Master's Thesis

Audio- and Visual Cues in Mixing and Mastering

Lukas Knoebl

Graz, April 18, 2016

Institute of Electronic Music and Acoustics Graz
University of Music and Performing Arts

Graz University of Technology

Advisor:
Msc. Ph.D. Georgios Marentakis

Assessor:
O.Univ.-Prof. Mag.art. DI Dr.techn. Robert Höldrich

Abstract

With the emergence of digital technology into the audio market, computer-based music systems have become an integral part for the creation of professional audio productions. Most software in that area is still operated with mouse and keyboard and often involves complex graphical user interfaces. The goal of this master thesis is to investigate, whether and to what extent the visualisation of audio-effect parameters can influence the mixing decisions of audio engineers. For this purpose, test subjects performed two common mixing and mastering tasks by using just the aural modality, just the visual modality or both modalities at the same time. The first experiment was designed to investigate whether visual cues affected test subjects when matching the loudness of various test signals. Similarly, test subjects were asked to match the spectral characteristics of two test signals by using an equalizer in a second experiment. Though software such as sequencers and audio effect plug-ins have become indispensable tools for most audio engineers, the influences of audio- and visual cues in mixing and mastering have hardly been researched scientifically. This thesis should give some first indications about the interaction and implications of aural and visual modalities in conjunction with common tasks in mixing and mastering.

Zusammenfassung

Seit dem Einzug der Digitaltechnik in den Audiobereich werden computerbasierte Musiksysteme für viele Arbeitsschritte einer professionellen Tonträgerproduktion herangezogen. Die Bedienung solcher Software erfolgt meist traditionell mit Maus und Tastatur und erfordert daher die Bereitstellung von oftmals komplexen grafischen Benutzeroberflächen. Das Ziel dieser Masterarbeit ist, zu untersuchen, ob und in welchem Ausmaß die Visualisierung von Audioeffekt-Parametern die Entscheidungen von Mixing- oder Mastering-Ingenieuren beeinflusst. Zu diesem Zwecke wurden Hörversuche entworfen, in denen die Testpersonen typische Mixing- und Mastering Aufgaben unter unterschiedlichen Testbedingungen durchgeführt haben. Im ersten Versuch sollten jeweils zwei Testsignale unter rein auditiven, rein visuellen sowie audio-visuellen Versuchsbedingungen in ihrer Lautheit angepasst werden. In einem zweiten Experiment sollte die Kernfrage der Arbeit auch in Bezug auf die Angleichung der spektralen Charakteristik zweier Signale untersucht werden. Trotz der breiten Anwendung computerbasierter Systeme in der Musikproduktion wurde der Einfluss der damit einhergehenden Visualisierungen auf die kreativen Entscheidungen Tonschaffender bisher kaum untersucht. Diese Masterarbeit soll erste Aufschlüsse über das Zusammenspiel von Sehen und Hören in Verbindung mit typischen Mixing- und Mastering Aufgaben liefern.

Acknowledgements

I would like to thank my supervisors O.Univ.-Prof. Mag.art. DI Dr.techn. Robert Höldrich and Msc. Ph.D. Georgios Marentakis for their persistent support and many insightful discussions. Thanks to DI Ph.D. Matthias Frank for a very helpful introduction to Pure Data.

My audio-engineering colleagues at the Graz University of Technology, who have truly made my university years a friendly and vibrant experience. Thanks for all the fun we have had in the last years. Special thanks to those who willingly shared their precious time and participated in my experiments, this work would never have been possible without their efforts.

Lastly, I thank my friends and loved ones, who have supported me through the entire process and inspired my passion for music. My family - Mum, Dad and Lydia - for their love, patience and for providing me with a peaceful environment to live, grow up and learn. Thank-you for inciting me to strive towards my goals and for ongoing support on the new roads I am taking.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Attributable to constantly increasing computational power, entire audio productions from recording to final mastering can nowadays be completed on a single personal computer [70]. Once recorded, sound can then be processed within the digital domain. Graphical user interfaces of audio processing tools, displayed on the screen and typically operated with mouse and keyboard, usually consist of faders, knobs and may even show certain features of the source signal, such as its waveform or its real-time frequency spectrum [42, 54]. Though software, such as sequencers and audio effect plug-ins, have become indispensable tools for most audio engineers, the impact of visual representations of audio-parameter changes on sound mixing has hardly been researched scientifically.

The goal of this thesis was to determine whether users of digital music production systems are affected by the presence of task related visual information. Do sound engineers act differently when the graphical user interface of an audio effect is shown compared to when they are mixing solely by ear? If yes, what are the differences and is it possible to quantify them?

In order to find answers to these questions, two separate experiments were designed and implemented. In the first part, subjects were asked to match the loudness of various test signals to a given reference signal. Similarly, as part of the second experiment, participants were required to remove a resonance from a signal by modifying the parameters of a semi-parametric equalizer. In both experiments, the task had

to be performed under several experimental conditions involving different sensory modalities and various degrees of visual bias.

