HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communication Engineering

Toni Hirvonen

Headphone Listening Test Methods

This Master's Thesis has been submitted for official examination for the degree of Master of Science in Espoo on September 10th, 2002

Supervisor of the Thesis Professor Matti Karjalainen

Instructor of the Thesis Markus Vaalgamaa, M.Sc.

HELSINKI UNIVERSITY OF TECHNOLOGY
ABSTRACT OF MASTER'S THESIS

Author: Toni Hirvonen

Name of the Thesis: Headphone Listening Test Methods

Date: September 10th, 2002 **Number of Pages:** 104

Department: Department of Electrical and Communications Engineering

Professorship: S-89 Acoustics and Audio Signal Processing

Supervisor: Professor Matti Karjalainen

Instructor: Markus Vaalgamaa, M.Sc.

This thesis introduces three subjective listening tests conducted to gain knowledge on listening test methods involving headphones. The purpose was to gain general understanding of the subject and also to find answers to more specific problems.

The possibility of simulating real-life devices with recorded and processed sound samples is an interesting possibility that could facilitate the test procedure. An attempt at this simulation was made here by utilizing artificial head recording and compensated headphone reproduction.

The test results showed significant differences between the simulation and the actual situation. The outlook, ergonomics etc. of the headphones had an effect to the sound quality evaluation. Thus the simulation method was not validated.

One of the goals was also to link objective measurements to the test subjects' preference of the devices. The flatness of the diffuse-field response seems to correlate somewhat with the subjective preference of the headphones.

In addition, commercial music as well as wideband and narrowband speech were investigated for their relationship in sound quality evaluation.

Keywords: Sound quality, headphones, listening tests, timbre

TEKNILLINEN KORKEAKOULU
DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Toni Hirvonen		
Työn nimi:	Kuuntelukoemetodit Kuulokkeilla		
Päivämäärä:	10.9.2002	**Sivumäärä:**	104
Osasto:	Sähkö- ja tietoliikennetekniikan osasto		
Professuuri:	S-89 Akustiikka ja signaalinkäsittely		
Työn valvoja:	Professori Matti Karjalainen		
Työn ohjaaja:	DI Markus Vaalgamaa.		

Diplomityö käsittelee kolmea subjektiivista kuuntelukoetta, joissa tutkittiin kuulokkeisiin liittyviä kuuntelukoemenetelmiä. Tavoitteena oli sekä ymmärtää aihetta yleisesti että tutkia tiettyjä kysymyksiä tarkemmin.

Todellisten laitteiden simulointi prosessoiduilla nauhoituksilla on kiintoisa mahdollisuus jota soveltamalla voitaisiin helpottaa kuuntelukoejärjestelyjä. Näissä kokeissa yritettiin tälläistä simulointia käyttämällä kompensoitujen kuulokkeiden kautta soitettuja keinopäänauhoituksia.

Testin tuloksissa näkyi merkittäviä eroja todellisen tilanteen ja simulation välillä. Kuulokkeiden ulkonäkö, käyttömukavuus yms. seikat vaikuttivat niiden äänenlaadun arviointiin. Näin ollen simulaationmenetelmää ei voitu validoida.

Kokeen tavoitteena oli lisäksi löytää yhteyksiä laitteiden mitattavien ominaisuuksien ja koehenkilöiden subjektiivisen preferenssin välillä. Kuulokkeiden diffuusikenttävasteen tasaisuuden havaittiin korreloivan jossain määrin subjektiivisen preferenssin kanssa.

Kokeessa tutkittiin myös kaupallisen musiikin sekä laaja- ja kapeakaistapuheen suhteellisia ominaisuuksia äänenlaadun arvoinnissa.

Avainsanat: Äänenlaatu, kuulokkeet, kuuntelukokeet, äänenväri

Preface

The work for this thesis was carried out at the Nokia Mobile Phones audio department and in the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology. Many capable persons from both locations offered their help and guidance throughout the whole process.

I would like to thank my instructor Markus Vaalgamaa, with whom I have worked closely to complete the thesis. He always offered ideas and help when they were needed. My teacher Juha Backman provided additional guidance and had an important contribution to the test design. I am also very grateful to my supervisor professor Matti Karjalainen.

Co-workers in Nokia deserve praise for assistance and smooth co-operation. The author collaborated with NRC SAS lab and NMP Salo acoustics lab personnel. To Gáetan Lorho from the former location and to Ossi Mäenpää from the later, I give special thanks. Nick Zacharov and Ville-Veikko Mattila form NRC Tampere offered valuable comments about the test arrangement and statistical analysis. Dr. Ville Pulkki from HUT was very helpful with practical arrangements involving the listening room used in the third test.

To all the above-mentioned and otherwise involved persons I am ever grateful for by doing this thesis, I have learned a great deal.

Espoo,

Toni Hirvonen

Table of Contents

List of Abbreviations

3GPP	3rd Generation Partnership Project
ADAM	Audio Descriptive Analysis and Mapping method
AMR-WB	Adaptive Multi-Rate Wideband codec
ANOVA	ANalysis Of VAriance
B&K	Brÿel and Kjær, manufacturer of acoustic measurement tools
DSP	Digital Signal Processing
DRP	Drum Reference Point, measurement point at the eardrum
ETSI	European Telecommunications Standards Institute
FIR	Finite Impulse Response
GSM	Global System for Mobile communications
GP2	Guinea Pig 2, listening test software
HATS	Head And Torso Simulator
HRTF	Head-Related Transfer Function
HUT	Helsinki University of Technology
ILD	Interaural Level Difference

IIR	Infinite Impulse Response
ITD	Interaural Time Difference
ITU-R	International Telecommunications Union, Radiocommunication sector
ITU-T	International Telecommunications Union, Telecommunication sector
MIDI	Musical Instrument Digital Interface protocol
MOS	Mean Opinion Score
MP3	Mpeg Layer 3
NMP	Nokia Mobile Phones
NRC	Nokia Research Center
PTF	HeadPhone Transfer Function
STI	Speech Transmission Index
SNR	Signal-to-Noise Ratio
THD	Total Harmonic Distortion

1. Introduction

The future holds interesting things for acoustics. The mobile phone industry is one of the areas in constant development; the arrival of the next generation standards doubles the bandwidth used in speech transmission and this alone presents new requirements for the devices. The cellular telephone is no longer seen as a mere speaking apparatus. MP3, radio, MIDI and sampled ring tones, games and other applications elevate the mobile device to the status of an entertainment system.

Constant evolution and increasing complexity make it hard to determine the *subjective quality* of these devices. Manufacturers want to know the reasons behind the personal preferences and perceived attributes of the customers. This is where *subjective testing* comes in. From acoustic point of view, the researcher performs *listening tests* for a group of *test subjects*. Ideally, when the researchers know all the *variables* that control the subjective experience of the customer, they can tell the designers how to modify the product in a desired manner. In practice, finding correlation with subjective test results and objective measurements is not an easy task.

1.1. Creating Simulated Listening Experiences for Listening Tests

For quite some time now, it has been a common dream of many scientists to find ways to create a virtual reality. Examining this concept merely from an acoustical point of view, the goal is to produce listening experiences as they would happen in real life. Successful simulation would eliminate the requirements for the actual sound source and the original listening environment. Ideally the listener must not distinguish the real sound source from the simulated situation. The quality to strive for in this case is naturalness. This is not necessarily same as personal preference.

Binaural theory states that this kind of authentic reproduction is indeed possible, provided that the reproduced sound pressure at the listener's eardrum does not differ from the real life sound pressure [1]. It is presumed here that the hearing experiences are not affected by other sensations, such as vision, even though this *ventriloquism* effect is sometimes evident [2]. Applications are usually examined with localization performance tests where the subject's ability to distinguish specific sound source locations with simulated sounds is compared to the performance with actual sounds. Whether this is a correct method of validation is another issue but so far localization has been the meter of authenticity for acoustical reality simulation.

When trying to simulate sound events a good starting point is to determine what causes the brain to determine the direction of the sound. The task is to find out what are the *spatial cues* that affect the listening experience. There are several binaural cues, such as the *interaural time difference* (ITD) and the *inteaural level difference* (ILD) but one important acoustic factor is the monaural *head-related transfer function* (HRTF). Determining HRTFs requires knowledge on how the subject's body shape (for example pinna, head and torso) affects the incoming sound. Several extensive studies have been made on HRTF measurements, for example [3].

One way to simulate spatial cues is to record the sound event with a human or an artificial head and use the recording with a playback device in an arbitrary location. The HRTF created with an artificial head i.e. *head and torso simulator* (HATS) is unfortunately for the time being found to be inferior to subject's own HRTF [4]. The HATS is however far more practical for recording purposes than an individual human head. One can record arbitrary sound events with it and use the recordings to give at least some illusion of spatiality. The applications of this recording technique are limitless. Especially, the idea of creating simulated test signals for listening tests has recently surfaced. Usually when testing audio devices etc. the test setup is quite extensive and difficult to move. By recording the necessary sound events with HATS the test could theoretically be reproduced at any location with minimal playback devices. This would greatly alleviate the burden usually involved with listening test arrangements.

Arguably the most convenient way for reproduction of HATS recordings is to use headphones. They offer for instance almost complete channel separation and independence of head movement. Their small size makes headphones easy to transfer. In addition, some isolation from environmental noise is also provided. There are nevertheless some issues involved with headphone listening that will be inspected in Chapter 3.

1.2. Bandwidth and Preference

As mentioned earlier, the forthcoming third generation mobile phone standard will include, among other things, an increase to speech bandwidth. Since the dawn of telecommunication the telephone has only transmitted speech in a frequency band of 0.3 – 3.4 kHz. This is usually referred to as *narrowband speech.* The bandwidth limitation causes speech to sound clearly unnatural. To remedy the situation ETSI and 3GPP have introduced a new coding algorithm, AMR-WB, to be used in third generation systems [5]. An AMR-WB codec performs coding in a frequency area of 0.05 – 6.4 kHz and adds frequencies up to 7 kHz. Thus with a typical telephone device the effective range will be approx. 0.15 – 7 kHz. This *wideband speech* is comparable to natural human voice and thereby offers significant improvement of sound quality over the old system.

There exists vast amounts of standards and recommendations that deal with measuring speech quality in telecommunication (see Chapter 2.) On the other hand, little research has been made with wideband speech. It is not entirely clear how the perceived quality, naturalness, and the intelligibility of speech are affected when the bandwidth is doubled.

A related question involves the concept of so called *preferred equalization* for given sound material. Some mobile phone models offer a group of equalization pre-sets for incoming speech. This allows users to modify the *sound color*, i.e. *the timbre* [6] according to personal preference. Some may like the warmth that emphasized low frequencies introduce while for sake of intelligibility, the middle and high-frequency area can alternatively be enhanced. Researchers in the industry are interested in finding out what kind of timbre people prefer when listening to narrowband or wideband speech. It is also interesting to compare preferences on speech material to those on commercial music material.

1.3. Scope of the Thesis

A series of listening tests and measurements were conducted to find answers to questions related to the previous discussion. The methodology of these tests will be discussed with more detail in Chapters 4, 5 and 6.

A formal specification of the problems studied in this thesis is:

- first, *to evaluate whether HATS recordings played with compensated, high-quality headphones can be used to substitute actual sound sources in listening tests.*

- second, *to study subject's preferences of sound color with narrowband speech, wideband speech, and music material*

- third, *to determine whether the preference order of the devices used in listening tests can be explained by measurable objective quantities of these devices*

- fourth, *to examine differences between music and speech and to determine if music could replace speech in listening tests.*

In the listening tests, subjects expressed their sound color preferences between devices while listening to different *sound samples*. The sound reproduction device for simulated sounds, i.e. recordings was decided to be a pair of high-end headphones. An additional idea also presented itself in the course of test planning; because music is generally speaking more interesting and entertaining to listen to than speech, why not replace speech with music in listening tests? This way the test subject could sustain interest more effectively to the listening task. Thus the fourth point was added in the list above.

1.4. Organization of the Thesis

Chapter 1 gives an introduction to the thesis. Background information and reasons why this study has been done are provided.

Chapter 2 discusses about measuring sound quality with subjective listening tests. Human characteristics as a test subject and general testing methods are also presented in a general manner.

Chapter 3 shortly introduces headphones as a special case of transducers. Issues involved with headphone listening are discussed.

Chapters 4 through 6 present the author's own work and results. The tests were done in three parts, all of which form a unity of their own

Chapter 7 gives a summary of the final conclusions and hypothesis along with suggestions for future work.

2. Subjective Testing of Sound Quality

By definition, subjective testing involves inquiring about personal experiences of an individual human. This makes the method rather laborious and in some ways more difficult than other types of measurements; especially so if a somewhat ambiguous thing like sound quality is the target of the research.

This chapter gives a summary of subjective testing in a generic manner. Various commonly-used methods are introduced and their possible shortcomings considered. Most of these methods are adaptable in general subjective testing but this thesis focuses in measuring the sound quality of audio devices. In addition, this chapter gives some ideas how to actually interpret the term “sound quality”. First however, the purpose is to present some properties of human beings as test subjects.

2.1. Human as a Test Subject

Measurement and classification of real life events is important because it makes the development and testing of theories and models possible. These models allow us to make predictions of future events and phenomena in various situations. But regardless of measurements and theories, it is impossible to predict all the factors that affect the observer’s individual experience. In consequence, there is a “gap” between objective measurements and subjective experiences.

Subjective testing is being utilized vastly in testing for example audio products. The main reason for this is that no artificial instrument or measuring device has yet matched the complex accuracy of human reception system. Although simple quantities, such as sensitivity@1kHz or *total harmonic distortion* (THD) are by no means useless, they tell little about the effect of for example a specific loudspeaker in person’s mind. The researcher can perform extensive objective measurements to a device but often a simple subjective comparison with other products will give more perspective about the sound quality. That being said, for a subjective test to have scientific value, several test subjects along with other preparations are required. This in turn means that subjective testing is a relatively resource-consuming method compared to simple objective measurements.

The reason behind using many test subjects lies in the uncertainty of humans when using them as a measuring instrument. To gain reliability, the researcher can reduce the noise in the measurements by repetition. This scientific approach is discussed further in Section 2.3. The purpose is first to deduce some possible reasons for uncertainty between individual responses of humans.

2.1.1. Human as an Individual Observer of Sound

Numerous theories about sensory reception have been introduced in the area of cognitive psychology. The most notable ones of these are summarized in [7]. A common interpretation is that observations are created by comparing incoming information to *inner models* and so creating an image of the outside world. This comparison is based on extracting features from the incoming information. The process has been described as highly interactive and inner models can supposedly change in the course of life. Some theories also involve “feedback loops” in the comparison system. In order to be able to perform a comparison, an observer needs some kind of repository for the incoming information. This function is carried out by memory which is usually divided to three parts: *Sensory, short-term* and *long-term memory* [7].

When dealing with auditory perception, the sensory memory is referred to as *echoic memory*. The “echo” of an auditory event is stored here before cognitive processing and classification take place. As an example of utilizing echoic memory, one sometimes asks a person to repeat the question just being asked and proceeds to answer before this happens. The question is tracked from the echoic memory and then processed because the *attention* of the listener was focused on something else at the time. Estimations on the length of the echoic memory vary depending on the study. A summary of these studies is presented in [8]. Based on the results, an estimate of the decay time of the echoic memory is approx. one second and the capacity is quite limited.

As an effect of attention, the information transfers from the echoic memory to the short-term memory. For example a phone number can be stored in the short-term memory for a short while if one focuses on remembering it. Short- term memory is also rather limited in time and amount of information it can preserve; even the smallest disturbance in focus can loose the information. Estimates of short-term capacity are again varying but in general, little information from it can be retrieved after 15 seconds.

Humans are also able to remember things that happened long time ago. This is explained using the concept of long-term memory, where information is transferred by *rehearsal* or via strong emotional experience. Rehearsing usually means repetition. Different theories describing long-term memory are dismissed here, except for its common division to *implicit* and *explicit memory.* The latter can be understood as a conscious attempt to retrieve information, whereas the former refers to subconscious processing. Implicit comparison can be understood as referring to the inner models. It must be emphasized that the strict division of memory to three specific blocks is merely a simplified model that has not been formally proved to be accurate.

From an auditory standpoint, the role of long-term memory is not very significant. A regular consumer rarely has reliable inner references that can be used to determine the sound quality. The reason for this is perhaps the dominant nature of vision in human reception system; hearing has not been needed as much as sight during the course of human evolution. Inner sound references have not developed properly and contingency has a large role in the outcome of an observation. One way to compensate this shortcoming is to utilize echoic memory. The test situation can be arranged so that the subject is able to compare the presented stimulus to the information received just a moment ago.

Even though generally not very evolved among average humans, the hearing resolution can greatly be upgraded by rehearsal. Musical pieces are remembered after a few times of listening and the voice of a familiar person immediately invokes associations. It is understandable from this point of view that musicians and audio professionals would theoretically be the best test subjects in listening tests. Their ears are trained to observe slight differences that often are crucial in listening tests. The term *expert listener*, as opposed to a *naïve listener*, is commonly used. This division is not universally valid but must be associated to a specific task; validity of the test subjects depends on the test itself. Experience is a difficult attribute to quantify. It has been demonstrated that musicians that are not interested in audio, are also not very good at determining sound quality of audio devices [9]. One way to determine the suitability of listeners to the task is to apply some form of pre-test *listener selection.* One this sort of method is presented in [10].

One might wonder why only certain people should be chosen as test subjects. If merely expert listeners are used in the test, is the distribution of people not incorrect being that test devices are usually intended for common usage? The modern view is that sound quality is thought to be universal in nature. The use of experts is justified by thinking them as nearly ideal observers that produce the same results as rest of the people, merely with less variation. People with experience know what to listen for and are able to analyze auditory events more precisely. ITU-T suggests in its method recommendation that especially when dealing with small differences, the test panel should consist of persons used to performing these kind of tasks [11].

One way to increase listener competence is to apply *training*. The subjects are familiarized especially with the upcoming task before the test. Training has been shown to increase listener performance significantly [12]. The scope of atraining session can vary depending on the resources and the time available. Researchers can for example merely introduce the subjects to the task by familiarizing their ears with the type of sounds used in the test. Best results are achieved with carefully prepared education material. All the sound variations in the actual test are to be made familiar during training. When training subjects for more difficult tasks, such as learning a *descriptive language* (see Section 2.4.), obviously more effort is required [13]. Certain procedures, for example audio descriptive analysis and mapping method (ADAM) by Mattila and Zacharov [13], involve both listener selection and development of a descriptive language. In any case training is useful when possible. Bech has concluded in his study that an experienced and trained listener equals seven regular listeners from a statistical standpoint [14]. This significantly reduces variation in results and thus reduces the number of subjects required. In practical smaller scale tests there is often no possibility to organize vast training sessions or listener selection. It is nevertheless recommended that the subjects are in some level familiarized with the task and finding differences between samples. *Biasing* the listeners' opinions in any way should however be carefully avoided. There are no "wrong" opinions about sound quality.

Two more details worth considering are the gender ratio and possible hearing impairments of the test subjects. ITU-R recommends that an equal number of male and female subjects should be used in subjective tests [11]. However both sexes have been shown to give very similar results when the subjects have similar social backgrounds [15]. Hearing impairments on the other hand are not a desirable quality among listeners because they only cause more noise, i.e. variation to the results.

2.1.2. Effect of Cultural Background on Perceiving Sound Quality

It was assumed in the previous section that sound quality is a universal attribute that all people agree on. It is the purpose of this section to examine and possibly disagree with the former assumption. The point is not to scientifically prove anything true or false, but rather to put forward certain issues and questions related to this area of study.

Human beings start to learn new things from the day they are born by perceiving information. This complicated process cultivates inner models in psychological, social and cultural levels through *conditioning*. Conditioning can happen in a simple sense of reward and punishment or by some more profound way. The mechanism joins individuals as a part of different communities again in many levels. The members of the group share same values and meanings. According to Sneegas, this kind of reasoning leads to the conclusion that we must separate two attributes for each other [16]:

- *perception*, which refers to individual's ability to perceive information from outside world.

- *preference,* which is affected by *urges* created by evolution and social and cultural *tendencies* depending on personal background.

Urges are presented here as qualities that are somewhat common to all humans whereas tendencies vary. Sneegas winnows these tendencies further; two main reasons for them are *fashion* and *cultural capital.* The later is a long-term property of individuals depending on upbringing and social and cultural surroundings, like described above. Fashion on the other hand is understood as rapidly, unpredictable change in values and preferences. There are number theories about fashion which are presented in [16].

So what does all this mean in terms of acoustics? It is the view of the author that during history, sound quality has been of little interest to others than audio professionals or enthusiastics. Presently, and more so in the future, "ordinary" consumers are more and more taking interest in their audio devices; for example, quite a few people own a home theater system nowadays and are somewhat familiarized with it. The concept of sound quality will perhaps change more according to fashion in the future. The effect of cultural capital is starting to show as people take more notice to the quality of audio devices they hear. Even now it is appropriate to ask whether for example 40-year old male engineers have the same sound quality preferences as 10-year old girls. Researchers doing listening tests must carefully consider the characteristics of the test subjects they use, as usually expert listeners are audio professionals who might be biased towards certain devices or sounds.

2.2. Sound Quality

So far the term “sound quality” has not been formally defined in this thesis. The concept is divided to a variety of subfields that are presented here. The division is based on a summary given by Karjalainen in [6]. Other viewpoints are also possible as there is no official or universal definition available.

