated in that portion of the test generated data that can be used to evaluate some of the above assumptions regarding the components of the variance that would impact the test design.

In Section 4, the result of the 63 Hz tone in the 63 Hz-centered band of noise is treated as a single value. This single number was arrived at by taking the mean of 3 repetitions of that test condition – two of the repetitions were part of the five extra staircases alluded to in that section. These repetitions used the same staircase parameters (50 trials each), so their results are assumed to be identically distributed. From Figure 17, the predicted σ_{SC} is seen to be around 1 dB for this number of trials.

Following the development given in Montgomery [31] (Sec. 3-5), a one-way analysis of variance (ANOVA) was performed on the repeated measures data. In this ANOVA, the 10 subjects were treated as a random effect with 3 repetitions each. The entrants of the ANOVA table were used to partition the components of the observed variance that came from within-subject effects and between subject effects. The resultant values are $s_{within} = 1.9$ dB, and $s_{between} = 3.8$ dB.[12]

Further, as it is difficult to produce confidence intervals on variance components for ANOVA analyses (ibid.), a more basic analysis was done in which the residuals from the per-subject means were treated as all coming from a single distribution. CIs are easily generated from the χ^2 distribution for the standard deviation statistic in this case [27]. This analysis gives s_{within} as 1.6 dB, with 95% confidence interval of [1.2, 2.4]. Given the method used, which does not take all of the degrees of freedom into account, this

---

[12] Here 's' is used to indicate a sample standard deviation, in contrast to the 'σ' values discussed above which are theoretical/true values.

value is expected to be low, and the CI size is expected to be small, though it overlaps with the value from the ANOVA analysis.

Discussion

Either measured value of s_{within} (1.6 or 1.9 dB) is significantly greater than the value assumed of $\sigma_{SC} \approx 1$ dB. The implication of this result is that there are more sources of variation affecting the outcome of an individual's performance on a staircase than simply the random nature of the staircase evolution. As stated in Section B.3, sources of variance add as the squares (Euclidean sum) of their standard deviations. If the true value is either 1.6 or 1.9 dB, then σ_{SC} as measured above is accounting for at least half of the variation observed, if not most of it. This is a significant result that seems to not be discussed in literature on the subject: this experiment measured listeners with little training and found that *at least half* of the observed intersubject variation is explained by the irreducible variation of the testing method. When naïve listeners are compared to experienced ones in studies, it is often found that there is not a significant difference between the performance of the two groups (cf., Gundy's quote above). This result offers an explanation – the plurality of the observed variance may simply be coming from the test methods used.

Possible sources of the variation beyond σ_{SC} include effects of learning/improvement over the course of the test day (which was the subjects' 3rd visit to the facility), and fluctuations in hearing acuity that may occur on a shorter scale than the thousand-trial stability discussed earlier (these fluctuations are typically referred to as 'lapses,' see Madigan and Williams [32]). Periodic shifts in attention could be natural, or they could be caused by the interleaving of the test conditions, as discussed and assumed benign below. The predisposition to lapse may also be a feature of some subjects and not others ([32], and references therein). It is interesting to note that, when s_{within} is computed for the 10 subjects individually, 5 subjects are seen to have $s_{within} < 1$ dB, and the other 5 subjects are seen to have $s_{within} > 2$ dB. There are no subjects for which s_{within} is between 1 and 2 dB. This may indicate that some subjects are prone to nonrepeatability, though it could as easily be a consequence of the low number of samples and/or of the idiosyncrasies of the distribution of σ_{SC}.

Finally, the measured value for the between-subject standard deviation component of 3.8 dB was smaller than the expected 5 dB assumed in Section B.3. One known effect is that masked thresholds tend to display smaller intersubject variance than unmasked ones (i.e., audiograms, minimum-audible field, etc.). Fastl reports that subject-dependent differences "almost completely disappear" when broadband (white) noise is used as a masker, but that tone-in-tone masking comes with a high degree of variability, especially for naïve subjects ([33], see Chapters 2 and 4). Perhaps this result lies in somewhat of a middle ground – the maskers used were relatively narrow bands of noise – and that accounts for the lower than expected (but nonzero) variance. It should be noted that many test conditions produced standard deviations closer to 5 dB, as can be seen in Figure 11 and Figure 14.

From the point of view of the overall design, this partition between the variance components does not produce a deleterious effect. The measured total standard deviation is 4.3 dB, as compared to the design goal of 5.1 dB. Given that Eq. (B.2) for the basic confidence interval uses only this value, and is not dependent on the partition of the variance, the confidence intervals produced by the data from this test should be largely as expected given the design. One thing to note is that the ANOVA partitioning is a subtractive process: positive error in one s-value puts a negative strain on the other [31].[13] Differences between a normal distribution (the assumption of ANOVA) and the true distribution of the staircase (which is known to be discrete, and therefore not normal) may cause a positive skew in the educed value of s_{within}, and hence a negative skew in $s_{between}$, while leaving the overall observed variance unchanged.

---

[13] Montgomery notes that, in some cases, negative values for s can arise during this partitioning. This is generally an indication that the ANOVA model is inappropriate for the dataset in one way or another.