The present thesis is divided into eight chapters. Chapter 2 outlines the historical evolution of audio production and describes the essential concepts in mixing and mastering. Further, the reader is presented with information about the modern mixing environment. In Chapter 3, established interface metaphors found in Digital Audio Workstations are discussed with emphasis on their suitability, ease of use, their drawbacks and their impact on sound mixing. Chapter 4 provides important information about cross-modal interactions and summarizes mechanisms related to selective attention between vision and sound. Chapter 5 provides a literature review about existing research related to the topic of this thesis. Next, the procedures and implementations for both the loudness and the spectral matching task are described in great detail in Chapters 6 and 7, respectively. Results are presented and discussed separately in Sections 6.3 and 7.3. Finally, main findings are summarized and an outlook for the future is sketched in Chapter 8.

Chapter 2

Audio Production

For a long time, recorded sound was stored on magnetic tape [40]. One of the main drawbacks of this medium was that referencing the original material was not possible. The audio existed only once and was locked to a certain tape position. Additionally, making a copy of the recording by re-recording it to another tape introduced a loss in quality as a consequence thereof. The digital domain did not only remove this restriction, as referencing the original file became possible with hard disc drives and digital audio tapes, but also provided an interface in order to exploit this technology [70].

2.1 The Evolution Of Mixing

In the early days of the recording industry back in the 1950s, mixing was almost non-existent due to the fact that there were generally just a few microphones involved even in large recording sessions. Furthermore, the recording format was only mono. All of this changed in 1955 with the process of Selective Synchronous recording (Sel-Synch), which allowed overdubbing of individual tracks. This was achieved by using the record head of the tape machine also as a playback head in a way that all tracks stay in synch [42].

Back in the days of tape recorders with up to eight tracks, it was common practice to reduce the number of tracks of related instruments. For instance, a multi-miked

drum kit was mixed down to a single stereo track whereas at the same time the original files were overwritten. However, with each submix also the noise of all the separate recordings added up, which effectively limited the possible number of tracks. Since the previous tracks could not be accessed any more, engineers had to continuously relate their current mixing decision to something which had not even been recorded yet, thus, mixing was an integral part of the recording process [42].

As technology evolved and more and more tracks could be recorded with a single machine, larger mixing consoles with automation- and recall functions had become necessary. In 1975, 24-track machines were already very popular, which again had great influence on the mixing philosophy. Many instruments were now recorded in stereo and the drums superseded the bass as the central element of a mix. The reason for this was that there were more microphone inputs and recording tracks available. Mixing engineers were now able to sculpt the sound of the drums more precisely, because instead of only three they could now use eight or even more microphones to record a drum kit [75]. However, in the late 1970s and 1980s digital recording was still very expensive [54].

The history of the digital era in commercial audio production began in the early 1980s, when the compact discs standards were introduced and the first CD players debuted [74]. Early digital audio workstations[1] ran on computers seriously underpowered for many traditional music production tasks and some developers compensated this with proprietary hardware-software solutions [70].However, digital systems developed rapidly, and with increased performance many of the original bottlenecks, like hard disk performance and access, have disappeared. Nowadays, computing power is well ahead of what is required for digital music production [70].

Since the turn of the century, the recording industry has changed quite dramatically. Not only have recording and production budgets been reduced, but also the number of big sound studios has decreased significantly [75]. They were often substituted by small project studios and as a consequence the so-called *Digital Audio Workstations* became more and more popular and made it possible for almost everybody to produce CDs at home and reduce the costs involved [54, 75].

[1] A digital audio workstation (DAW) is commonly known as a computer with the appropriate hardware and software needed to digitize and edit audio [75]

Nowadays, the computer based digital audio workstation can deliver exceptional sound quality, technical capability and flexibility at the fraction of a cost of earlier 48-track digital mixing consoles. A properly outfitted computer or laptop can now function like a sophisticated recording studio and by using special software one can achieve studio-quality sound. The possible number of tracks which can be recorded and played back simultaneously is essentially only limited by the power of the computer. Various types of digital audio effects are available at very low cost for many platforms, ranging from classic digital equalizers to accurate models of legendary audio processing hardware units. All of them can be inserted on a track with a single mouse-click and all settings can be stored and recalled at any time. Such a system is used in most of today's professional recording, broadcasting and production studios and can provide great flexibility, simplify the engineer's workflow and speed up the mixing process [54].

With today's technology, it is possible to edit and playback hundreds of tracks in a single project. Current DAWs provide the flexibility to undo all mixing decisions at any point, as the original recordings are always accessible [42]. Most audio software includes functions to directly compare several different settings of audio effects against each other or the unprocessed file, and everything can be altered until the final mixdown is printed [70].

2.2 The Mixing Process

Audio mixing is an important part of the video and music production chain where dozens of individual recordings have to be combined to a single recording (the mix) [16].

Izhaki [42] describes mixing as "a process in which multitrack material – whether recorded, sampled or synthesized – is balanced, treated and combined into a multichannel format", or in a less technical definition as "a sonic presentation of emotions, creative ideas and performance".

