Aalborg University

4th Semester Acoustics

The Role of Compression in the Subjective Evaluation of Music Quality

Authors:
Mathias Bødker Borup
Paul Zabbal

Supervisor:
Rodrigo Pizarro Ordoñez

June 3, 2014

Department of Electronic Systems
Acoustics
Fredrik Bajers Vej 7
Phone: +45 9940 8710
http://www.es.aau.dk/sections/acoustics

Title:
The Role of Compression in the Subjective Evaluation of Music Quality

Topic:
Master Thesis

Group:
1063

Participants:
Mathias Bødker Borup
Paul Zabbal

Supervisor:
Rodrigo Pizarro Ordoñez

Pages: 82

Appendices: 3

Project period:
Spring Semester 2014

Handed in:
4^{th} of June 2014

Synopsis:

The motivation for this master thesis is to investigate which impact compressed music has on subjective listening quality. The role of loudness is investigated and program loudness is used to normalize and eliminate it as a factor. The compressed excerpts for the listening test are generated using fixed parameters except for the threshold. The listening test is performed using three different degrees of compression; uncompressed, -12 dBFS and -24 dBFS. Four attributes are evaluated for four different genres. Results from the test show that subjects have difficulties to distinguish the three degrees of compression. The audible cues for the compression are present in the loudness differences and therefore removed by the normalization.

Foreword

This master thesis and the results presented within, are developed under the Department of Acoustics and Audio Technology at Aalborg University during the period from the 1st of February to the 4th of June 2014. This report is based on the proposal posed by supervisor Rodrigo Ordoñez regarding *The role of loudness and dynamic range in the evaluation of music sound quality.*

We would like to thank TC Electronic for the invitation to show us their facility and to Thomas Lund and Esben Skovenborg for inspiration and ideas to the project approach, Prof. Brian C. J. Moore for joining a discussion about the perception of loudness, Claus Vestergaard Skipper for technical support and Roger Langvik for providing uncompressed music.

Mathias Bødker Borup **Paul Zabbal**

Contents

1 Introduction and Initial Problem **1**

2 The Role of Loudness **3**
2.1 Loudness and Loudness Level . . . 3
2.2 Perception of Loudness . . . 4
2.3 Masking Effect and Auditory Filters . . . 6
2.4 Physical Description of Music and Its Characteristics . . . 8
2.5 Loudness in Music . . . 10
2.5.1 Loudness Normalization . . . 10
2.5.2 Measuring Loudness . . . 11
2.5.3 Calculation of Loudness in Music . . . 12

3 Compression Parameters and Algorithms **15**
3.1 Data Compression Algorithms in Music . . . 15
3.1.1 Lossless Compression . . . 15
3.1.2 Lossy Compression . . . 16
3.2 Dynamic Range Compression . . . 16
3.2.1 Main Parameters of a Compressor . . . 16
3.2.2 Static Curve . . . 17
3.2.3 Dynamic Behavior . . . 19

4 Internet Questionnaire Survey **23**
4.1 General Considerations . . . 23
4.2 Design of Questionnaire . . . 24
4.3 Results . . . 25

5 Test Design **31**
5.1 Questionnaire about the Hearing . . . 31
5.2 Audiometric Test . . . 32
5.3 Otoacoustic Emission Measurements . . . 33
5.4 Listening Test Design . . . 33
5.4.1 Previous Studies . . . 33
5.4.2 Listening Panel . . . 34
5.4.3 Program Material . . . 34
5.4.4 Attributes . . . 35
5.4.5 Listening Conditions . . . 36
5.4.6 Test Method and User Interface . . . 36

6 Results and Data Analysis **39**
6.1 Subjects and Response from the Questionnaire . . . 39
6.2 Audiometric Test and TEOAE Measurements . . . 39
6.3 Listening Test . . . 41

7 Conclusion and Discussion **51**
7.1 Project Summary . . . 51
7.2 Discussion . . . 53

A Measurement Journal **59**

B Questionnaire - Listening Habits **67**

C Questionnaire - Hearing **71**

Chapter 1

Introduction and Initial Problem

The evolution in personal music players (PMP) has left the user with a variety of possibilities. Starting from the first vinyl records, the evolution of playback platforms took a big step with the introduction of the CD and digital format, allowing more mobility in music. Nowadays, mobile phones and other PMPs are widespread allowing almost unlimited music especially amongst adolescents. A combination of the mobility and the large amount of music makes the music listening part possible in almost every environments, including the ones with noisy backgrounds. To be able to hear the music even in noisy environments, the obvious reaction is to increase the playback volume to block out any background noise.

A commercial solution to make music more audible under noisy conditions is to increase the loudness by compressing the music. The solution has been used more frequently and with higher degree of compression in the last few decades, leading to a growing concern in today's music industry. This issue is referred to as "The Loudness War", which also corresponds to the belief that "louder is better". Are listeners just accepting the available quality, or is the assumption that louder is better in listener preference really valid? As a consequence of this belief, the dynamic range in modern music is decreasing. Studies have clearly illustrated the reduction in dynamic range throughout the decades. Figure 1.1 shows the tendency of the decline in dynamic range values from 1980-2010.

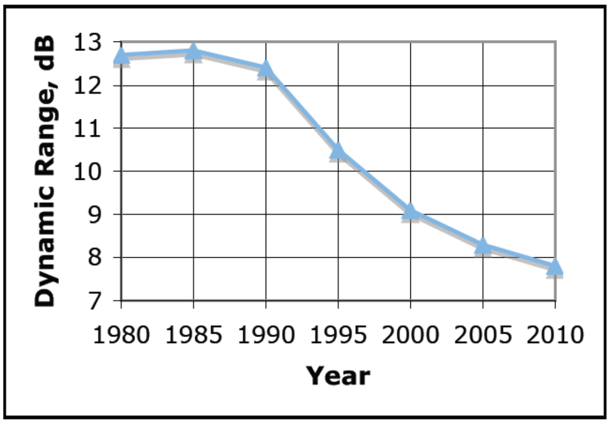


Figure 1.1: The measured average dynamic range from albums listed on The Unofficial Dynamic Range Database [Vickers, 2010].

Even though this compression issue involves all broadcasting domains e.g. television and movies, it mostly concerns the music industry and the radio. It is ironic that with the increase of processing power and equipment, the thought of the better audio quality is obvious but might not be valid. The loudness war initiates two problems:

1. Are the consequences from high compressed audio acceptable regarding subjective listening quality?
2. Compared to relatively uncompressed music, is high compressed audio more fatiguing?

So far, few technical papers have focused only on the subjective quality evaluation of high compressed audio. Taking nowadays use of PMPs into account, thorough studies need to be performed in order to support the concerns. Furthermore, high compressed audio is proclaimed to be more fatiguing due to e.g. fast-acting compression or lack of variation in loudness, reducing the rest periods. Little experimental evidence linking high compression to listening fatigue has been done, because of the difficulties to measure it.

This project will focus on the first problem initiated by the loudness war and investigate which impact compressed music has on subjective listening quality. For the evaluation, a listening test is chosen for this purpose. Whether the compression factor has an impact on listening fatigue is left for further study and will not be included in this project. [Vickers, 2010].

Chapter 2

The Role of Loudness

In order to understand the role of loudness and its importance in compression, a study including the basis parameters of loudness is introduced. The aim of this project is to study only the impact of the compression in music quality, so the purpose of this chapter is also to eliminate loudness as a significant factor for the listening test. The situation is more complex for studying loudness in non-stationary sounds like music than for stationary sounds like pure tones. This chapter begins by looking into the perception of loudness for pure tones. After a description of music and its characteristics, recommendations and methods for measuring loudness in music are presented with respect to find a method to eliminate loudness as a significant factor.

2.1 Loudness and Loudness Level

Loudness is one of the most researched auditory sensations and can be retraced up till [Fletcher and Munson, 1933]. The term loudness is used to describe the subjective perception of strength or powerfulness of a sound. Describing the loudness of a given sound is typically ranked on a subjective "in-head" scale from quiet to loud. The unit for loudness is sone. A doubling in sone will correspond to an approximate doubling in perceived loudness. This measure is not to be confused with loudness level which for a given sound is the Sound Pressure Level (SPL) in dB of which a 1000 Hz reference tone is adjusted to be perceived equally loud. The unit for loudness level is phon. Both of sone and phon units are only used to describe loudness for pure tones. Figure 2.1 shows the relation between the sone and the phon scales, based on a great number of loudness comparisons. It is noted that a doubling in perceived loudness is not considered as a doubling for phons but corresponds more to a 10-phon increase in loudness level. A rule of thumb for many daily life sounds is that to perceive a doubling in loudness, a 10 dB increase is needed. This rule can also be observed with Steven's power law introduced in the following section. [Poulsen, 2005a]

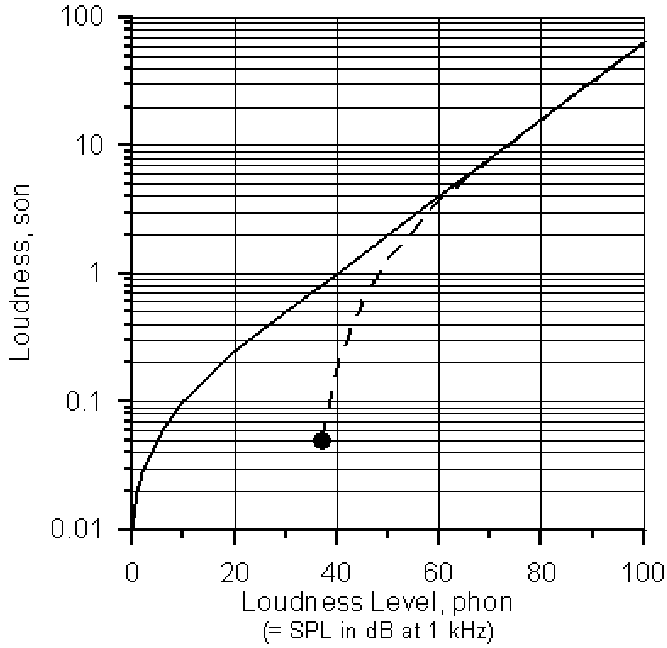


Figure 2.1: The relation between the sone and phon scales. The solid line represents normal hearing and the dashed line is for cochlear hearing loss [Poulsen, 2005a].

2.2 Perception of Loudness

The human hearing is complex and non-linear, it is both frequency and intensity (or pressure) dependent, which affects the perception of loudness. The human hearing is limited and sound is audible only in the frequency range between 20 Hz and 20 kHz. The sound level has also its importance, below a certain level called the hearing threshold nothing is audible. The threshold is subjective and also frequency dependent [Poulsen, 2005a].

The approximation of the perception of loudness in function of the intensity is described by Steven's power law in Equation (2.1), where I represents the physical magnitude and L the perceived sensation. K is a subjective constant and the power n is 0.3 for loudness perception.

$$L = KI^n \tag{2.1}$$

As mentioned previously, perceived loudness is frequency dependent. The equal loudness contours from [ISO-226, 2003] illustrate the difference of loudness perception according to the frequency and can be seen in Figure 2.2.

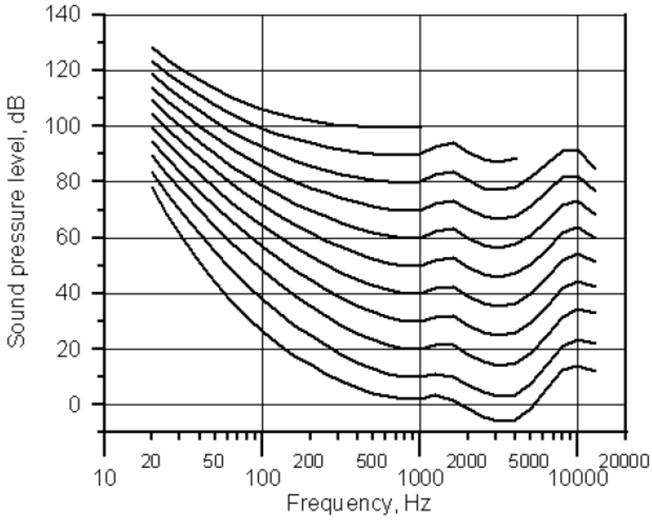


Figure 2.2: Equal loudness contours for SPL in dB, over the frequency spectrum [Moore, 2013].

The contours are measured for pure tones with a great number of persons with normal hearings and is realized with a paired-comparison method with a 1000 Hz tone. Figure 2.2 is only valid for pure tones which are presented one at the time and can not be used for complex signals. From the plot, the mid- and high frequencies have a larger dynamic range than the lower ones. As the sound pressure level increases the equal loudness level contour gets flatter with less differences between frequencies.

Another factor that is also relevant in respect to loudness perception is the duration of the sound. A phenomenon is the build up time that appears with short duration sounds of less than a second. A short sound is perceived less loud than the same sound with same SPL but longer duration. This is called temporal integration and one example of its representation is observed in Figure 2.3. The build up in time will be perceived as an increasing loudness up till the stabilization which in this case happens around 600 ms. Numerous experiments have been completed using loudness comparisons between short tones, and results show that the temporal integration is not linear compared to the level of the tones. [Poulsen, 2005a]

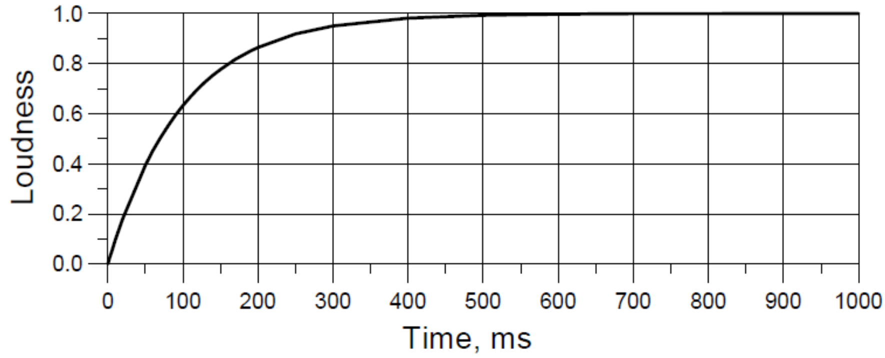


Figure 2.3: An example of the temporal integration of a tone over time, from 0 to 1 for loudness perception [Poulsen, 2005a].

2.3 Masking Effect and Auditory Filters

This section focuses on the phenomena which occur when two sounds are present at the same time, and how that affects the perception of loudness. The masking effect is defined as the process when the threshold of audibility for one sound is raised by the presence of another one, it is expressed in decibels. The masking is more dominant for sounds having same frequency components or similar to the masking sound. The human hearing has a frequency selectivity, which gives the ability to separate pure tones from complex sounds and is important for the perception of e.g. music. The frequency separation is located along the basilar membrane. Masking occurs when the frequency selectivity is not accurate enough to distinguish between the sound and the masker.

Both [Moore, 2013] and [Poulsen, 2005a] agree on the clear differentiation between frequency selectivity and frequency discrimination, the ability to distinguish two tones close in frequency. The frequency selectivity is described by a set of overlapping band-pass filters, called the auditory filters. Each location of the basilar membrane corresponds to a limited bandwidth. Experiments in order to observe the effects of the filters were performed by Fletcher (1940) and consisted in playing a pure tone simultaneously with noise of a bandwidth centered to the tone frequency. The listener was assumed to use the auditory filters to focus on the tone and reduce the noise. The results show that the wider the noise bandwidth gets the harder it is to hear the tone. An experiment is performed in [Moore, 2013] with a 2 kHz sinusoid played along with a noise with a bandwidth centered on the same frequency. Figure 2.4 represents the results of this experiment.

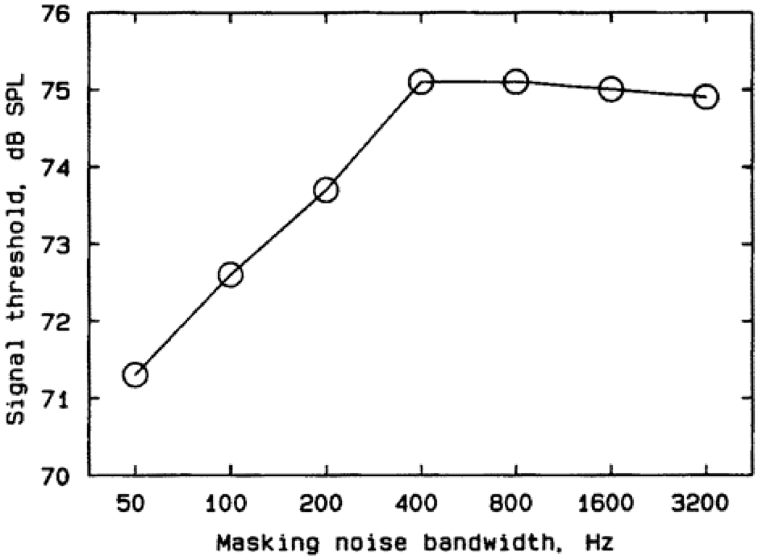


Figure 2.4: Threshold to detect a 2 kHz sinusoid plotted as a function of the bandwidth of a noise masker centered at the same frequency [Moore, 2013].

The threshold of the tone is increasing along with the noise bandwidth until it stabilizes. The bandwidth where the increasing stops is called the critical bandwidth. It stabilizes due to the fact that only the components in the noise that pass through the filter have an effect in masking the tone. When the noise bandwidth exceeds the auditory filter, the effect has reached its maximum. Figure 2.5 represents the different measurements from a variety of papers in order to obtain the different centered critical bandwidth in function of their equivalent rectangular bandwidth. These are based on the assumption that the auditory filters are rectangular, which is not the case.

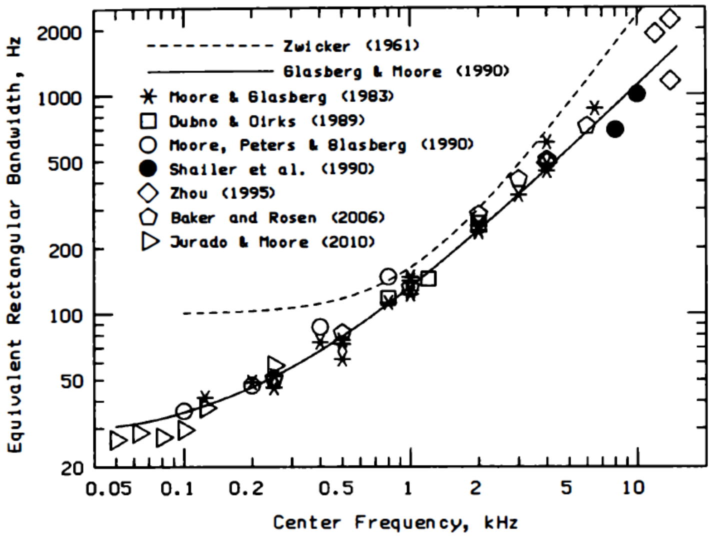


Figure 2.5: Different measurements of the critical bandwidths in function of frequency, also the value of ERB of the auditory filter. [Moore, 2013].