B.5 Other Considerations

The rest of this appendix concerns itself with two other effects, the consideration of which played a part in the design of this experiment. The discussion here may seem excessive in its treatment of effects that either do not seem to play a significant role in this experiment, or effects that may play a significant role but it is argued that they can be ignored. This section hopes to capture – in one place – information that was distributed throughout the relevant literature and took a significant amount of effort to bring together during the design of this experiment. It is hoped that this documentation may serve future researchers well in situations where answers to these questions in the form of institutional knowledge is not readily available.

Interleaving of Test Conditions

The test protocol used calls for presenting the subject with sessions of 3 interleaved staircases of different test conditions. While interleaving staircases is common practice in laboratory situations – it prevents the subject from 'figuring out' the staircase algorithm – it is usually done with two staircases that use the same test condition (combination of signal and masker). While there are some notable exceptions, such as Glasberg and Moore [21] (Experiment 4), there are also canonical reasons to think that interleaving might produce unwanted effects. This section reviews those considerations.

First, it should be noted that the interleaving of heterogeneous staircases can only be accomplished when using an nAFC task with $n > 2$. With a 2AFC task, a subject is presented with the signal at a high level at the outset of a staircase, and is expected to 'remember' this target stimulus throughout the run. If various test tones and masking conditions are mixed, it is highly possible that a subject would be able to differentiate between the two intervals, but not know what the difference was (i.e., "I know these are different, but I don't know which 'contains the signal,' so I'm still guessing 50-50."). With a 3AFC (for example) version of the same question, even though the wording of the question may be the same, a subject would be able to determine which of the 3 intervals contains the signal as long the interval is audibly distinct (it is not necessary for them to recall the target stimulus).[14]

Another consequence of using 3AFC is that the subject's decision criterion (to borrow the parlance of signal detection theory) is now a measure of the perceptual difference between the intervals, instead of, for instance, simply a measure of energy. This leads to a negation of the strict concept that subjects are responding to the energy within an auditory filter alone, and that the filter in question is the only one operating at any given moment. While this may seem unsettling, given that the premise of Experiments B and C are to measure auditory filters by making assumptions about their function as simple linear time-invariant systems (see the assumptions listed in [34]), it should be noted that this is not a new concept. Authors since the late 1980s have been aware that simple energy detection is, for the most part, not what subjects are using for their cues in these types of detection experiments, even if energy detection models provide the most consistent predictors of subject performance across a large range of detection tasks (see, for instance, Buus' discussion in Chapter 89 of [35]).

Subjects who participated in Experiment B completed both a 2- and a 3AFC version of the same test condition in order to determine whether the above effect was significant. Unfortunately, the methodology used and the amount of data collected was far below the level that would be needed to draw conclusions as to the existence (or not) of an effect. Given that many psychophysical effects of a similar nature produce

---

[14] It is important to note that this task is still nAFC, and has *not* become an 'Oddity' or 'Triangle' test (see [22], Chapter 9). In that procedure, the subject is tasked with locating the 'odd' interval, but the two equivalent intervals can either be both noise, or both signal and noise. In this way, the oddity procedure generates a significantly nonlinear relationship between the underlying PMF and that measured by the test – perhaps a complication to be avoided – and can be cumbersome to explain to subjects.

effect sizes of 1 to 3 dB (e.g., the use of feedback after each trial ([24], Appendix III)), a significant effort would be required to satisfactorily resolve this question.

A second comment on the interleaving approach is illustrated by a relatively obscure psychoacoustic task called the probe-signal method. Using this protocol, a subject is conditioned over time to expect a certain tone in some noise. By changing the frequency of the tone ('probing,' perhaps) on a very small number of trials, an experimenter can reveal perceptual boundaries that resemble auditory filters. That is, a subject can be conditioned to listen using only one auditory filter, and the performance of that subject on off-condition trials can be thought of as another way to explore the shape of these filters (the most pertinent example seems to be Wright and Dai [36], who indicate that the origin of the method dates back to the 1960s). The usage of this approach has declined over time, likely because of its obvious extreme inefficiency.

It is important to entertain the idea that subjects can become conditioned to use particular listening strategies over time. This may include auditory filter selection, but could also be extended to other components of the listening strategy. Regarding auditory filter selection, one should not expect to see a huge effect in these experiments from the interleaving of staircases, as all interleaved staircases targeted the same auditory filter. If there are more complexities to what constitutes a 'listening strategy' than auditory filter selection, there is a chance that interleaving staircases could lead to a decrease in performance. However, given that the derived auditory filter shapes are the result of one condition measured *relative* to another, it seems unlikely that this would have a large effect on the results, as this effect would have to impact one test condition differentially to do so.

In summary, given the discussion of both effects above (2- vs. 3AFC, and subject conditioning), added with the fact that Jurado's results agree with those shown in this work, it is likely that the effect of interleaving test conditions played a minor role in these tests.