The most basic task in mixing is to balance the levels of the individual recordings throughout a song by applying a volume change to each of the tracks. The mixing

process involves both technical and creative aspects. The former are usually not directly concerned with the sound but relate to technical problems or routines, for instance track-layout and organization of audio files in the DAW, checking for phase issues, editing (performance corrections), restauration and clean-up. The creative process comprises all tasks that are necessary to craft the mix, like equalizing a vocal track, adding reverb or using compressors to control the dynamics of certain instruments. This requires a great amount of concentration, thus it is suggested to clearly separate this process from the technical tasks. Any distraction which can break the creative flow should be avoided [42].

Basically, the mixing process is affected by whether the production process involves recorded music or sequenced music. From start to finish, every professional audio production of recorded music typically consists of five different stages: Songwriting, arranging, recording, mixing and mastering [42]. Theoretically, each task of the production chain can be carried out by different people, though sometimes the mixing engineer will also be responsible for the recordings [42, 75]. Mixing is also largely dependent on the recording and arrangement stages, for example on the choice of stereo-microphone techniques or on the number of involved instruments. For instance, the mixing engineer has to overcome frequency, panning and time domain gaps in sparse instrumentations, and on the other hand the challenge in a very dense arrangement is to create space for each individual instrument [42].

While recording engineers are concerned with capturing the sound of an instrument as good as it can possibly be, the mixing engineer is responsible for combining many of those individual sounds. Sometimes these tasks can contradic each other, as a recorded instrument might sound great when played in isolation, but it might mask another instrument in the mix. In such a case, the mixing engineer has to filter some frequencies in order to make it work better in the context. Therefore, in many audio productions a single person (the producer), who is experienced in all three areas, is responsible for anticipating the mix in advance. This person usually gets involved early, helps with the arrangements and observes the recording process [42].

In contrast, the production process of sequenced music is even closer related to mixing. The functionality of current digital audio workstations allows mixing as the song evolves. Producers commonly select samples and effects based on how well

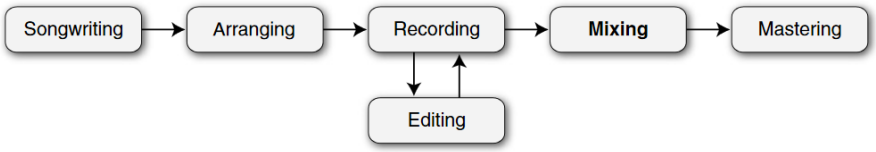


Figure 2.1: Common production chain of recorded music. Reprinted from [42]
.

they fit into the existing composition which makes the mix an integral part of the project file. Therefore, sequenced music is in many cases already pre-mixed even before it arrives at the dedicated mixing stage [42].

Figure 2.2: Common production chain of sequenced music. Reprinted from [42]
.

Unless they were involved in earlier production stages, most mixing engineers start their mix by getting to know the material. They either listen to a previous rough-mix from the artist or they simply turn all tracks to full volume and listen through the whole song a few times. The goal is to get an impression of the mood and emotion of the song and how it should sound like after mixing [42, 75]. In other words, the goal of this process is to create a clear vision about the final version. In most cases the mixing engineer starts working on a project alone but usually collaborates closely with the producer and/or artist once the basic framework of the mix is accomplished [42].

The process of mixing also involves some milestones, for instance bringing the mix to an adequate level for the beginning. In such a state, the result might still be far away from being perfect, but here the goal is to eliminate any major issues, such as problems with the raw tracks. The next milestone could be making the mix distinctive, or in other words, memorable by means of interest, emotion or any other sonic property. The final milestone is often concerned with stabilizing the mix and involves listening in different locations and at varying levels in order to make sure that it translates well to other rooms, sound systems and listening positions [42].

Owsinski [75] distinguishes between the following three relevant steps during the mixing process: Figure out in which direction the song develops, build and underline

the groove and find and emphasize the most important element.

Most successful mixing engineers also think in three dimensions: They want to make sure that all frequencies are represented adequately, that the mix has depth (front-to-back separation) and a proper stereo-panorama (left-to-right panning) [75].

The first dimension evolves from the understanding of what sounds right and this is related to a certain reference, for example successful audio productions which are commonly considered as High-Fidelity references among experienced mixing engineers. The goal is to achieve clarity and transparency, sparkling highs and a powerful bass.

The effect of depth can be accomplished by integrating atmospheric stereophonic sounds into a mix. This is usually achieved by adding reverb and delay, ambient microphones or even by making use of the crosstalk between different microphones.

The third dimension affects the width of the stereo-image, the spatial separation of individual instruments and provides the possibility to create interesting soundscapes.

There are six elements which are considered necessary to obtain an outstanding mix [75]:

- Balance: A proper relation of levels of the individual instruments
- Frequency range: All frequencies have to be represented adequately
- Panorama: Placing a musical element in the stereo-field
- Dimension: Adding spaciousness to musical elements
- Dynamics: Controlling the volume progression of a track or an instrument
- Interest: Emphasize the most important elements

In audio productions where it is crucial to capture a musical performance as purely as possible (like it is the case with classical music or jazz), it is probably sufficient to consider the first four elements. Otherwise, dynamic and interest have become very important factors in modern music [75].