The critical bandwidth and the masking can have a great influence in the perception of loudness in music, where multiple sounds are present.[Moore, 2013] [Poulsen, 2005a]

2.4 Physical Description of Music and Its Characteristics

To expand the study of loudness perception from pure tones to music, first a physical description of music and its characteristics are depicted. Music is a cultural art and has an extensive variety of forms. Some universals components are still observed cross-cultural like the use of instruments or singing. The main parameter of music is the structure, where the repetitions are one of the most fundamental elements [Gebauer and Vuust, 2014]. Even though the human ear can perceive high frequencies, the average music or speech is mostly focused on the mid-frequencies; for example, a classical piano can play in a range from 27.5 Hz to 4186 Hz [Müller et al., 2011].

Music consists of time-varying signals and belongs therefore to non-stationary sounds. It is identified by a few main components; *pitch*, which is the distinct fundamental periodicities, *rhythm*, which is the regularity of temporal structures, *timbre*, which is the variety of sound characteristics, and *polyphony* which is the overlapping of musical sound sources [Müller et al., 2011]. All of those elements, combined together form music and have specific characteristics. Instruments create fundamental periods which are described as harmonic series of sinusoids at multiples of a fundamental frequency [Prasad and Prasanna, 2007]. To illustrate this fact, Figure 2.6 represents two different elements: on top is a plot of a 440 Hz tone, represented in time on the left and in frequency on the right. On the bottom is the same presentation but for a E_4 note played on a piano.

The frequency representation of a sine is called a Dirac-function and is represented by a specific value, a peak, at a certain frequency. Looking at the spectrum of the note played by the piano, it contains also several peaks of different values, indicating the fact that the musical note contains different sine functions. In the frequency representation, the peaks are located approximately at 330 Hz, 660 Hz, 990 Hz etc. The tone at 330 Hz is in that case the fundamental frequency and the other ones are all its multiples, which are also called the overtones.

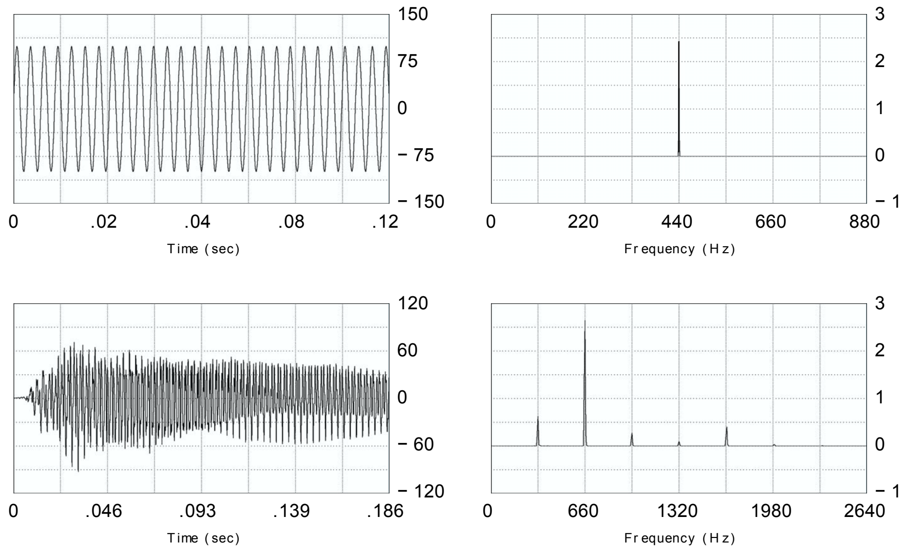


Figure 2.6: Time and frequency representation of a 440 Hz tone on the top, and the note E_4 on the bottom.

Furthermore, another element to take into account is the pitch circularity. It is the phenomenon when a tone is played and then shifted an octave up, it will sound the same as the previous one but higher in frequency, or a higher pitch. In [Shepard, 1964], this phenomenon has been studied thoroughly. It comes from the fact that the shifted tone has the same structure, meaning the same fundamental frequency and overtones, and its spectrum is stretched toward higher frequencies.

Figure 2.7 represents the time representation and the spectrogram of an extract of a danish pop song, *Glass* by Mø. The spectrogram shows time on the x-axis and frequency on the y-axis, and the intensity of the color corresponds to the amplitude. The most significant frequencies are observed in the low- and mid-range. The song has only few frequency components above 15 kHz.

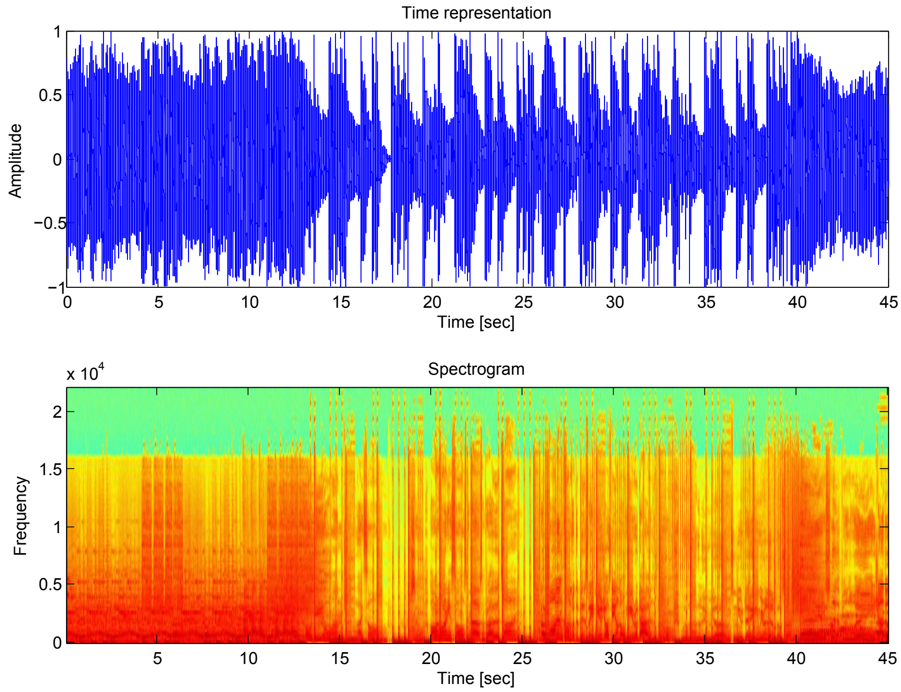


Figure 2.7: The time-representation of 45 seconds of a pop song (Glass by Mø) and its spectrogram.

2.5 Loudness in Music

Now that the physical description of music has been introduced, the assessment of loudness in music is presented in this section. First the loudness normalization problem is introduced. Then the standards for both stationary and non-stationary sounds are presented with the perspective of finding an accurate method to calculate loudness in music.

2.5.1 Loudness Normalization

Loudness variations across different songs or programs are an issue encountered by everyone. The situation can happen for example when a PMP is in "shuffle mode" and plays songs between different artists or albums. Sometimes considerable loudness differences are audible and the user is forced to reduce or increase the level of one of them. These loudness differences are an issue also in broadcasting, e.g. for TV or radio where loudness differences between channels or between programs and commercials can be very significant. The issue has recently been taken into account by countries and companies, for example Apple proposes with iTunes an application to match loudness between all the songs. In the USA, a bill called the CALM Act has been approved in 2010 in order to forbid the commercials to have an average loudness louder than the program material [Lund, 2012a]. Based on loudness measurements and the concerns about the missing loudness matching, [EBU-R128, 2011] recommends loudness normalization. Both the recommendation [EBU-R128, 2011] and [Lund, 2012a] state that the peak normalization for loudness matching is inconsistent, it does not necessarily match the loudness between different programs and it reduces the dynamics. The next part will introduce different loudness measures, some of them which can be considered for loudness normalization.

2.5.2 Measuring Loudness

First thoughts of measuring loudness was to implement filters corresponding to the ears' sensitivity, like the A- and C-filter, which are present in sound level meters. It matches indeed the perception of loudness in terms of the physical structure of the human ears but does not include the psychoacoustic impact on the perception of loudness. It will not give valid results corresponding to loudness. For pure tones, the standard [DS/ISO-532, 1975] suggests two procedures to measure loudness which are both based on results of psychoacoustic experiments. This standard is updated by [DIN-45631, 1991] and [ANSI-S3.4, 2007], which introduce computer program to measure the loudness of stationary sounds. It only requires as an input the third-octave band spectra and specify whether these are obtained in free- or diffuse field conditions. [Hugo Fastl and Straubinger, 2009] [Poulsen, 2005a]

For music, the assessment of loudness is more complex and includes a variety of different measures. While [Vickers, 2010] focus on the term dynamic range, and its reduction due to compression, it may not be the best term to use. The following describes the most common measures for loudness in music.

Dynamic Range is the difference in loudness between the softest and loudest passage in a piece of music. The unit of dynamic range is in dB. The work from [Deruty and Tardieu, 2014] states that dynamic range should be well-defined since it may vary in the literature.

Program loudness is the integrated loudness over the duration of a program and thereby describes the average loudness. The standard uses *Loudness K-weighted Full Scale* (LKFS), which is an absolute measure where 1 LKFS is equal to 1 dB. Since 2006, and revised in 2012, the algorithms and recommendations from [ITU-R.BS.1770, 2012] has been used for measuring program loudness. [EBU-R128, 2011] recommended in 2011 that program loudness should be used to normalize programs to a target level of -23 *Loudness Unit Full Scale* (LUFS), which is identical to LKFS bot the to notations are found in to different standards.

Loudness Range or LRA, is a statistical measure of the difference in loudness level between the soft and loud parts of music, movies or any broadcasting program. It is measured in *Loudness Units* (LU), which is a relative measure used as a 0 dB reference for a normalization target level. 1 LU is equal to 1 dB. The algorithm is originated by TC

Electronic [Skovenborg, 2012] and is described in the standard [EBU-TECH3342, 2011]. LRA is supplementary to the main audio measure, program loudness, and includes a time-window to calculate momentary- or short-term loudness, depending on the length of the window.

True-Peak Level is developed as a solution from the issue initiated by the sample-peak level. The level of the analog signal may be higher than the sample levels. A true-peak meter has the aim of estimating those in-between levels, referred to as true peak values. This is done by increasing the sample rate by at least four times. The algorithm to calculate true-peak level is described in the standards [ITU-R.BS.1770, 2012] and is measured in *decibel True Peak* (dBTP). [Lund, 2012a]

Peak-to-Loudness Ratio or PLR measures the true peak level of a track relative to normalization. PLR is measured in dB. [Lund, 2012b]

Crest Factor is described as the ratio between the peak values and the RMS value, and is used to describe the dynamic behavior of a program. It is dimensionless and widely used. The crest factor is not perceptual; during a song, small peaks or short quiet passages may provoke major changes in the crest factor which might not be perceived in the same way by the listener. Generally, the lower crest factor the less dynamic or the more compressed the program is. [Zölzer, 1997]

2.5.3 Calculation of Loudness in Music

To calculate loudness, two different measures are implemented in MATLAB. The two measures are LRA and program loudness. Both measures are based on the same algorithm stated in [ITU-R.BS.1770, 2012]. As indicated from the standard unit LKFS, the algorithm includes a K-weighting filter to match the subjectiveness with objective measures. For the LRA, the implementation is also based on the recommendations from [EBU-TECH3342, 2011]. The first steps are to apply the k-weighting filter and then perform loudness calculations based on a time window. The time constant is the only variable and determines the statistical distribution of loudness in music. This time window can be used to observe three different time scales; momentary loudness [0.4 s], short-term loudness [3 s] and program loudness.

Figure 2.8 shows an excerpt of 40 s of an uncompressed folk-song. On top is the time presentation and below are loudness distributions for three different time windows. The smaller the constant is, the more fluctuations are present.

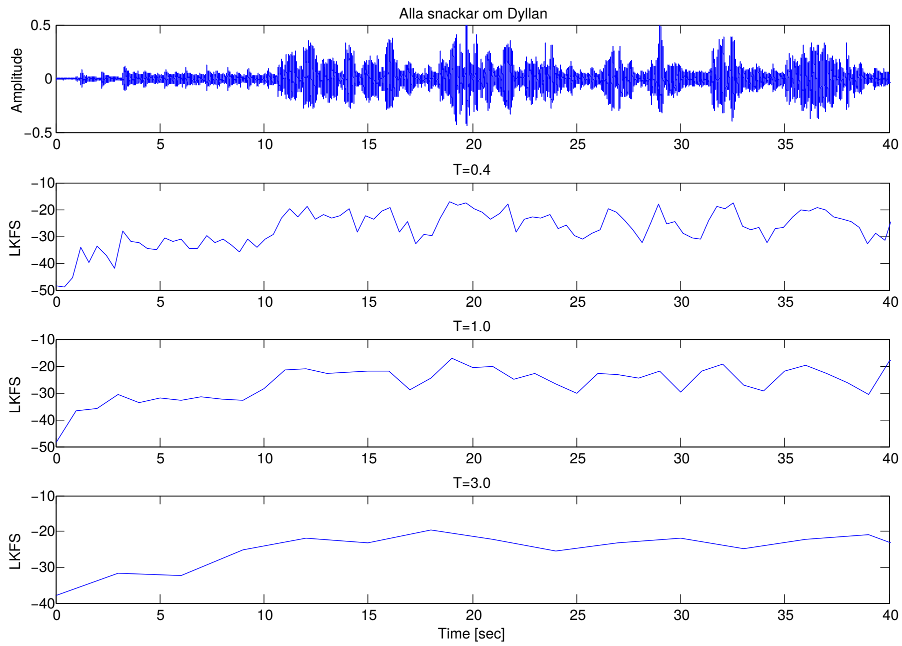


Figure 2.8: Calculation of loudness distribution in LKFS for different time windows.

For each different time constant, the calculation of the LRA is performed using the loudness distribution as an input. The LRA is the difference between the 10^{th} and 95^{th} percentile of the cumulative distribution. The results are provided in Table 2.1. As it can be observed, depending on the time window the loudness range is modified modified.

Time window	0.4 sec	1 sec	3 sec
LRA [LU]	15.2	13.6	11.4

Table 2.1: Results of the LRA.

Choosing a proper time window to calculate the LRA depends on the program. Examples are given in [Skovenborg, 2012] where different types of programs are used such as; movies or radio broadcasting of news and music. For speech, the loudness is quite homogeneous and does not require to show all variations of loudness in order to perform the LRA calculation, therefore the window is selected to be of 3 seconds or more. In music, the proper window depends mostly on the structure of the songs.

The other measure of interest is the program loudness. As mentioned earlier, program loudness describes the integrated loudness for the whole duration and is therefor a measure of average loudness. The same calculations are performed as for the distributed loudness, without the use of a time window. The obtained value for the same excerpt is -23.1 LUFS.

An issue brought by these results is that the loudness is only described by a single value even though a program can vary from few seconds to hours. Especially compared to the loudness distribution, where the variations can be more pronounced as seen from the plot in Figure 2.8.

Summary

Loudness plays a main role in the perception of music and involves a variety of psychoacoustics phenomena, which complicate the loudness assessment in music. For broadcasting and music, a brief view into the diversity of loudness measuring methods has been introduced. For choosing a measure to normalize loudness, the length of the program material in the listening test needs to be considered. Two measures have been implemented to calculate loudness in music, both representing the subjective perception.

The standards recommend the target for a normalization level to be a single LUFS value. Those presented measures are subject to criticism, which is also stated during a group discussion session with Professor Brian C. J. Moore, who emphasized the importance of a more descriptive loudness measure for longer music durations, as one value in that case will not present all the variations significantly.

Chapter 3

Compression Parameters and Algorithms

The compression can represent two different functions; data compression and dynamic range compression. Both will be introduced but the thesis focus on the dynamic range compression and should be clearly distinguished from the data compression. First part of the chapter presents different data compression algorithms used in the music industry to reduce data storage, for e.g. PMPs. The following part includes a study of the dynamic range compression to understand the use and effect, in order to produce excerpts for the listening test. A comparison between uncompressed and compressed music are presented to illustrate the impact of the compression. Finally, different existing compressors available on the market are introduced.

3.1 Data Compression Algorithms in Music

For compressing data to reduce digital storage, two main groups of compression algorithms are defined; lossless- and lossy compression. The most widely used format is *Waveform Audio File Format* or WAV, which is capable of storing very high audio quality and used especially for CDs. High audio quality is a criterion for many audiophiles but comes with a price regarding storage consumption. This may not be a main concern for some users where the listening part is the main activity. For others who may listen to music from a PMP in order to block out noise from background environments e.g. during transport, a more important concern might be storage capacity on the PMP. The two main groups of compression algorithms are presented in the following part. [JISC, 2014]

3.1.1 Lossless Compression

In lossless compression, data is reduced to save storage space but the original data can be fully reconstructed from the compressed data. This concept can be related to the

ZIP file formats, which can reduce storage of a file on a computer and fully reconstruct the data. In audio, a well known lossless compression type is the Free Lossless Audio Codec or FLAC format, which is already being used by artists and music distributors in the industry. Data can be reduced to 50-60 % of the original copy. Lossless compression applies to the user who wants a combination of less data storage and high quality music. Other formats that use lossless compression algorithms are e.g. Apple's Lossless Audio Codec or ALAC which is supported by iOS devices from Apple, Audio Lossless Coding or MPEG-4 ALS and Monkey's Audio. [Coalson, 2013] [Bobulous, 2009]

3.1.2 Lossy Compression

For lossy compression, the main goal is to reduce the maximum amount of data. As the name indicates, the lossy part comes from the need of discarding data from the original copy. The most popular compression format is the MPEG-1 Layer III or MPEG-2 Layer III, better known as MP3. Based on a compression algorithm, the original audio file is reduced to around 1/10 of the original data size, saving significant amounts of memory. The disadvantage is that the high degree of compression comes with a cost in sound quality. However, the trademark of the MP3 format is that the quality of the compressed data is relatively close to the CD. It comes from the statement, that the compression algorithm is based on psychoacoustic models. The algorithm discards all data outside the audible frequency range of the human hearing. Another psychoacoustic phenomenon also included in the algorithm is the masking effect. All redundant data from the psychoacoustic analysis is compressed with low accuracy or non at all. The reduced storage requirements make the MP3 formats ideal for streaming music, which has an increasing tendency amongst music listeners.