The traditional acoustics has over the years been involved with *concert hall acoustics* and *noise quality.* The former has traditionally been the most respected as well as the most demanding area of the whole acoustical canvas. Even nowadays with modern methods there is no easy way to design a good concert hall. When dealing with noise quality the purpose is to diminish the sound because unwanted noise has shown to cause psychological detriment.

Speech quality must be dissociated as a completely own category because it closely involves the concept of intelligibility. It is hard to associate the same quality with for example music or noise. There are many objective measurements created to describe intelligibility, such as the STI value. The MOS value on the other hand is a more generic measure. It is used for example to determine audio codec speech quality in GSM systems.

Even more modern approach to sound quality is *product sound quality*. This means that the sound emitting from a commercial product must be integrated with the purpose of the product and serve the whole as well as possible. This does not necessarily mean that the sound level should be as low as possible; for example car engine sounds can be informative when applied properly.

This thesis focuses on the perhaps most widely known variety of sound quality, namely on *audio sound quality.* Traditionally, the abbreviation *hi-fi* (high fidelity) is affixed to audio devices. Originally, hi-fi refers to audio reproduction that is natural, i.e. similar to the actual sound sources. This definition is somewhat old-fashioned since it is presumed that there actually exists a real sound source to which reproduction can be compared. It is more suitable to think audio devices as sound transforming instruments which prepare the sound to users liking. Again it must be pointed out that naturalness, intelligibility and perception are all separate issues. Nowadays hi-fi is more like “exaggerated brilliance” rather than “pursuit for reality”. There are also some objective measurements which can be applied here, like *signal-to-noise ratio* (SNR) and frequency response but in the end they tell little about the subjective experience. Toole has proposed that in this area objective acoustical measurement techniques are the least evolved [17]. It must be always remembered that the final route of the sound consists of the device, the listener and the path between them. In conclusion, subjective experiences are very difficult to measure by other means than listening tests.

In any case, the term “sound quality” always needs some target for which it is assigned. In this thesis the word “device” is used in a generic manner, describing mainly audio equipment but also for example codecs or other products mentioned in this section. As mentioned, the range of this thesis however, is limited to audio sound quality.

2.3. Theory of Subjective Testing

As discussed earlier, there are no devices that can match the human perception system in sensitivity or accuracy. The problem now becomes how to read results from human mind. As *neuropsychological research* continues to develop new methods for monitoring brain activity, we are starting to comprehend the mysteries of the mind [7]. Sufficient to say however, that at present the operation of the neurological system remains a mystery. This is why the human perception mechanism is presented as a “black box” (Figure 1) where a *stimulus* is fed [6]. A *description* of the event is obtained at the output of the system. The actual *response* to the stimulus remains within the black box, i.e. it cannot be extracted from the system. This is of course unfortunate when measuring sound quality. The researcher has to find some way to derive the response from the (possibly inaccurate) description.

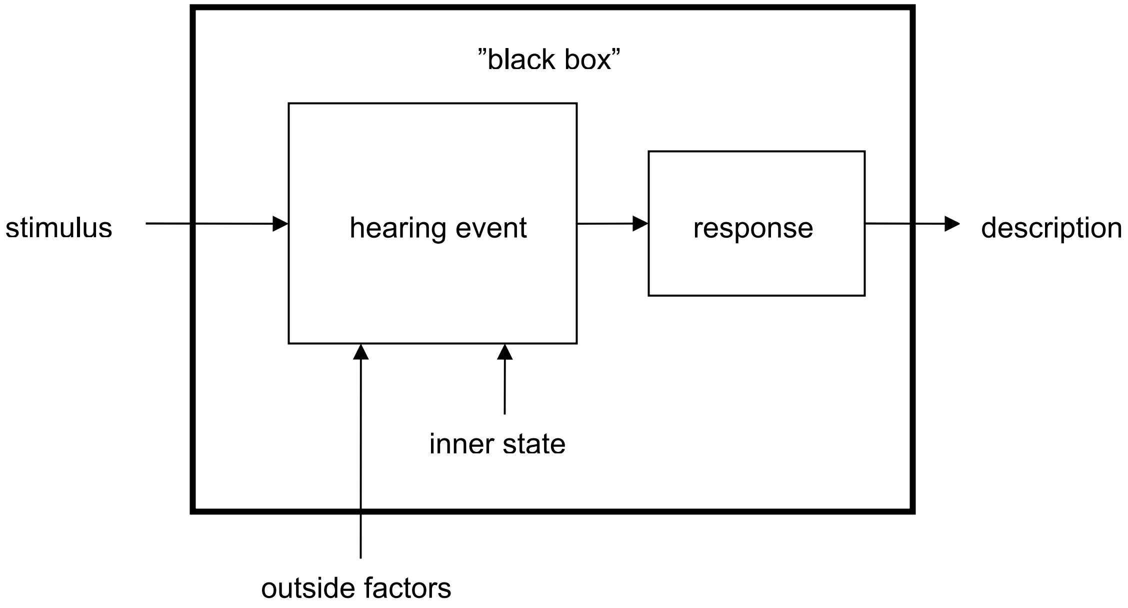


Figure 1. A simplified model of human hearing. Response can only be investigated through description. Inner state of the subject and random factors are unknown.

In Figure 1 human reception system is compared to a measuring instrument with given amount of uncertainty. When this uncertainty is random, the results can be *averaged* to obtain lower SNR. The information starts to stand out from the noise as the number of results increases. This is why scientifically valid subjective tests use several subjects. When the *confidence interval* of results is too large, the results give no useful information. ITU-R recommends that using c. 20 subjects is sufficient in simple tests examining sound quality.

In the 1980’s, subjective tests divided experts to two camps [18]: The other side claimed that subjective testing simply does not work. Obtained results are in no direct way correlated to the *audible difference*s. The ones who spoke for subjective testing believed that the results do tell something about true sound quality if the test conditions are carefully controlled and all the variables have been accounted for. This way the test is scientifically valid and useful information can be extracted from the results. If the variables are not controlled, so to speak, there is no certainty what caused the results. So the modern view is that subjective listening tests do work, if done properly. Toole proposes that an ideal listening test should produce results that [17]:

- are reproducible at different times and places, with different listeners
- reflect only the audible characteristics of the product or parameter under examination
- reveal the magnitude of audible differences or a measure of absolute values on the appropriate subjective scales.

These are the goals to strife for when planning a test. Some possible means to achieve these objectives are presented in the following sections.

2.4. Common test arrangements

The purpose of this chapter is to give some idea on how listening tests are done in practice. In the following segments more precise descriptions of test planning and execution are presented. This can be done better when some of the various test types are familiar. The scope of this thesis is to examine audio sound quality of a specific device group, namely headphones. Some of the test types presented here are not closely related to the problem at hand, but rather meant to be used for other types of measurements. It is however meaningful to give wider perspective of the topic before focusing on a narrow sector. This section is based on information given in [6] and in [19].

The simplest of all test arrangements are arguably *threshold measurements.* The task is to resolve whether given stimulus causes a specific hearing event. Threshold values are divided to two types: *Absolute threshold* measurements examine if the sound is registered at all, whereas a *relative threshold* tells if the difference between two sounds is detectable. For example, the hearing threshold is investigated with the former method. One common arrangement is the *ABX-test* where the subject is asked: “Which one of A and B is the same as X?” This is a so called *forced choice test.* The subject can also participate actively to the test by *adjusting* the sample until the difference compared to the reference is audible. It must be noted that thresholds are obtained from the results by applying statistical methods, for example taking a median of the values.

When proceeding to more sophisticated test arrangements, the simple “yes” and “no” answers are not sufficient anymore. As the number of samples to be compared increases, simple *paired comparisons* would take much time. For this reason, several samples are presented at the same time. All samples can be *ranked* by some attribute or a stimulus can be *classified* by assigning some value from a *scale* to it. *Indirect scaling* means that the values of the scale are not comparable with each other. A simple this kind of application is the *nominal scale* where stimulus is given verbal labels such as “dark” or “nasal”. When using a *direct scale*, the goal is to specify the mutual order of the samples and also the magnitude of the gaps between them. This is a more demanding task for the subjects than the previous methods. When using *numerical scales,* the point of origin, i.e. zero value, can be specified or omitted; some statistical methods require this to function properly. Common numerical scales are for example 1 – 5 or 1 – 10, with one or zero decimal accuracy. The most used MOS scale uses integer number values 1 – 5.

As discussed earlier, test subjects usually have a lot of variance between individual results. In case of numerical scales, one subject might only give *grades* from 2.0 to 4.5 while another one uses the whole scale. To remedy this situation, a number of *reference points* can be implemented by designating certain grade values to one or more items in the test. This way other samples can be compared to the reference. Numerical values can also be labeled with nominal values using *anchor points*. Table 1 shows one application of using anchor points.

The problem with nominal scales and labels is that the adjectives used might bear different meanings to different people. It is as if the subjects use different languages. In order for nominal attributes to be universal, various descriptive languages have been developed mostly in wine and food industry. This requires a special training session where the persons involved learn to use the common language, much like in usual everyday communication. Only this time the vocabulary is limited to fewer words or descriptions. There are few descriptive language methods to be used for audio testing use, with the exception of ADAM (see Section 2.1.1).

Impairment compared to reference	Grade
Inaudible	5.0
Audible but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

Table 1. ITU-R five-grade impairment scale.

2.5. Planning Subjective Tests

The inspiration of scientific research is usually a problem or a question that needs to be answered. It is reasonable to choose the best possible method to test the *hypothesis* presented. As mentioned earlier, the use of subjective testing is often appropriate when investigating sound quality. All the necessary parts of the procedure must be carefully planned beforehand. The purpose of this section is to give an overview of the important issues when devising listening tests.

2.5.1. Objective of the Test

This question is undoubtedly thought out before deciding to use the listening test methodology. The researcher has a clear idea of the question at hand. The objective of the test is however a totally different issue. The original problem is usually a vast theoretical one which involves several scientific subfields. For example a question like "What is good audio sound quality like?” is far too extensive for one test. It is preferable to define and outline the problem more and use result fragments to build a larger picture. The problem should be investigated in smaller parts, for example “How does the magnitude response of a loudspeaker correlate with personal preferences in sound quality?”

When the problem is specified, the next step is to formally state the *null hypothesis* about the test’s outcome. Null hypothesis is a statistical term that is used to label the hypothesis studied in the test. The goal of the test is to determine whether the null hypothesis can be stated to be significantly incorrect. Usually in scientific research the level of significance is 95%, i.e. if the null hypothesis is determined correct, the probability of error is 0.05. Null hypothesis can be for example: ”There are no audible differences in the magnitude response of the devices A and B in the test conditions used.” It must be noted that even if the results show that the difference does not exist, it does not mean that the differences are not there. Null hypothesis can never be proven indubitably correct as there is always some amount of uncertainty involved. The only result that actually tells something certain is that the null hypothesis is deemed wrong and differences between the devices have been proven to exist [20].

2.5.2. Listening Test Variables

Toole has listed the variables involved with listening tests in a generic manner [21]. Some or all of these can affect the outcome of the test in addition to the investigated parameter i.e. the *dependent variable.* Because of this the researcher must be able to control the other variables when the goal of the test has been decided. This way a proper testing method can be found for a given situation. Toole divides his investigation to two areas: The physical variables caused by test location and implementation and the psychological variables associated with the test subject.

The primary physical variable is the *listening room* where the test takes place. The listening experience depends on the properties of the room such as volume, decay time etc. The environmental conditions should be sufficiently stabilized. One way to simulate *free-field* conditions is to use an *anechoic chamber.* One important factor is also the amount of *background noise* in the room. ITU-R has specified the tolerances of a proper listening room in [11].

The physical *location* of the subject and the devices in the listening room must additionally be considered. The loudspeakers and the listener must be sufficiently far from the walls. The listener must not be too near to the loudspeaker. Positioning may affect the sound color and the reflections since different frequencies have different directivity. One way to avoid these effects is to use headphones as a sound reproduction device. But while the dependence on location is eliminated, there are other factors involved with headphone listening (see Chapter 3).

Perceived *loudness* is one of the most important features of aural stimuli. Many other perceived attributes of the sound are dependent on it. Especially when measuring sound quality it is imperative that all samples have equal loudness [11]. Otherwise, subjects tend to bias towards louder sounding samples. Additionally, the properties of audio devices, such as distortion, are dependent on the output level. Therefore the level should remain same throughout the test. Of course, the level is not the same thing as loudness and using some type of *loudness model* is preferable.

Choosing the *test material* to be used is problematic. Even though there are some recommendations, the researcher is usually obligated to determine the stimuli used. When testing the sound quality of audio devices, a secure alternative is to utilize the type of signals that are usually listened through the devices, i.e. commercial music. In practice, audio equipment are often tested with music and speech codecs with speech. Some test signals make it easier to detect audible differences; for example distortion can be more easily spotted with music that has a broader spectrum than speech. Listening to “mere” speech can additionally be more wearisome than music listening if the test is very long; music is understandably more entertaining than speech or a noise signal. There should also be some variability, for example different speakers or various music styles. With commercial material the quality is often an issue; especially old music samples are often rather degraded [18].

ITU-R suggests [11] that *critical audio material* is necessary for effective comparative testing of audio transmission systems. Here critical material is such that it can reveal the limitations of a system under test, which means that it includes "component samples which specifically challenge each system under test – though not necessarily at the same time" [22]. Some properties of a good test signal intended for subjective audio quality research are proposed in [11] and [22]. Namely it should:

- be potential broadcast content
- not distract a subject from the task of evaluation
- be normalized for loudness
- not include specially contrived material to "break" a particular system
- represent a significant range of broadcast material
- include mainly broadcast material
- originate from a high fidelity source, preferably CD quality (stereo format, 16 bit, 44,1 kS/s per channel).

At this time it is good to point out that a listening test sample used in the test is not necessarily the same thing as a *test item*, which referrers to the target of evaluation at some instant. For example, the same sample can be played through many systems thus creating several test items.

All the electrical equipment used in the test also create uncertainty factors of their own. New devices require some amount of *burn-in* before the components are "settled down". This time is usually between 24 and 48 hours. The *stabilization* of the devices also takes a few minutes after the power is turned on. Usually the devices are left on for the duration of the experiment. Overdriving the equipment should carefully be avoided unless this is precisely the purpose.

Human perception system was discussed about in Section 2.1. The physiological hearing system is introduced thoroughly in [6]. Toole bases most of the psychological and physiological variables in his paper to learning ability of humans. Familiarity with the listening room helps subjects to better disregard the coloration effects caused by it. Experience with listening tests and more specifically with the task at hand eases the evaluation. It is noteworthy that some are statistically speaking better test subjects than others. Normal hearing ability, as opposed to hearing impairment, is preferable and usually translates to smaller variation in the results. The subjects should use the rating scale somewhat similarly; this eases the statistical analysis. One of the most important variables is the *objectivity of a subject,* which should always be preserved. If for example a person recognizes the device by brand the grading could be hopelessly biased.

2.6. Implementation of Listening Tests

When the goal of the research and the dependent variable have been established, the test can be implemented using an appropriate method that eliminates as many of the other variables as possible. It must be remembered that in order to obtain an accurate answer, one must ask the right question. Classifying perceptual attributes is not an easy task for anyone. It should be made certain that all subjects know what they are expected to do before the test. The consistency of these directions to all the subjects is essential [17]. As mentioned earlier, the training session should be as comprehensive as possible with the resources available.

When choosing a method for audio sound quality assessment, there is a temptation to use a wide and accurate numerical scale. A lot of information about the interval magnitudes is theoretically gained this way. A wide scale however makes the analysis of the results harder to perform. A paired comparison with a few devices is simple to implement and a rather “safe” alternative, especially when the resources are limited. It must also be remembered that without the use a common descriptive language, a mere preference order obtained from the test tell very little about the properties of individual devices or why the results were the ones obtained.

In order for the test to be scientifically valid, the subjects must not know what exactly they are listening at a given instant. The devices must be protected from any kind of associations, visual or otherwise. For instance, loudspeakers can be hidden behind a curtain. This procedure is called *single-blind testing.* In addition to this, many authors recommend that the test is done *double-blind*. This means that the test implementer has no knowledge of the test item order so there can be absolutely no favoring [11] [20] [21].

Even if the subject is unfamiliar with the product brand, there is still a strong possibility that *external qualities*, such as the outlook and the ergonomics of the devices have an effect to the results. Toole and Olive investigated the biasing effect of visual perception in loudspeaker sound quality evaluation [15]. They discovered that big and visually appealing loudspeakers received significantly worse grades in blind tests compared to the situation where the device was visible. The smaller good-quality loudspeakers’ grades behaved contrary to this. The result was clear: Vision is the most dominant of human senses and it can bias sound quality assessment.

People can distinguish even very small nuances with good test arrangement. The sample must not be too long or the amount of information to be handled increases too much. ITU-R recommends that samples should be 10 – 25 seconds of length, though even shorter sounds could be used [8]. The listener should be able to stop the sound reproduction at will.

The best way to find the differences is to compare the test items with each other. A common arrangement is to allow the subject freely listen to all the items one at a time and *switch* from one to another. The statistical efficiency of a test run is increased if several items are compared simultaneously with each other. A simple paired comparison is an easy task but often insufficient if multiple comparisons need to be conducted. The switching technology also presented a problem before the modern ages; *A/B switches* used in listening tests should be quiet and have zero delay so that the echoic memory is not disturbed. Today the use of mechanical switches should be avoided altogether [20].

In the actual test the subjects are required to focus their attention to the task at hand. There must be no inappropriate activity such as reading, eating, drinking, watching television, smoking, talking etc. During long tests the attention is deemed to wander and frequent breaks are in order. Discussing the test during these breaks is not advisable. All the necessary details about the test and the devices can be revealed after the test.

2.7. Test Results Processing

Preparing and executing listening tests is laborious and time-consuming. The attention given to the previous parts does not pay off unless the procedure after the test is similarly thorough. Results processing and analysis is imperative for the test to be valid in scientific sense; usually little conclusions can be made based on the raw data alone. A proper statistical analysis presents the results clearly with a certain amount of uncertainty. Usually the *significance level* is 0.05, as in there is 5% chance of error in the results. This level determines what is *statistically significant* and what is not.

The main issue of listening tests in general is the amount of audible differences between the test items. According to Lipshitz and Vaderkooy, the differences truly exist if a properly executed double-blind test shows statistical biasing in evaluating the difference [18]. According to Leventhal, two specific cases of error in listening tests can be determined based on this [23]:

- type 1 error: inaudible differences are concluded to be audible
- type 2 error: audible differences are concluded inaudible

The goal is of course to avoid both of these error types. Understandably the designers of the devices (i.e. the engineers) are more concerned in the first type, whereas the customers or “audiophiles” are worried about the second type. The researcher can determine that the subjects cannot perform at the 0.05 significance level and thus conclude that no audible gaps exist. Nevertheless audiophiles are sometimes certain that the differences are clearly detectable. Leventhal has deducted that the traditional listening tests arrangement measures if the probability for type 1 errors is below 0.05, whereas the probability for type 2 errors can in many cases be surprisingly high. He recommends that the type 2 error risk should be taken into account and presents a method to do so. This way the “fairness” of the test can be determined. This procedure is however seldom implemented.

There are ways to calculate the statistical significance of the results. A *statistical t-test* can be conducted between two *groups*. A group is an outtake from a *population*, for example the grades of one device in a listening test. If the t-test shows no significant deviation between two groups, it can be concluded that the groups emerge from the same population, i.e. there are no audible differences between them. Unfortunately the t-test has its shortcomings: As the number of groups increases, the quantity of pairs to be compared grows exponentially. If there are for example 10 devices, the researcher must perform 45 t-tests. In addition, the t-test does not take into account the factors within the group and therefore does not utilize the whole test data.

An improvement to the t-test is the *ANOVA* procedure [24]. With it, all the test data can be studied with one analysis. ANOVA examines the similarity of all the groups and subgroups. The null hypothesis is dismissed if one of these differs from the others. Furthermore the analysis reveals which factors cause the differences. For example if the analysis shows that all the factors: LOUDSPEAKER, SUBJECT and LOUDSPEAKER*SUBJECT are significant, then the following conclusions can be made: 1) Different loudspeakers receive different grades, 2) Subjects’ grades are in different statistical *distributions* and 3) Subjects give different grades to different loudspeakers.

The idea is then to explain all the significant factors. The data should fulfill certain criteria if the ANOVA is used [25].

The data can also be examined in terms of confidence intervals. The averages and standard deviations of the grades for each loudspeaker are calculated. The standard deviation is then used to calculate the confidence interval usually at 95% level. When the confidence interval is very wide, the validity of the test can be questioned. The subjects themselves can also be examined if the test had repeated trials. Subjects that gave grades illogically can be then removed before continuing further with the analysis.

ITU-R has issued instructions on how to *normalize* the results if the scale used has no anchor points [11]. The following method can also be used to eliminate the effect of the test subject from the results. All the subjects are assumed to respond similarly to the test procedure. This means "loosing" the SUBJECT factor in ANOVA as all the grades are normalized to have same average and standard deviation. The purpose is to ease the comparability of the grades. This procedure is described by Equation (1):

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s \qquad (1)$$

where:

Z_i :	normalized result
x_i :	grade of subject i
x_{si} :	mean of grades for subject i in session s
x_s :	mean of grades for all subjects in session s
s_s :	standard deviation for all subjects in session s
s_{si} :	standard deviation for subject i in session s .