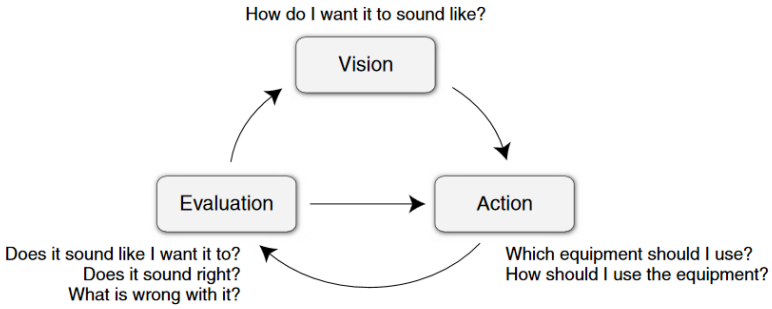


Figure 2.3: The iterative process of mixing. Reprinted from [42]

.

Izhaki [42] describes the creative part of mixing as an iterative process containing three major steps, which are illustrated below. Firstly, the mixing vision is primarily concerned with the question 'How do I want it to sound?'. While the novice engineer will often manipulate the sound by trial-and-error, the professional engineer has a clear sonic visualization of how the outcome should sound like instead. Secondly, the imagination of the desired sound leads to the mixing action, which involves the actual sound shaping procedure. The engineer will choose suitable tools and certain methods in order to achieve the desired result. Lastly, this result is evaluated and compared to the initial idea. In this way, the mixing vision manifests over time and the iterative process continues until the engineer is satisfied.

With their enormous amount of flexibility, current DAWs definitely support the correlative process of mixing, which may benefit from the iterative coarse-to-fine approach illustrated below [42]. While a certain group of instruments may sound good in isolation, some settings may have to be refined each time a new track is added to the mix. This suggests starting with rather coarse adjustments and then working and spending more time on the details as the mix progresses. It usually takes a few iterations for the engineer to get to the final settings as most attention is given to the last stage of the process, in which only subtle changes are made [42].

This whole iterative concept is largely consistent with Norman's [73] seven-staged model of interaction [70]. Accordingly, Norman's model consists of three major stages following the initial goal: *Intention*, *action* and *evaluation*, which are further subdivided in different aspects. The process starts with the goal, which implicates something to be achieved. This goal is then translated into an intention to perform actions so as to achieve it. Moreover, it has to be specified which sequence of actions

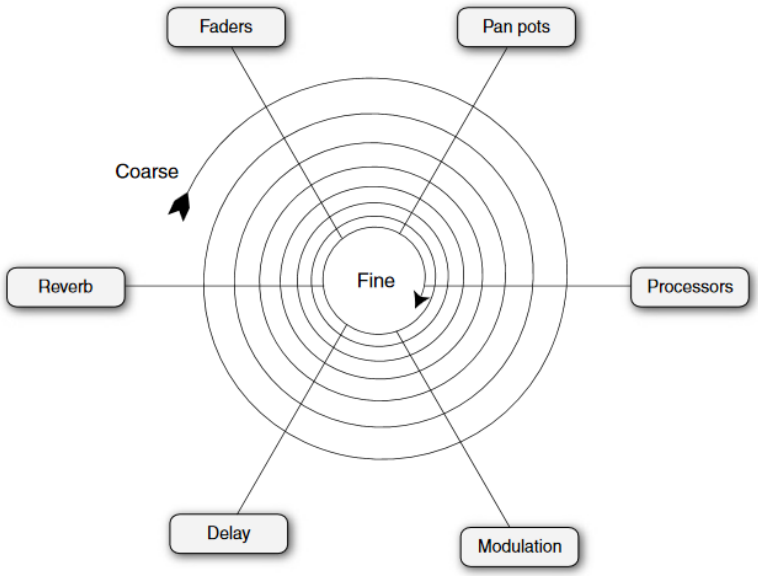


Figure 2.4: The iterative coarse-to-fine approach while mixing. Reprinted from [42]
.

is suitable in order to satisfy the intention. What follows is the aspect of execution where the physically possible actions are performed. Next, the stage of evaluation is formed by the perception of the world, its interpretation and the evaluation of the outcome in comparison to what was intended originally. Norman's seven stages of interaction do not exactly claim to be a complete psychological theory, they rather form an approximation model [73].

As illustrated in Figure 2.5, the seven stages of interaction are the following: It starts with *forming a goal*, followed by the process of execution, which is divided into *forming the intention*, *specifying the action* and emphexecuting the action. Lastly, the stage of evaluation consists of the *perception of the state of the world*, *the interpretation of the perception* and *the evaluation of the interpretations* [73].