Other examples of lossy compression formats are Advanced Audio Coding (AAC) which is the successor of the MP3 and is also based on the same compressing models, Ogg Vorbis and WMA. [Hendrik Böhne and Sussek, 2011] [Bobulous, 2009]

3.2 Dynamic Range Compression

Dynamic range compressors are a complex type of audio effect. Multiple numbers of designs can be used to reduce the high peaks in a signal, in order to obtain a smaller dynamic range. Some of the different existing designs can be found in [Massberg and Reiss, 2012]. This section introduces the main parts of a compressor and two different types; static- and dynamic compressor, where examples are realized in MATLAB. For this section, the literature from [Zölzer, 1997] is also used.

3.2.1 Main Parameters of a Compressor

A compressor depends on a different sets of parameters which all have an influence on the output. Some of these parameters can be automated and processed to act the best depending on the input and the user's preferences. For music compression, the settings of the compressor may vary between songs and genres. The main equation controlling the compressor is stated in Equation (3.1).

$$Y_{dB}(n) = X_{dB}(n) + G_{dB}(n) \tag{3.1}$$

The variables are in logarithmic levels where $Y_{dB}(n)$ and $X_{dB}(n)$ represent respectively the output and the input of the compressor. The heart of the compressor is the weighting level, $G_{dB}(n)$, which attenuates the peaks of the input signal. As mentioned, the compression is controlled by different parameters. The *threshold*, T, sets the starting level of where the compressor has to start reducing the input signal. The *ratio*, R, controls the input/output ratio for signals trespassing the threshold level. The *knee width* controls the bend of the response curve, by offering either a hard or soft compression. The *make-up gain*, M, can be added to Equation (3.1) to match output and input levels. The final factors are the *attack*- and *release times*. The attack time defines how quick the compression attenuates to the level set by the ratio when the threshold is trespassed. The release time is the time it takes to get back to the normal level when it goes back below the threshold. These parameters must be carefully set to avoid instantaneous compression, which may introduce distortion.

3.2.2 Static Curve

The first step is to implement the main part of the compressor, which includes the threshold, ratio and knee width. Equation (3.2) is based on a feed-forward compressor design. [Massberg and Reiss, 2012]

$$y_{dB} = \begin{cases} x_{dB} & \text{if } x \leq T \\ T + \frac{x_{dB} - T}{R} & \text{if } x > T \end{cases} \tag{3.2}$$

The output is the same as the input below the threshold, but above this level it will decrease taking the ratio into consideration. The feed-forward design is implemented as an example instead of the feed-back design, because most modern compressors are based on it. Figure 3.1 represents a static curve with different ratios but with the same threshold.

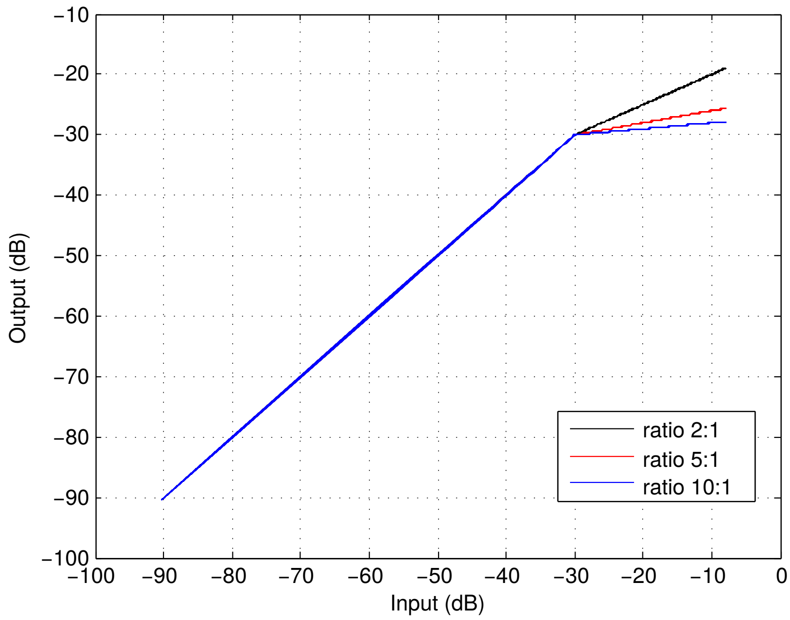


Figure 3.1: A static compressor with a -30 dB threshold and different ratios.

If the ratio is infinite, then the static curve is called a *limiter*. For this example, the curve is implemented using hard knees for all three ratios. To change the curve into a smoother transition, a soft knee can be implemented instead. Equation (3.3) shows the function of the soft knee.

$$softknee(x) = \begin{cases} 0 & \text{if } x < -\log|width|/2 \\ \frac{1}{2\log|width|}\left(x + \frac{\log|width|}{2}\right)^2 & \text{if } -\log|width|/2 \leq x < \log|width|/2 \\ x & \text{if } x \geq \log|width|/2 \end{cases} \tag{3.3}$$

From the soft knee function, a weighting level can be found in Equation (3.4) and used to determine the output in Equation (3.1).

$$G_{dB} = \left(\frac{1}{R} - 1\right) \cdot softknee(X_{dB} - T) \tag{3.4}$$

The second degree polynomial equation used for the period when the input signal is in between the threshold and the width is enough to provide a smooth transition of the compression. When the width is equal to zero, the compressor is going to be linear as before as it can be seen on Figure 3.2. Otherwise with the increase of the width the smoother the transition is but of course it must be limited to not take a long time to reach the input level.

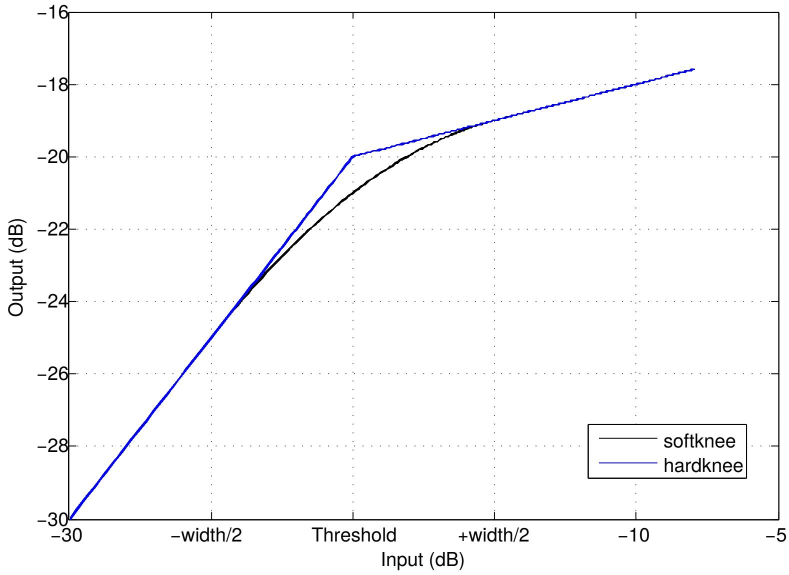


Figure 3.2: The soft knee represented as a function of the input with a 10 dB width along with the hard knee.

3.2.3 Dynamic Behavior

These previous parameters compose the static part of the compressor and now the parameters composing the dynamic behavior of the compressor is presented. These are the attack- and release times and the level measurements. The following shows one way of calculating the level measurement parameters, which includes peak detector (3.5) and RMS level (3.6) [Massberg and Reiss, 2013].

$$y_{peak}[n] = \sqrt{\max\left(x^2[n], \alpha \cdot y_{peak}^2[n-1] + (1-\alpha) \cdot |x^2[n]|\right)} \tag{3.5}$$

$$y_{RMS}[n] = \sqrt{\alpha \cdot y_{RMS}^2[n-1] + (1-\alpha) \cdot x^2[n]} \tag{3.6}$$

where $\alpha = \exp(\frac{1}{\tau \cdot fs})$. In Figure 3.3, the level measurements are illustrated for a 100 Hz sinusoid which amplitude varies over time with 0.3 s of silence before another tone starts. The amplitude of the tone is modified throughout the process in order to see how quickly the changes can be seen. As this is a simple example, the detector has an instant attack time and only a defined release time, τ. In this case τ is set to 0.1 s.

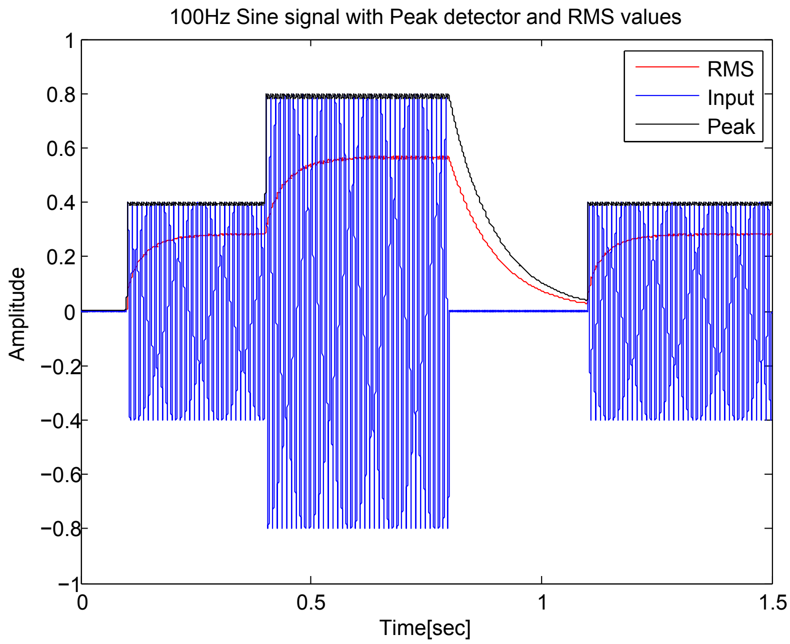


Figure 3.3: The peak detector and RMS values for a 100 Hz sine signal with a 0.3 second silence at 0.8 second.

As it can be observed, the release time does not react instantly on the peak detector and the RMS value. The smaller the τ is the faster the change is going to be taken into account. The flowchart in Figure 3.4 gives an overview of the final implemented dynamic range compressor.

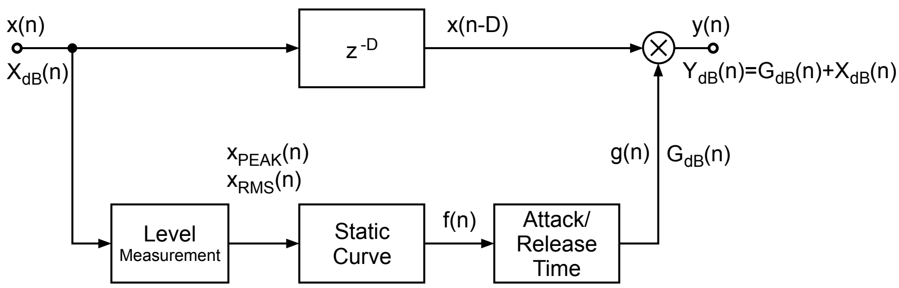


Figure 3.4: The dynamic range compressor [Zölzer, 1997].

The first thought for this project is to use this implementation of a compressor, in order to produce the excerpts for the listening test. After a discussion at TC Electronic the decision is made to use an existing compressor, as the quality of the excerpts depends on the compressor. The aim of this project is to test impact of the compression itself, and not the quality of the compressor. From the literature, designing a good compressor can be a complex task and require a full study. Instead the Finalizer 96K from TC Electronic is used for the compression. The compressor offers adjustable parameters in three bands, meaning that it can act separately on low-, mid- and high frequencies. This function is not used in this project as the compression will act the same in all three bands with the exception of the attack and release times, which are set as default values. Figure 3.5 shows a 5 s example of a pop-song. On the top, a time presentation of the uncompressed version and below a compressed version.

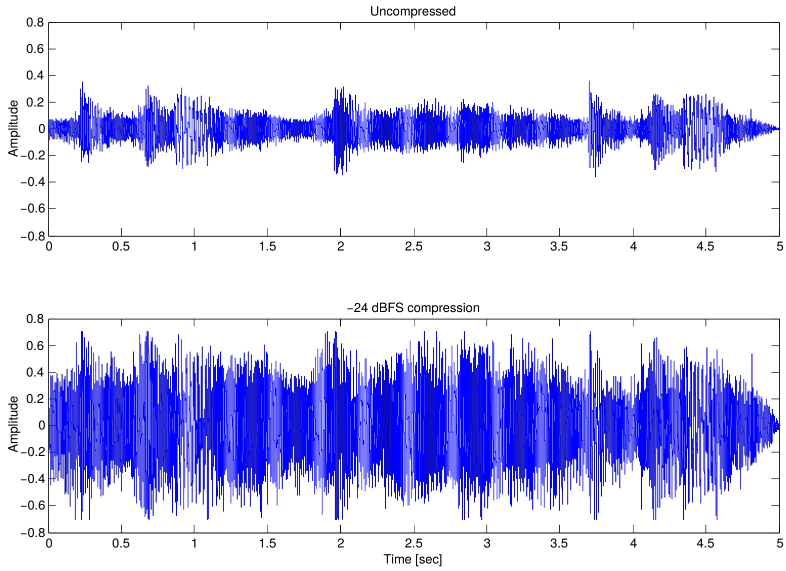


Figure 3.5: A 5 s example of a pop-song in its uncompressed version on top and compressed to -24 dBFS below.

The parameters configurations for the compressed one is a the threshold of -24 dBFS (decibel Full Scale), a ratio of 3.2:1. The default values for the attack time in the low-, mid- and high frequency bands are respectively 30 ms, 20 ms and 20 ms. For the release time, the values are respectively 0.7 s, 0.5 s and 0.3 s. As seen from the example, the differences between soft and loud parts for the compressed version are smaller due to the compression. There is no loudness matching between the two and the average loudness for the compressed version is generally higher. The two time representations are different with less dynamic between the soft and loud parts for the compressed one. The listening test aims to define if these differences in dynamics are perceptual after the loudness is normalized.

Summary

In this chapter, both data compression and dynamic range compression are introduced. These two terms must be distinguished. An implementation example of a dynamic range compressor is described but not used in this project. Instead, the Finalizer 96K is used to compress program material for the listening test.

Chapter 4

Internet Questionnaire Survey

Based on the assumption from [Vickers, 2010], compressed music might satisfy best the PMP users. Compressed music is rated higher in quality by passive listeners, meaning that the environment and the conditions have an impact on quality. The time spent on listening to music via PMPs is assumed to be high supported by the substantial music library and the durability of the battery. To support this assumption, a questionnaire is performed with a study in how widely PMPs actually are used and under which listening conditions. General considerations and final design of the questionnaire are described leading to a presentation of the results, which are used as an inspiration for the design of the listening test.

4.1 General Considerations

The starting point for the design of the questionnaire is defining the goals, and express them clearly. The questionnaire can be developed and related directly from the goals. The motivation behind this survey is two-fold; 1) to collect information about PMPs popularity and the listening habits for those who listen to music from a PMP, and 2) to assemble a listening panel for the listening test. The results from the survey should be able to answer or at least give information about the following questions:

- Are PMPs a preferred listening platform?
- How much time is spend on listening to music from them?
- Are PMPs used in noisy environments?
- Is the listening level correlated to the level of the background noise?
- Are PMPs used for active or passive listening?
- Are PMP users bothered by music quality?

The questionnaire is composed based on the literature from [Walonick, 2010]. First the method to obtain information needs to be considered. An Internet survey is a cost efficient method compared to a face-to-face interview, as social media like Facebook and LinkedIn can be used to distribute the questionnaire. Many uses the Internet to get information, so the process is familiar to most people. It can be biased by the demographic profile as the possible respondents do not represent the general population but in this case the assumed focus group is the generation which uses the Internet on a daily basis. The risk of a low response rate is considered, it typically varies between 10-90 % so a way of maximizing the response rate is preferred. For that, the length of the questionnaire plays an important role. Long and time-consuming questionnaire may cause the respondents to give up and thereby lowering the response rate. The efficiency of an Internet survey excludes visual cues from the respondent and in general offers little answer flexibility. The flexibility can be improved by including comment boxes in the questionnaire, allowing the respondents to qualify the answers.

In order to reach as much people as possible, an Internet questionnaire survey is used as the method to collect the information. It is realized on a HTML script and with the service FormMail, which is a free web-based HTML service and one of the advantages of using it is that after respondents submit the answers, an e-mail is formed including all of the answers. It eases the analysis, especially in the case of high response rate.

4.2 Design of Questionnaire

The following states and describes the considerations and motivation behind the questions:

- **Do you listen to music from a portable music player?** To get an estimate of the number of PMP users. The respondents are also asked to choose which device they mostly use in case several different PMPs are used, together with brand and model of both the device and the headphones used. Finally, the respondents are asked how often they listen to music from the PMP and the preferred listening level.

- **Where do you spend most time listening to music?** To get information whether or not most of the listening time is spend in environments that might require background noise to be blocked out. The subjects are also asked to specify whether they consider themselves as an active or passive listener. Finally, they have to estimate the background noise in the specific listening place.

- **Have you ever been bothered by the sound quality of a song?** In order to find out if the respondents actually pay attention to the music quality. Asking directly for quality issues may indicate if it is a noticeable problem. If the answer is yes, the respondents are asked to define the issue.

- **Which genre(s) of music do you listen to?** To get information about which genres the respondents listen to. The answers can be useful for selecting the music material for the listening test.

- **Would you be interested in participating in a listening test?** This is for screening subjects to a listening test. If the respondents accept to participate, they are informed that the will be contacted.

The questionnaire includes a short cover letter in order to inform the respondents of the short length of the questionnaire, in an attempt to maximize the response rate. A copy of the final questionnaire is included in Appendix B.

4.3 Results

After distributing the questionnaire on Facebook, LinkedIn, Twitter and student e-mails, the collection included a total of 88 completed responses. The average age of the respondents is 24 years ranging from 13-50 years. Only seven are above the age of 30. Most of the respondents are students, only 17 are not, which is an expected result since the student e-mails and social media are used to distribute the questionnaire. The responses from the questionnaire are analyzed in respect to the defined goals using MATLAB. A short summary of early results is listed in Table 4.1.