3. Headphones and Hearing

The subjective tests presented in this thesis investigate the subjective audio sound quality of various headphones. Headphones are a special case of transducers, “the most controversial and elusive components of the electroacoustic transmission chain”, as characterized by Poldy [26]. It is therefore necessary to discuss some of the unique issues involved with headphone listening before proceeding to the actual tests. There are many related themes which are not examined in this chapter, for example the mechanical and electrical properties of headphones and transducers. For these issues, see [26] and [27].

3.1. General Properties of Headphones

This section is largely based on a more comprehensive presentation by Poldy [26]. The issues related to the background of this thesis are presented shortly.

3.1.1. Structure and Types of Headphones

Headphones produce a sound field in a relatively small volume, as opposed to loudspeakers which create a propagating sound field around the listener. However, the differences in mechanism do not necessarily change the subjective experience, since according to the binaural theory the ear is merely a pressure detector. The fact that headphones always have some *leakage* is more relevant, especially in the low frequencies.

A headphone consists of two *earshells* connected by the headphone band. The main parts of the earshell are the *cup* which creates the volume around the ear and the *transducer* where electrical signals are transformed into sound waves. Transducer principles commonly used in headphones are isodynamic, dynamic i.e. moving-coil, electrostatic and electret. The source of the sound which moves inside the cup is called the *membrane.*

Any electrical analogs are omitted here, as well as the other structural issues. These include the effect of cup vibration on the frequency response as well as sound insulation, among others.

Table 2 specifies the categories to which most headphones are associated based on their structural properties. This division is given by ITU-T in [28]. As can be seen, there are five main groups from which all can be open-back and four closed. Openness referrers to an intentional leakage built in the back of the earshell. This way the bass response of the device can be controlled better. Closed-back headphones also have some amount of leakage, since a perfectly airtight design would be too clenching.

Earshell type	Open back / Closed back
Circum-aural	both
Supra-aural	both
Supra-concha	open only
Intra-concha	both
Insert	both

Table 2. Headphone types based on structure.

Circum-aural headphones have a relatively large coupling volume so that the ear is more or less inside the earshell. This design offers the best bass response since the leakage through the headphone cushions is minimal, unless otherwise intended. However, since human anatomy varies somewhat, the amount of leakage may not be the same for everybody. The former results in coupling variations and therefore the frequency response cannot be defined precisely. To remedy this problem, some amount of controlled leakage can be integrated to the design. The leakage can happen through the cushion or, as mentioned earlier, be accomplished by an open-back structure. Knowing the amount of leakage more precisely causes the frequency response vary less. Of course, increasing the leakage means that the bass frequencies must be boosted more to equalize the effect.

Unlike their circum-aural equivalents, supra-aural headphones do not surround the whole ear. The flat cushion, which can be quite large, is placed on top of the ear and sound is reproduced through the cushion. The supra-aural headphones are predominantly smaller and lighter than circum-aural models. What is gained in comfort is lost in hi-fi; according to Poldy: “The reproducibility of the given frequency response of a supra-aural headphone is lower that of its circum-aural counterpart, due to the relative ambiguity of the positioning of the earpiece.” [26] Supra-concha headphones are similar to supra-aural but have smaller cushions, covering only the concha. This type is often associated with portable devices.

Intra-concha headphones are inserted at the entrance of the ear canal and are supported by the cartilage of the concha. They are understandably very small and portable compared to the previous types. On the other hand, there are some disadvantages: Due to the size, the bass frequencies cannot be reproduced at the same level as in the larger designs. Also, the larger models may feel uncomfortable to some people. Nevertheless, intra-concha headphones are used in entertainment audio devices.

An insert headphone offers the tightest design in terms of leakage; they are placed directly inside the ear canal. These models are usually needed in professional situations where a good insulation from external sound is required. Expensive models include individually manufactured insert couplers into which the sound reproduction device is planted.

3.1.2. Headphone Applications

As this thesis is focused on audio sound quality, headphones are also inspected mainly from this point of view. The reproducibility of the bass response is one of the high-fidelity criteria. Generally it can be stated that all headphones fulfilling this criteria are open to some extent, except those with fluid-filled cushions [26]. However, headphones do not fit in the original concept of hi-fi because of the reasons discussed in Section 3.2. Nowadays it is not maintained that the reproduced sound should be indistinguishable from the original. Again, preference and naturalness are different issues, although some hint of the original hi-fi definition is still maintained among the audiophiles. When naturalness is not an issue, it can be stated that headphones offer the same level of sound quality and accuracy as high-end loudspeakers with far lesser cost.

In communications, the most important attribute is usually intelligibility. There is no need for accurate reproduction of bass frequencies; speech intelligibility is not affected by bass cut but rather improved by it. The present telephone bandwidth (0.3-3.4 kHz) is sufficient for speech transmission. This allows the use of intra-concha and insert headphones with these applications. Such communicational devices include for example hearing aids and speech ear monitors.

As mentioned in the introduction, the headphones are often used to play back sounds intended to simulate spatial events, i.e. *binaural signals*. One way to create binaural signals is to record actual sound events with an artificial head, HATS. Because headphones offer a good channel separation in noise insulation, they are a valid choice for this purpose. Theoretically, after some compensation, the reproduction should be perfect. Spatial hearing and other related issues are discussed in Sections 3.2. and 3.3.

Headphones can also be used to actively insulate unwanted sounds. Some devices employ an active circuit that reduces loud noises but boosts speech. Other special-purpose headphones are for example underwater headphones and audiometry headphones.

3.2. Headphone Listening Issues

The purpose of this and the following section is to give an overview of the present knowledge about headphone acoustics and the problems involved with it. This section focuses on specific listening issues. To understand these better, a brief summary of human spatial hearing mechanism is given.

3.2.1. Spatial Cues

The tool used extensively with spatial research in the past is the *spherical model* of the human head. Here the head is assumed to be a perfect sphere with two point-like pressure detectors symmetrically on both sides. In the beginning of the last century, Lord Rayleigh presented a theory of localization where two spatial cues are used: ILD difference and ITD [29]. The hearing uses mainly the latter at low frequencies. At approx. 1.25 kHz and above the phase information becomes ambiguous and the same ITD can be associated with many locations. Luckily, the head starts to shadow the higher frequencies and ILD information can be used to determine sound location. Figure 2 illustrates these basic cues.

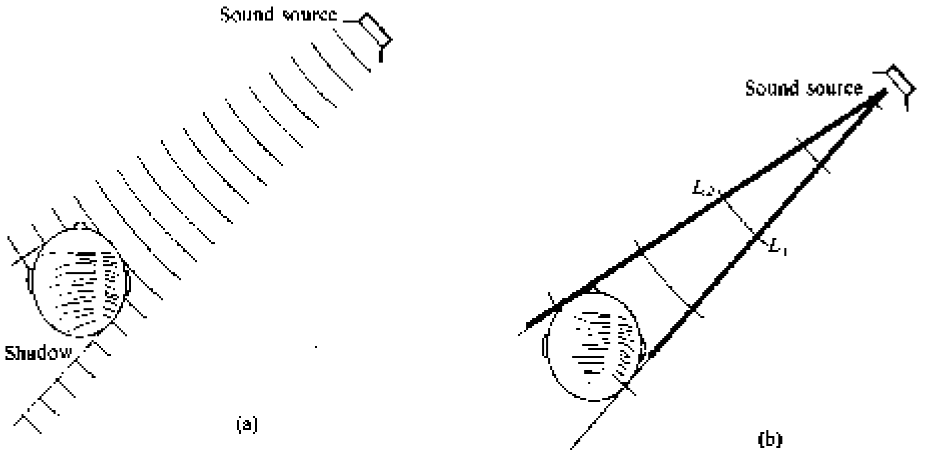


Figure 2.a) ILD caused by head's shadowing and b) ITD caused by different distances L_1 and L_2 (from [30]).

Later research has shown that the IDT can be divided into three types [31] [32]: 1) the onset flank ITD, used to localize brief impulses, 2) Delay of the fine structure (for example zero crossings), important for sinusoidal components of the signal and 3) envelope delay for complex waveforms. From these 2) and 3) are ongoing whereas 1) is transient.

The spherical model is still a valid tool but it does not explain all the aspects of localization. Figure 3 shows the so-called cone of confusion that illustrates the problem of Rayleigh's theory: Multiple locations can give same ILD and ITD information. The real asymmetric human head, as well as the rest of the body, cause spectral coloration to the incoming sound depending on the direction. This linear distortion or filtering effect is described by the HRTF. Since the HRTF is a monaural cue, it must be calculated to both ears separately. Combining the ITD information to the HRTF theoretically produces enough information to simulate sounds so that the binaural theory requirement is valid. In practice things are somewhat more complex.

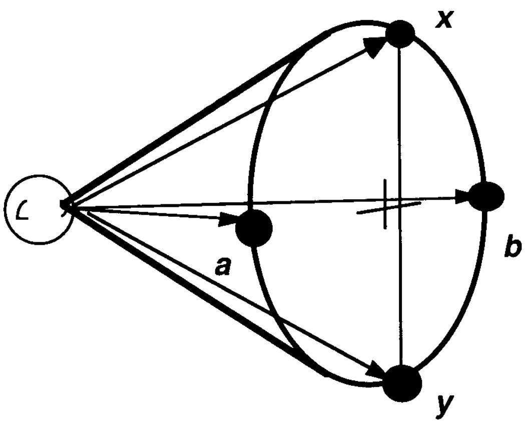


Figure 3. Cone of confusion. Sounds coming to the ear from all four points x, y, a and b give the same ITD and ILD. The head in the left is assumed to be spherical. (from [33]).

During real-life listening, human beings are able to move their heads. Thus there is much more information for which to base perceptions. This makes a major difference compared to basic binaural reproduction where the sonic information does not react to head movement in the correct way. When using headphones, head movements do not alter the sound at all. Unless some kind of head tracking system is used, the realism is helplessly diminished. Another difficult aspect of sound localization is the collaboration of hearing with visual perception, which is also not achieved in normal reproduction. It is uncertain to what extent human hearing is "designed" to localize sound events altogether;

throughout human evolution the vision has been used to confirm the actual location of the sound source. As mentioned, the vision can affect the hearing but randomly vice versa.

3.2.2. Inside-the-Head Localization

When headphones are fed with “normal” signals, the usual experience is that the sound appears to be coming from inside or near the head. If the listener is asked to describe the sound location, usually the task is reduced to determining lateral displacement; the sound presents itself on the axis going through the head from one ear to another. Hence the term *lateralization* is used in contrast to three-dimensional localization [32]. Commonly, lateralization is mentioned when referring to the fact that when listening to regular stereo signals with headphones, the sound seems to emerge inside or near the head. This is especially true for signals intended for loudspeaker listening, such as commercial music. For many listeners inside-the-head localization feels unnatural and even tiresome after a while.

There is not one definite explanation on why lateralization occurs. The phenomenon itself is ambiguous. Sounds that convincingly appear outside the head when listened with headphones can be created easily. The problem is to reproduce sounds that emerge near the median plane where the front-back discrimination is not reliable [32]. Inside-the-head localizations can be achieved with loudspeakers as well, as long as major head movements are not allowed [34].

Over the years there have been theories on the inside-the-head localization. Most of them are based on the assumption that there is something profoundly unnatural with the listening conditions achieved with headphones. The spatial cues are somehow in contradiction with each other; the fact that the acoustical signals do not fit a familiar pattern is causing confusion [35]. Major factors that may cause inside-the-head localizations are the head movement issues as well as the lack of natural reverberation [36]. Furthermore, the ILD and ITD give no reason to localize the sound elsewhere than inside the head during headphone listening [6].

With modern day DSP, it is possible to implement various routines that offer some relief to the problem of in-head localization. A common procedure used by so called *spatial expanders* is to add reverberation to achieve similar effect as with loudspeaker listening.

3.2.3. Other Headphone Characteristics

Some further headphone characteristics are presented shortly in this section, based on [26]. These are outside of the thesis' scope but nevertheless worth mentioning in order to emphasize the uniqueness of headphone listening in general.

The *missing 6 dB effect* concerns the fact that up to 10 dB more SPL is required for headphones to produce a similar loudness sensation as loudspeakers. The effect begins at 0.3 kHz and increases with decreasing frequency [37]. This phenomenon has given rise to much discussion about the nature of hearing; it could be related to a "perspective illusion" that causes the objects further away sound louder [38]. In this way the missing 6 dB effect would contradict the assumption that the ear is merely a pressure detector.

The audibility of monaural phase distortion with headphones is also a troubling phenomenon [26]. With loudspeakers the harmonic distortions are more audible and phase is not perceived so accurately. The situation is another way around with headphones for reasons not completely known. The reverberation could have an effect here as well.

3.3 Design Goals for Headphones

Typically hi-fi loudspeakers as well as other devices in the audio signal path aim at maintaining a flat frequency response. This way the device itself does not change the timbre of the sound and the theoretical hi-fi criteria of naturalness are preserved. The user can equalize the sound afterwards *in situ* to preference.

Headphones however are a special case among audio devices. During headphone reproduction the headphone replaces the whole setup, including loudspeakers and the listening room. As explained in the previous section, the listening situation is by nature rather unnatural so the question arises whether a flat frequency response is really the best design goal. Because headphone listening is experienced by majority as strange or even exhausting, preserving these sensations is not preferable. Instead, some way to simulate "traditional" listening conditions and lessen the unnatural effects could lead to an improvement in subjective sound quality.

3.3.1. Free-Field Calibration of Headphones

Free-field calibration process was published by Villchur in 1969 [39]. The basic idea of this procedure is to compare the HRTF measured from a given position in the free field, usually in front of the listener, to the measured *headphone transfer function* (PTF) at the same point. The measurement point is usually chosen to be the ear canal entrance. If the two responses are close to each other, the timbre of the headphones should sound as if there would be a "real" sound source in the free field. Thus the listening experience would theoretically be more natural. However, free-field calibration does not take into account other spatial cues than the monaural HRTF.

3.3.2. Diffuse-Field Calibration of Headphones

In 1986, Thiele proposed an improvement over the free-field calibration called *diffuse-field calibration* [38]. As a theoretical basis, he presents an *association model* describing human hearing. This model is described in Figure 4.

The model includes the familiar linear filtering of the incoming sound caused by the head and the body, i.e. the directionally dependent HRTF. This is represented by the symbol M in Figure 4. The brain uses the spatial cues such as ITD and reverberation time to determine the location of the sound source. Based on this elaboration with inner models, the hearing system creates the inverse filter M^{-1} used for the auditory signal. For natural hearing, $M \cdot M^{-1}$ equals 1 so that the original sound is reconstructed at the form-determining stage.

The problem with headphones is that the sound is localized inside the head arguably regardless of linear filtering; the other spatial cues are responsible of this. Thiele's theory is based on accepting this fact and equalizing the sound accordingly. With free-field equalization M and M^{-1} are not equal. The task is to create a HRTF of the sound field that localizes itself inside the head and use it in place of M. The solution lies in diffuse field, i.e. a sound field where all directions are equally probable for the incoming sound. This causes the sound to localize in the middle of the head.

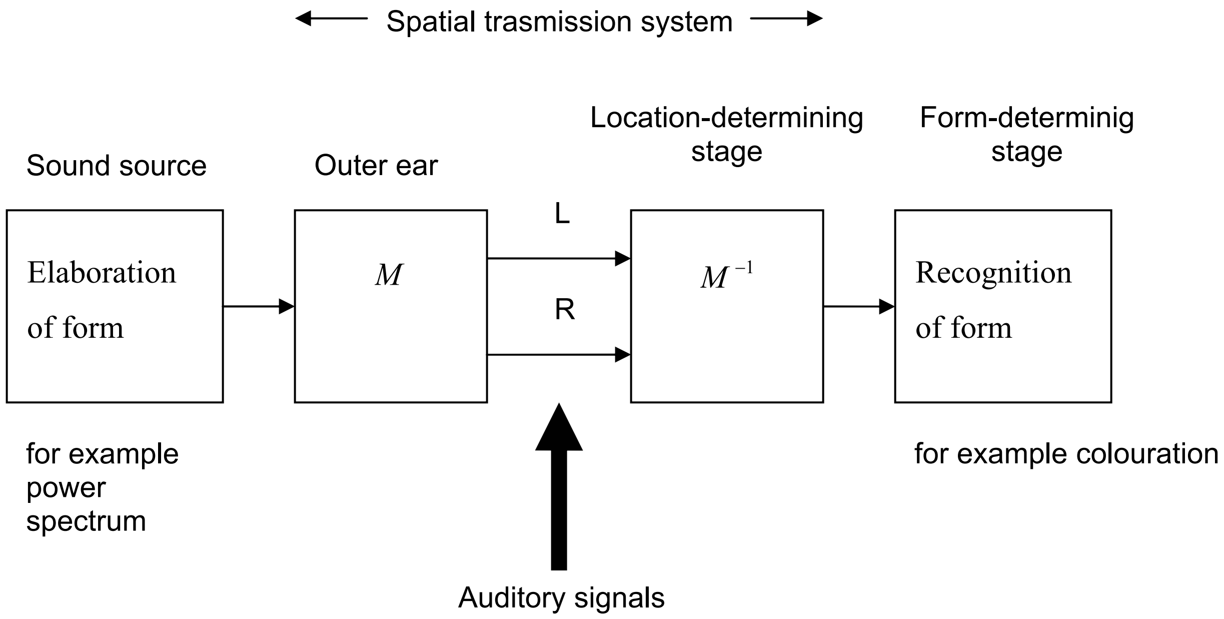


Figure 4. Association model for spatial hearing (from [38]).

In conclusion, diffuse-field calibration procedure consists of measuring the HRTF in a diffuse sound field and comparing the result to the PTF measured at the same point. Thiele has tested his theory with subjective preference tests and the diffuse-field equalization prevailed over its free-field comparison. ITU-R has determined that studio monitor headphones should have a flat diffuse-field response within small tolerances [40].

3.3.3. Design Criteria for Headphones

The two calibration/equalization methods presented in the previous sections are the most common implemented in practical headphone design. To conclude this chapter, a general method for headphone design by Møller *et al* is presented formally in Equation (2) [41].

$$PTF = RFHRTF \qquad (2)$$

where:

$RFHRTF$: head-related transfer function in the reference sound field

The principle is no different to the methods so far, only the actual equalization goal is not specified. According to Equation (2), the measured HRTF of the goal sound field *RF* should be the same as the measured PTF of the headphone. The measurement point in both cases should be the same. Ideally this procedure is done individually but in practice HATS is often employed.

4. First Test – Real Headphones

A series of listening tests done by the author and various other people are presented next. There are a total of three tests; the first two form a unity by testing real headphones and HATS recordings, respectively. The third experiment tested both, using only speech material. The first two tests can be seen as a pilot research to the third, where the method was perhaps the most advanced. ANOVA is employed only to the third test results. The subject of this chapter is the first test, done during summer 2001.

4.1. Purposes of the Test

This test studied the subjective sound quality of eight different headphones using music and speech samples. The goals were:

- first, *to evaluate user preferences of headphone sound color, i.e. timbre with narrowband, wideband, and music material*

- second, *to determine whether music excerpts can be used instead of speech samples in subjective sound quality tests, i.e. how well do these correlate.*

The first goal helps gaining some general perspective on headphone sound quality. The initial presumption was that subjective sound quality correlates well with price of the device. The second goal could have offered a minor improvement to the comfort of subjective testing from the subject's standpoint as music is more entertaining than speech. Music also offers a wider spectral content and thus challenges devices more. Hypothetically, if the grades given to speech and music are similar in this test, this kind of replacement could be considered.

4.2. Experimental Method

The idea of the test was to grade various headphones with different samples based on subject's personal preference of sound color, i.e. timbre. A total of eight headphones were placed at the listening room along with the test interface. The subject graded eight headphones on a scale from 0.0 to 10.0. Switching between devices was done manually by the subject while the sound sample was playing in all headphones. Thus there were eight test items on each trial. One subject performed this grading with ten different samples, i.e. did ten trials. The subjects were instructed to base their judgment only on timbre and not on any other qualities such as background noise, comfortableness, or outlook of the headphone.

4.3. Test Variables

There are two main variables in the first test, these being the headphones used and the sample type. The following sections specify these variables more precisely.

4.3.1. Headphones Used in the Test

The list of the headphones used in the test is as seen in Figure 5 and in Table 3. Prices in Table 3 are examples from some audio stores in Finland. (The prices include a 22% income tax.) The goal was to choose headphones that are somewhat different in terms of timbre so that they could be distinguished more easily. Three intra-concha models were also included along with the more traditional headphones.

The objective measurements of the headphones' magnitude responses were done in NMP Salo audio laboratory by Ossi Mäenpää. Measurements were performed in an anechoic room using B&K HATS model 4128C and Audio Precision workstation. Results were processed with Audio Precision Inc‘s APWIN software. The graphs from these measurements can be seen in Appendix A. One should keep in mind that these measurements are not very reliable in higher frequencies, as the type 3.3 ear used in

HATS is not specified for frequencies above 8 kHz. In addition, the AKG K-240 magnitude response lacks low end frequencies because of poor fit to HATS' ears. The real human ear is covered more effectively by this model.