In many cases, the seven stages of interaction form a loop. However, a poorly designed system, such as a system with an overly complex graphical user interface, may cause unnecessary iterations through the different stages [70]. According to [73], the major problems for users result from the gulfs that separate mental states from physical states, or in other words, the mismatch or differences between person's intentions and the actions provided by the system. Such gulfs can for instance be created by a slow system response or unclear layouts and input methods [70]. On the other hands, gulfs can be kept small when the system satisfies the expectations of

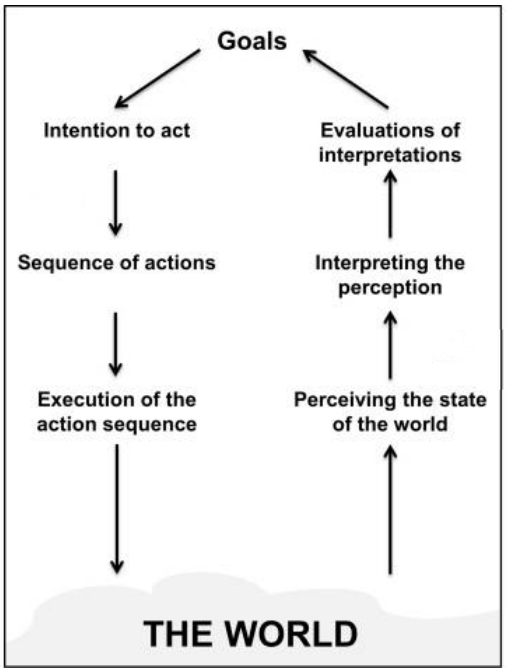


Figure 2.5: Norman's seven-staged model of interaction . Reprinted from [73]

the user and provides information about its state that is easy to understand. Good designs stand out due to a consistent conceptual model, comprehensible mapping and continuous feedback [73].

2.3 The Mastering Process

Mastering is a creative and technical intermediate step between mixing and duplication for distribution. The process of Mastering converts a collection of songs into a finished, professional sounding record, which translates well to a variety of playback systems. It ensures that the songs on an album sound cohesive in terms of tone, loudness and timing [46, 75].

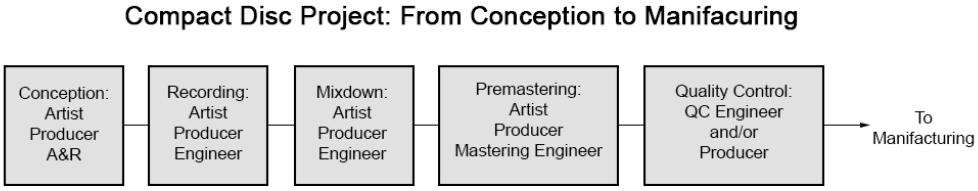


Figure 2.6: Major steps in Compact Disc production [46]

The early generation of mastering engineers (sometimes also referred to as *trans-*

fer engineers) became necessary in 1948 and was first and foremost responsible for transferring sound from magnetic tape to vinyl, which was quite a difficult process [75, 88]. In case the audio level was too low, the record was noisy and if it was too high, there was a risk of damaging the disc. Soon, they developed a method to increase the loudness of a vinyl by using equalizers and compressors, and producers as well as artists recognized that indeed those records sounded louder (and better) on the radio than others. This awareness consequently broadened the area of responsibility of the mastering engineers as they were also tasked to creatively affect the final sound of a record [75, 88].

The mastering workflow is typically comprised of editing, clean-up, levelling, processing and output to the final medium [46]. These procedures are usually applied either to the whole stereo-mix or individual subgroups (e.g. all drum instruments, all vocals, all guitars on separate stereo tracks), the so-called stems [46]. Therefore, mastering influences the overall dynamic and spectral balance of the recorded musical performance [88]. Generally, a mastering engineer will use similar tools as a mixing engineer, such as compressors, equalizers or reverberators. However, dedicated mastering studios will often use the very specific hi-end audio hard- and software in order to make the recordings sound bigger and louder [75].

According to Hodgson [39], the mastering engineer is responsible for tuning the acoustics of the mastering environment, setting up an appropriate monitoring, sequencing the order of tracks and adjusting the length of the pauses between them, applying fade-ins and fade-outs, equalization, managing perceived loudness and optimizing the master for different mediums. As described by Shelvock [88], all these areas concern musical taste.

2.4 The Digital Audio Workstation

Technological progress has influenced the recording industry in many ways, and while some vintage analogue audio processors are still very sought-after, especially by prestigious recording studios, computer systems have gradually superseded analogue devices. Regardless of whether it is a large scale professional or a small project studio, in most cases Digital Audio Workstations (DAW) form the core element of

any audio production facility [70].

Digital Audio Workstation is a term that describes a multifunctional computer-based audio system, equipped with a professional sound card and specialized software to handle typical audio production tasks, such as multitrack recording, editing, mixing, signal processing, MIDI sequencing and integrating virtual instruments and synthesizers [40, 70, 82]. Although sometimes the term *Digital Audio Workstation* refers solely to the audio software running on computer hardware, a DAW can only exist as a combination of hardware and software [70]. Typically, such a system also provides external high-quality analogue-to-digital and digital-to-analogue converters. Extensive, multifunctional DAW software is available from numerous different developers, the most popular being Steinberg Cubase and Nuendo, Avid Pro Tools, Apple Logic, Cockos Reaper, Ableton Live, Magix Samplitude and Reason [82].