Question	**Response**
How many use a PMP?	$\approx$ 92 %
How much time are respondents listening to music from PMPs?	$\approx$ 13 h/week
How many are active listeners?	$\approx$ 28 %
How many are bothered by music quality?	$\approx$ 75 %
How many are willing to participate in a listening test?	$\approx$ 42

Table 4.1: Results from questionnaire in percent based on 88 respondents.

The results of the questionnaire show that as many as 92 % of the respondents listen to music from a PMP, where 79 % of them use smartphones, 16 % use music players and 5 % use tablets. The questionnaire also focuses on which kind of headphones is typically used with PMPs. The results of the popularity amongst different types of headphones are shown in Figure 4.1.

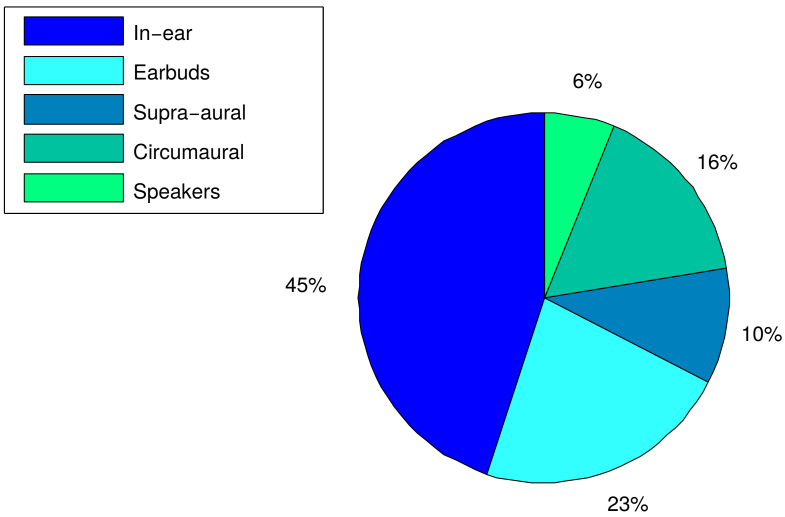


Figure 4.1: Distribution of used headphones illustrated by a pie chart.

In-ears headphones are the most popular ones with 45 %, earbuds are second with 23 %, supra-aural are used by 10 %, circumaural are used by 16 % and 6 % use speakers instead of headphones. These results may come in handy when it comes to selecting equipment for the listening test. In Table 4.2 are the most used combinations of PMPs and headphones based on the responses. Since apple devices are highly used in this study, PMP brands are divided into apple device or non-apple device. Headphone brands are listed at either matching, meaning headphones that comes with the used PMP or other, meaning any other brand that does not come with the used PMP. Only the ones who answered brand and model boxes for both PMP and headphones are included (66).

PMP brand	Headphone brand	Score/responses
Apple device	Matching	21
Apple device	Other	17
Non-apple device	Matching	7
Non-apple device	Other	21

Table 4.2: List of most used combinations of PMPs and headphones. The headphone brands are listed as "matching"; the ones that comes with the PMP and "other"; any other brand that does not come together with the PMP.

A total of 21 respondents using an apple device use the matching apple headphones that comes with the PMP and 17 use a combination of other headphones along with an apple device. 7 of those who are not using an apple device use matching headphones, whereas 21 use a combination of another headphone brand.

In order to find an average time spend on listening to music from PMPs for each respondent, the responses are divided into two parts including; the ones who listen on a monthly basis (9 respondents) and the ones who listen on a weekly basis (63 respondents). 9 of the PMP users did not fill out this part, as some questions were not mandatory to submit. Looking only at the monthly users shows an average of approximately 2 hours pr. week, whereas the average result for weekly listeners shows approximately 14 hours spend each week on listening to music from a PMP. Including both monthly and weekly listeners gives an estimated listening time of 13 hours pr. week. To evaluate the use of PMPs in noisy environments, the respondents are asked for listening location but only the one where most listening time is spent. Figure 4.2 is a chart of these locations.

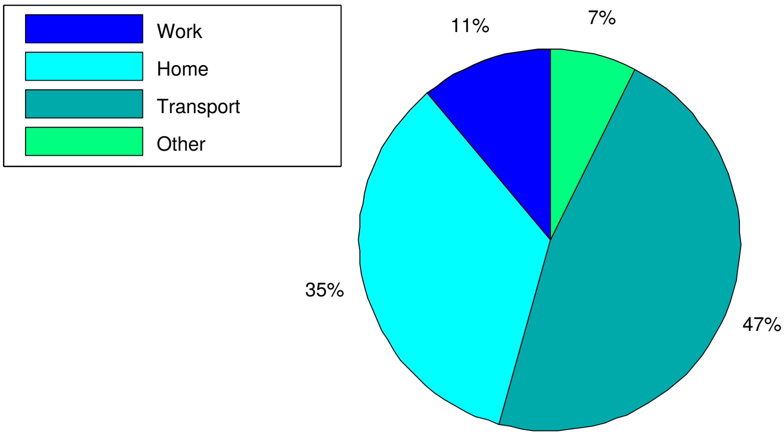


Figure 4.2: Distribution of listening locations where PMPs are mostly used.

Almost half of the respondents, 47 %, are mostly listening to music during transportation, 35 % are listening at home, while 11 % are listening at work. The last 7 % are listening at other locations than specified, where during workout or exercising are repeated in the answers. The respondents also answer how loud an estimation of the background noise at that specific location is. The results are shown in Figure 4.3 together with the preferred listening level in Figure 4.4.

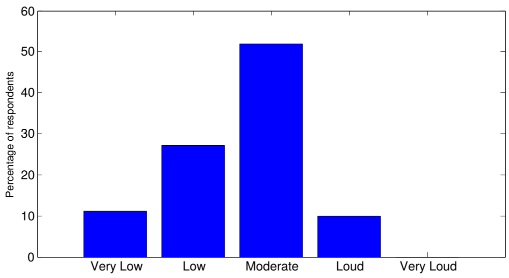


Figure 4.3: Background noise estimated by the respondents at the listening location.

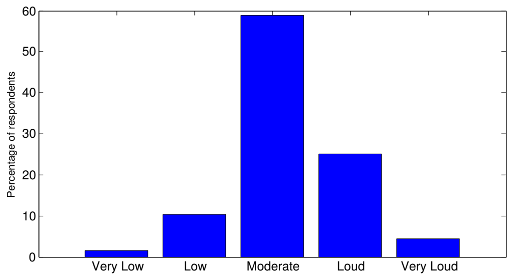


Figure 4.4: Respondents estimation of preferred listening level.

Moderate level is mostly chosen for both preferred listening level, almost 60 %, and estimated background noise, approximately 53 %. For the listening level only a few consider it to be very low or very loud, 9 % consider it to be low and 25 % consider it to be loud. For the background noise about 11 % think it is very low, 26 % report it as low and about 10 % for loud. Non of the respondents considered the background noise to be very loud. To analyze whether there is a correlation between the background noise and the listening level, a scatterplot is shown in Figure 4.5.

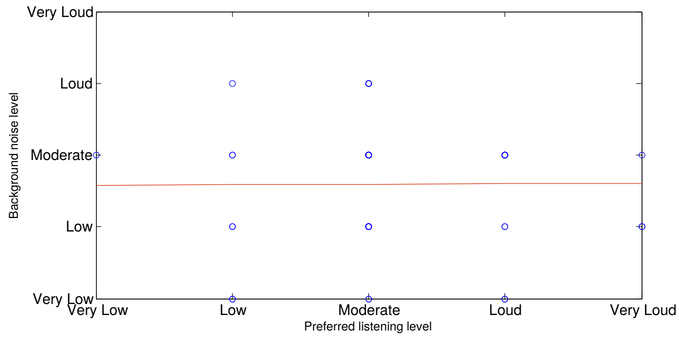


Figure 4.5: The scatterplot of the preferred listening level in function of the background noise, the density of the circle color represents the number of answers.

The plot shows no linear progression. For loud background levels the listening level is not necessarily loud. Some subjects who prefer a loud listening level, are mostly listening to music in very low background levels, and the opposite, low listening level in loud background noises. Also an interesting part is that all the respondents answering that they are in loud background noise level considerer themselves as passive listeners.

Out of all the respondents who uses PMPs 28 % consider themselves as active listener, but despite this low number 75 % of the respondents who use PMPs have at some point been bothered by the sound quality of a song. The respondents which are bothered by music quality are almost evenly divided by active and passive listeners, being 70 % of the actives and 78 % of the passives. When asked to the issue of what caused the irritation, 21 % of those who where bothered did not answer illustrating the difficulty of the task. Several explained the issue by bad quality, annoyance and missing clarity. Other observed issues are; missing depth, low dynamic range, compression, instruments are pressed together or problems with the bass, too low voices, aliasing in high frequencies and volume changes between tracks. Also, 9 subjects complained about the quality of listening to music on YouTube or other streaming services. The following states a list of words used in the answers to describe issues due to quality in music:

- scratchy
- sizzle
- unreal
- blasting

- blurry
- clipping
- disturbance
- flat
- saturation
- crackling

These words can be useful for the listening, to select reliable attributes in order to describe the subjective perception of quality in music. Another concern for the listening test is which music material to present. The results of preferred genres are shown in Figure 4.6, including only the five most popular ones.

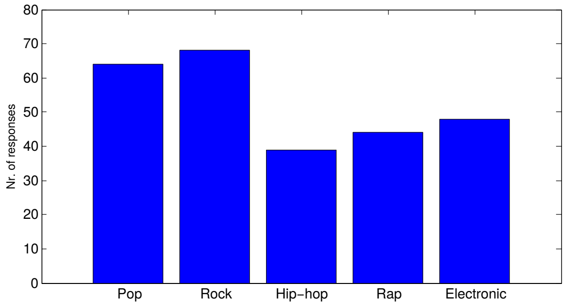


Figure 4.6: Preferred music genre.

The most popular genres among the respondents are; pop (64), rock (68), hip-hop (39), rap (44) and electronic (48). The rest of the genres are evenly preferred amongst the respondents, but none of the remaining genres scored higher than 25.

Summary

Based on the results from the questionnaire it turned out that 92 % of the respondents use PMPs, which supports the belief that PMPs are widespread. It is also supported by the amount of time spend on PMPs. An average of 13 hours is spend each week, almost two hours per day not including weekends.

Results show that PMPs are mostly used during transportation or at home. Only 28 % are active listening to music at this location, meaning that the greater part is not fully paying attention to the music. The assumption that music is used to block out environmental noise can be strengthen by this result. Additionally, the in-ear headphones are the most popular used by almost half of the respondents. The in-ears are extended to the ear canal, which in some cases are preferred in order to block out environmental noise. This study only included one listening location, where some respondents may use the PMPs in different locations.

No correlation between preferred listening level and background noise was found. Preferred listening level was asked in a more general case, whereas the background noise was an estimation at a specific location. Both are subjective estimations which can be a difficult task. Both arguments may be factors in the results of the correlation.

75 % of the respondents had at some point been bothered by the sound quality of a song. Some of the explained reasons can be used for considering attributes for the listening test. The purpose of the questionnaire was also to recruit subjects to form a panel for the listening test. This part was achieved as 42 of the respondents agreed to participate. Some of the questions were not mandatory to submit the questionnaire, in a few cases the respondents did not answer to all questions.

Chapter 5

Test Design

The previous studies are used to construct a listening test, in order to evaluate the subjective quality of compressed music. Before the test is performed, a hearing assessment of the subjects is realized. It includes three different parts; a questionnaire about the hearing, audiometric measurements and Otoacoustic Emission (OAE) measurements. These parts are introduced in this chapter. The design of the listening test is described. Finally, the results are presented and analyzed.

5.1 Questionnaire about the Hearing

This part is the first step of the listening test. The questionnaire is performed in the same format as the Internet survey but with the purpose of getting more information about the hearing of the subjects. This time the subjects have support in case they have issues understanding the questions. The questionnaire is composed of the following:

- **Have you suffered from or experienced hearing issues?** Those first two questions are present to have a rough estimate of the antecedents problems that the subjects may have faced. Different issues are listed so the respondents can clearly define them.
- **Have you taken medicine or other drugs affecting your hearing or had an ear operation?** To get information if any of the mentioned could impact the subjects performance in the listening test.
- **Is there someone in your close family that has or has had a hearing disorder?** It is known that there is a higher risk to have hearing loss if other family members are suffering from impairments.
- **Do you go to nightclubs or concerts?** The different places are well-known to have a high sound pressure level which can be harmful. It is to verify if people attending too many of those events have a higher chance of showing signs of hearing loss. The subjects have also to estimate for how long and how often these events are attended.

- **Are you a musician?** The purpose of this question is to observe if people more experienced to music have different ratings for quality of songs than others. It is only an assumption that playing an instrument is a significant factor and is going to be verified only in the results of the listening test. The number of years of practice and which instruments are also part of the question.

5.2 Audiometric Test

An audiometric measurement is used to determine the threshold of hearing and it is widely used for screening subjects. A variety of methods exists to perform an audiometric measurement but two are described in the standard [ISO-8253-1, 2010] using pure tones: the ascending and the bracketing method. An audiometer is used, which contains an automatic threshold determination for the ascending method. Since this is available and mostly used for audiometers, the ascending method is chosen [Poulsen, 2005b]. Both ears are tested separately and the procedure is described in Appendix A. The methods unit is decibel Hearing Level (dBHL) which is the unit for determining threshold in audiograms. The 0 dBHL is the softest sound that a person with normal hearing is be able to detect at least 50 % of the time. The method is using specifically calibrated headphones.

Figure 5.1 represents an example of the ascending method with steps of 5 dB. The choice of the 5 dB step is mainly based for practical reason, a lower step is time consuming but more precise. For this project, only a rough estimate of the subject's hearing threshold is wanted. For that reason, only the octave bands from 125 Hz to 8 kHz are measured. Usually, a limit of 20 dBHL is used to discard subjects. In this project, all subjects are performing all the steps of the listening test even if this limit is violated, because the impact of compression is observed for all listeners even for the ones who have a higher hearing threshold. [Poulsen, 2005b]

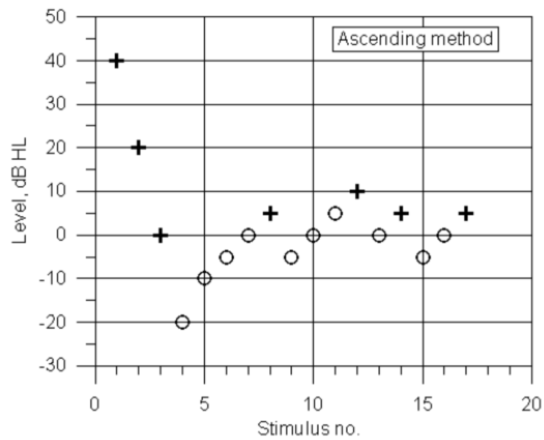


Figure 5.1: Example of the ascending method with steps of 5 dB [Poulsen, 2005b].

5.3 Otoacoustic Emission Measurements

A second test for the hearing assessment is the OAE measurement. It is performed in order to gain more insight about the subjects' hearing. Otoacoustic emissions are sounds which are generated from within the inner ear, and can be recorded in the ear canals to measure the health of the cochlea. Two common types of OAE measurements are the Transiently-Evoked Otoacoustic Emissions (TEOAE) and the Distortion Product Otoacoustic Emissions (DPOAE). Performing a TEOAE measurement is a quick procedure and is a widespread method to check for hearing loss in newborns. The TEOAE measurements are performed on each subject. [Kemp, 2002]

The TEOAEs are measured according to different protocols but the one chosen for this measurement is performed with the help from [Ordonez et al., 2012] and made using the non-linear differential stimulus block. The procedure of the test and the parameters are described in Appendix A

5.4 Listening Test Design

This section describes the design of the listening test and includes; previous studies, listening panel, program material, attributes, listening conditions and test method. Considerations and final choices for the test design are inspired from [ITU-R.BS.1116-1, 1997].

5.4.1 Previous Studies

Two different listening tests have been performed to assess quality of music under influence of the compression. The oldest one, [Croghan et al., 2012], tested the influence of the dynamic range compression by only modifying the threshold and keeping the other parameters fixed. The threshold is varying of steps of -4 dB from uncompressed to -24 dBFS and two different music genres are tested: rock and classic. Two different conditions are tested also: equalized and unequalized loudness. The equalization is performed according to Moore's model and a RMS matching. The method for the design of the listening test is a paired-comparison between two different stimuli. Four different attributes are asked to the subjects: the overall loudness, the dynamic range, the pleasantness and the preference.

[Walther-Hansen and Hjortkjær, 2014] is the most recent study and is comparing uncompressed and remastered music. The different degree of compression is calculated by evaluating the difference between the peaks of the two versions, the program loudness is also measured. Two different attributes are asked in a paired-comparison form about the preference of one of them and about the depth. This thesis is not similar to this listening test because the remastered song is not only about dynamic range compression and other effects are usually added, the conductors did not have a full control of the excerpts.

The main difference between these previous studies and this project is that only the impact of the compression is tested.

5.4.2 Listening Panel

As a part of the questionnaire survey, a purpose is to recruit subjects for the listening test. A total of 42 of the respondents agreed to participate. In order to inform and schedule the subjects, an e-mail with specific times is sent to each of the 42 subjects. All replies is put in a final schedule for the listening test.

5.4.3 Program Material

A folder including seven tracks of uncompressed music is provided by a Swedish studio. The tracks are unmastered and recorded in a music studio and includes a mix of genres. A mix of genres is preferred in the listening test, as the compression will have different impact depending on the genre. From the questionnaire survey, results showed that rock and pop genres are highly preferred. An excerpt from a rock song is used in the test so that the subjects have something to relate to a weekday listening situation. The rock excerpt is picked out where the drums a very dominating. The rest of the excerpts for the test is composed of a classic guitar, flutes from a symphony orchestra and a violin. It gives a total of four different excerpts with different dynamics. All of the excerpts used in the listening test are re-sampled to 48 kHz to agree to the filters used in [ITU-R.BS.1770, 2012] to calculate the program loudness and the LRA. The length of each excerpt is approx 15 sec.