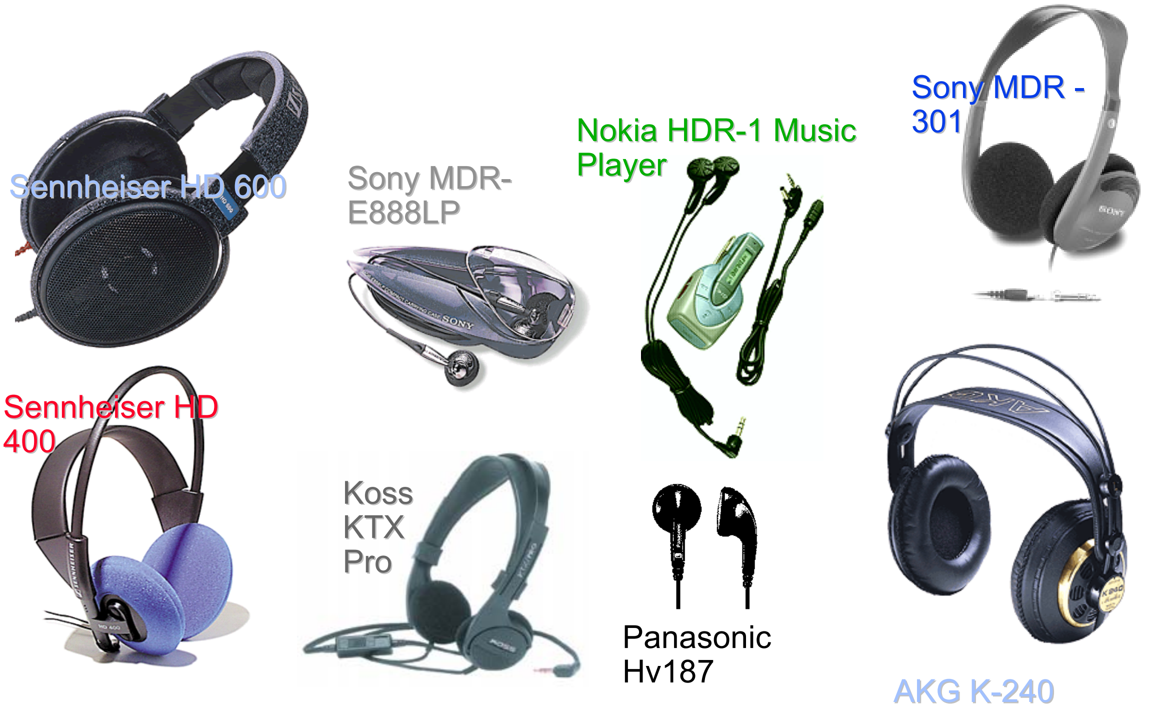


Figure 5. The headphones used in the test.

4.3.2. Sound Samples Used in the Test

The sample used in listening tests should have certain properties as mentioned in Section 2.5.2. These recommendations acting as guidelines, it was decided that sound samples should include narrowband coded speech as well as wideband coded speech. Additionally, both male and female speakers should be used. Codecs implemented were AMR 12.2 kbit/s for narrowband (i.e. the GSM EFR codec) and AMR-WB 23.05 kbit/s for wideband. The subjects listened to the same speakers in wide- and narrowband. The differences and similarities between the present and future telephone bandwidth could this way be investigated.

Manufacturer	Type	Price (Euro)	Type
Sennheiser	HD600	303	Circum-aural, open back
AKG	K-240	122	Circum-aural, semi-open back
Koss	KTX Pro	50	Supra-concha
Sony	MRD 301	33	Supra-concha
Sennheiser	HD400	42	Supra-aural
Nokia	HDR-1 Music Player headphones aka Mulan	N/A	Intra-concha
Sony	MDR-E827G	93	Intra-concha
Panasonic	RP-HV147 player headphone	20	Intra-concha

Table 3. Details of the headphones used in the test.

After the test additional useful procedure was considered: Additional bass attenuation via filtering could have been applied to the speech samples because the actual sound coming out of a phone speaker does not have the lower frequencies our sample had. These low frequencies are typically filtered off before speech codecs. This procedure was implemented in the third test.

Since the subjects participating in the tests were both Finnish and non-Finnish it was agreed that the spoken samples should be in English and in Finnish so that the non-Finnish subjects would listen to English samples and Finnish-speaking subjects would mainly listen to the Finnish samples. One male and two female speakers from the NMP Acoustics Platform/AQUA Speech database were chosen to provide the English samples. For the Finnish samples we used one female and two male speakers from the NATC Multi-Lingual Speech Database for Telephonometry. Each speaker provided one sample from a total of six speech samples, three of which constituted the Finnish set and three the English set. The two codecs were used on all these samples. It was agreed that a subject should listen to same speech samples wideband coded and narrowband coded so that possible differences between the two formats could be detected.

It was also agreed that sound quality evaluation based merely on speech samples would not be sensible since speech occupies the whole audible spectrum only partly. Due to this, various unprocessed music samples were included in the test. According to the recommendations in Section 2.5.2., these samples should represent as broad range of different music styles as possible. Table 4 lists the music chosen for this study.

The samples were edited to be approx. 30 seconds of length using Syntrillium's Cool Edit Pro. Speech samples were looped from 10-second segments. All sound files were in .wav -format CD quality (16 bit, 44100 Hz, stereo). Speech samples were obviously upsampled from their 8 and 16 kHz sampling rates. This was done to satisfy software's demand for similar samples. The speech samples were originally monophonic so they had to be split to both channels, i.e. they became "double mono". The music samples were stereophonic.

	Description	Source
Set 1	Classical Instrumental	Jean Sibelius, Symphony no.3
	A Cappella vocal group	King's Singers, Chanson d'Amour
	Electronica Instrumental	Tokyo Eyes Soundtrack, Follow This Cam
	Rock Male Vocal	David Bowie, Rebel Rebel
Set 2	Pop Male Vocal	The Police, Every Breath You Take
	Rock Female Vocal	Tokyo Eyes Soundtrack, Eye to Eye with You
	Jazz Instrumental	Miles Davis, So What
	Classical Female Vocal	Renata Tebaldi, Suicidio!

Table 4. Music samples used in the test.

4.4. Test Subjects

It was decided that Nokia employees from NRC Ruoholahti and from NMP Salo could be considered as ‘professionals’ and be used as expert listeners. A total of 20 male and female subjects, 10 from Ruoholahti and 10 from Salo, participated in the first test. These persons have a strong professional audio background and/or musical experience. There were no reported hearing impairments. Due to a busy schedule and the apparent professionalism of the listening group no systematic listener selection procedure was used. It was also assumed that adding training cases could present a risk of listener fatigue and loss of concentration, since the test was already quite time-consuming. Although in practice the subjects used in these tests can no doubt be relatively objective, it was clear that there would be biasing toward certain things. For example the Sennheiser HD600 model is widely used among the test subjects and perhaps also considered to be “good” headphones. Furthermore some models used were certainly recognized by the subjects in the first test although the brand and model names were hidden.

4.5. Test Setup

4.5.1. Test Sites

The listening test was conducted in Nokia facilities in NRC Ruoholahti and NMP Salo sites. In Ruoholahti, the test took place in a normal office room and in Salo the listening room of the acoustics laboratory was used. The test in Ruoholahti was done in an office room where background noise was minimized by removing all but the necessary electrical equipment from the room. Background noise varied depending on the number of people passing the listening room but was generally at same low level as in Salo. The listening test itself and its interfaces ran on a Silicon Graphics Octane workstation (SGI). The main unit, being rather noisy, was wired so that it could be located outside the listening room during the tests so that only the monitor was inside the room while testing.

In Salo the test was performed in a listening room of Acoustic laboratory. Due to cabling problems the SGI workstation was set up in a same room as listeners causing a bit of background noise. However the noise level was decreased by acoustic damping material built around the workstation. In the listening place the SPL of background noise was below 30 dBA (the minimum of the dB-meter's scale), i.e. relatively quiet. Noise level right next to the SGI main unit was roughly 35-38 dBA. There was also some low-frequency hum detected. This was eliminated using an isolation transformer.

4.5.2. Test Arrangement

The headphones were laid on the table and the subjects manually switched between them. The same sample played simultaneously from all the headphones creating eight test items per trial. The test environment from Ruoholahti is seen in Figure 6.

The test was organized such that each subject listened to three samples of narrowband speech, the equivalent three wideband speech samples, and either one of the two music sets (see Table 4). The order in which these three sessions were presented varied between subjects. Each test also had a second permutation where the order of the samples in the session was changed. Altogether, there were twelve different sessions by permutations of test material and playing order, from which each subject underwent three. The sessions took about 20 minutes each. A break of few minutes was kept between the sessions. This was done to prevent fatigue from affecting the grades.

The question which was presented to the test subjects was simply: "Sound quality of headphone x". The form of the question had been under heavy consideration prior testing but "sound quality" seemed like a sensible option; this way the subject would not try to act as a spectrum analyzer (a question like "Audio quality of …") or on the other hand he or she would not be too general taking into account outlook, loudness, etc. (a question like "Your preference of …"). At the beginning of the test subjects were verbally instructed to base their judgment on sound color preference and not on perceived loudness, headphone outlook, etc. The concept of timbre was also discussed briefly with each subject. Some nominal anchor points from “very bad” to “very good” were also used (see Figure 10). Their function was however not predominant; the subjects were asked to use grades on a scale from 1.0 to 10.0 as they felt suitable. Later the results were normalized.

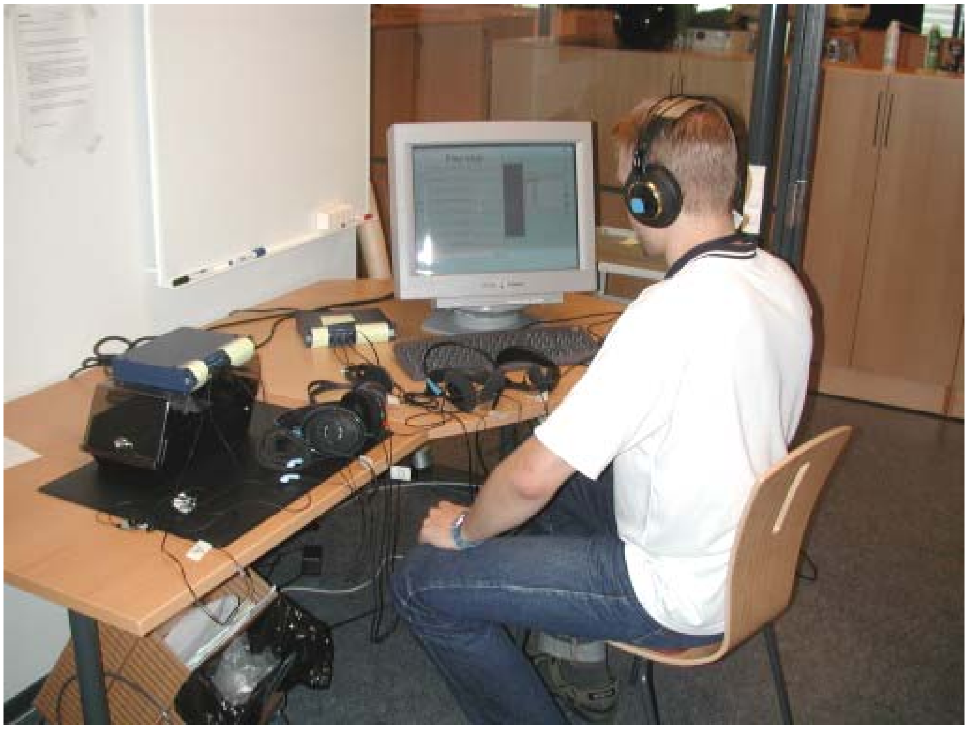
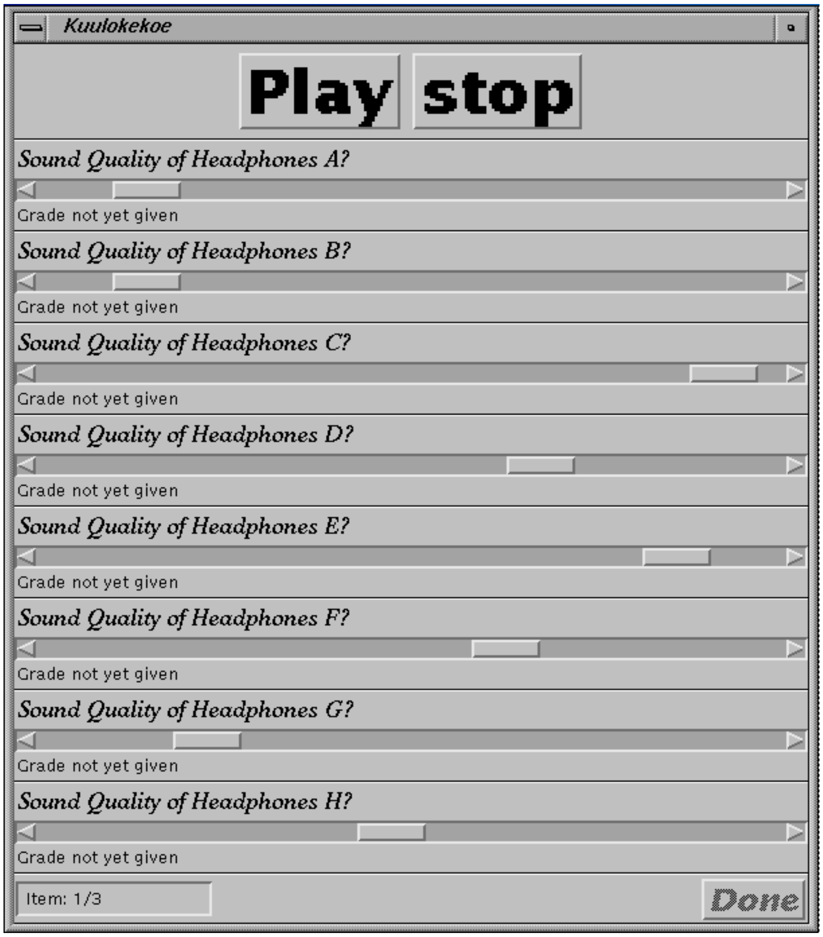


Figure 6. First test setup from Ruoholahti.

The test interface was done with Guinea Pig 2 (GP2). It is software created for NRC’s versatile listening test needs [42]. The GP2 graphical interface is shown in Figure 7. The length of the samples was approx. 30 seconds, after which the subject had to press “Play”- button to hear it again. After all eight grade sliders in the test window had been moved, the subject could move on to the next music sample by pressing the “Done”- button. The subject could also adjust the overall volume level in all headphones by changing SGI’s Analog Out output volume with a slider.

The subjects were instructed prior to the test to analyze each sample separately. The time used per sample nevertheless usually decreased as the test progressed. This indicates that subjects learned to predict the sound quality of each headphone and perhaps did not analyze the headphones case by case so much as in the beginning.



Sound Quality of Headphones A?

Grade not yet given

Sound Quality of Headphones B?

Grade not yet given

Sound Quality of Headphones C?

Grade not yet given

Sound Quality of Headphones D?

Grade not yet given

Sound Quality of Headphones E?

Grade not yet given

Sound Quality of Headphones F?

Grade not yet given

Sound Quality of Headphones G?

Grade not yet given

Sound Quality of Headphones H?

Grade not yet given

Item: 1/3

Figure 7. The test window of GP2 in the first test.

The .wav -files were played by GP2's sound player (programmed in C++) through SGI's Analog Out port. The signal then went to two Symmetrix 304 headphone amplifiers, which were wired in series. Each amplifier has four outputs with individual output volume controls and a master volume adjustment. These controls were used when calibrating the loudness to equal levels (see next section). All eight headphones were positioned on table next to the SGI's display. The brand labels and model names printed in the headphones were covered in order to avoid prejudice towards a certain brand. After the test the subjects commented the devices and the experiment in general. All subjects filled out a question form in which the headphones were graded based on nominal attributes. These were outlook and comfortableness of the device as well as an estimate of its price.

4.5.3. Loudness Alignment

Aligning the loudness levels of all headphones to be equal was particularly important because even a slight difference in levels would benefit the louder sounding headphones. The importance of loudness matching has been pointed out in earlier investigations, for example in [43]. A B&K HATS model 4128C with type 3.3 ears was used to transfer the SPL produced by the headphones into loudness measuring tool by Olli Tuomi (NRC/Tampere) running under Mathwork's Matlab technical computing software in a laptop. This tool uses Moore's subjective loudness model [44].

Various signals were tried out for calibrating, for example 1 kHz sine wave and looped speech, but the final measurements were done with "artificial speech-like noise", which was actually white noise filtered to roughly match the spectrum of average speech [45]. The filter used is specified in Table 5. Filtering was done in CoolEdit Pro. The Moore loudness value of each headphone output was adjusted in the preamplifiers until they were all within tolerance (circa 80 +/- 1 phones).

lower -3 dB (2nd degree Butterworth)	upper -3 dB (1st degree Butterworth)
0.1 kHz	0.7 kHz

Table 5. The -3 db points of the Butterworth filters used for white noise to create "artificial speech".

Although the volume levels of the headphones were calibrated and adjusted with Moore's loudness model, headphones fit to people's ears differently and thus have varying subjective loudness level from person to another. Especially this was noticed in preliminary listening with the in-ear-headphones when test setup was built.

Loudness calibration between different samples was determined to be not so important since the test would be done one sample at a time and in theory the test items would have the same loudness level in all headphones. There were however significant differences in sample sources in terms of loudness because of various recording techniques used (for example compression) and on account of the fact that the music samples have a spectrum unlike the "pseudo-speech" we used. Due to this samples were subjectively adjusted to have approx. same loudness level by adjusting the GP2's sound player output. Wide- and narrowband speech samples were left at original digital levels while music samples were attenuated several decibels compared to original signal level in CD depending on the source.

4.6. Results

4.6.1 Headphone Preference

Figure 8 shows the results of all subjects from both sites normalized and averaged in four cases: all excerpts, wideband-, narrowband- and music samples. Equation (1) was used for the in-between-subjects normalization. This causes the absolute grades to be lost. The most important aspect of the results is now the mutual order of the headphones' grades. The 95% confidence interval is also visible at the top of the grade bars.

The headphones are shown in decreasing order in terms of overall grade average, although Mulan's grades are practically at the same level as the Sennheiser HD400's. One should also consider the 95% confidence intervals; headphones in the "middle cast" are in principle all inside the same grade gap and should not necessarily be placed in any particular order. It should be emphasized that the mutual preference order of the headphones is the attribute one should mainly concentrate to when examining these results, since the grade normalization was performed.

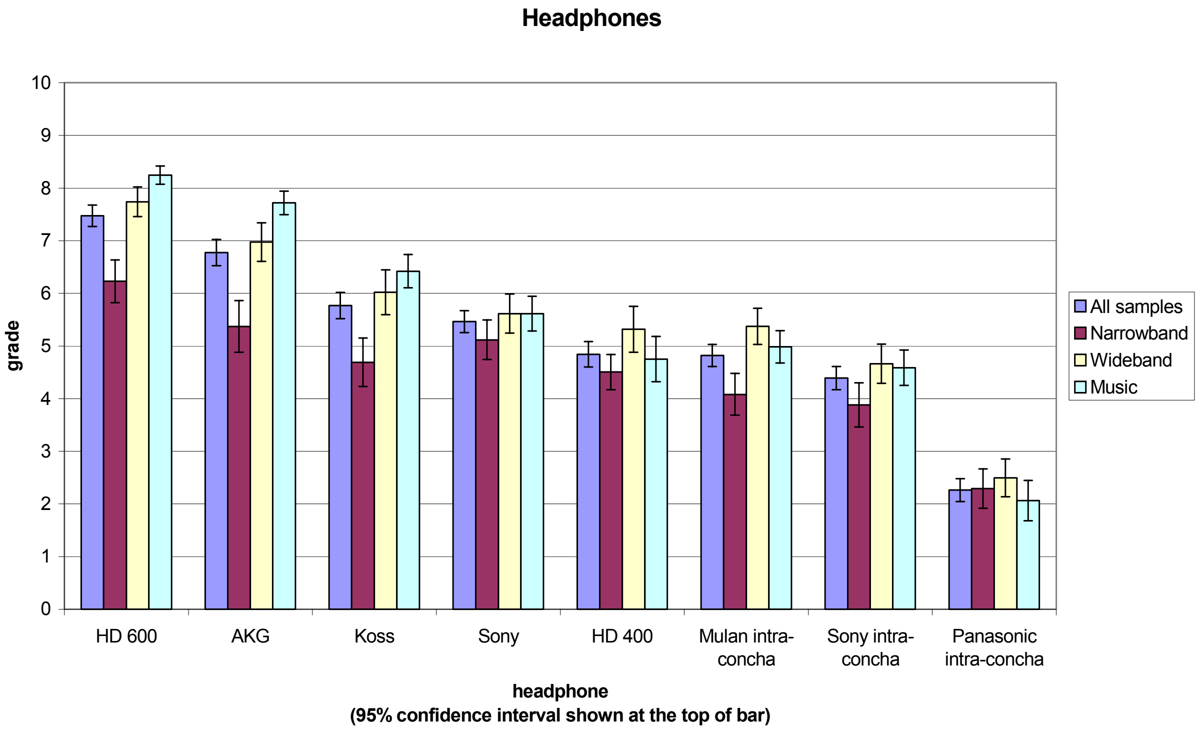


Figure 8. Normalized headphone grades of all 20 subjects with different headphones. 95% confidence intervals are visible at the top of the grade bars.