Subjective Naiveté

It is important to review the topic of subject experience, or lack thereof, and how it may impact the results of this test. Clearly, the test design calls for the use of large numbers of subjects, and a premium is being put on the time of the facility and of the experimenters, meaning that training of an individual subject will be minimal.

Typical psychoacoustic testing employs subjects that go through long training periods in order to produce subjects that have some 'sufficient experience.' However, evidence in the literature is mixed on whether there is a significant effect or not (though one has to be careful as different authors' definitions of 'inexperience' and 'naïve' varies). In cases where it is, the magnitude of this effect is on the order of 1 dB, both in an absolute sense and in an addition to the observed variance between blocks of trials.[15] In the latter case, this should be a benign effect given the expected intersubject variation.

In terms of absolute performance, the derivation of auditory filter shapes is a relative calculation between performances on various staircases, so absolute effects, especially on the order of 1 dB, should not be an issue. Further, in the case of subjects learning listening strategies over time (i.e., over the course of the test), the Latin-square ordering of the conditions should turn this effect into a random one.

The evidence is for a similar magnitude of effect for the use of feedback to the subjects (telling them if they were correct or not after their response). In terms of an extreme naiveté, i.e., not being capable of the psychophysical task, subjects included in Experiments B and C will have already gone through the absolute threshold screening in Experiment A. Subjects displaying large variances in their response during that screening experiment were disqualified from further involvement.

---

[15] See, as an example among many, [37]. There are also discussions of related issues in both [23] and [24].

Appendix C   Analysis of deviation of measured thresholds

Based on the lowest frequencies of Experiment A (Figure 7), it can be concluded that measured tone thresholds (in the absence of ambient noise) are approximately 2 dB below the ISO standard median value. This deviation from the standard is likely due to the choice of psychophysical test method (adaptive 3AFC).

Frequencies above 40 Hz

It appears that the ambient noise is affecting the measured thresholds. For example, the expected threshold at 63 Hz is ~36 dB (ISO median value – 2 dB), but the observed threshold is ~ 41 dB. If the measured ambient noise is passed through the average auditory filter shape centered at 63 Hz (the inverse of the PTC in Figure 11), the ambient noise level in the EER is found to be 40 dB.

Based on the power spectrum model of masking in an auditory filter, and the notion that there is an "equivalent auditory system noise" (EASN) that is always present, we can write:

$$\text{Expected threshold in the absence of ambient noise } = \; 36\,\text{dB} \; = \; \text{EASN} + k \qquad\qquad \text{(C.1)}$$

in which k is $10 \log_{10}$ (S/N) necessary for detection, and

$$\text{Observed threshold} = 41\text{dB} = 10\log_{10}(10^{\frac{\text{EASN}}{10}} + 10^{\frac{\text{Ambient}}{10}}) + k = 10\log_{10}(10^{\frac{\text{EASN}}{10}} + 10^{\frac{40}{10}}) + k \qquad \text{(C.2)}$$

Solving Eqns. (C.1) and (C.2) yields a value of 36.5 dB for EASN and a value of -0.5 dB for k. A k value near zero is not unexpected since published values of k [10] range from zero to -2 or -3.

A similar analysis can be performed at 50 Hz. The measured threshold was 46 dB and the expected threshold is 41.5 dB (ISO median value – 2 dB). When the EER ambient noise spectrum is passed through the average 50 Hz auditory filter (Figure 14), the noise level is found to be 45.5 dB. Applying the same procedure as described above yields a value of EASN of 44.5 dB and a value of -3 dB for k.

Lower Frequencies

At a lower frequency, e.g., 31 Hz tone, the ambient in the lowest band is unchanged (45.5 dB), but the expected threshold (and also EASN) is much higher, i.e., EASN >> ambient noise. Thus, ambient does not affect the measured threshold.

REPORT DOCUMENTATION PAGE

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
1-01-2019	Technical Memorandum	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Characterization of Low Frequency Auditory Filters	
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Rafaelof, Menachem; Christian, Andrew W.; Shepherd, Kevin P.; Rizzi, Stephen A.; Stephenson, James H.	
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
	664817.02.07.03.02.03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
NASA Langley Research Center
Hampton, VA 23681-2199	L-20983

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
National Aeronautics and Space Administration
Washington, DC 20546-0001	NASA
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
	NASA-TM-2019-220120

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified-
Subject Category 71
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The purpose of this study is to characterize auditory filters at low frequencies, defined as below about 100 Hz. Three experiments were designed and executed. The first experiment measured 36 subjects' hearing threshold for pure tones (at 25, 31.5, 40, 50, 63 and 80 Hz) in "quiet" conditions. The second and third experiments measured psychophysical tuning curves (PTCs) that describe auditory filters with center frequencies of approximately 63 and 50 Hz. The latter is assumed to be the lowest (bottom) auditory filter; thus, sounds at frequencies below about 50 Hz are perceived via the lower skirt of this lowest filter. Measured PTCs were found to be very similar to other recently published data.

15. SUBJECT TERMS

Audibility; Auditory filters; Human aural detection

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk (email: help@sti.nasa.gov)
U	U	U	UU	49	19b. TELEPHONE NUMBER *(Include area code)* (757) 864-9658