As stated in Chapter 3, the Finalizer 96k is used for compression. The input to the Finalizer is peak normalized to -6 dB [Bregitzer, 2009]. The parameters in the compressor are fixed for the attack and release times, same default values are used as in the example in Section 3.2.3. A compression ratio which does not introduce any distortion to the output is wanted. Different ratios are tested to find a limit before any audible distortion is present. The limit shows to be a ratio of 2.5:1 and is not changed during the compression. Leaving these parameters fixed, the only variable in the compression is the threshold which determines the level of compression. To find out which thresholds to use for the compression, first thought is to perform a test with threshold variations of -4 dB steps from uncompressed to -24 dBFS. In this way, any insignificant threshold that may be too difficult for a subject to distinguish from another will be eliminated. After listening to the different threshold compressions, it is considered too difficult to distinguish compressions in -4 dB steps. The solution is to use the two extremes which are the uncompressed version and -24 dBFS, which is the highest level of the compressor. To have a midpoint between the two extremes, a compression of -12 dBFS is also included.

To only test the impact of the compression, the loudness in all the excerpts needs to be matched. From the implemented example in Section 2.5.3 both program loudness and

LRA are calculated for all of the excerpts. The -12 dBFS and -24 dBFS compressions are normalized to the program loudness values of the uncompressed version. It is normalized by finding the ratio of the uncompressed excerpt and the compressed one, and this ratio is applied to the latter. The algorithm runs in a loop, for each iteration the loudness values are converging until the stop criteria is reached. This stop criteria for the loudness matching is when the ratio of the uncompressed and the compressed excerpt is 0.01. An overview of all the excerpts and the corresponding loudness values for LRA and program loudness before and after normalization is given in Table 5.1. The differences between soft and loud parts in the excerpts are illustrated by the LRA, which decreases with high degree of compression.

Excerpt	**Thresh. [dBFS]**	**LRA [LU]**	**Unnorm. [LUFS]**	**Norm. [LUFS]**
Drums	Uncompressed	2.29	-20.9533	-20.9533
Drums	-12	2.24	-15.6698	-20.9392
Drums	-24	2.07	-9.9972	-20.9406
Flute	Uncompressed	6.15	-16.7509	-16.7509
Flute	-12	5.96	-11.4610	-16.7391
Flute	-24	2.96	-8.9678	-16.7374
Guitar	Uncompressed	8.78	-26.2063	-26.2063
Guitar	-12	8.60	-21.1522	-26.1881
Guitar	-24	6.19	-15.7581	-26.1867
Violin	Uncompressed	5.48	-19.4527	-19.4527
Violin	-12	5.34	-14.1556	-19.4398
Violin	-24	3.08	-10.1363	-19.4355

Table 5.1: The 12 excerpts used for the listening test along with the LRA and program loudness values before and after normalization.

5.4.4 Attributes

Finding relevant dimensions of sound quality can be a difficult task. In a listening test, attributes describing sound quality needs to be fully understandable for the subjects. The questionnaire survey includes a part where respondents have to describe in words, what could be the cause of missing quality in music. A study from [Gabrielsson and Sjögren, 1979] revealed a number of sound quality dimensions, based on experiments including music played over loudspeakers and headphones. Considering results from this study and the results from the questionnaire, the following four attributes are used in the listening test; *depth*, *clearness*, *disturbing sounds* and *quality*. The meaning of *depth* can be related to lack of dynamics due to the compression. It can be related to the word *flat* from the response of the questionnaire. The depth in music depends also on the genre and the instruments, a rock song might have low depth compared to a classic piece of music. *Clearness* covers the feeling of presence, with expressions like *pure*, *clean* and *rich in details*, in contrast expressions like *diffuse*, *muddy*, *blurry*, *noisy*, which all can be related to the words from the questionnaire. *Disturbing sounds* is the presence of artifacts in the excerpts. Finally, *quality* relates directly to the overall rating of the subjective quality. As mentioned earlier, quality in music perception is complex but by forcing the subjects to rate the quality of different compressions may show interesting results.

5.4.5 Listening Conditions

To ensure optimal listening conditions, the test is performed in a listening cabin. A camera is installed to monitor the subjects. The test is performed on a Samsung laptop and the subjects is listening to the test excerpts in Beyerdynamic DT990 headphones, which are preferred over loudspeakers. Results of the questionnaire showed that the majority uses headphones in daily listening situations. Headphones are chosen so the subjects can relate the test situation as much as possible to a daily listening situation.

5.4.6 Test Method and User Interface

The GUI for the test is a MATLAB program, a screenshot of the main page can be seen in Figure 5.2. Included in the test are the three factors; 1) attributes, which appears in the text at the top of the screen, 2) excerpts, which are not visible for the subjects and 3) compressions, which each are represented by a play button. This method is chosen so the subjects can compare the three different compressions by focusing on only one attribute and one excerpt at a time. The test is composed of four attributes, for each attribute the four different excerpts are played, which gives a total of 16 trials.

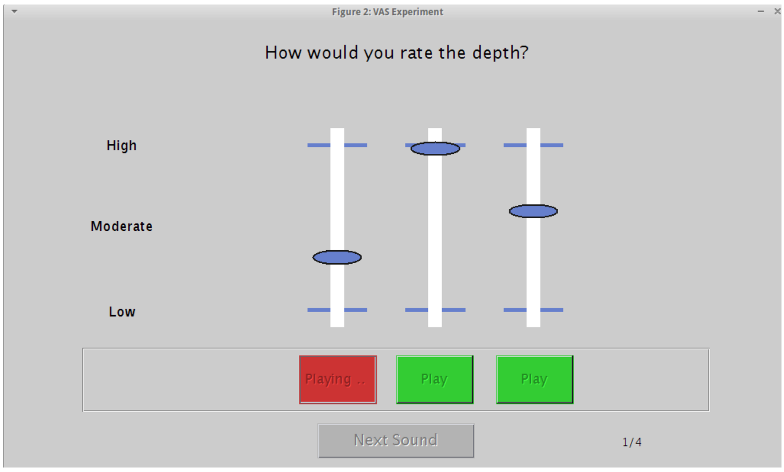


Figure 5.2: Screenshot of the test user interface.

For all 16 trials, the subjects' task is to rate each of the three different compressed versions for the given attribute described in the top text. For every trial the three different compressions appear in a random order. The attributes are fixed, meaning that they appear in the same order for all the subjects. For each attribute the four different excerpts appear in a random order. After playing the excerpts, the subjects can use the slider to rate the three different ones on a visual analog scale from 'Low' to 'High'. When playing a track, the green play button turns red and the subjects have to listen to the entire excerpt. The subjects can repeat the excerpts as much as wanted, but at least all three ones have to be played in order to move on to the next trial.

The test is constructed with an introducing familiarization phase. The first attribute is *depth* so before starting the test, the subjects are introduced to the meaning of this attribute through the familiarization. This includes two examples of an excerpt with low depth and one with high depth, respectively. It is noted that none of the excerpts from the listening test appear in the familiarization phase. During the test the subjects have a document with words that should help understand the meaning of the attribute, and also if the subjects should forget during test. The subjects are offered breaks during the familiarization phase in-between each attribute.

Chapter 6

Results and Data Analysis

All results for the hearing assessment and the listening test are presented in this chapter. The results of each part are examined and a thorough analysis of the results from the listening test is performed.

6.1 Subjects and Response from the Questionnaire

For the listening test, 21 subjects are a part of the listening test. The first subject is only a pilot test to be sure that the different parts are working correctly and that the overall time is not too long. The results for the pilot test are not being treated in the following part, therefore the finale subject number is 20 composed by 9 females and 11 males. All the subjects are students between the age from 18 to 26 except one who is 50. From the questionnaire about the hearing, none of the subjects reported major issues or frequent symptoms that could indicate hearing impairments. Four subjects have suffered from tinnitus a few times, but since it is not a permanent issue those subjects are still treated like the rest. Five respondents play an instrument, three of them for more than five years and more than 10 years for the two others.

6.2 Audiometric Test and TEOAE Measurements

As previously described, the hearing level is considered as normal for this project in the range from -10 dB(HL) to 20 dB(HL). Only two subjects violate those boundaries: the first reaches 30 dB(HL) at 8 kHz, which in this case is approved. The other one, who is also the older person of the listening test, is at all frequencies above 20 dB(HL) reaching a maximum of 75 dB(HL) at 8 kHz. Because of the high hearing levels and the fact that the age of this subject differs a lot from the rest, this subject is not considered in the experiment. For the following results the panel is composed of 19 subjects. The mean hearing levels for all the subjects are shown in Figure 6.1. The curves show slightly lower levels for the right ear, except at 2 kHz.

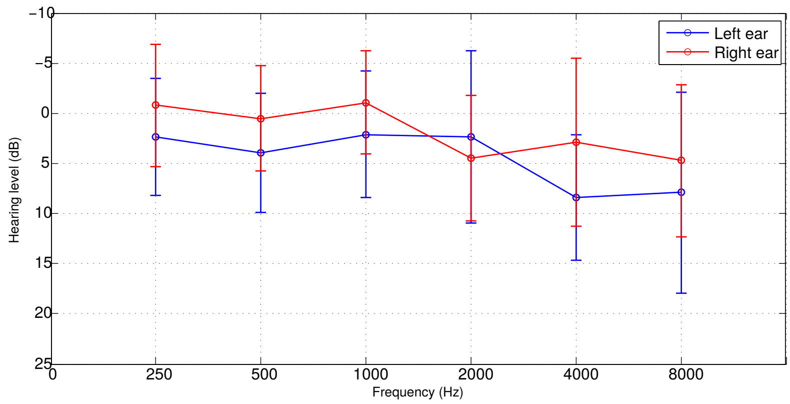


Figure 6.1: Mean and STD of the hearing levels for all 19 subjects, calibrated according to [ISO-389, 1975].

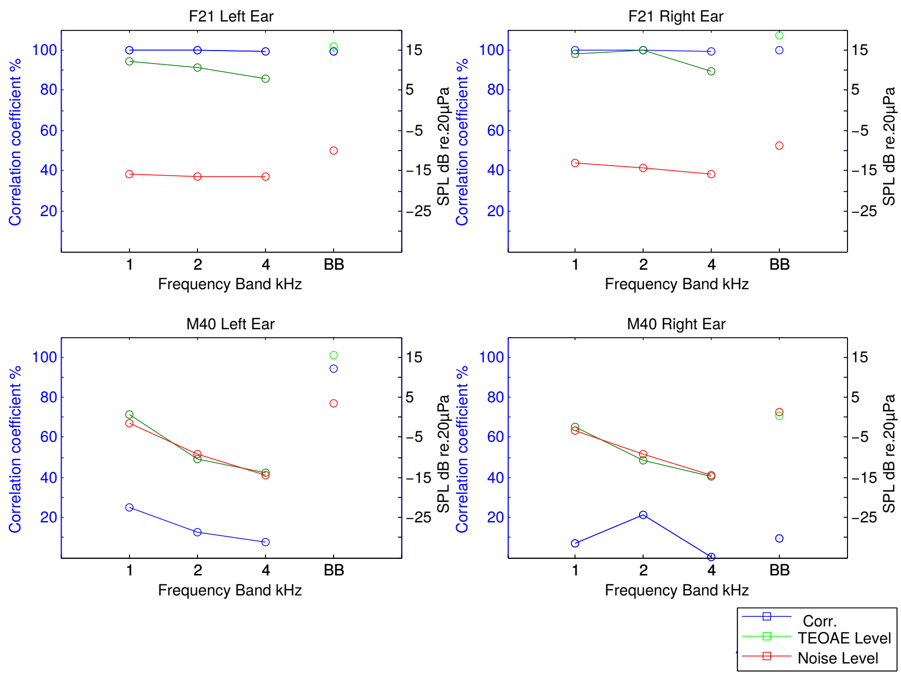


Figure 6.2: TEOAE of both ears for two subjects.

For the TEOAE measurements, a MATLAB script is used to interpret the response of the subjects by plotting the TEOAE, the noise and the correlation for the broad-band signal (BB) and in the three octave bands centered at 1, 2 and 4 kHz. It is noted that for one subject the TEOAE is not measured due to a technical issue during the session. The results of the TEOAE measurements are presented in Figure 6.2 for two subjects and both ears. The left scale on the y-axis relates to the correlation (blue graph) and the scale on the right represents the TEOAE and noise levels. The top ones have a fine

response when looking at both the correlation and the TEOAE levels, with only a low decrease at 4 kHz for both ears. The bottom ones are the results of the rejected subject, showing low values for the TEOAE at 1 kHz band and even lower for the 2 and 4 kHz band. The correlation is around 20 % and even lower for some bands.
The mean TEOAE responses are shown in Figure 6.3. The left ear is a bit lower which is also observed for the results of the audiometric test in Figure 6.1. The tendency of the curves is that the levels decrease with respect to higher frequencies. The average results of the left and right ear corresponds to the average results from the audiometric test.

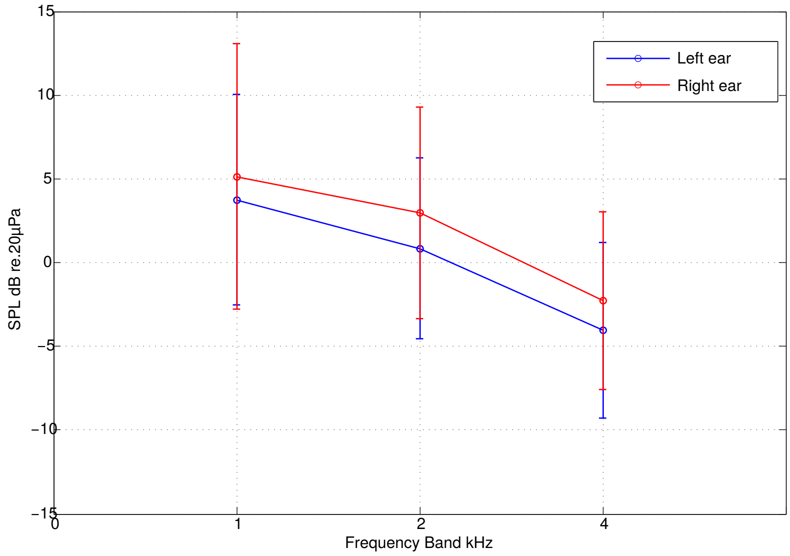


Figure 6.3: Mean and STD of TEOAE of both ears for all subjects.

6.3 Listening Test

All the results are listed in a .txt file and imported in the statistical analysis program 'R'. It contains the factors: *attributes*, *compression* and the excerpts listed for the results as *genres*. First results for each compression are plotted. Figure 6.4 shows the mean ratings across subjects and genres along the y-axis and compared to each attribute on the x-axis. The plots show only small variations in compression for the four attributes. Figure 6.5 shows the mean rating across subjects and attributes, but compared to each genre on the x-axis. The plots show slight increase in ratings for the flute when looking at the three compressions, whereas the violin has a small decrease looking at the compressions.

Since three different compressions are tested, the data is analyzed using a repeated measures ANOVA. First step of the data analysis is to perform a three-way ANOVA in order to test for differences between the three factors; compression, attributes and

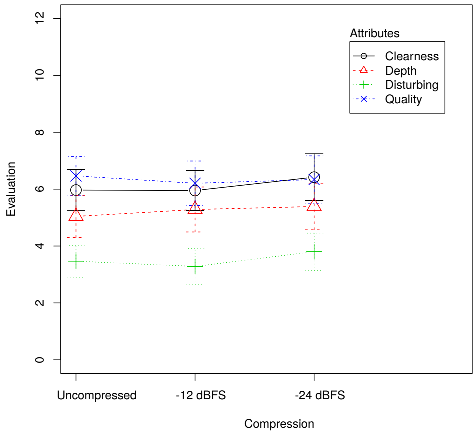


Figure 6.4: Mean ratings and confidence interval with a level of 95 % across genre and subjects, for each compression.

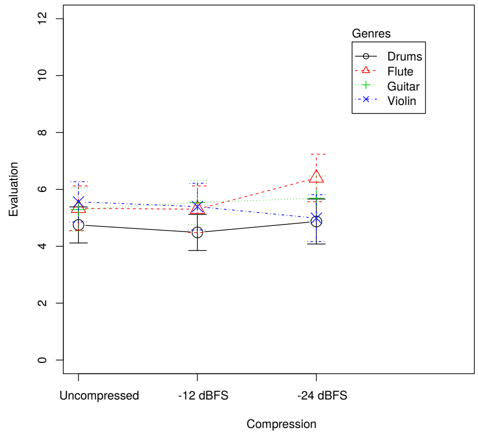


Figure 6.5: Mean ratings and confidence interval with a level of 95 % across attributes and subjects, for each compression.

genres. and is performed on the data from all 19 subjects, with a significance level of 0.05. Result of the ANOVA shows that the attributes (F=19.37, p=1.18e-08) and the genres (F=3.661, p=0.0178) are significant factors. However, the ANOVA also shows an interaction between the attributes and the genres (F=2.821, p=0.00291) which increases the risk of a false positive test.

To observe tendencies for the first significant factor, attributes, Figure 6.6 shows the plots of mean ratings across subjects and genres. The three curves represent each compression, and the variation between those are small for each attribute. From the figure it is seen that the ratings for *disturbing sounds* are lower than the rest, especially the difference to *quality* can explain why the result of the ANOVA for attributes is significant. To observe tendencies for the second significant factor, genres, Figure 6.7 shows the plots of mean ratings across subjects and attributes. The variations across genres show no obvious outliers, even tough the ratings for drums are a bit lower.

The *disturbing sounds* is removed from the attributes and a new three-way ANOVA test is performed, with same conditions as the previous one. The results show that attributes are still a significant factor (F=5.772, p=0.00669), so the variance of *disturbing sounds* itself is not the only one causing the significance for the attributes. Because of this result, the attribute *disturbing sounds* is still part of the analysis and present in the following results. To analyze the significant factors in more details, two-way ANOVAs are performed starting for each attribute, with compression and genres as factors and a significance level of 0.05. The results for *depth* and *quality* show no significant factors. For *clearness*, the genres show to be a significant factor (F=5.794, p=0.00165) and for *disturbing sounds*, the compression shows to be a significant factor (F=2.48, p=0.098) but

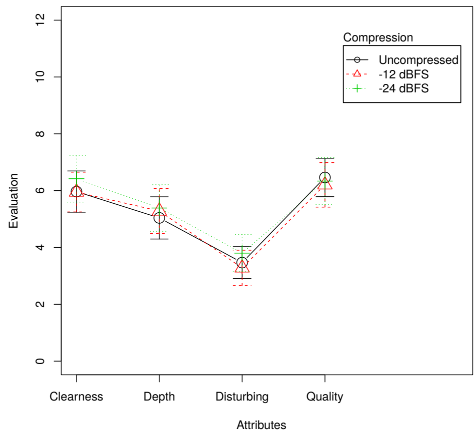


Figure 6.6: Mean ratings and confidence interval with a level of 95 % across genre and subjects, for each attribute.