4.6.2. Correlation of Speech and Music Samples

As established before, one of the objectives of these tests was to find correlation between the results of coded speech samples and unaltered commercial music excerpts. In case such resemblance is to be found, music samples could theoretically be used in place of speech in future tests. Furthermore, it was speculated that some music samples would act more similar with speech excerpts than others.

Comparing narrowband speech and music tentatively yielded deviations so great that no further inspection was done on that area. Narrowband coded speech was thought to deviate more from natural sound than wideband coded speech and thus it perhaps cannot be compared to music so well. Wideband speech was however studied further.

Comparison was done between the grades of a music excerpt and the wideband grades as a function of headphone listened. Each music sample was compared separately to the speech. An error value was calculated for each case using Equation (3):

$$e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(d_i - \frac{1}{n}\sum_{i=1}^{n}(d_i)\right)^2} \qquad (3)$$

where:

$d_i = x_{wbi} - x_{mii}$: difference between wideband grade averages and grade averages of an individual music sample

$n = 8$: number of headphones.

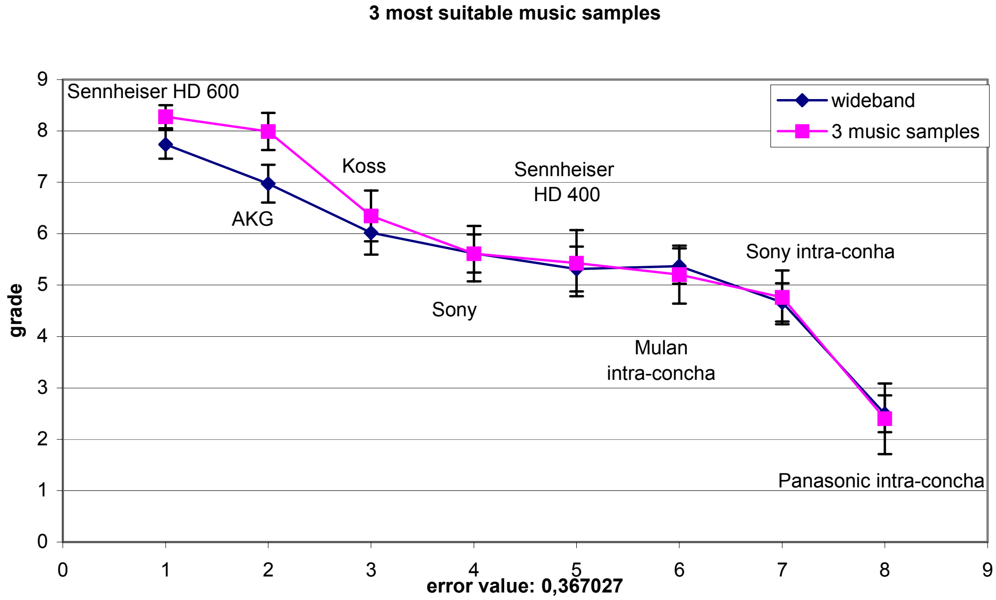


Figure 9. Three best-suited music samples compared to wideband speech excerpts. These samples were Symphony 3 by Jean Sibelius, Chanson d'Amour by the King's Singers and Rebel Rebel by David Bowie.

As it turns out these error values differ somewhat for each sample and seems that some music samples are indeed more suitable to the purpose of replacing wideband speech excerpts in listening tests. In Figure 9 three music samples with smallest error value were jointly compared to the wideband average grades. The two curves overlap quite nicely with a few exceptions. The error value is also relatively low.

4.6.3. Estimates of External Qualities of Headphones

After the test, each subject was asked to fill a form where the headphones were graded by some verbal attributes. The intention was to assess how the outlook etc. of the devices was thought to be. These results could be used to explain possible differences between the real headphones and simulated recordings. The results given in Figure 10 will be investigated more in Section 6.7 and in Section 7.1 where final conclusions are presented.

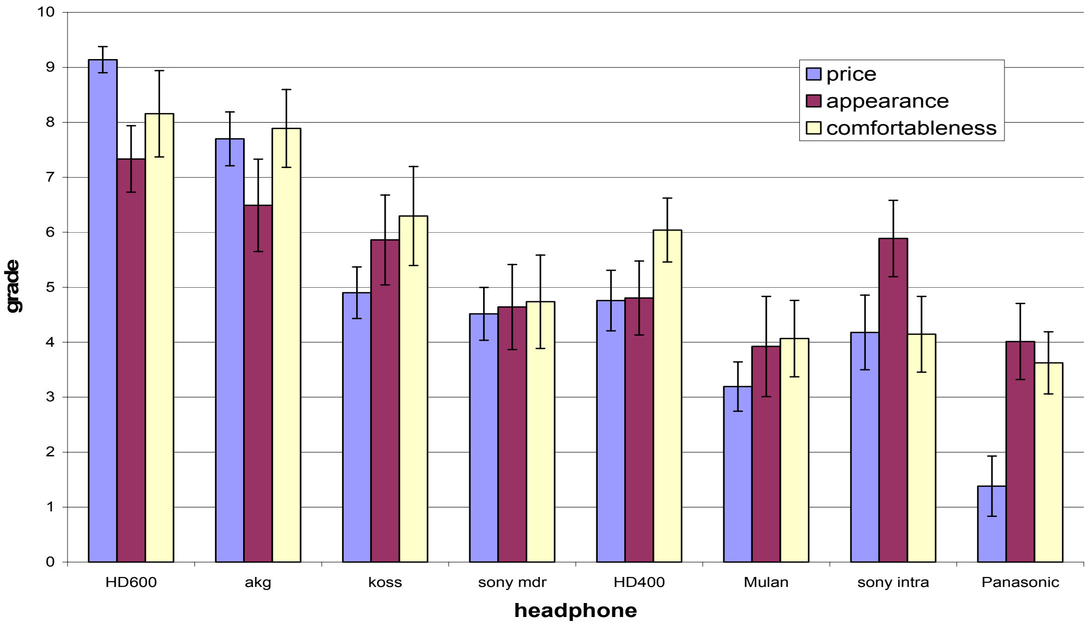


Figure 10. Normalized external attribute grades of all 20 subjects with different headphones. 95% confidence intervals are visible at the top of the grade bars.

The verbal attributes considered here are estimate of price, appearance, and comfortableness of the device. The subjects gave grades to all headphones on a scale from 1 to 10 by these attributes. The results were normalized according to Equation (1). The ranking order in Figure 10 is quite similar to general headphone preference in this test, although the 95% confidence intervals are quite large.

4.7. Discussion

4.7.1. On Headphone Preference

The general ranking order of the headphones was somewhat guessed before the test was conducted; The Sennheiser HD600 were thought to be the "best", The AKG K-240 were almost certainly the "second best" etc. As a hypothesis, the intra-concha headphones were predicted to fall behind the ordinary ones mainly because of their limited bass response. (See Appendix A for headphone frequency responses.) This was especially true for Panasonic intra-concha model, which was almost always placed to be the worst one in the tests. A little surprising was the performance of the other two intra-concha phones; both did fairly well and the Mulan intra-concha headphones were appreciated more than the pricier Sony MDR-E827G in-ear-phones. Sony MDR 301 were generally liked more than predicted. The Sennheiser HD 400's emphasis on higher frequencies was often thought to be distracting. This model has a boost on frequencies from 2 kHz to 4 kHz. On the other hand, this property, along with the attenuated middle frequency area, possibly increased intelligibility of the speech samples. Almost the exact opposite of the former was the Koss KTX Pro model whose bass emphasis probably gave rise to a large difference between the music grades and narrowband grades given to it.

The bass response of the headphone was probably one of the qualities that had the most substantial influence on the results. It seems that bass boost was appreciated when listening to music samples but on the other hand it apparently reduces the intelligibility of speech somewhat. This was especially true for Koss KTX Pro and AKG K-240 models, whose bass response is the most boosted. The subjects often complained about the lack of bass response with the intra-concha headphones. Possibly for that reason the differences between sample categories were not so large with these models. Other qualities, such as distortion should not be excluded as possible reasons for low preference, even though no measurements of these were made.

The intra-concha headphones actually challenged the other headphones in terms of subjective sound quality; the Mulan and Sony intra-concha received equal grades with the regular lower-end headphones. On the other hand, some subjects found it hard to compare the intra-concha and the regular models because the “soundscapes” were so dissimilar between the two. The intra-concha headphones have furthermore one important property to consider: They fit people’s ears differently. Sony model, being the largest intra-concha, was told to be too large by some subjects, while some (sometimes the same) subjects complained that the Mulan and Panasonic models are too small and they do not stay inside the ear. This variation in ergonomics, which has an effect to the acoustical performance of the headphones, possibly translates to the subjective sound perception as well. A difference in the experienced loudness of the intra-conchas was found between subjects, even after the careful loudness calibration (explained in Section 4.3.). Amplifier output was however not adjusted after the calibration because opinions on the sound volume were not unanimous.

The mutual order of the headphones was considered the most important feature of the results and it remains with few exceptions the same when comparing the sample categories. The narrowband grades deflect the most from the others, which could be explained with the following speculation: Narrowband coded speech is in itself so unnatural sounding that it cannot be compared to the other to sample types. Wideband coded speech has a frequency range so wide and quality so good that it can almost be considered to be “natural” speech. The music samples, being unprocessed, fall into same “natural” category.

4.7.2. Replacing Wideband Speech with Music

As mentioned before, narrowband samples received many accusations for sounding unnatural. The usual complaint was that the grading with narrowband was harder. Figure

6 also suggests that narrowband coded speech differs too much from music to be comparable with it; at least the high-end headphones show discrepancy beyond the confidence intervals. The wideband samples on the other hand received similar grades to music.

All three music excerpts in Figure 9 can be considered lacking bass somewhat, at least compared to other samples. The type of music where bass is not overly emphasized might be best suited to approximate wideband speech. One should keep in mind that no normalization was performed to the two curves in Figure 9. In this case they would overlap even more. The error value calculated with Equation (3) considers the normalization with respect to average but not to standard deviation. AKG K-240 is the only headphone whose grades do not fit the confidence intervals.

Despite this result the opinions against this kind of replacement grew stronger after further consideration. Although this idea of replacing wideband speech with music in listening tests is of some academic interest, it is probably not very helpful after all unless the test is quite long or exhausting for the subject. There are also other reasons that speak against practical applications: The method only applies for good quality devices that produce little linear or nonlinear distortion and background noise. In addition, the attribute investigated in the test must be related to sound quality and not to for example intelligibility.

5. Second Test – HATS recordings

The second test presented here can be seen as a direct successor to the first one. It was done during late 2001/early 2002. The test is similar to the first one with the exception of using HATS recordings to replace actual headphones. In essence, the attempt was to create simulated aural experiences similar to reality.

5.1. Purposes of the Test

The goal here was to examine:

- first, *HATS recording process in general*
- second, *could real sound sources in subjective sound quality tests be replaced by HATS recordings of the same sound sources, played back through compensated high-quality headphones, i.e., do external qualities of the headphones affect the sound quality evaluation.*

One of the themes in this thesis is to find redundancies in listening tests. As subjective testing consumes both time and resources, savings on either area would be desirable. The comfortableness of the test subject can also always be increased. In the first part the idea was to replace speech with music. If HATS recordings could replace the actual devices, the benefit would be even more drastic; the subject could do the test with one pair of headphones and no extensive equipment setups would be required. Adequately processed test signals could provide the stimulus and more control over the sample could be attained. Previous studies however, have found that external factors, for example outlook, affect the perceived sound quality of loudspeakers [15]. With headphones the ergonomic issues are also relevant in addition to appearance so the validity of the method was initially uncertain.

5.2. Experimental Method

The headphones and samples used in the previous test were employed to create recordings that attempt to simulate the actual aural experience. The recording device was a B&K HATS model 4128C placed in an anechoic chamber. The recordings were played pack with one pair of Sennheiser HD600 headphones whose frequency response had been compensated based on objective measurements. The test setup now allowed for a fast computerized switching between the test items and a truly double-blind procedure. The external qualities of the headphones were extracted and the subjects could base their judgment on sound quality only. Otherwise the test was similar to the first. The correlation between the two test results implies the validity of this kind of recording-replacement method.

5.3. Test Variables

Essentially this test was similar to the first one so the test variables are also the same: The test sample and the headphone. As mentioned, this time the physical headphones were replaced by recordings that attempt to simulate the audible qualities characteristic to the original devices. Simulation was accomplished by recording the samples used in the first test through HATS and applying some post processing. The processed samples were played through one pair of high-quality headphones (Sennheiser HD600) during the test. This headphone was compensated in the post-processing stage. The whole procedure is described in the following sections.

5.3.1. Recording process

Headphone recordings were done in NMP/Salo acoustics laboratory with help from the audio lab staff, especially Ossi Mäenpää and Jukka Kiljunen. The samples used in the first test were converted to an audio CD and played through a CD player and a Symmetrix 304 headphone amplifier during recording. A B&K HATS model 4128C was placed in an anechoic chamber so that background noise was in theory eliminated. Each headphone was in turn placed on HATS with type 3.3 ears and fed with all samples from the CD player. Because of the different impedances, the headphones were fed with different signal levels based on the loudness measurements made in the previous test. To improve signal-to-noise ratio, maximum level that did not cause non-linear distortion in the headphones was used for each headphone.

In the beginning of this task one of the goals was to investigate HATS recordings in general to gain practical knowledge on this area. As always in practical cases, some problems were discovered in the recording process. First of all the recordings had some background noise; apparently the anechoic chamber we used is not totally isolated after all. On the other hand, any recording technique is always bound to introduce some noise to the sample. Another problem was that the samples in the sample CD were not level aligned. In consequence the speech samples, which were at a lower level than the music, produced more background noise. In the future tests it would be preferable to do level calibration prior to recording and feed all the samples at the same maximum possible level to the HATS.

The second problem was that the right and left channels of the recorded samples were at different levels. This causes spatial deviation during playback as the sound lateralizes somewhere else than in the middle of the head. This in turn can distract the listener because comparing two sample items with different spatial locations can be difficult. The reason for the phenomenon is suspected to be the HATS and headphone transducers themselves; it is quite hard to place headphones to the artificial head so that both phones lay identically in/on the ears. The worst cases were the Mulan and Sony intra-concha headphones that generally had 4 dB and 3 dB differences between the channels of the recorded samples, respectively. The intra-concha type phones are most difficult to place identically whereas smallest difference between channels was detected in supra-aural headphones, such as HD 400 (approx. 1.5 dB). Furthermore, the elements of headphones are perhaps not equal in terms of sensitivity and frequency response.

5.3.2. Samples processing

Samples were processed with Cool Edit Pro and Matlab software. The former software was used to divide the continuous recording of each headphone to different samples, some simple gaining and listening to the incomplete samples. Matlab was utilized for equalizing, i.e. filtering as well as loudness calibration with NRC's loudness tool with some additional self-made functions.

The first task was to segment the continuous recordings to similar samples (approx. 30 seconds each) as in the first test. The next step was to calibrate the loudness of the samples. This was also done in Matlab using the dynamic loudness of a sample as a basis for calibration. Dynamic loudness is the level that is exceeded 10% of a time in the sample. The sound sample was divided in frames of 1024 samples (i.e. 23.2 ms) and the loudness tool was used to calculate the dynamic Moore's loudness value for each frame. There was also a 50% overlap to the previous frame. Thus the loudness-per-frame information of the sound sample was obtained and from this, the dynamic loudness (as well as other loudness measures) could be calculated. Some comparison between different loudness measuring methods could also be done based on this information.

There is no standardized method for calculating the loudness of a transient signal but the dynamic loudness seems to give fairly good results in aligning purposes based on the subjective loudness impressions from the tests. Based on the dynamic loudness values samples were boosted or attenuated so that the eventual loudness variations were below 1 dB as calculated by the previously described system.

After this the samples were filtered with "inverse-Sennheiser HD 600-to-HATS/DRP" FIR filter created with Matlab from the HD 600 frequency response measured with HATS. The HD600 average response from both channels was used for the filter and both channels of the samples were processed with it. The purpose of this was to remove the united HD 600 and HATS transfer function from the chain so that a listener would only have a specific PTF altering the sound in each sample case. The reason this was done after the loudness calibration was because in theory the real-HD600-and-the-head-of-the-listener – block and the filter used here are the same and in theory cancel out each other.

The HD600 compensation described above was rather basic and it was felt that the method could be improved. For this reason no further details of the procedure are presented. A better compensation was used for the third test described in Chapter 6.

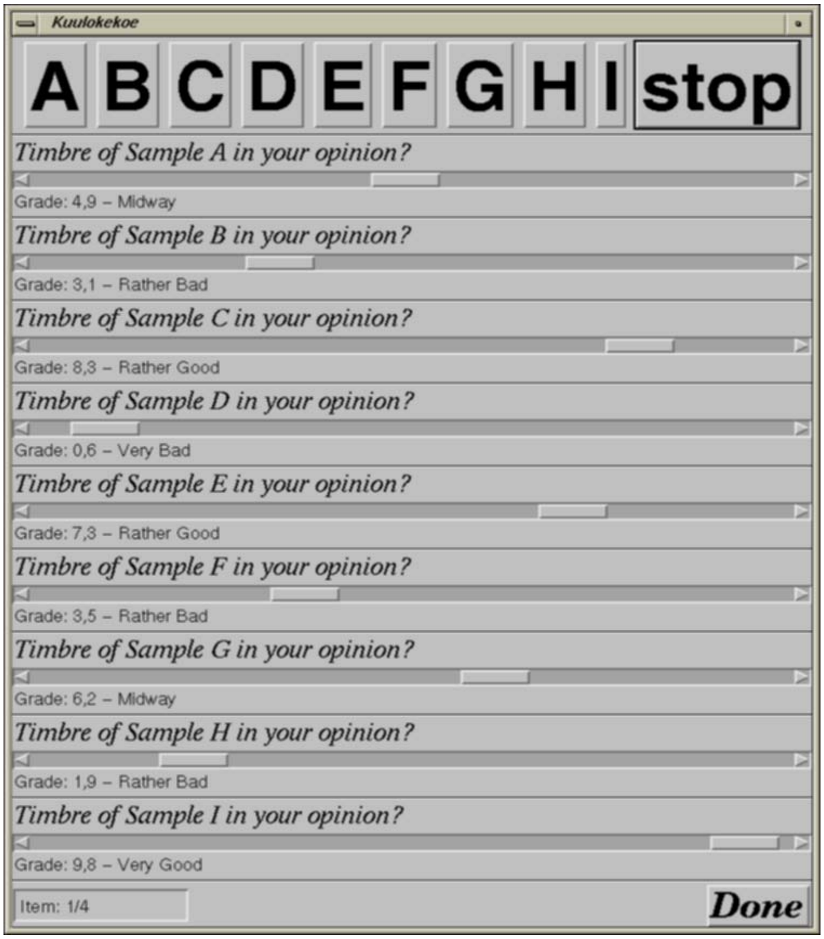
Regarding background noise in the recordings, it was agreed that speech samples could be filtered on the part that exceeds the bandwidth of the speech. In music samples the noise was not so audible Thus, speech samples were filtered with 20th order low-pass FIR filters created with Matlab so that the spectral content of the speech was unaffected. The -3 dB points for narrow- and wideband filters were 4 kHz and 7 kHz correspondingly. This procedure removed some of the noise.

The difference in stereo channel levels was a more difficult subject. If one channel is boosted or attenuated too much, the spectral differences might be overemphasized. That is why we preferred to leave most of the samples to their unequal channel levels, although they sounded somewhat different spatially. In some cases, however the effect was so severe that some alignment between channels was done. Since the speech was in "double-mono", the samples should have same relative spectral content in each channel. This way it could be checked that the alignment did not cause too much mutation in the sample. The channel-aligned samples were recorded with Mulan intra-concha, Sony intra-concha and Sony MDR headphones. This implies that the HATS ear is not ideal for recording intra-concha or small headphones.

5.4. Test setup

Eight subjects from NRC Ruoholahti who did the first test were also recruited for the second test with HATS recordings. No training session was arranged but the subjects were again instructed verbally on what to do prior to the test. The subjects listened to the recordings of the same ten samples as they heard in the first test, this time with only one pair of Sennheiser HD 600 headphones equalized to be “transparent” for the listener. Theoretically the samples should simulate the actual headphones. Only now the subject did not have to change between headphones, but could instantly switch from one test item to another. The external qualities of the headphones had no effect on grading this time, only the simulated sound quality.

Test was done in a regular office room with background noise minimized same way as in the first test. GP2 was again used for test implementation. The test window is shown in Figure 11.



Timbre of Sample A in your opinion?

Grade: 4,9 – Midway

Timbre of Sample B in your opinion?

Grade: 3,1 – Rather Bad

Timbre of Sample C in your opinion?

Grade: 8,3 – Rather Good

Timbre of Sample D in your opinion?

Grade: 0,6 – Very Bad

Timbre of Sample E in your opinion?

Grade: 7,3 – Rather Good

Timbre of Sample F in your opinion?

Grade: 3,5 – Rather Bad

Timbre of Sample G in your opinion?

Grade: 6,2 – Midway

Timbre of Sample H in your opinion?

Grade: 1,9 – Rather Bad

Timbre of Sample I in your opinion?

Grade: 9,8 – Very Good

Item: 1/4

Figure 11. The test window of GP2 in the second test. The nominal anchor points used in the first and the second test are visible.