It is reasonable to assume that most modern digital audio workstations originally evolved from MIDI sequencer software, as their core functions are still comparable [71]. Thus, it is not surprising that they are sometimes referred to as sequencers. However, DAWs also offer a lot of features that are not directly connected to sequencing [70] and their success and advantages are based on the combination of the essential functions from preceding, specialised devices to a single system [25, 70].

Generally, one can distinguish between *Integrated* and *Computer-based* DAWs. The former consist of a single device compromising a mixing console, a control interface and a digital interface. Though they were more popular back in the times when computing power was insufficient for serious audio work, Integrated DAWs made a comeback with the rise of touchscreens and embedded devices. On the other hand, Computer-based DAWs generally consist of a personal computer hosting a sound card with analogue-to-digital and digital-to-analogue converters and a powerful digital audio editing or sequencing software, like the ones described above. Usually, the software also provides the graphical user interface to allow for audio recording, editing and playback [82].

Moving to the digital domain has simplified many common tasks that are rather tedious in the analogue environment, for instance editing or copying material. It is remarkable that digital music production systems can fit on a single desktop and that

Figure 2.7: Photo-realistic graphical user interface of Wave's CLA-2A compressor plug-in. It is modelled after the hardware unit LA-2A from Teletronix

most of the audio work can be done by using only a minimum of external hardware, considering that analogue recording studios usually consist of many different devices connected to a large mixing console [70].

When inspecting particular elements of the interface in recent DAWs, it is evident that certain conventions evolved from a strong connection to the industry history [70]. A key principle in the interface design of digital audio workstations is that parallels are drawn between software interface elements and analogue devices. Consequently, a specific function in the software is often implemented by analogy with the procedure to achieve the same result with a hardware device. A volume control representing the familiar channel fader of a mixing console is a good example of this concept [70]. While the use of mappings and metaphors in form of physical analogies are strongly recommended in traditional usability literature [73], such models can on the other hand arguably prevent new users from developing an effective understanding of a system [35, 48].

2.4.1 Audio Plug-ins

The term *audio plug-in* refers to self-contained effects that can be used to extend the core functionality of a host-program [70, 82]. Their main task is to receive an input audio signal and apply some processing to it. In most cases, the output signal is passed back to the host-software. However, some types of plug-ins, like spectrum-analysers and level-meters, do not necessarily generate an output signal and yet others, for instance synthesizers, only receive Musical Instrument Digital Interface (MIDI) messages instead of audio signals [82];

Basically, audio plug-ins can be divided into two categories, namely instruments and effects. The former class includes virtual instruments, synthesizers or any other kind of signal generators. On the other hand, effect plugins can be used to process the dynamics or the spectrum of the incoming signal or to create additional sounds according to the input. Common audio manipulation tools like compressors, equalizers or reverberators fall into the latter category [70].

Furthermore, it can be distinguished between insert effects and send effects, depending on their position in the signal chain. As opposed to insert effects, a send effect is typically placed on a separate channel [70]. Then, the auxiliary outputs from many other arbitrary channels can be routed to this channel and the combined audio signal gets processed by the plug-in. This method is commonly used for reverberation or modulation effects, because it allows controlling the amount of *dry* and *wet* signals with at least two distinct faders [75].

Several industry-standard plug-in forms guarantee compatibility with many different DAWs. Most audio plug-ins are written in C or C++ programming languages, as for example *Steinberg*'s Virtual Studio Technology (VST). It is supported by nearly all professional audio software, as well by *Apple*'s AudioUnit (AU) and *Digidesign*'s Real-Time AudioSuite (RTAS) as their newer *Avid* Audio Extension (AAX) format are all equally popular and provide similar functionality [82].

Users can change the parameters of an audio plug-in through a graphical interface that is either automatically provided by the host-software or created by the plug-in developer [82].

Chapter 3

Interface Designs For Audio Processing

The design of current DAW software still mimics the analogue environment and thus is closely related to analogue devices, such as mixing desks and hardware audio processors, both functionally and visually [70].

Some theses, studies and developments are concerned whether such analogy restricts the usability of digital audio workstations. In this context, possible solutions and modernised approaches have been suggested by many authors. Some of them will be shortly addressed in the following of this thesis.

For decades the common personal computing interaction paradigm has been based on controlling a graphical user interface with mouse and alphanumeric keyboard, and therefore the same interface conventions found their way into computer-based digital audio workstations [70]. However, an interface is not necessarily limited to a graphical control surface [80]. Instead, new techniques are used in order to accomplish those tasks with different methods. Thus, an increasing number of concepts for tangible or auditory interfaces have been developed in recent years and some of them proved to be equally suitable for human-machine interaction. A number of works indeed provide evidence that user interfaces without visual interaction are feasible for audio processing [24, 54, 78, 81].

While auditory interfaces are evidently impractical in context with sound-mixing,

most developments in this area focus on gestural interaction methods [70].