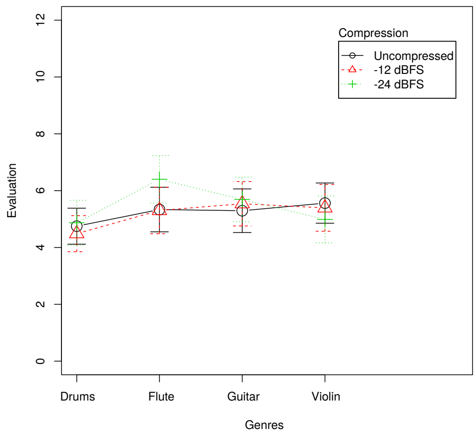


Figure 6.7: Mean ratings and confidence interval with a level of 95 % across genre and subjects, for each genre.

at the lowest level possible. Also there is a significant interaction (F=2.351, p=0.0358) between the compression and the genres for *disturbing sounds*.

Two-way ANOVAs are also performed on each genre, with compression and attributes as factors and still with a significance level of 0.05. The results for *drums* show no significant factors. For *flute* the compression shows to be a significant factor (F=3.527, p=0.0399) but the attributes show to have higher significance (F=14.97, p=3.26e-07). For both *guitar* and *violin* the attributes have a significance of (F=6.706, p=0.000627) and (F=10.91, p=1.04e-05) respectively. To go more into details in the study of the results, one-way ANOVAs are performed on the data with only one factor; compression. First, all the data is divided into groups of each attribute and each genre in order to perform a one-way ANOVA on the total 16 pair combinations. The results indicate that the compression is still only a significant factor (F=4.64, p=0.016) for the attribute *disturbing sound* in pair with the *flute* which is also observed with the two-way ANOVA.

In order to validate the repeated-measure ANOVA, the dataset is checked for sphericity for the three level of compression. Mauchly's test is commonly used to prove sphericity and is included as a standard function in MATLAB and R. The method is applied on the 16 combination pairs from the previous one-way ANOVA, and the variance is calculated in each level of compression for all subjects. The null-hypothesis of the Mauchly's test is that the variances in-between the compressions are equal and a p-value calculating the probability of this condition is computed. The closer the p-value is to 1 the higher the probability of sphericity, if it is less than 0.05 then the sphericity is considered as not tenable. Table 6.1 represents all the p-values from the Mauchly's test and as it can be seen, for five conditions the sphericity is rejected. It is also observed that most of the p-values are very low and on the border to get rejected. Especially for *clearness* and *disturbing sounds*, the values are very low and four out of five rejections are present for these two attributes.

	Depth	**Clearness**	**Disturb**	**Quality**
Drums	0.1961	*0.0080*	*0.0000*	0.0976
Flute	0.8856	0.0932	0.0247	0.2208
Guitar	0.1220	0.1260	0.1907	*0.0406*
Violin	0.1961	*0.0042*	*0.0000*	0.3915

Table 6.1: Sphericity check of all attributes and genres with a given significance level of 0.05. Italic p-values mean that the assumption of sphericity is not tenable.

Because of the violations from the sphericity test, a correction method needs to be applied in order to obtain a valid repeated measure ANOVA or another choice is to perform a multivariate analysis. From the literature, the Greenhouse-Geisser method is a common used correction method, and is applied in this case. The method adjusts the degrees of freedom in the ANOVA, in order to produce a more accurate p-value. First step in the method is to estimate an ε value. This value indicates to which degree sphericity has been violated. If sphericity is perfectly met then this value is one and the further the value gets below one, the worse the violation [Field, 2012]. The lower bound for ε is decided from the number of levels, k, on the repeated measure factor and is calculated in Equation (6.1).

$$\varepsilon_{lowerbound} = \frac{1}{k-1} = \frac{1}{3-1} = 0.5 \tag{6.1}$$

In this case the number of levels is 3, and the lower bound of ε is 0.5. So the estimated ε for each violation must be between 0.5 and 1. The example for the corrections method is given for the *quality* of the *guitar*. The compression is not a significant factor for this combination, but since the sphericity is violated a new corrected p-value is computed. The estimated ε value is calculated in MATLAB by the Greenhouse-Geisser method and applied on the results from the one-way ANOVA for this combination, where F=0.13, the degrees of freedom are 2 and 54 and p=0.8779. The ε value is 0.723 and gives new degrees of freedom 1.46 and 26.03 which yields a new p-value of 0.8130. So after the corrections, the compression is still not significant for this combination. Corrections are performed for all five violations, but it does not change the fact that the compression is not as significant factor.

The subjects are split into groups to check if subjects with musical experience or a better hearing show different results in detecting the compression. From the questionnaire about the hearing, five subjects have played an instrument for at least five years, which might be a benefit when listening for details in music. Comparing the results for a three-way ANOVA of only the subjects playing an instrument and the results for the three-way ANOVA of all the subjects, shows similarities with the attributes as a significant factor (F=3.54, p=0.0482), the genres as a significant factor (F=2.746, p=0.0879) and a significant interaction (F=3.090, p=0.00193). As the results are not showing any different information and the fact that there is a bias coming from comparing groups of different sizes, a more detailed analysis for the group that plays an instrument is not performed.

Instead, results from the TEOAEs are included in order to roughly divide subjects into groups with different hearings. By looking at the correlation and the TEOAE levels for 1, 2 and 4 kHz, six subjects have both high correlation and high TEOAE levels with only a small decrease in the 4 kHz band. These six subjects distinguish from the rest, and will from now be referred to as a group of 'good hearing'. Seven subjects are difficult to point out from the rest, as both correlation and TEOAE levels varies in the mid range of the scale, without being in neither the high or low end. The last six subjects have either low values for the correlation or the TEOAEs, or large drops in levels at 2 kHz or 4 kHz. From this criteria, the response of all six subjects distinguish from the rest, and will form a group that will be referred to as a group of 'bad hearing'.

Figure 6.8 and 6.9 represent the average TEOAE for both ears for all the subjects and the two different groups in frequency bands centered in 1, 2 and 4 kHz. The average levels for the group of good hearing are higher in all three bands, compared to the average of all subjects. For the group of bad hearing the levels are lower in all three bands.

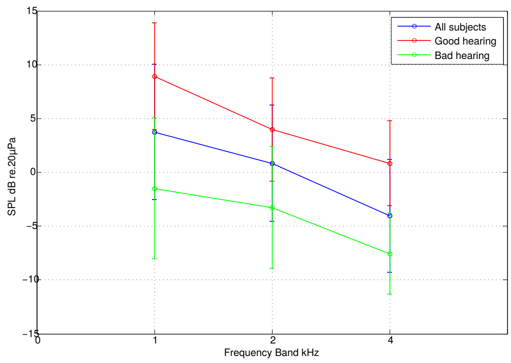


Figure 6.8: Average TEOAE for left ear of all subjects and the two groups with good and bad hearing including the STD.

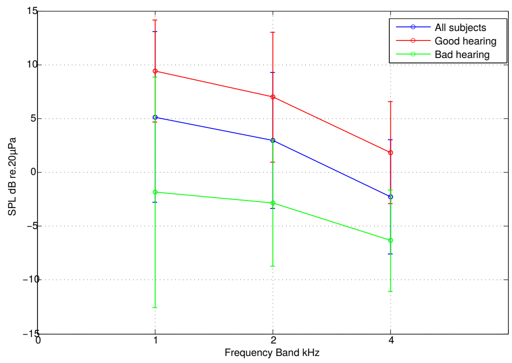


Figure 6.9: Average TEOAE for right ear of all subjects and the two groups with good and bad hearing including the STD.

Same comparisons are illustrated for the audiograms in Figure 6.10 and 6.11. For both left and right ear, the group with good hearing have lower average hearing levels than the group of bad hearing. The hearing levels are also lower than the ones for all subjects. Only violation is at 8 kHz where the group of bad hearing are lowest for both left and right ear. The purpose is now to compare the results from the two groups, by repeating the procedure that is performed for all the subjects, processing only the subjects from the two groups.

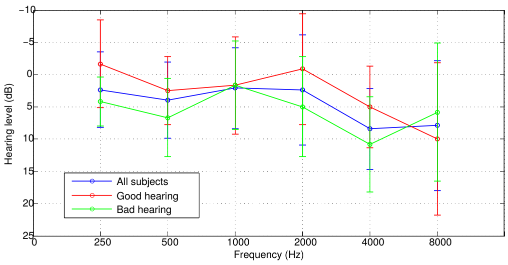


Figure 6.10: Audiogram with mean and STD for left ear of all subjects and the two groups with good and bad hearing, respectively, calibrated according to [ISO-389, 1975].

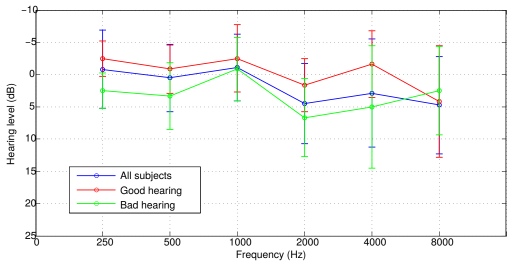


Figure 6.11: Audiogram with mean and STD for right ear of all subjects and the two groups with good and bad hearing, respectively, calibrated according to [ISO-389, 1975].

Performing a three-way ANOVA test for the group with good hearing shows that only the attributes are a significant factor (F=9.692, p=0.00158). The interaction between attributes and genres is in this case still a significant factor (F=2.018, p=0.0391). For the group of bad hearing, the three-way ANOVA shows that the attributes are a significant factor (F=4.149, p=0.0251), the genres are a significant factor (F=4.117, p=0.0257) and the interaction between attributes and genres is also significant (F=3.436, p=0.000597). In Figure 6.12 and 6.13 are observed the plots of the group with good hearing and bad hearing only for the significant factor, the attributes. Starting with the group of good hearing the graphs have common points as for the case of all the subjects in Figure 6.6, but for all attributes the most compressed one scores the highest mean ratings. For the bad hearing it is harder to find a clear tendency, but noticeable is the depth and quality where the most compressed ones score the lowest ratings.

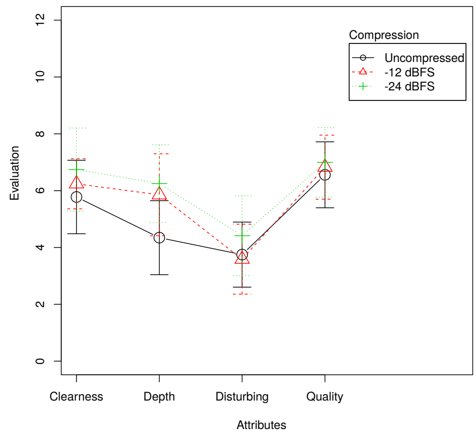


Figure 6.12: Mean ratings and confidence interval with a level of 95 % across genre and subjects from the good hearing group, for each attribute.

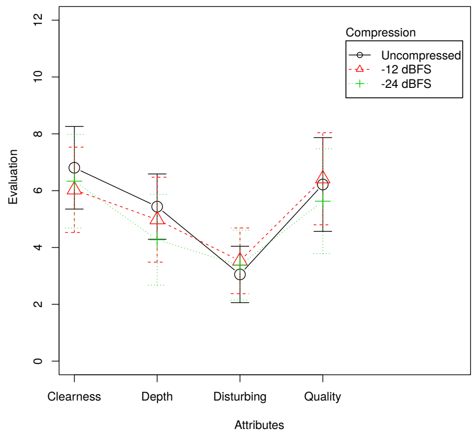


Figure 6.13: Mean ratings and confidence interval with a level of 95 % across genre and subjects from the bad hearing group, for each attribute.

As previously, two-way ANOVAs are performed on each attribute with genres and compression as factors. Also, two-way ANOVAs are performed on each genre with attributes and compression as factors. The results of the eight ANOVAs do not report compression as a significant factor, for both groups. Instead, one-way ANOVAs are performed on each combination pair of the attributes and the genres for both groups. The results of the one-way ANOVAs show no significance for the compression in all cases, for both groups.

The sphericity test is performed also for the two groups of subjects, starting with the ones with good hearing. The p-values are significantly higher with only two rejections shown in Table 6.2. The rejections are still present for *clearness* and *disturbing sounds* but with higher values compared to Table 6.1 for all the subjects.

	Depth	**Clearness**	**Disturb**	**Quality**
Drums	0.3801	0.3244	*0.0431*	0.3333
Flute	0.0904	0.1088	0.1011	0.2901
Guitar	0.9799	*0.0264*	0.2415	0.6368
Violin	0.7429	0.5919	0.1351	0.2808

Table 6.2: Sphericity check for subjects only with good hearing. Italic p-values mean that the assumption of sphericity is not tenable.

The sphericity test is also performed on the group with bad hearing, this is shown in Table 6.3. The sphericity test shows five violations, including all values for *disturbing sounds* and for *clearness* for *violin*.

	Depth	**Clearness**	**Disturb**	**Quality**
Drums	0.0665	0.5418	*0.0043*	0.3765
Flute	0.3343	0.3019	*0.0049*	0.1757
Guitar	0.3413	0.1136	*0.0286*	0.7532
Violin	0.2787	*0.0302*	*0.0116*	0.5190

Table 6.3: Sphericity check for subjects only with bad hearing. Italic p-values mean that the assumption of sphericity is not tenable.

Same correction method is performed on the violations for the two groups. In all cases, the corrections do not change the fact that the compression is not a significant factor.

A bias is noticed in the data. Some zeros are present in the dataset without corresponding to the subjects' ratings, and are noticed only after the experiment. The false-zeros come from the graphic interface, when reading the values of the rating scale. Table 6.4 represents the number of zeros found in each subjects' ratings even though some of them might not be misreadings but actual ratings. The total of zeros is 121 for the 19 subjects and are evenly distributed across attributes, genres and compressions. It represents 13.3 % of the total number of ratings, which is 912.

Subj.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Zeros	1	2	9	0	8	11	2	15	7	16	0	4	2	18	1	0	15	6	4

Table 6.4: The number of zeros in the subjects' ratings.

As it can been seen from the table, the repartition is not even across subjects and therefore is slip into groups. Indeed, some have a numerous amount of false-zeros compared to others but never more than 18 out of a total of 48 answers. To discard the subjects with a high number of zeros a limit of 4 is set, which results in a new group of 10 subjects. A new three-way ANOVA with the same factors as previously is performed and the results show compression to be significant (F=3.222, p=0.0637), attributes (F=13.16, p=1.75e-05), and genres (F=4.592, p=0.0101), which are similar results to the first analysis for all subjects. The attribute disturbing sound has still a lower rating than the others, a new ANOVA is performed without it. The results show only the attributes to be significant (F=3.17, p=0.0662) and genres (F=4.619, p=0.00983).

The decision is made to discard the zero of the answers for all the subjects and plots of the new results are shown in Figures 6.14 and 6.15 for attributes and genres respectively compared to the compression. By comparing those plots with Figures 6.4 and 6.5, the mean ratings are higher due to the absence of zeros but the tendencies are similar. In order to verify the attributes and check for correlation, a scatterplot is performed in Figure 6.16 without taking into account the zeros. As it can be observed, the disturbing sounds is rated usually low compared to the other attributes and most of the values are below the linear curve for the quality.

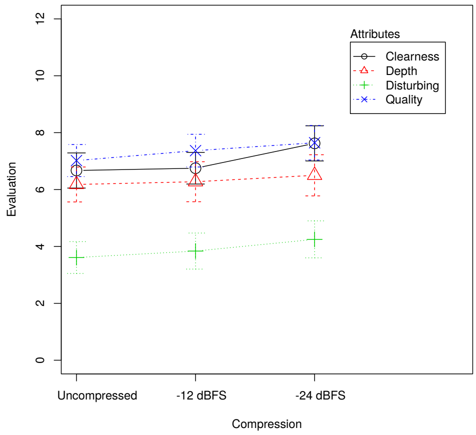


Figure 6.14: Mean ratings and confidence interval with a level of 95 % across genre and for all subjects without including the zeros.

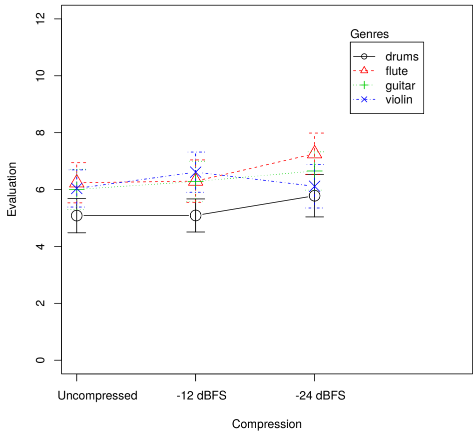


Figure 6.15: Mean ratings and confidence interval with a level of 95 % across genre and for all subjects without including the zeros.

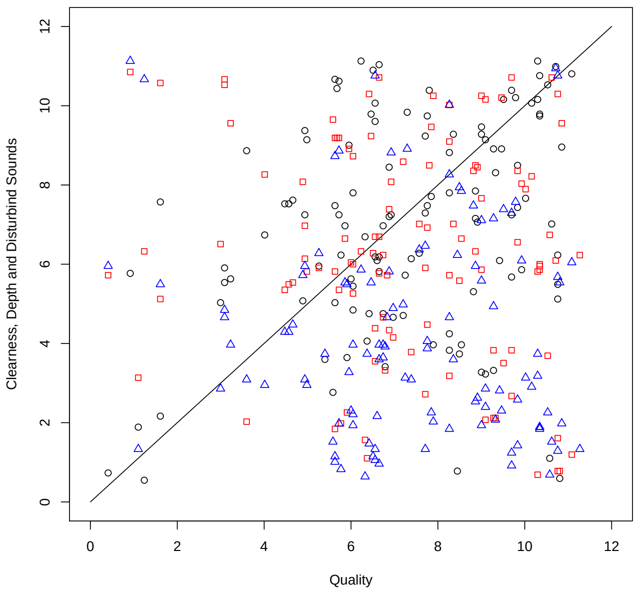


Figure 6.16: The attribute quality compared to the three others: depth in blue, disturbing in red and clearness in black.