The interface had eight buttons, one for each headphone recording of the sample. An additional ninth button was used to play the original unprocessed sample. This constituted the ninth test item, which should theoretically be the same as the "recorded-and-processed-HD600"- item. Comparing these two some idea about the validity of the recording method could be attained. Thus each case had nine items, whereas the first test had eight headphones. Toggling the buttons played the corresponding test items. The order of the headphone recordings was randomized so that the order was different on each sample, i.e. the order of test items was random. The test was otherwise similar to the first test; grade scale, breaks between sessions etc. were similar.

This time the term "sound quality" was replaced with "timbre" in the question so that the listeners would not focus their attention to background noise or spatial location of the sound. In the beginning of both the first and the second test the subjects were instructed the same way: To base their judgment on sound color preference and not on anything else. The actual literal question did not have so much weight in the grading process.

5.5. Results

Figure 12 shows the results from the second test. Grades of all eight subjects were again normalized with the method introduced in Equation (1).

The new aspect in this test was the "ninth item", an unprocessed sample played back through HD 600 headphones. Comparing the average grades with the recorded HD600 – item, some difference is found but the two dominate competition along with the AKG recordings.

The rest of the headphones have slightly changed their order compared to the first test. Mulan and Sony intra-concha headphones' grades have risen and Koss and Sony grades have lowered relatively. Panasonic intra-concha phones are again the worst, even more clearly this time.

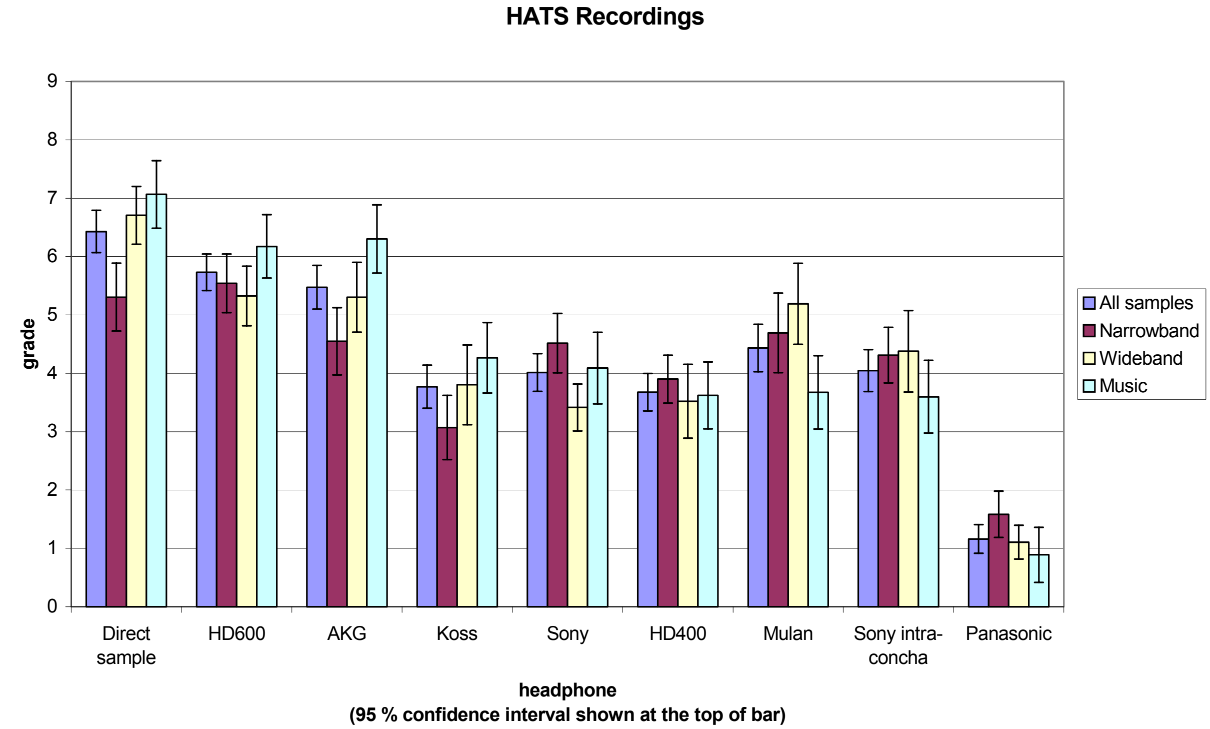


Figure 12. Grades of all eight subjects with different recordings. 95% confidence intervals are visible at the top of the grade bars.

The “shape” of different sample categories within headphones has also remained quite comparable to those of Figure 8; the mutual order between sample categories’ grade averages has mostly remained the same, especially with the two higher-end headphones.

5.6. Discussion

This section is largely hypothetical in nature. Little statistical analysis is performed because of the reasons discussed below.

Overall grades given without normalization were lower in the second test where HATS recordings were used as stimuli. The reason for this was probably the quality of the recordings. There was some background noise due to insufficient sound insulation of the recording booth and low playback levels compared to microphone noise. Some spatial deviations were also present in the recorded samples. Although minimized, it undoubtedly caused unnatural sensations. This can be verified when examining the "direct sample" and HD600 recordings grades in Figure 11. Theoretically these two should be the same but some difference can be seen among the category averages.

Another reason might be the lack of visual and physical reference. When performing blind tests, Toole found that expert listeners constantly gave lower grades than in sighted tests [15]. The lack of visual support leaves the listeners uneasy and perhaps afraid that they might give "wrong" grades. Interesting is that both tests had one common feature; The HD600 (referred as the "direct sample" in the second test) headphones were fed with unprocessed samples in both tests. As the non-normalized grades in the second test were lower for this item, Toole's theory is supported.

After the previous examination, the grades were normalized and the confidence intervals were calculated. As the number of subjects was only eight, the confidence intervals increased compared to the first test. This is generally not desirable and conclusions based on these wide intervals should be drawn cautiously.

Since normalization was done, the mutual order of the headphones is most important in Figures 8 and 12. When comparing the two pictures, the two "hi-fi headphones" (HD600 and AKG) did well in both tests. The Mulan intra-concha model was surprisingly good in the first test compared to more the expensive Sony intra-concha. In the second test they did even better. Mulan frequency response (see Appendix A) and other characteristics would be an interesting subject for further research dealing with preferred equalization. Since only eight subjects participated, the headphone grades in the "middle class" are rather ambiguous and no clear rank order can be determined there. Seems that the two best and the one not-so-good headphones were easy to categorize, but the other devices are quite difficult to rank based on sound quality alone.

All and all, the intra-concha grades were relatively higher and regular headphones' grades were correspondingly lower in the recordings test. There is probably some visual bias towards intra-concha models in terms of sound quality. In fact, most subjects recognized the Panasonic model from the recordings and that explains some of the inferior grades given to it in the second tests; Panasonic was collectively judged to be the "loser" of the first test.

A nice feature in the results is that the mutual grade order of the different sample categories within a given headphone model is rather similar in both first and second test. This supports the HATS recording method on its own part, only the confidence intervals are somewhat large.

At this point no further analysis was made based on the second test results. The confidence intervals were wide because only eight subjects were used and this makes the statistical analysis, although achievable, more difficult. A more important reason was the quality of the recordings; it was felt that the background noise and the spatial deviations should be minimized in order to examine the recording method properly. Many subjects found it hard to examine sound quality when these distracting factors were present. Furthermore, the HD600 compensation could be improved as well. Therefore based on this test, the method is in author's opinion not validated; the grades received by the "direct sample" and the HD600-recording differ at least in the wideband category beyond the wide confidence intervals. These two items should be similar before the validity of the HATS recordings is examined further. Because of these conclusions, a third test was devised.

6. Third Test

A third test which combined the methods used in the previous two tests was conducted during spring 2002. The reason for this was that the recording method used in the second experiment described in Chapter 5 was found not to be optimal for the purpose. Furthermore the number of subjects in the second test was limited to eight. This time 21 subjects were used. ANOVA was also used to extract the significant factors from the results.

6.1. Purposes of the Test

The goals of this test resemble those of the previous experiment. Some additional analysis methods were employed to study:

- first, *could an improved HATS recording and headphone reproduction method yield an adequate simulation of actual headphones with speech material to be used in subjective testing of sound quality*

- second, *whether the preference order of the headphones tested can be explained by measurable properties (for example frequency response, distortion) of the headphones.*

The first goal is basically the same as in the second test. This time however, the recording and processing of the test samples was done again with some additional improvements. The test was also simplified somewhat based on the observations made in the previous experiments. The preference order of the headphones and the speech-music correlation was examined earlier and there was no further need for similar research. To reduce the complexity of the test only narrow- and wideband speech were used as stimuli. The two most "radical-sounding" headphones were removed leaving six models to be graded.

These simplifications made it possible for the subjects to do one two-part test where both real headphones and recordings were examined.

The objective measurements made to the headphones are for the time being limited to frequency response. The preferred equalizations and the optimal design goals for headphones were examined. Some further analysis on this area can be done in the future.

6.2. Experimental Method

This test used methods based on the previous experiments, with slight modifications. All subjects did two sessions; in one they graded the real headphones similar to the first test and in the other, the processed HATS recordings as in the second test. The sessions and the break between them constituted the whole test. Some additional training and discussion compared to the previous tests was implemented before each session. The scale and interface were similar to the other experiments presented here. During the preparation stage, the frequency responses of the headphones were calculated. Especially the HD600 correction was this time done more precisely based on numerous measurements with several headphones of this model. Again, the results from both sessions were examined for similarities and ANOVA was employed for both sessions separately.

6.3. Test Variables

6.3.1. Headphones Used in the Test

Table 6 lists the six headphones used in this test for both regular listening and recording. These are the same devices as in previous tests, with the exception of two models removed: The Sennheiser HD400 and the Panasonic intra-concha were now absent. The latter model caused some problems in the second test when many subjects reported that they recognized and recalled the headphone from the recordings. This arguably caused the grades received by this model to decrease because of associations with bad sound quality as the Panasonic was universally thought to be the “loser” in the first test. This distinctiveness broke the double-blind condition. For this reason the decision was made to remove the headphones that stood out from the others. Panasonic and HD400 were the obvious choices since the Sennheiser model had a distinctive middle-frequency resonance (see Appendix A). In addition, the number of permutations decreased as some headphones were removed; the subject could now do both sessions in the same test.

Manufactu rer	Type
Sennheiser	HD600
AKG	K-240
Koss	KTX Pro
Sony	MRD 301
Nokia	HDR-1 Music Player headphones aka Mulan
Sony	MDR-E827G

Table 6. The headphones used in the third test.

6.3.2. Samples Used in the Test

Investigating the recording and simulation method with narrow- and wideband speech was deemed to be more important than with unprocessed music samples. The future applications of the technique would mainly involve speech samples which are limited to below 8 kHz. Commercial music bandwidth often exceeds this limit and as mentioned, the B&K HATS is not specified for these high frequencies. As a result, music was excluded from this experiment. Furthermore this omission made the test procedure less time-consuming.

Two male and two female speakers were chosen from the NATC Multi-Lingual Speech Database for Telephonometry to provide the samples. All four speakers used in this test were Finnish. As before, each speaker provided a 10 second segment of speech. One male and one female speech sample were coded with the AMR 12.2 kbit/s narrowband codec and the other two with AMR-WB 23.05 kbit/s wideband codec. This time no distinction was made between speakers and the subject listened to different speakers in narrow- and wideband. This was also the only factor by which the samples were sorted in the statistical analysis.

The actual codec does not filter out the low frequencies normally non-audible in mobile phones, as discussed in Section 4.3.2. This is why additional filtering before the coding/encoding process seemed appropriate. Table 7 specifies the filters used for this purpose. Filtering was done in CoolEdit Pro. The goal was to simulate telephone bandwidths more realistically.

Codec	lower -3 dB (4th degree Butterworth)	upper -3 dB (4th degree Butterworth)
AMRNB	0.25 kHz	3.5 kHz
AMR-WB	0.125 kHz	7 kHz

Table 7.The -3 db points of the 8th degree Butterworth filters used for speech samples prior to coding/encoding process.

After the coding/encoding the samples were aligned for loudness using the dynamic loudness meter with the method described earlier. This was done again for the recordings, as described in the next section. Next the speech samples were again upsampled from their 8 and 16 kHz sampling rates. The 10 second segments were looped so that the final samples were 30 seconds of length. All sound files in the test computer were in .wav -format CD quality (16 bit, 44100 Hz, stereo). Instead of true stereo however, the samples were “double-mono”.

These samples were used for both sessions of the test. No further processing was done for the part with real headphones; only the devices themselves were calibrated and adjusted for loudness (see Section 6.5.3.). To create the simulations for the recording test, a similar process as in the second test was required. This process is described in the next two sections.

6.3.3. Recording process

The recording environment used in this test was the small anechoic room in the HUT Laboratory of Acoustics and Audio Signal Processing. During the recording process done in the previous test, the background noise proved to be a problem. This time the noise did not cause problems due to the better insulation and fewer personnel. In addition, the samples had been aligned by the loudness tool to be at the same level before the recording. This allowed for maximum recording levels without distortion from the headphones to be used.

A B&K HATS model 4128C was placed in an anechoic chamber and each headphone was carefully fitted on it by turns. The four test samples had been converted to an audio CD. A Sony CD player with Symmetrix 304 headphone amplifier was used to play the samples through headphones. Files were saved in a Macintosh computer in CD-quality .wav-format.

Previously the HATS had been fitted with type 3.3 ears made from hard material. The type 3.3 is an anatomically realistic replica of the ear but the hard material prevented the AKG headphone model to cover the ear as well as it does cover the real human ear. For this reason type 3.3 ears made from softer material were used this time. The use of softer ears accounted for somewhat more realistic coupling of the headphones. Still, the AKG model did have some leakage when placed on the HATS. In addition, since the ears are made from the same mold, the use of softer material did not remedy a particular problem involving intra-concha headphones.

The spatial deviations on the recordings were the most serious problem at the last test. To correct this problem, the frequency responses of both channels of each headphone were compared prior to recording. The headphone was fitted on the HATS and both earshells' responses were measured with MLSSA system mounted on a desktop PC. MLSSA uses MLS technique to acquire impulse responses and allows for two measurements to be compared in the frequency domain. After a few iterations, the responses of the non-intra-concha earshells matched within 1.5 dB in the area of 0.1 - 5 kHz. The recording was done immediately after this without moving the headphones. The only exceptions were the intra-concha models, which were more difficult to fit similarly on the HATS ears; especially the Mulan model suffered from poor coupling to the ears. It seemed that the left and right ear of the B&K HATS were slightly different in shape so that the coupling of the intra-concha headphones to the left concha was tighter than to the right one. Intra-concha earshells' responses differed as much as 3.5 dB dB in the area of 0.1 - 5 kHz, despite of numerous re-fittings. The channel-comparison measurements of the headphones can be seen in Appendix B.

Along with the four speech samples, speech-like noise similar to the one used in the first test was also recorded through all headphones (see Table 5). Power spectral density estimate for these signals was calculated in Matlab. The headphone responses for both channels in Appendix B were calculated by reducing the original artificial speech spectra from the measured spectrum.

6.3.4. HD600 compensation

For this test, a new HD600 compensation filter was created. The aim was to measure several HD600 headphones and derive a more generic correction. The results were also used in another project. Three pairs of HD600 were measured with the MLSSA system described earlier. Each earshell was measured three times and the headphones were re-fitted between the measurements. The correction was done separately for both channels so that the nine responses per channel were used to design a two-channel inverse-HD600 IIR filter. This filter was used for the second part of the test.

6.3.5. Samples processing

As mentioned, the samples were aligned by their dynamic loudness prior to recording. For the first part of the test with real headphones, only these aligned original samples were needed because of the headphone loudness alignment (see Section 6.5.3).

For the second part where the HATS recordings were to be evaluated, some further preparation was needed. After recording, the obtained samples were again aligned by loudness tool because of the different headphone responses. Only after this, the HD600 compensation filter was used for the samples. This way the alignment would theoretically be preserved when the HD600 headphone and the compensation cancel out each other during the test.

The recordings were done so that the spatial deviations caused by the channel differences were minimized. There was a clear perceivable improvement on this part compared to the previous recordings. It was decided that no post-processing was needed for the recorded samples on this part. The recordings were however low-pass filtered with Matlab to remove some of the noise induced by the recording process. The filters used were similar to those presented in Section 5.3.2.

6.4. Test Subjects

The 21 subjects used in this test were university students from HUT. All of them had completed at lest some of the laboratory's acoustics courses and some had a professional background in audio and acoustics. The gender ratio was rather unilateral since only one female subject was tested. This was deemed to be an insignificant factor, based on Toole's results [15].

It was decided that a brief training session would be held prior to both parts of the test. Here the concept of timbre was specified and some training samples were presented in order to familiarize the hearing with the tasks ahead. These samples contained segments from all the test samples. No grading was required; rather the idea was to “warm up” the subject’s hearing. It was emphasized that the grading should be based on timbre alone and not on background noise or other factors. Training was done separately for both sessions. Possible hearing impairments were also inquired and none of the subjects reported one.

The first test suffered somewhat on “brand biasing” since many of the subjects recognized the Sennheiser HD600 model. Although there were a couple of such cases here, this time the phenomenon was relatively rare and is not considered important. Nevertheless, the general outlook and other external features of the headphones affected the evaluation during the first part of the test.

6.5. Test Setup

6.5.1. Test Site

The listening tests were done in the listening room of the Laboratory of Acoustics and Audio Signal Processing at HUT. The room offered somewhat ideal conditions for headphone listening and provided a smooth test interface. The image from the SGI monitor was projected to a video screen from outside the room. All the electrical equipment, e.g. amplifiers and computers were located in an isolated control room next to the listening room. Only the headphones, a keyboard and a mouse were placed inside with the listener. Details about the design and objective measurements of the listening room can be read from [46]. Sufficient to say, that there were no problems related to the background noise etc. distractions of the listening site.

6.5.2. Test Arrangement

The headphones were laid on a table and the brand indicators were covered with tape. Letters A through F were used to indicate the devices. The subject controlled the test interface with a mouse and a keyboard. The listening site aimed to be as ergonomic as possible so the subjects would not have to move more than absolutely necessary.

The test arrangement here was similar to the first two tests. The methods from these were combined in the two sessions that the subjects underwent. In one session the four samples were listened through the actual headphones and grading was performed based on timbre preference. The subject manually switched between the headphones. The second session was done solely with the HD600 headphones and the switching of the test items was done seamlessly via computer. The test was done with no time limit so the subjects could switch between the items as many times as they liked.

The four narrow- and wideband samples were listened in both sessions. The order of these samples varied with subject and session. Additionally, the order of the test items was randomized in the second session so the subject could not associate a certain timbre with any one letter. There was some concern that the subjects might show learning effects from previous sessions so the effect of the session order was investigated with ANOVA. Half of the subjects listened to the real headphones first and the recordings second and vice versa.

The GP2 was used again to implement the test interface. The grading scale was again 0.0 – 10.0 with the anchor points used before. The test window was similar to that used in the second test (see Figure 10). The difference was the number of test items: In the real-headphones session there were six representing the headphones. In the recordings session an extra item similar to the one used in the second test was added. This was again the unprocessed sample fed to the HD600 headphones. By comparing the “direct sample” to the recording simulating HD600 headphones, the quality of the recording process could be investigated.

The loudness of all the samples had been aligned according to their dynamic loudness by the Matlab loudness tool. A similar alignment was also done to the headphones. The measurements with speech-like noise made in the first test were used to align the final listening level to 80 +/- 1.5 phones. The listening level was constant for all subjects. It must be noted that individual differences in human anatomy make it difficult to align the actual subjective loudness of the headphones, especially the intra-concha models'. There were some remarks considering the alignment but the overall opinion was that they sounded equally loud.

The average time used per session was about half an hour. There was a five-minute pause between sessions. With the two five- to ten-minute training sessions, the total test time per subject was approx. one hour 15 minutes.

6.5.3. Statistical Experimental Design

The results from the two sections of the test were analyzed separately. The experimental design used in the ANOVA was as follows. In the session with real headphones the factors are headphone (HEADPHON i.e. H - six levels), sample type (S_TYPE i.e. S - two levels) and test session order (ORDER i.e. O - two levels). The headphone factor obviously derives from the six different devices. Sample type denotes whether the sample listened to is narrow- or wideband. In the analysis presented here, no further distinction of the samples (for example male-female speaker) was made. The effect of session order emerges from the fact that some people listened to the recordings first and real headphones in the following session and vice versa. The analysis for the recording session results was otherwise similar except the headphone factor had seven levels because of the use of an additional direct sample. The complete ANOVA model for the experiments is given in Equation (4):

$$Rating = \mu + H + S + O + H * S + H * O + S * O + residuals \qquad (4)$$

where:
μ: mean of all grades.

All factors are considered fixed and up to second order interactions have been included. Note that the ANOVA was used for results normalized according to Equation (2). This is the reason why “subject” is not included in the model as a factor. It is assumed that all the subjects are equal. Any deviation from this is included in the residuals.

6.6. Results

6.6.1. ANOVA Results

The data from the headphones section and recordings section were analyzed separately. A normalization of test subjects was performed for these sections according to Equation (1). All the subjects were used for examination. A univariate ANOVA model given in Equation (4) was then used for analysis. The ANOVA tables are presented in Appendix 3.

The significant ($p<0.05$) factors were:

- Headphone session: HEADPHON, S_TYPE, HEADPHON*S_TYPE, HEADPHON*ORDER, S_TYPE*ORDER
- Recordings session: HEADPHON, S_TYPE, HEADPHON*S_TYPE

6.6.2. Comparison of the First and Second Session Grades

Averages of normalized grades from both sessions are presented in Figures 13 and 14. The two sessions were analyzed separately. 95% confidence intervals calculated from standard deviations are visible at the top of the grade bars. The results of all 21 subjects were used for both calculations.