Gestural control means that control signals are extracted from a performance (e.g. from hand movement), whereas the performer does not have physical contact with the instrument. A well-established application of gestural sound mixing is the conducting of a classical orchestra, where the conductor controls the dynamics of the orchestra or instrumental groups within the orchestra with specific hand movements [24]. Examples for an early representative of gestural instruments include the Theremin [60]. A similar technology was used by Martinez et al [60] to control MIDI parameters of a VST plug-in. Drossos et al [24] proposed a prototype gestural interface for real time multitrack stereo mixing, incorporating gestures similar to the hand movements of an orchestra conductor. For this purpose, several sensors, buttons and micro-controllers were used. Subjective evaluation measurements of this experiment have shown that controlling the spatial position of a source via the proposed system is not as satisfactory as controlling the gain of the sources. However, the majority of the test subjects found that the proposed gestural interface enhanced the artistic expression and user experience.

Small-sized portable computers have become extremely popular in recent years [70]. According to Gartner [30] the worldwide combined shipment of new devices like tablets and smartphones is estimated to reach 2.5 billion units in 2015, and they are still forecast to increase in quantity in the near future. Most of them are equipped with touchscreens, thus, leading to new interaction methods with a closer user-content connection compared to the traditional mouse and keyboard interface paradigms of desktop devices [70]. For example, recent multi-touch devices allow adjusting multiple controls simultaneously and individually, in a similar way as it is possible when moving faders on an analogue mixing console [70]. However, gestural interfaces also have some disadvantages: For instance, they are criticised for causing increased fatigue [13] in comparison to the mouse. Furthermore, the touch interaction offers only limited accuracy as touching a glass surface with fingertips does not allow for fine control, especially in conjunction with small user interface elements [70, 77]. Although the desktop computer arguably lacks in ergonomics, the precision and speed of the combination of mouse and keyboard is therefore still considered to be superior for protracted tasks [70].

3.1 Interface Conventions in Digital Audio Workstations

Though the impact of technology on music had changed the way music is performed, recorded and composed quite dramatically, it is surprising then that since the first mixing consoles had been manufactured in the late 1950s, very little has changed in their design. For each channel, the audio travels through a channel strip, consisting of a panning knob, various signal processing units and a fader. This design was later reused in digital mixing consoles and found its way into the virtual mixers of today's Digital Audio Workstations [81].

Due to their increasingly high number of features, current Digital Audio Workstations have become very complex recording systems [32]. Most of them include multi-track recording and playback, editing, mixing, MIDI and signal processing [44]. Additionally, they provide the capabilities of synthesizers, samplers, sequencers and mixers, all in a single environment [20]. The workspace usually consists of an editing window (showing the waveforms of the various sound files on a time axis), transport controls (play-, stop-, forward-, return- buttons), editing tools, menu bars and a configurable, a dynamic virtual mixer with routing and panning options, inserts, sends, equalizers and faders [44]. In this way, the mix is divided into separate channel strips, which may facilitate direct visual comparisons of parameters and reduce the cognitive load involved in navigation, thus leaving more resources for critical listening tasks [69]. However, as the number of channels increases, it is argued that such channel strip representations are getting impractical as soon as the interface exceeds display dimensions and requires scrolling [67, 69].

Additionally, those systems usually provide a software interface that allows third party developers to add code to the DAW. Such programs can be inserted as plug-ins and are designed to emulate signal processors and similar functions [44]. The graphical user interface designs of such audio-plug-ins are in many cases based on their hardware prototypes [54].

Pardo et al [76] argue that complex graphical user interfaces can be especially discouraging for inexperienced users because controls are either modelled on preexist-

ing analog tools or they describe the parameters of the algorithm used to alter the sound instead of describing how sound is actually perceived. Potential users without technical expertise can easily get overcharged by the large range and complexity of parameter settings. This may disrupt the creative process and distract attention from the music itself [84].

Also, the handling of knob-based plug-ins via keyboard and mouse involves some major drawbacks. Firstly, only one parameter can be accessed at the same time, and secondly, hardware faders, knobs and meters are considered to offer better ergonomics and better control [54].

3.2 Influence Of Graphical User Interface Design

Efficient interaction design exits in two directions: Its basis is ultimately formed by how the machine might respond to the user and the user to the machine [83].

Modern information and communication technologies, such as used in DAWs, can support human abilities, but unfortunately some interface designs can also be responsible for distracting the user's attention [90]. Unlimited track counts, multiple effects-plug-ins and the need to navigate through several different windows have resulted in increasingly complex graphical user interfaces, which have been criticized for having potential consequences for the successful mixing of audio [68].

According to Fichtner [28], who investigated the influence of interface design on music composition, the support of today's standard audio software for the creative phase is "provisional and far from optimal". Current DAWs also feature many novel and visually complex audio analysis tools and frequent navigation is necessary in order to access different channels and parameters [68]. It was suggested that the user's attention gets focused on the visual display rather than on aural modalities and that a poorly designed interface can even impede the user's ability to make accurate adjustments [67]. Recent research in human vision has found that merely a few properties of only a few items in a scene can be given attention at any time. Thus, it is important for display design to present information in a way that it is compatible with the limited attentional capacity [83].