Chapter 7

Conclusion and Discussion

7.1 Project Summary

To make a subjective evaluation of the impact of the compression in music quality, the decision was made to perform a listening test. First step to achieve this goal was to study the loudness which plays a significant role in music compression. The listening test would focus only on the compression, eliminating loudness as a factor. Therefore the study of the loudness was performed in Chapter 2 to find measures for assessing loudness in music in order to normalize between different degrees of compression. The LRA was used as a measure to obtain more information about the impact of the compression between soft and loud parts. Program loudness was chosen for normalizing loudness across compression. The normalization gave fair results since no loudness differences were detected in the pilot tests.

The first thought was to implement a compressor and use it for generating the excerpts for the listening test but after a discussion at TC Electronic, the decision was made to use an existing compressor. The purpose of the project was to test the impact of the compression and not the quality of an implemented compressor. The Finalizer 96K from TC Electronic was used to generate the excerpts. The decision was made after an implementation was already realized in MATLAB. Descriptions and examples of the main parameters of this implemented compressor were included in Chapter 3.

The assumption from [Vickers, 2010] is that the use of compressed music is indeed to satisfy the PMP users, who may listen to music in noisy environments. Therefore an Internet questionnaire survey was performed to study the listening habits and with the purpose to recruit subjects for the listening test. Chapter 4 included the design and results, which confirmed that PMPs are widespread as 92 % of the respondents are using one. An average of 13 hours pr. week is used on PMPs in noisy environments like during transport, preferred by most respondents. No correlation was found between listening level and background noise. Only 28 % of the respondents considered themselves as active listeners in the preferred listening location, indicating that many use PMPs as a secondary activity to block out the background noise. 75 % of the respondents had at sometime been bothered by quality but only some managed to put words on the issues. This indicates that the quality is noticed by many respondents but to express its issues in words is a difficult task. 42 of the respondents agreed to participate in the listening test.

The design and procedure of the listening test was described in Chapter 5. The test phase included two parts; a hearing assessment and the listening test. First part was divided into three, including a questionnaire about the hearing, an audiometric test and TEOAE measurements. The audiometric test was used to reject one subject and the questionnaire and TEOAEs were used to group the subjects in the result analysis. The response from the questionnaire did not give arguments to reject subjects but a group of subjects with musical experience was formed. A frequency band analysis of the TEOAEs were used to group subjects into the ones having good responses and low responses.

The listening test material included two different compressions with fixed parameters and a threshold of -12 dBFS and -24 dBFS, and an uncompressed. The four attributes were tested; depth, clearness, disturbing sounds and quality. Four different genres were used and consisted of a rock song with dominating drums, a flute, a guitar and a violin. The method chosen for the listening test was a three way comparison of the different degree of compression.

The results of each part of the test were presented and analyzed in Chapter 6. By use of repeated measures ANOVA for 19 subjects showed no significance for the compression. Eliminating attributes and genres separately as factors and repeating the ANOVAs showed only low significance for compression in few cases. The analysis was performed for a reduced pool by using the group of musicians and the groups resulting from the TEOAEs. The ANOVAs showed similar results for the compression. A bias was noticed from the GUI, which replaced the actual ratings with zeros in about 1/10th of all the ratings. This bias was consistence between attributes, compressions and genres but not for the subjects. A new group was found, allowing only subjects with a maximum 4 numbers of zeros in the evaluations. Similar results were observed.

Because all the different analysis show similar results, a conclusion can be drawn. The three different degrees of compression are a difficult task for the subjects to distinguish. Compression with fixed parameters and changing only the threshold shows not to be a significant factor in respect to quality for a naive listener, when loudness normalization is applied. Audible cues are present in the loudness due to compression but disappear with loudness normalization.

7.2 Discussion

The discussion involves the design and decisions for the listening test, where different choices could have been made. First for the excerpts, the threshold for the compression is limited by the equipment but a higher level could have been used if there are no artifacts. A test was considered with different thresholds with steps of 4 dBFS, in order to observe if subjects could distinguish between them. Using the same excerpts and performing a paired-comparison test could be another method to confirm the results that audible cues for the compression are removed with the normalization of loudness. In this project the ratio is set at a fixed limit without introducing artifacts. This parameter could have been modified along with the threshold to observe the different impact of the compression depending on the genre.

To ease the evaluation for the subjects, other alternatives for test method could have been used. An example could be the three alternative forced choice method (3AFC), in which the subjects would only have to choose the preferred one instead of rating all three.

The choice of attributes were based on the study from [Gabrielsson and Sjögren, 1979] and the results from the questionnaire about listening habits, in order to use proper words for evaluating quality. Different attributes could have been used. Especially disturbing sounds showed to have lower means than the rest of the attributes and a correlation between low ratings of disturbing sounds and high ratings of quality was observed.

The test panel was composed of average listeners, for whom the degree of compression was difficult to detect. Another procedure could have been to include longer training sessions for the subjects to get familiar with the test or include only professionals from the music industry already familiar with compression. Both might have a better detection of the compression, which could have shown different results.

Bibliography

ANSI-S3.4. *Procedure for the computation of loudness of steady sounds.* ANSI, 2007.

Bobulous. Lossless and lossy audio formats for music, 2009. URL `http://www.bobulous.org.uk/misc/audioFormats.html`.

Lorne Bregitzer. *Secrets of recording.* Elsevier, 2009.

Josh Coalson. What is flac?, 2013. URL `https://xiph.org/flac/`.

Naomi B. H. Croghan, Kathryn H. Arehart, and James M. Kates. Quality and loudness judgments for music subjected to compression limiting. *Journal of the Acoustical Society of America*, 2012.

Emmanuel Deruty and Damien Tardieu. About dynamic processing in mainstream music. *Journal of the Audio Engineering Society*, 2014.

DIN-45631. *Calculation of loudness level and loudness from the sound spectrum.* DIN, 1991.

DS/ISO-532. *Method for calculating loudness level.* Dansk Standardiseringsråd, 1975.

EBU-R128. *Loudness normalisation and permitted maximum level of audio signals.* European Broadcasting Union, 2011.

EBU-TECH3342. *Loudness range: a measure to supplement loudness normalisation in accordance with EBU R 128.* European Broadcasting Union, 2011.

Andy Field. Repeated measures anova. *www.discoveringstatistics.com*, 2012.

Harvey Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *The Bell System Technical Journal*, 1933.

Alf Gabrielsson and Håkan Sjögren. Perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America*, 1979.

Line Gebauer and Peter Vuust. *Music intervention in health care.* Danish Sound Innovation Network, 2014.

David Hammerschmidt Robin Helm David Hoga Julian Kraus Jakob Rösch Hendrik Böhne, Rene Gröger and Christian Sussek. Subjective audibility of mp3-compression artefacts in practical application. *Results of Praktikum Musikpsychologie, student work at Hamburg University*, 2011.

Florian Völk Hugo Fastl and Martin Straubinger. Standards for calculating loudness of stationary or time-varying sounds. *Inter-Noise 2009, Ottawa, Canada*, 2009.

ISO-226. *Normal equal-loudness-level contours.* International Organization for Standardization, 2003.

ISO-389. *Standard reference zero for the calibration of pure-tone audiometers.* International Organization for Standardization, 1975.

ISO-8253-1. *Audiometric test methods - part 1: Pure tone air and bone conduction audiometry.* International Organization for Standardization, 2010.

ITU-R.BS.1116-1. *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.* International Telecommunication Union, 1997.

ITU-R.BS.1770. *Algorithms to measure audio programme loudness and true-peak audio level.* International Telecommunication Union, 2012.

JISC. Uncompressed audio file formats, 2014. URL `http://www.jiscdigitalmedia.ac.uk/guide/uncompressed-audio-file-formats`.

David Kemp. Otoacoustic emissions, their origin in cochlear function, and use. *British Medical Bulletin*, 2002.

Thomas Lund. The calm act and cross-platform broadcast. *National Association of Broadcasters (NAB) Las Vegas*, 2012a.

Thomas Lund. Loudness wars and sensation. *Loudness summit*, 2012b.

Dimitrios Giannoulis Michael Massberg and Josha D. Reiss. Digital dynamic range compressor design - a tutorial and analysis. *Journal of the Audio Engineering Society*, 2012.

Dimitrios Giannoulis Michael Massberg and Josha D. Reiss. Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 2013.

Brian C.J. Moore. *An introduction to the psychology of hearing.* Brill, 2013.

Meinard Müller, Daniel P. W. Ellis, Anssi Klapuri, and Gaël Richard. Signal processing for music analysis. *Journal of Selected Topics in Signal Processing*, 2011.

Rodrigo Ordonez, Dorte Hammershøi, Carsten Borg, and Jan Voetmann. A pilot study of changes in otoacoustic emissions after exposure to live music. *Journal of the Audio Engineering Society*, 2012.

Torben Poulsen. Acoustic communication - hearing and speech. *Lecture Notes from DTU*, 2005a.

Torben Poulsen. Psychoacoustic measuring methods. *Lecture Notes from DTU*, 2005b.

Bhanu Prasad and SRM. Prasanna. *Speech audio image and biomedical signal processing using neural networks*. Springer, 2007.

Roger Shepard. Circularity in judgments of relative pitch. *Journal of the Acoustical Society of America*, 1964.

Esben Skovenborg. Loudness range (lra)-design and evaluation. *Journal of the Audio Engineering Society*, 2012.

Earl Vickers. The loudness war background speculation and recommendations. *AES Convention*, 2010.

David S. Walonick. *A selection from survival statistics*. StatPac, Inc., 2010.

Mads Walther-Hansen and Jens Hjortkjær. Perceptual effects of dynamic range compression in popular music recordings. *Journal of the Audio Engineering Society*, 2014.

Udo Zölzer. *Digital audio signal processing*. John Wiley and Sons Ltd., 1997.

Appendix A

Measurement Journal

The measuring journal for the listening test is specified in this appendix, including equipment, procedure and results for the audiometric test, the TEOAE measurements and the actual listening test.

Equipment

Table A.1 contains all the equipment used for all the test setup.

Type	AAU-number	Brand	Model
Audiometer	33968	Madsen Electronics	Orbiter 922
Probe driver preamp	75584	ER	-
Sound card	64682	Edirol	-
Headphones	2036-70	Beyerdynamic	DT990
PC	64688	Fujitsu Siemens	-
Laptop	-	Samsung	-

Table A.1: Full list of equipment used for the test.

Procedure

The testing phase includes four steps, which are the same procedure for every subject. Each subject is granted with an ID number, in order to structure the information and keep it confidential. The procedure of each step is described in the following.

Questionnaire

First step is for the subject to answer a questionnaire similar to the one for listening habits, but this time about the hearing. The subject enters an ID number, and fills out the questionnaire. If there are any questions during the process, support is available at the time to clear out any doubts.

Audiometric Test

After the subject has submitted the answers, an audiometric test is prepared in the listening cabin. The subject is instructed in the procedure of the test before starting. The test is controlled outside the listening cabin in a control room. First a single tone is presented as a level expected to be certainly audible, e.g. 40 dBHL. The subject has the task to always answer if the tone is audible by clicking on a button. The level is reduced with 10 dB until the subject can not hear the tone. Then the level is increased of 10 dB. Same procedure of decreasing the level again but this time the steps are only of 5 dB. If it is inaudible for the subject again, the level is increased with 10 dB and the decreasing steps following this are of 5 dB. This procedure continues until the subject reaches 3 times the same level which is going to be defined as the threshold. The results of first the left ear then the right ear are marked in an audiogram together with the subject's ID and participant number.

TEOAE Measurements

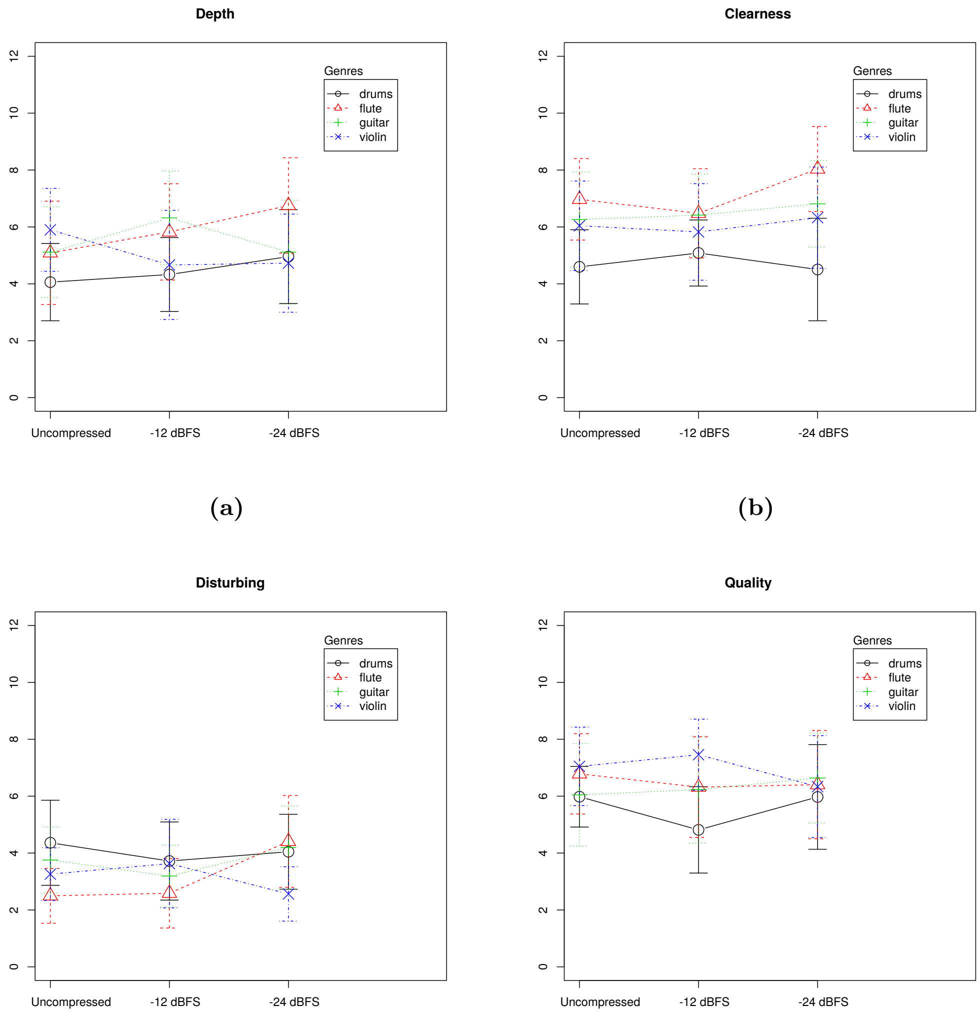
Next step is for the TEOAE measurements. The probe is placed inside the left ear of the subject, who is told to sit still and not to breath too deeply nor to swallow during the measurements. It also possess a tool to check the fit of the probe in the ear canal by sending a short impulse at 80 dB, if the response does not correspond the probe position is corrected. The system uses a sampling frequency of 48 kHz giving a TEOAE response of 960 samples (20 ms) with a frequency resolution of 50 Hz and a bandwidth from 500 Hz to 6 kHz. For each TEOAE measurement 500 repetitions of the stimuli block are recorded. The rejection threshold for the noise is set at 40 dB SPL. The measurements last about 1 minute and are performed for both ears. From the control room, the measurements are saved along with the subject's ID and which ear is measured.

Listening Test

Last step is the actual listening test. The procedure is described in Section 5.4.6.

Results

The results from the listening test include plots of the mean ratings for all 19 subjects, for only the group with good TEOAEs and for only the group with bad TEOAEs. In each case the mean ratings are plotted for each attribute and each genre and compared with each different compression, including the confidence interval of 95 %. Figure A.1 shows the mean ratings for all 19 subjects for each attribute.

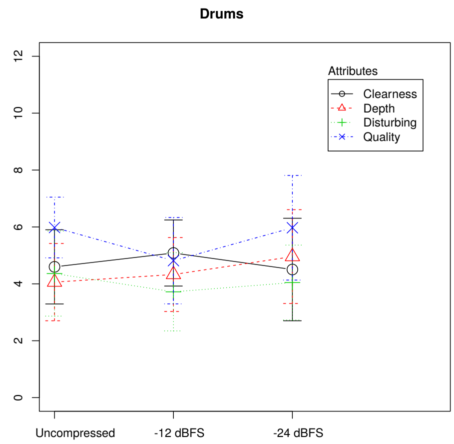


(a) (b)

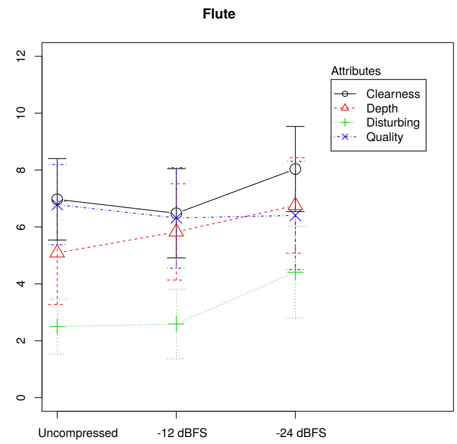
(c) (d)

Figure A.1: Mean ratings of all 19 subjects for each of the attributes. On the x-axis is the three different compressions.

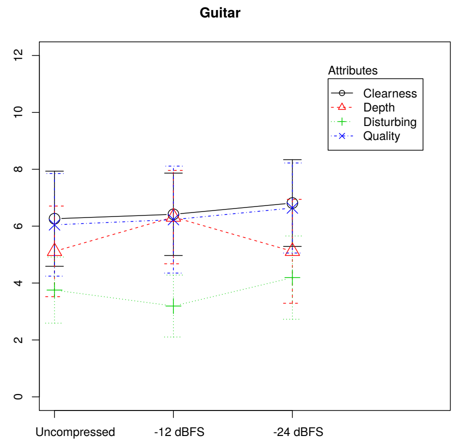
Figure A.2 shows the mean ratings for all 19 subjects for each genre.



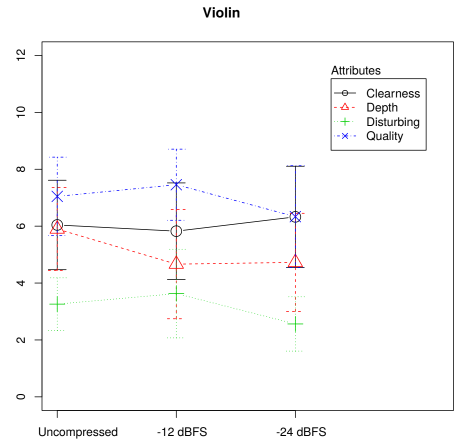
(a)



(b)



(c)



(d)

Figure A.2: Mean ratings of all 19 subjects for each of the genres. On the x-axis is the three different compressions.