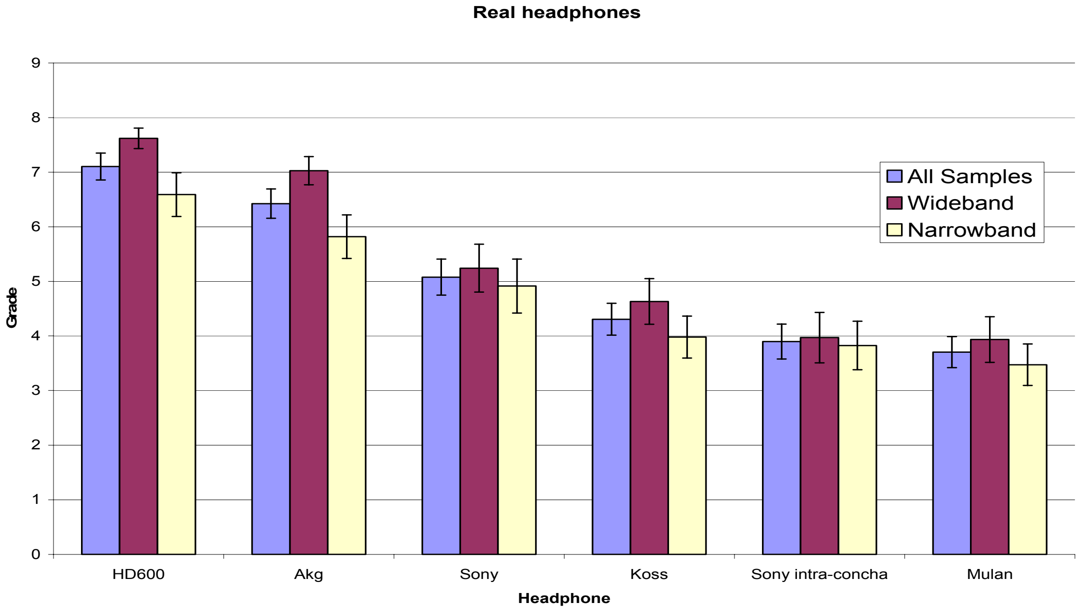


Figure 13. Grades of all subjects with different headphones in the third test.

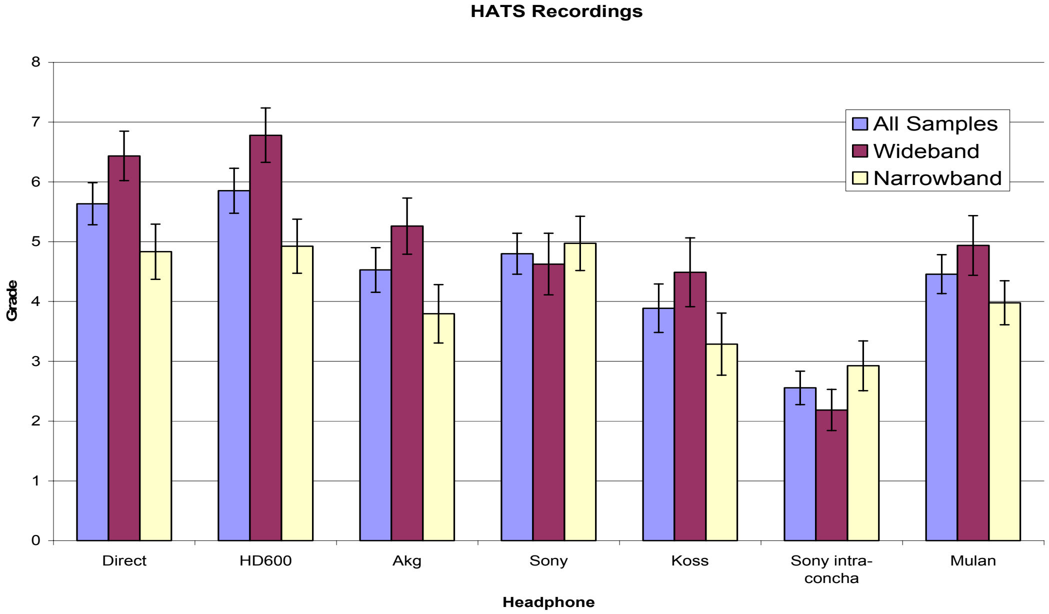


Figure 14. Grades of all subjects with different recordings in the third test.

6.7. Discussion

In this section some possible background for the results is presented. The reasoning is somewhat at a speculative level.

6.7.1. ANOVA Main Effects

From the three main effects, HEADPHON and S_TYPE were significant in both parts of the test. The former indicates that different headphones were given grades at different distributions, i.e. the headphones as well as the recordings were thought to vary in terms of timbre. This fact is of course what the whole test design is based on.

The significance of the S_TYPE factor indicates that narrowband and wideband speech are regarded to be different from each other. A similar observation was made in the two previous experiments; based on subjects' comments the device comparison was more difficult with narrowband speech. This was not surprising due to narrower bandwidth. Wideband speech was considered almost natural while narrowband speech in more discerned.

The significance of a factor in ANOVA is increased when the associated F-value grows. When examining the F-values of the main effects in Appendix 3, several suggestions of visual biasing can be seen. The HEADPHON F-value is much larger in the real headphones session than with the recordings. It is assumed that the subjects felt more insecure in the HATS recordings session without any cues helping them to associate timbres with certain devices. The recordings session was done double-blind with no visual or ergonomic references. In addition, S_TYPE is more significant with recorded samples where the subjects perhaps paid more attention to sound.

The ORDER factor alone was insignificant in both sessions. It must be noted that this refers only to the lack of fatigue or other similar effects caused by the test procedure. The order of sessions produced significant interactions with other factors in the real-headphone part of the test.

6.7.2. HEADPHON*S_TYPE Factor

The interaction between headphone model and sample type produced a significant effect in both tests. This can be interpreted by stating that different headphones received different grades depending on if the sample was wideband or narrowband. Based on F-values, the influence was more severe in the HATS-recording session, whereas in the headphones session the effect was almost insignificant. Grades given to different headphones with narrow- and wideband speech are visible in Figures 13 and 14.

As Figure 13 shows, only the higher-end headphones HD600 and AKG have deviation outside the confidence intervals among the two sample categories in the headphones-session. With other headphones the difference is within the intervals. This accounts for the small F-value. On the other session shown in Figure 14 more severe differences between headphones are noted. Especially the Sony intra-concha model has higher grades with narrowband samples, whereas the trend with other headphones (except for the other Sony model) is clearly the opposite.

The reasons for this dissimilarity of headphone recordings are at the time speculative. Perhaps the subjects felt insecure when grading the devices based on the narrowband samples with their limited spectral range. Many subjects did comment about the difficulty of the recordings-session and it was admittedly demanding. With wideband samples more distinction of timbres could be made and so the grade averages are spread out more than corresponding narrowband marks in Figure 14.

During the real headphones-session, the visual reference was probably used in grading causing the more similar results between sample categories. This biasing theory is supported by the fact that in the first test, the Sony intra-concha received high grades on their appearance (see Figure 10). It might explain why the grades of this model have decreased relatively in the HATS recordings-session.

6.7.3. HEADPHON*ORDER Factor

While the previous interaction surfaced in both parts of the test, this and the following section focus merely on the session with real headphones. Only the results from this part are considered. The F-values associated with these two factors are not large but nevertheless some justification to their significance can perhaps be found below.

The significance of factor HEADPHON*ORDER indicates that the subjects graded different headphones dissimilarly depending on had they done the recordings-session before this session. Figure 15 shows all the average grades given to the headphones depending on the sequence of the test. The effect is not so severe since all the variations are within the 95% confidence intervals visible at the top of the grade bars.

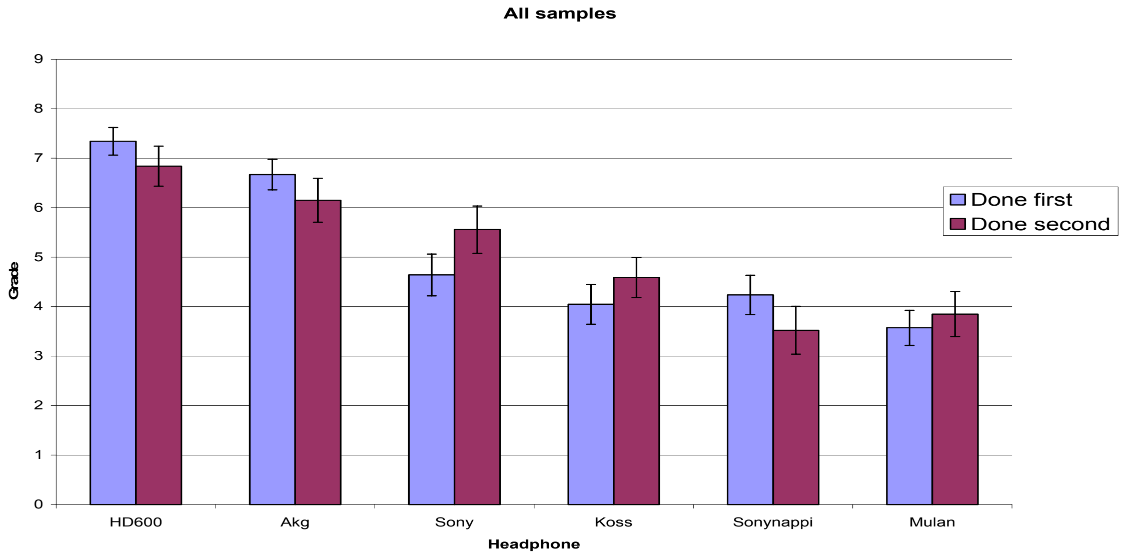


Figure 15. Grades of all samples with real headphones, depending on the order of the test sessions.

On the other hand a reasonable hypothesis is that the subjects who did the recording session first did evaluate the real headphones based more on their actual sound quality. The arguably difficult task of finding audible differences without visual reference on the recorded samples could have conditioned subjects to base their judgment more on hearing. In other words they trained their ears more than the group who did the real headphone-section first. This deduction is supported by Figure 15 where the grades of visually attractive and expensive-looking headphones, for example Sony intra-concha, received lesser grades when evaluated in the second session. The more “earthly-looking” headphones, for example Sony MDR model, were better received in the same situation.

So again the visual biasing effect can be seen to affect the results. An extended training session could perhaps have lessened the significance of the test session order.

6.7.4. S_TYPE*ORDER Factor

This factor had also some significance in the session with real headphones. It can be seen from Figure 16 that the subjects generally gave higher grades to the wideband samples if they had done the recordings-session prior to this part. The narrowband samples were respectively given lesser grades. This again alludes to the learning effect invoked by the recordings-session. Subjects who had done the other part beforehand based their grading more on what they heard and not on external qualities of the headphones. This is why they probably “had the courage” to gave grades at wider intervals. Again, this factor is of minor significance because the discrepancies in Figure 16 are within the confidence intervals.

6.7.5. Comparison between Recorded HD600 and Direct Sample

The recordings-session had two items that were theoretically similar: The recorded and processed HD600 and the unprocessed sample listened directly through the HD600 headphones used during the session. When examining Figure 14, it can be seen that the grades received by these two are similar within the 95% confidence intervals. There was a clear improvement compared to the recordings in the second test as the spatial deviations were minimized. Thus the recording process utilized in this test is deemed to be sufficient for the purpose of the test. This does not imply that the recordings were ideal in terms of background noise, distortion etc.

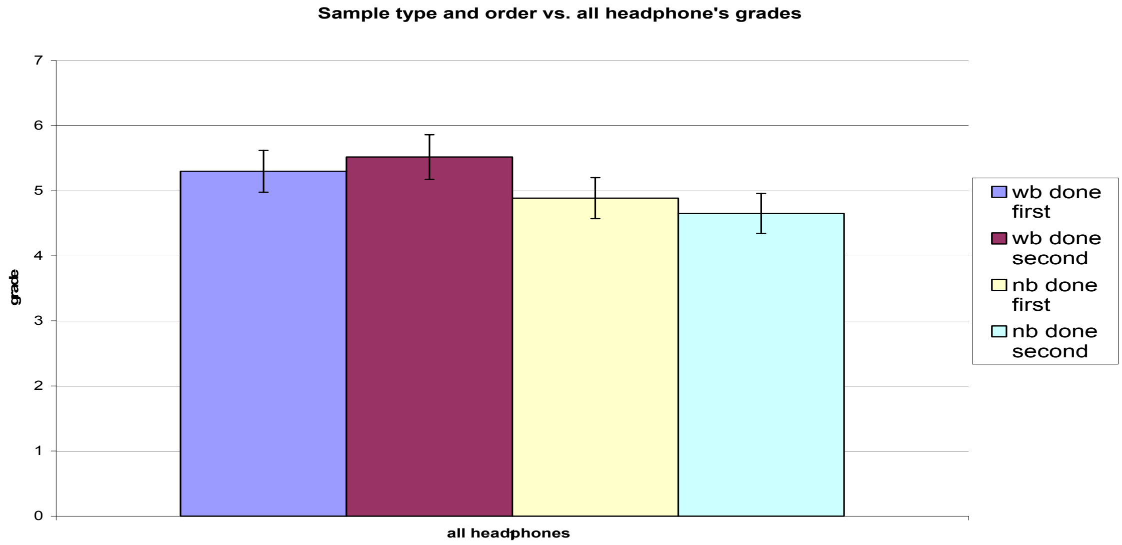


Figure 16. Grades of all headphones in different sample categories, depending on the order of the test sessions.

6.7.6. Comparison between Two Test Sessions

Comparison with the two session results can be made to determine whether the HATS recordings were evaluated to have similar sound quality as the real headphones. The results in Figures 13 and 14 are normalized by the respective session so the absolute grades are not a very important feature here; the mutual rank order of the headphones with different samples is the most prominent aspect that should be considered.

It can be seen that the Mulan intra-concha was appreciated more highly in the recordings-session. The headphone received similar grades as the Sony intra-concha during the session with actual headphones but in the recordings-part the grades have risen beyond confidence intervals. The other models have also some deviations between sessions but the effect is most radical between the intra-concha models. This dissimilarity along with the visual biasing effects discussed in previous sections give basis for stating that external qualities of headphones affected the evaluation of sound quality in this test. This is also evident when examining the external quality evaluation done in the first test (see Figure 10); at least Sony intra-concha's grades were apparently influenced by its outlook. Toole's results about visual biasing obtained in blind/sighted loudspeaker tests are thus supported [15].

In general, the differences between sample categories within headphones are more drastic in the recordings-session. This was discussed in previous sections. Also interesting is that all headphones except the both Sony models were given better grades with wideband samples than with narrowband samples in the recordings-section. No further hypothesis is presented to explain this at this point; examining the frequency responses of these headphones might give some clues on this topic.

6.7.7. Diffuse-Field Responses of the Headphones

The measurements shown in Appendix B were corrected with diffuse-field values given by B&K to obtain the headphone diffuse-field responses. These curves are given in Appendix D for channel averages of the headphones. It must be remembered that the HATS ear is not specified for frequencies over 8 kHz so above this the curves are speculative.

According to Thiele's theory discussed in Section 3.3.2, the flatter the diffuse-field magnitude response, the more natural the headphone should sound. It can be seen that the HD600 achieves the flattest response (within 7 dB in the region of 0.05 – 3.5 kHz). The other "high-end" model, AKG is also quite near that. The other headphones are somewhat inferior to these two with Sony intra-concha having perhaps the “worst” diffuse-field response. The measurements approximate the preference order given in the subjective tests, especially in the double-blind recordings-session. Even when using speech as stimulus, the preference of timbre seems to correlate with the flatness of the diffuse-field response.

7. Summary

The purpose of this chapter is to give answers to the questions presented in Section 1.3 based on the results of the test results obtained. Some of the conclusions are at this point hypothetical. In addition, ideas for further analysis are presented.

7.1. Conclusions

One of the most important goals was to determine the suitability of HATS recordings to replace real headphones in listening tests. The recording method was optimized for the third test. The test results with the HATS recordings played back through a pair of equalized high-quality headphones had similarities with the results obtained using real headphones but there is still some significant difference in the results. The reasons for this lie behind biasing effects caused by the outlook of the devices and other external qualities such as ergonomics. The method can not be validated based on these tests.

It seems that the ”preferred spectrum” of sound depends strongly on the type of sound itself. When the subjects in this test listened to music, boosted lower frequencies were usually regarded as a good thing but emphasized bass can reduce the intelligibility of speech. These tests were conducted in a noise-free environment and in presence of background noise, a more high frequency-boosted sound could be preferred. The intra-concha headphones are perhaps not ideal for music reproduction mainly due to their inadequate lower frequency response.

The preference order of the tested headphones was somewhat as expected. The Sennheiser HD 600 and AKG K-240 were generally regarded as the best ones. The Mulan intra-concha received significantly better grades compared to the Sony intra-concha model with HATS recordings than in real headphone listening. The Sony intra-concha model was regarded as attractive in the question-form evaluations of the first test. Thus the theory about visual biasing in audio listening is supported. The diffuse-field response seems to be a valid measure for headphone sound quality at least when spatial expanders or similar methods are not used.

The concept of using music samples instead of wideband speech excerpts in listening tests proved not to be practical. Although according to the first test the concept seems to be viable, it is most likely that in other test conditions (for example in a noisy environment) the method would not work. It seems that bass reproduction is a key factor since too heavily boosted low end probably discriminates the music and wideband speech more. Narrowband speech differs from the above stimulus types significantly and evaluating audio sound quality with it is a difficult task. No further investigation is going to be made regarding this topic since the method was found to be merely of academic interest. It is recommended that speech is still used in listening tests when appropriate.

Detailed analysis of the third test results showed some interesting effects caused by the test session order and the type of listened sample. The subjects' grades were influenced depending on had they done the recordings-session before the real headphones-session. Arguably the hearing was "trained" more if there was no visual reference in the first session. This in turn implies that the actual training session before the test could have been extended. The significant difference between wide- and narrowband coded speech was again encountered in the third test results. Based on subjects' comments, wideband speech is more suitable for sound quality evaluation of audio devices.

7.2. Future Work

Subjective testing provides the experimenter with results that can be analyzed extensively. The material obtained from these tests can also be used in the future applying various methods for information extraction.

There is an intention to study more on how objective measurements can be linked to subjective preference of sound quality. So called “preference mapping” method is used to find correlation between subjects’ grades and the measured magnitude response of the headphone. This is one of the techniques that will probably be employed in the future on these results.

The HD600 compensation filter design has been under some investigation. A series of verification measurements are being conducted to determine the authenticity of the method.

References

[1] Møller, H. Fundamentals of Binaural Technology. *Applied Acoustics*, vol. 36. 1996.

[2] Shinn-Cunningham, B *et al*. Auditory Displays. In Gilkey R. and Anderson T. (Eds.) pp. 611-663. 1997.

[3] Riederer, K. A. J. Head-Related Transfer Function Measurements. M.Sc. thesis. Helsinki University of Technology. 1998.

[4] Hammreshøi, D. Fundamental Aspects of the Binaural Recording and Synthesis Techniques, *Audio Engineering Society preprint, 100th Convention*. 1996.

[5] 3GPP TS 26.190, Version 5.0.0, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Speech Codec Speech Processing Functions; AMR Wideband Speech Codec; Transcoding Functions. 2001.

[6] Karjalainen M. *Kommunikaatioakustiikka.* Otamedia OY. 2000.

[7] Eysenck, Michael W. and Keane Mark T. *Cognitive Psychology: A Student's Handbook; 4th ed.* Psychology Press. 2000.

[8] Precoda, Kristin and Meng, Teresa H. Subjective Audio Testing Methodology and Human Performance Factors. *Audio Eng. Soc. 103rd Convention.* Preprint No. 4585. 1997.

[9] Gabrielson, A. and Sjogren, H. Percieved Sound Quality of Sound Reproducing Systems. *J. Acoust. Soc. Am.* Vol. 65. 1019. 1979.

[10] Mattila, Ville-Veikko and Zacharov, Nick L. Generalized Listener Selection (GLS) Procedure. *Audio Eng. Soc. 110th Convention*. Preprint No. 5405. 2001.

[11] International Telecommunications Union. *ITU-R Recommendation BS.1116*. 1994-1997.

[12] Olive, Sean E. A Method for Training Listeners and Selecting Audio Material for Listening Tests. *Audio Eng. Soc. 97th Convention*. Preprint No. 3893. 1994.

[13] Zacharov, Nick. Measurement and Analysis of Perceptual Attributes. *Measurements and modeling in acoustics and audio: Seminar in acoustics, Helsinki University of Technology, Laboratory of Acoustics and Signal Processing, spring 2002*. Otamedia Oy. pp.145-168. 2002.

[14] Bech, Søren. Listening Tests on Loudspeakers: A Discussion of Experimental Procedures and Evaluation of the Response Data. *Audio Eng. Soc. 8th International Conference.* 1990.

[15] Toole, Floyd E. and Olive Sean E. Hearing is Believing vs. Believing is Hearing: Blind vs. Sighted Tests, and Other Interesting Things. *Audio Eng. Soc. 97th Convention*. Preprint No. 3894. 1994.

[16] Sneegas, James E. Aural Acuity and the Meaning of Sound Quality: A Cultural Approach. *Audio Eng. Soc. 83rd Convention*. Preprint No. 2489. 1987.

[17] Toole, Floyd E. Listening Tests – Turning Opinion Into Fact. *Audio Eng. Soc. 69th Convention*. Preprint No. 1766. 1981.

[18] Lipshitz, Stanley P. and Vanderkooy, John. The Great Debate: Subjective Evaluation. *Audio Eng. Soc. 65th Convention*. Preprint No. 1563. 1980.

[19] CCITT, ITU. *Handbook on Telephonemetry.* Geneva. 1992.

[20] Risch, Jon M. A User Friendly Methodology for Subjective Listening Tests. *Audio Eng. Soc. 91th Convention*. Preprint No. 3178. 1991.

[21] Toole, Floyd E. Subjective Evaluation: Identifying and Controlling the Variables. Prensented in *Audio Eng. Soc. 8th International Conference*. 1990.

[22] Lucent Digital Radio Inc. Selection of Audio Samples & FM Processing. *http://www.nab.org/SciTech/Dab/Ldrappendixb.pdf*. 20.8.2002.

[23] Leventhal, Les. Type 1 and Type 2 Errors in the Statistical Analysis of Listening Tests. Vol. 34. No. 6. 1986.

[24] Milton, J. and Arnold J. *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences , 2nd Ed* . McGraw-Hill. 1990.

[25] Ruherford, A. *Introducing Anova and Ancova: A GLM Approach.* Sage Publications Ltd. 2000.

[26] Poldy, C. A. Headphones. *Loudspeaker and Headphone Handbook.* Focal Press. pp. 493-574. 1994.

[27] Backman, J. Study material for course: "S-89.104 Sähköakustiikka". Helsinki University of Technology, Laboratory of acoustics and Signal Processing. Edita Prima OY. 2002.

[28] International Telecommunications Union. *ITU-T Recommendation P.57*. 1993.

[29] Rayleigh. L. On our perception of sound direction. *Philosophical Magazine* 13. 1907.

[30] Rossing, T. D. *The Science of Sound.* Addoson-Wesley Publishing Company. 1990.

[31] McFadden, D and Pasanen, E.G. Lateralization at High Frequencies Based on Interaural Time Differences. *J. Acoust. Soc. Am.* Vol. 59, 769. 1976.

[32] Blauert, J. *Spatial Hearing.* MIT Press. 1983.

[33] Begault, D. *3-D Sound for Virtual Reality and Multimedia.* Academic Press Professional. 1994.

[34] Toole, F. E. In-Head Localization of Acoustic Images. *J. Acoust. Soc. Am.* Vol. 48, 943. 1970.

[35] Plenge, G. On Differences between Localization and Lateralization. *J. Acoust. Soc. Am.* Vol. 56, 944. 1974.

[36] Toole, F. E. Acoustics and Psychoacoustics of Headphones. Prensented in *Audio Eng. Soc. 2nd International Conference*. 1984.

[37] Musson, W. A. and Wiener F. M. In Search of the Missing 6 dB. *J. Acoust. Soc. Am.* Vol. 39, 465. 1952.

[38] Thiele G. On the Standardization of the Frequency Response of High-Quality Studio Headphones. *J. Audio. Eng. Soc.* Vol. 34, No 12. 1986.

[39] Villchur E. Free-Field Calibration of Earphones. *J. Acoust. Soc. Am.* Vol. 46, 1527. 1969.

[40] International Telecommunications Union. 1969. *ITU-R Recommendation BS.708*. 1969.

[41] Møller, H. *et al*. Design Criteria for Headphones. *J. Audio. Eng. Soc.* Vol. 43, No 4. 1995.

[42] Hynninen J. and Zacharov N. Guinea Pig – A Generic Subjective Test System for Multichannel Audio. *Audio Eng. Soc. 106th Convention*. Preprint No. 4563. 1999.

[43] Gabrielson A. and Lindström B. Perceived Sound Quality of High-Fidelity Loudspeakers. *J. Audio. Eng. Soc.* Vol. 33, No. ½. 1985.

[44] Moore B. *et al*. A Model for the Prediction of Thresholds, Loudness and Partial Loudness. *J. Audio. Eng. Soc.* Vol. 45, No. 4. 1997.

[45] Pearsons K. *et al*. Speech Levels in Various Noise environments. *U.S. Environmental Protection Agency, Washington D.C.* Report EPA-600/1-77-025. 1977.

[46] Järvinen A.. Kuunteluhuoneen suunnittelu ja mallinnus. M.Sc. thesis. Helsinki University of Technology. 1999.

Appendix A: Headphone Measurements from First Test

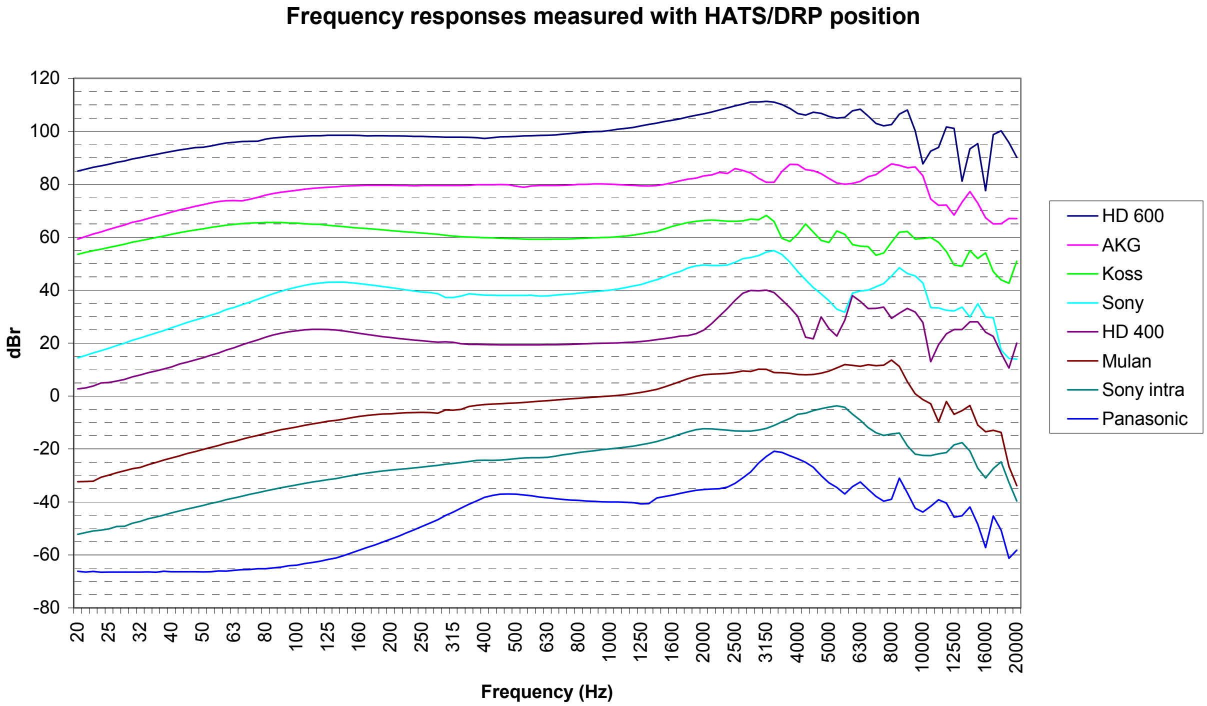


Figure 17. Magnitude responses of the test headphones measured with B&K HATS. Curves are shifted to have 20dB difference to the previous curve at 1 kHz.

The measurements in Figure 17 were done by Ossi Mäenpää in NMP Salo Acoustics laboratory with B&K HATS. Note that AKG K-240 response is possibly lacking low end because type 3.3 hard ears were used.

Appendix B: Headphone Measurements from Third Test

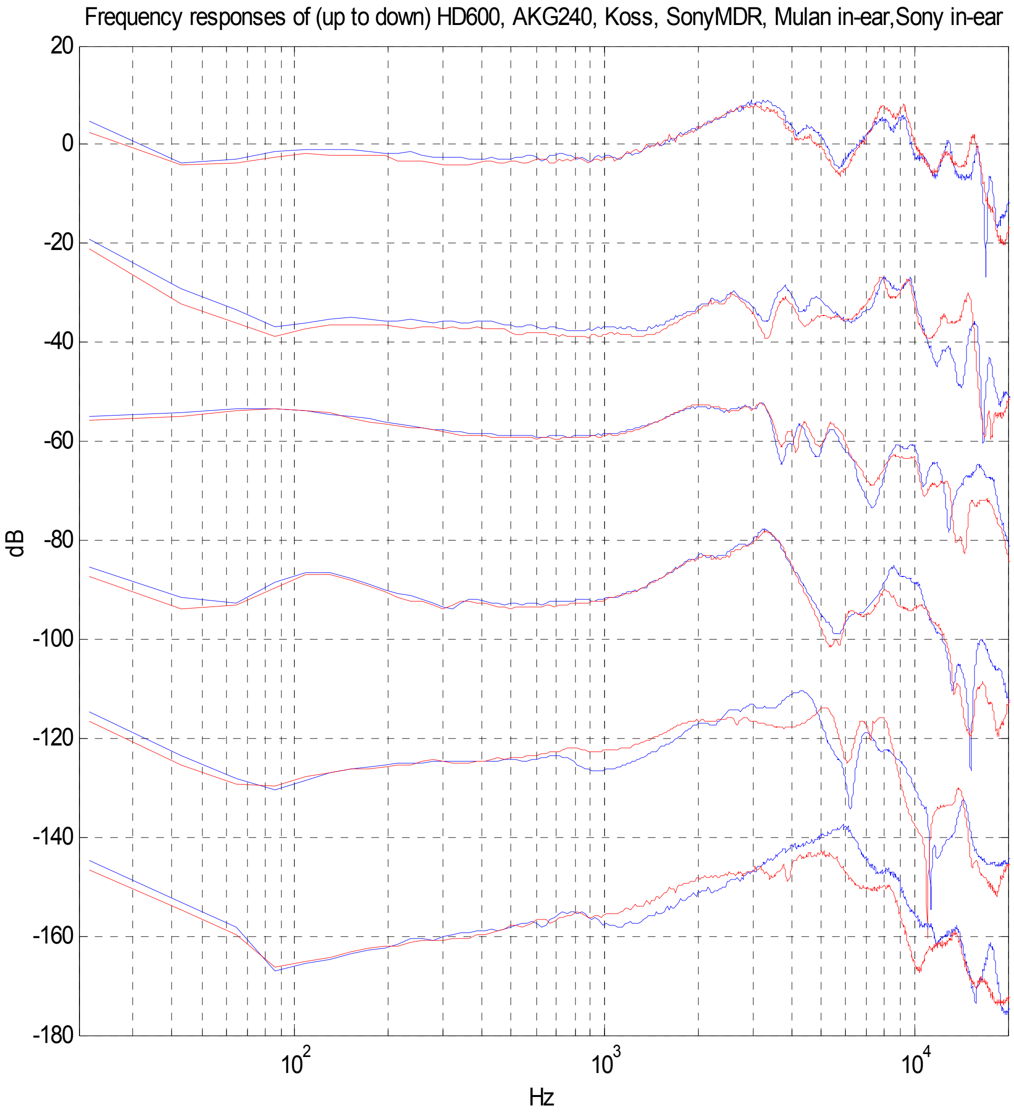
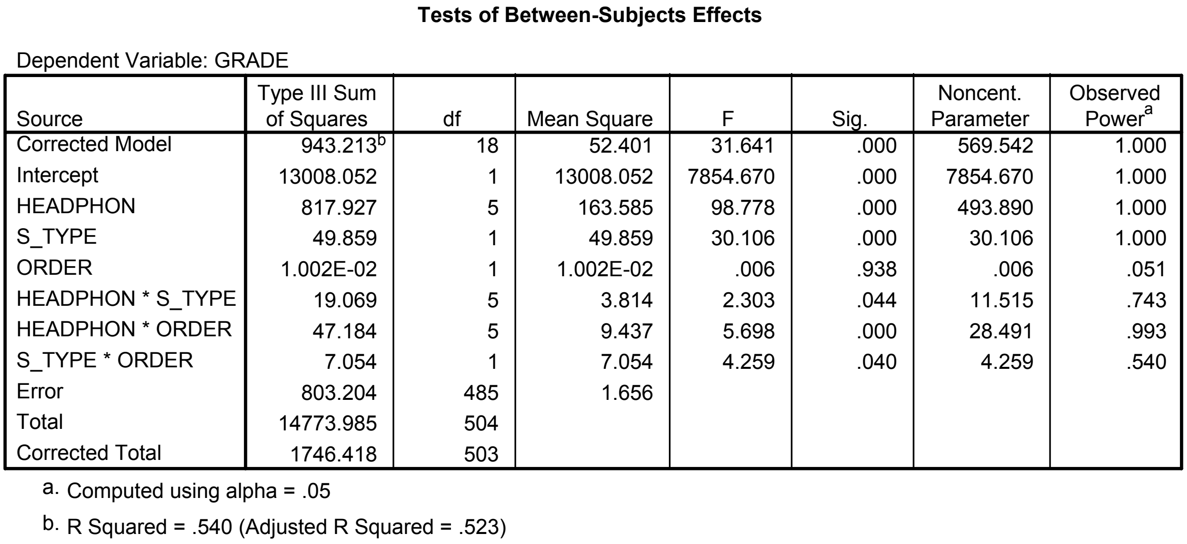


Figure 18. Magnitude responses of the test headphones calculated from recorded pseudo-speech signals. Two channels are shown in blue and red for each device. The curves of different headphones have been shifted to achieve comparability.

Measurements in Figure 18 include both channels of the headphones' magnitude responses. The differences between channels indicate the imperfectness of the recordings. Most notable deviations can be seen with the intra-concha headphones. Nevertheless, the recording process was determined to be sufficient for the test purposes. The low end frequencies below 0.1 kHz on the curves are not reliable as some responses show an unnatural bass boost caused by processing.

Appendix C: ANOVA Tables



Tests of Between-Subjects Effects

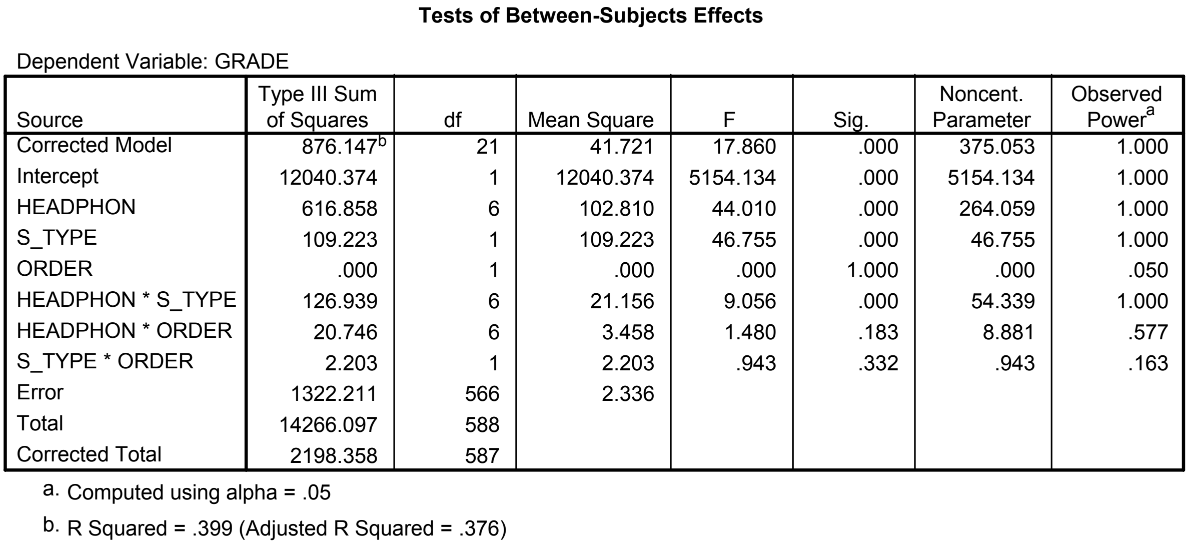
Dependent Variable: GRADE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
Corrected Model	943.213[b]	18	52.401	31.641	.000	569.542	1.000
Intercept	13008.052	1	13008.052	7854.670	.000	7854.670	1.000
HEADPHON	817.927	5	163.585	98.778	.000	493.890	1.000
S_TYPE	49.859	1	49.859	30.106	.000	30.106	1.000
ORDER	1.002E-02	1	1.002E-02	.006	.938	.006	.051
HEADPHON * S_TYPE	19.069	5	3.814	2.303	.044	11.515	.743
HEADPHON * ORDER	47.184	5	9.437	5.698	.000	28.491	.993
S_TYPE * ORDER	7.054	1	7.054	4.259	.040	4.259	.540
Error	803.204	485	1.656				
Total	14773.985	504					
Corrected Total	1746.418	503					

[a] Computed using alpha = .05

[b] R Squared = .540 (Adjusted R Squared = .523)

Figure 20. Real Headphones session results ANOVA table.



Tests of Between-Subjects Effects

Dependent Variable: GRADE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
Corrected Model	876.147[b]	21	41.721	17.860	.000	375.053	1.000
Intercept	12040.374	1	12040.374	5154.134	.000	5154.134	1.000
HEADPHON	616.858	6	102.810	44.010	.000	264.059	1.000
S_TYPE	109.223	1	109.223	46.755	.000	46.755	1.000
ORDER	.000	1	.000	.000	1.000	.000	.050
HEADPHON * S_TYPE	126.939	6	21.156	9.056	.000	54.339	1.000
HEADPHON * ORDER	20.746	6	3.458	1.480	.183	8.881	.577
S_TYPE * ORDER	2.203	1	2.203	.943	.332	.943	.163
Error	1322.211	566	2.336				
Total	14266.097	588					
Corrected Total	2198.358	587					

[a] Computed using alpha = .05

[b] R Squared = .399 (Adjusted R Squared = .376)

Figure 21. HATS recordings session results ANOVA table.

Appendix D: Headphone Diffuse-Field Responses,

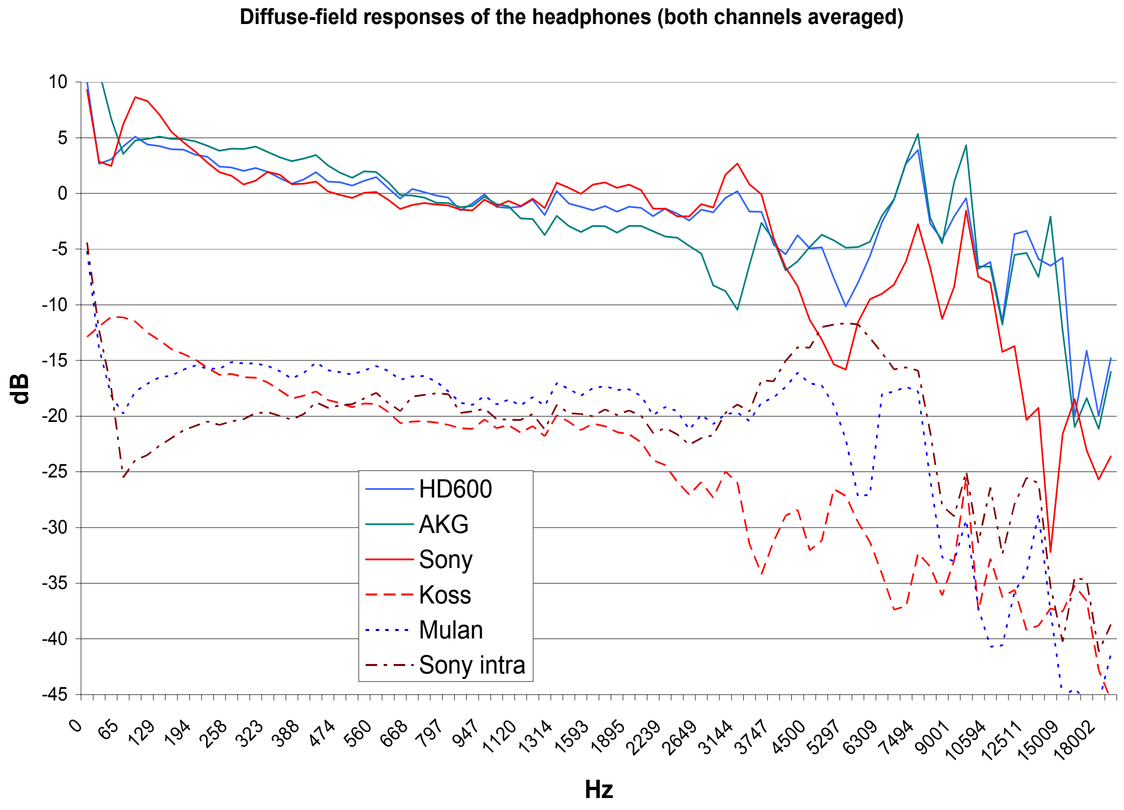


Figure 22. Diffuse-field corrected responses of the headphones used in the third test. The curves have been shifted to achieve comparability.

The measurements shown in Appendix B were equalized on a third-octave scale using diffuse-field information provided by B&K. The flatness of the curves is theoretically regarded to be desirable for natural sound. The curves in Figure 22 are not reliable for frequencies below 0.05 and above 8 kHz.