Figure A.3 shows the mean ratings for the group with good hearing and for each attribute.

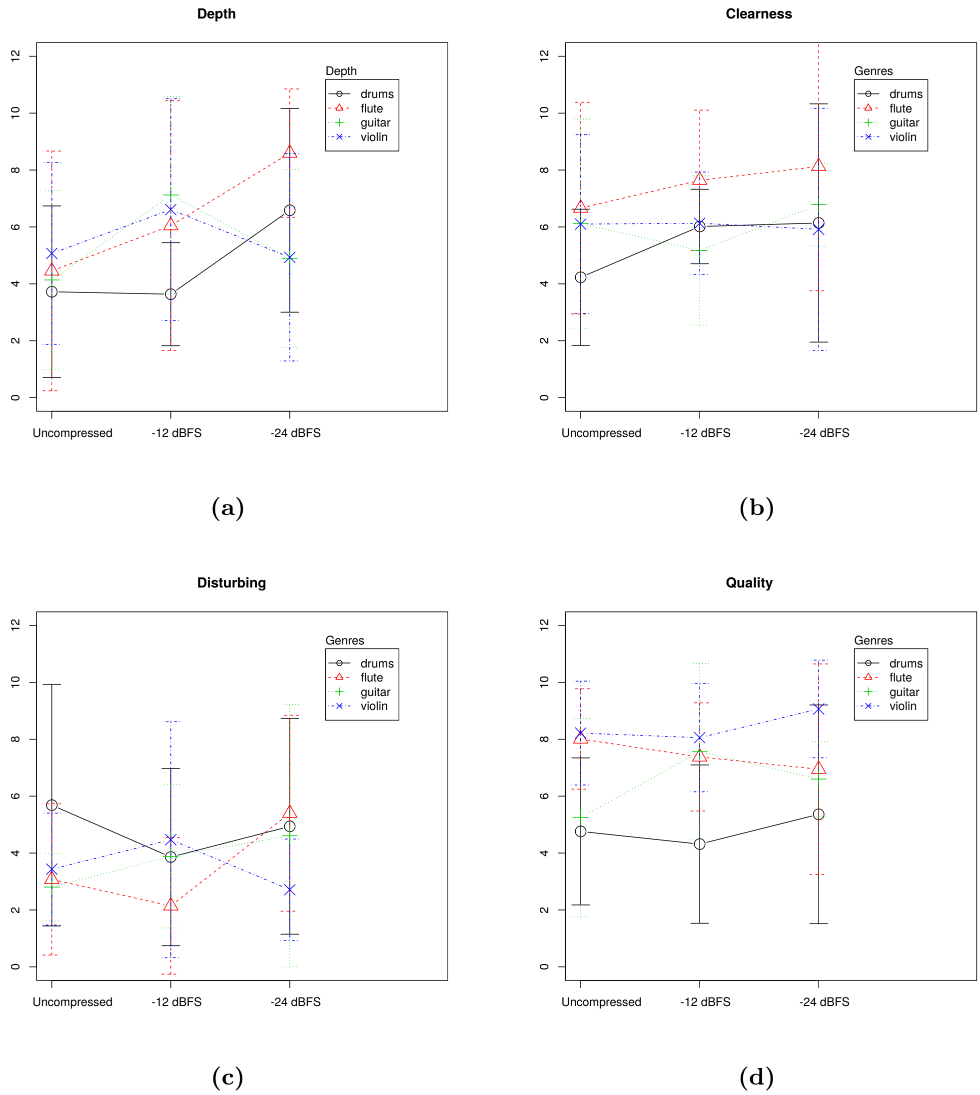
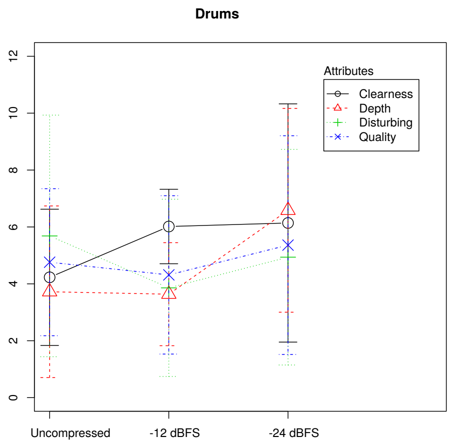
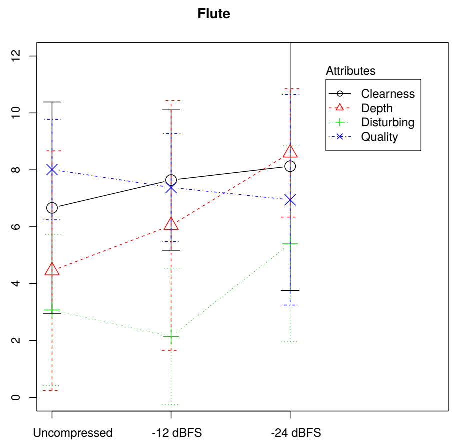


Figure A.3: Mean ratings of the group with good hearing for each of the attributes. On the x-axis is the three different compressions.

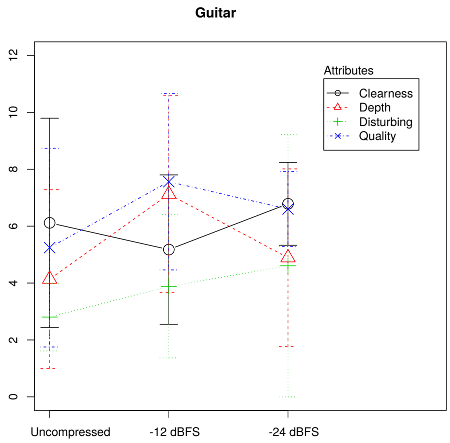
Figure A.4 shows the mean ratings for the group with good hearing and for each genre.



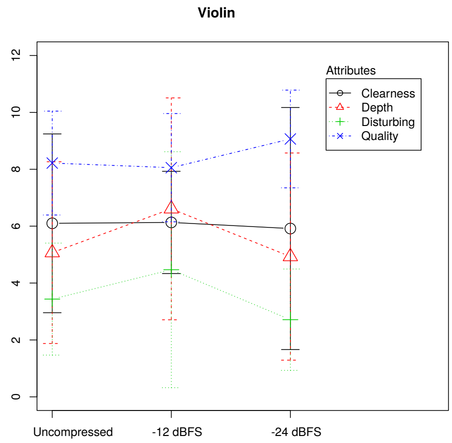
(a)



(b)



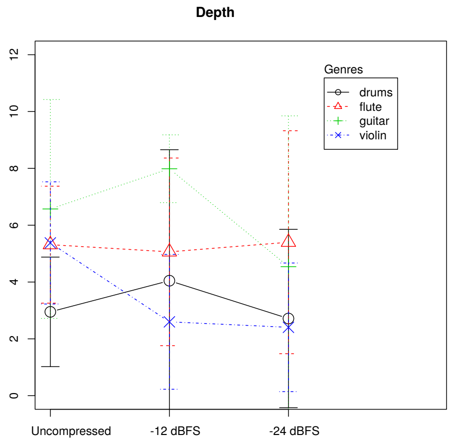
(c)



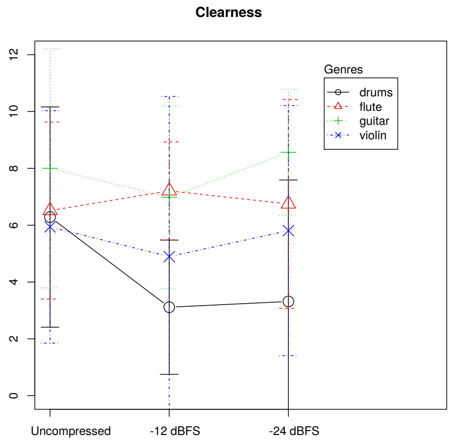
(d)

Figure A.4: Mean ratings of the group with good hearing for each of the genres. On the x-axis is the three different compressions.

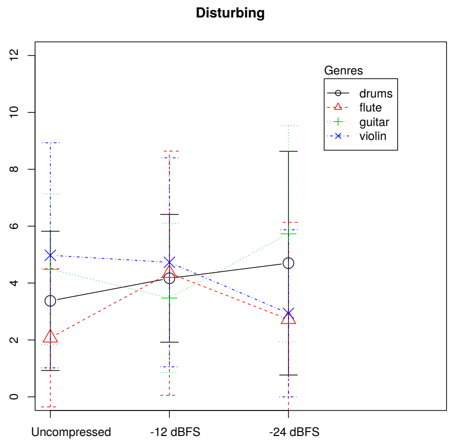
Figure A.5 shows the mean ratings for the group with bad hearing and for each attribute.



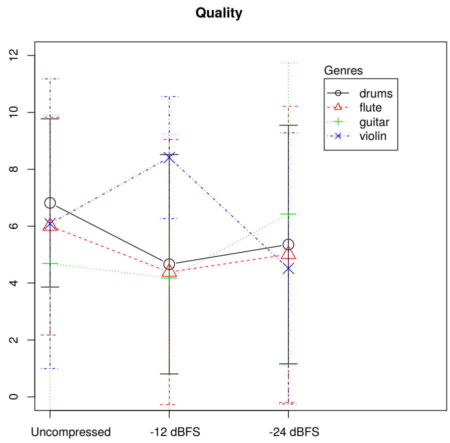
(a)



(b)



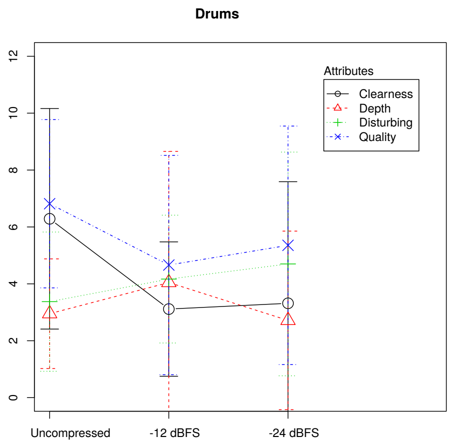
(c)



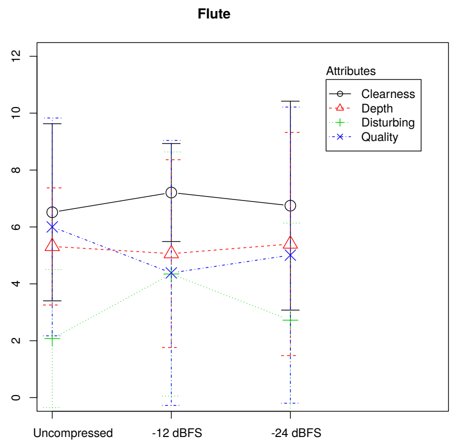
(d)

Figure A.5: Mean ratings of the group with bad hearing for each of the attributes. On the x-axis is the three different compressions.

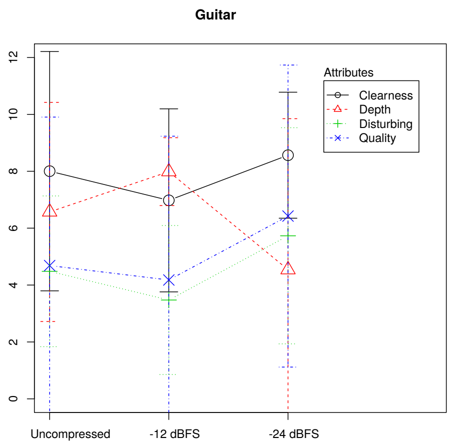
Figure A.5 shows the mean ratings for the group with bad hearing and for each genre.



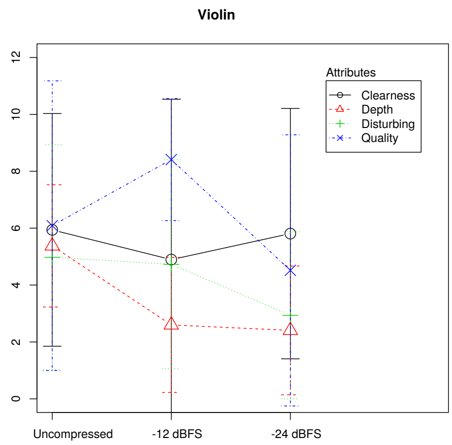
(a)



(b)



(c)



(d)

Figure A.6: Mean ratings of the group with bad hearing for each of the genres. On the x-axis is the three different compressions.

Appendix B

Questionnaire - Listening Habits

The questionnaire used for getting information about listening habits for PMP users is included in this appendix.

Questionnaire about your listening habits

Personal Informations

Name:

Birth date: Day Month Year

Gender: ◯Female ◯Male

e-mail:

What is your occupation/profession (e.g. student, office worker etc.)?

1. Do you listen to music from a portable music player? (smartphone, music player, tablet etc.) ◯**yes** ◯**no**

What kind of portable music player do you mostly use?
◯Smartphone (iPhone, Android phone etc.)
◯Music player (MP3 Player, iPod, CD player etc.)
◯Tablet (iPad, Windows Surface etc.)

which brand and model?

What kind of headphones do you use?
◯Earbuds (Placed in the outer ear but not in the ear canal)
◯In-ear (Placed inside the ear canal)
◯Circumaural (Placed around the ears completely surrounding them)
◯Supra-aural (Placed on the ears rather than surrounding them)
◯Speakers

which brand and model?

How often do you listen to it?

An example to answer the next question:
You listen to music while biking every weekday to work and back home for 30min each way and you have done this for 5 months, then you have to click on: Weekly, 5 days, 1 to 2 hour and 2 to 6 months.

How often?	Days	Hours	**For how long?**
◯Weekly	◯1	◯1/2	◯less than a month
	◯2	◯1 to 2	◯one month
◯Monthly	◯3	◯2 to 4	◯2 to 6 months
	◯4	◯4 to 6	◯6 to 12 months
◯Annually	◯5	◯6 to 8	◯more than a year
	◯6	◯8 to 10	◯more than 5 years
	◯7	◯10 to 12	◯more than 10 years

How loud do you listen to it?

◯very low ◯low ◯moderate ◯loud ◯very loud

2. Where do you spend most time listening to music?

◯ during transportation ◯ at home ◯ at work
◯ other:

There are two main ways of listening to music: active and passive. The first one represents listening to music while being fully focused only on this, whereas passive listening is while doing other activities in the mean time like working or exercising.

At that place, how do you listen to the music?

◯ active ◯ passive

How loud do you consider the background noise at that location?

◯ very low ◯ low ◯ moderate ◯ loud ◯ very loud

3. Have you ever been bothered by the sound quality of a song?

◯ yes ◯ no

How would you define the issue?

4. Which genres of music do you listen to?

☐ Pop ☐ Disco ☐ Jazz ☐ Funk ☐ Classical
☐ Soul ☐ Rock ☐ Heavy Metal ☐ Gothic
☐ Hip-Hop ☐ Rap ☐ Country ☐ Folk ☐ Electronic

☐ other:

5. Would you be interested in participating in a listening test later this spring at AAU?

◯ yes ◯ no

Thank you for your answers!

Submit Clear

Appendix C

Questionnaire - Hearing

The questionnaire about the hearing used as a step in the hearing assessment of the subjects is included in this appendix.

Questionnaire about your hearing

Subject nr:

1. Have you suffered from:	**How often?**				
	never	a few times	several times	often	always
Tinnitus					
Sudden hearing loss					
Otitis media (Middle ear infection)					
Excessive ear wax					
Drain					
Pain due to loud sound					

2. Have you experienced:	**How often?**				
	never	a few times	several times	often	always
Problems understanding speech					
Problems understanding phone conversations					
Problems understanding what is said on television					
Headache from loud sound					
Pain or discomfort in airplanes					

3. Have you taken medicine or others drugs that affect your hearing?
yes **no** **don't know**

Which medicine?

How often?
once several times often always

4. Have you had an ear operation?
yes **no** **don't know**

Which operation and when:

5. Is there someone in your close family that has or has had a hearing disorder?
yes **no** **don't know**

If yes, please specify:

6. Do you go to nightclubs? ◯ yes ◯ no			
How often?	Days	Hours	**For how long?**
◯ Weekly	◯ 1	◯ 1/2	◯ less than a month
	◯ 2	◯ 1 to 2	◯ one month
◯ Monthly	◯ 3	◯ 2 to 4	◯ 2 to 6 months
	◯ 4	◯ 4 to 6	◯ 6 to 12 months
◯ Annually	◯ 5	◯ 6 to 8	◯ more than 1 year
	◯ 6	◯ 8 to 10	◯ more than 5 years
	◯ 7	◯ 10 to 12	◯ more than 10 years
Do you use hearing protection? ◯ yes ◯ no ◯ sometimes			

7. Do you go to concerts? ◯ yes ◯ no			
How often?	Days	Hours	**For how long?**
◯ Weekly	◯ 1	◯ 1/2	◯ less than a month
	◯ 2	◯ 1 to 2	◯ one month
◯ Monthly	◯ 3	◯ 2 to 4	◯ 2 to 6 months
	◯ 4	◯ 4 to 6	◯ 6 to 12 months
◯ Annually	◯ 5	◯ 6 to 8	◯ more than one year
	◯ 6	◯ 8 to 10	◯ more than 5 years
	◯ 7	◯ 10 to 12	◯ more than 10 years
Which genre: ☐ rock ☐ classical ☐ jazz ☐ electronic ☐ RnB ☐ pop ☐ heavy metal ☐ reggae ☐ other: ______			
Do you use hearing protection? ◯ yes ◯ no ◯ sometimes			

8. Are you a musician? ◯ yes ◯ no			
Which instruments do you play? ______			
How often?	Days	Hours	**For how long?**
◯ Weekly	◯ 1	◯ 1/2	◯ less than a month
	◯ 2	◯ 1 to 2	◯ one month
◯ Monthly	◯ 3	◯ 2 to 4	◯ 2 to 6 months
	◯ 4	◯ 4 to 6	◯ 6 to 12 months
	◯ 5	◯ 6 to 8	◯ more than a month
	◯ 6	◯ 8 to 10	◯ more than 5 years
	◯ 7	◯ 10 to 12	◯ more than 10 years
Do you play in/with/at: ☐ rock band ☐ orchestra ☐ jazz band ☐ big band ☐ home ☐ studio ☐ pub ☐ work ☐ other: ______			

Do you use hearing protections?

◯ yes ◯ no ◯ sometimes

Thank you for your answers!

Submit Clear