In the Gestalt theory, the central idea is that the human visual perception is based on the concept of organisation, a cognizance which is very important in many fields. Some of its principles are directly applicable to user interface design [70]. for instance the well-known Gestalt law of perceptual grouping, and visual attributes like proximity, similarity or closure have to be considered for successful interaction design [49, 70].

There are strategies that help users to keep concentration, such as reducing short-term and working memory load. This can be accomplished by practical display design that provides quick and easy access to information needed in order to make a decision. Generally, any kind of workflow disruptions should be avoided as they can undermine short time memory. Therefore, compact interfaces that reduce scrolling and evade confusing dialog boxes are more effective, even if more information is displayed on the screen. Consistent terminology, layout, color and sequences of action can also help to improve performance. If multiple windows are required, effective coordination is necessary, so that selections in one window produce appropriate results in other windows [90]. Another strategy relies on the assumption that the capability of visual attention can be improved by reducing the amount of items simultaneously shown on a display [83].

Furthermore, it is suggested to minimize the saliency of items which are irrelevant for completing the current task. For instance, the amount of visual clutter can be lowered by grouping elements together on the basis of common alignment. In this way, data can be organized to improve the ease and effectiveness of attentional selection [83].

It was shown that a large amount of visual detail can also influence the perception of auditory information [67]. For instance, Macdonald and Lavie [58] investigated the cross-modal effects of visual perceptual load. In this study, participants made a discrimination concerning line color or line length with subtle differences in both high- and low- visual load conditions. The test has shown that the presence of a simultaneously presented brief pure tone was not noticed by 79 % of the participants in the high-visual-load condition, which is significantly more than in the low- load condition. In a similar study [22], trained pilots failed to notice alarm signals while handling complex graphical user interfaces. Thus, visual information processing

may interfere with the sensitivity for concurrent auditory alarms which is why the authors suggest to temporarily reduce the visual load of such GUIs to redress this problem [67].

Chapter 4

Cross Modal Interaction

The perception of our daily environment involves a strong relationship of information obtained through multiple senses, and the ability to combine such information from the different sensory modalities [99] is an area which has been extensively researched within the recent years [15]. In order to achieve a coherent picture of the external world and to interact with the environment, the human brain combines information from different senses [14].

For example, it was shown that our cross-modal abilities help us to hear and feel more acutely if we simultaneously look in the direction of the non-visual event [15]. We have little difficulties to assign a speech signal to the lip movement of a certain person, even if there are many other persons speaking at the same time [99]. The input of both sensors is then integrated by our cross-modal system in order to improve speech recognition [72].

4.1 Cross-Modality: Vision and Sound

Generally, humans use mainly information obtained from visual and auditory modalities in order to enhance the identification and localization of objects [72]. However, information within the various channels is usually not independent [99]. The present thesis is concerned with the simultaneous presentation of visual and auditory information, thus, it is important to review some findings on cross-modal integration and

the mechanisms related to selective attention between both modalities.

4.1.1 Spatial Domain Effects

Visual information does not only provide the main input from our surrounding environment, but also dominates auditory inputs in the perception of spatial parameters [99, 104]. One common example of this phenomenon is known as the McGurk effect [65]. It describes the interaction of visual and auditory modalities in speech perception and how the visual information of a speaker's lip movement can influence the way a sound is perceived.

Moreover, it has been shown that vision and hearing work together when it comes to localization judgements. For instance, the determination of the horizontal angle of the sound source is more accurate in the presence of vision. On the other hand, auditory spatial information can improve the localization and angular discrimination of a visual target, as well as reducing the reaction time to it [14].

The probably most popular and intuitively accessible example of spatial multimodal interaction is the ventriloquism effect, referring to a skilled ventriloquist who manipulates his or her voice in a way that it appears as it is coming from a dummy. The words are produced without any visible articulation or lip movement, but instead the mouth of the dummy is agitated in synchrony with the speech. The situation is similar to the effect when watching television, where the voices seem to come from the actors on the screen rather than from the actual position of the loudspeakers [2]. In this example, vision dominates what the ear is conveying, and the perceived position of the stimulus is not consistent with the position of actual source [15]. The notion that under certain conditions one sensory modality dominates the others, could be explained by the modality appropriateness hypothesis. It is assumed that the dominant sensory is determined by the highest reliability and suitability for perceptually coding a certain stimulus feature [99, 104]. However, a recent study on the ventriloquism effect by Alain and Burr [2] has shown that in case of distorted or blurred visual input the auditory modality gains more weight, so that the perceived localization can be expressed as bimodal integration. The authors state that "for less blurred stimuli, neither sense dominates and the perception follows

the mean". They state that "the precision of bimodal localization is usually better than either the visual or auditory unimodal presentation", which can be explained by the optimal combination of the information obtained from both senses.

Interestingly, equivocal results have been found about the influence of visual bias on the localization in depth [99]. The auditory distance perception is known to be rather imprecise while the physical distance is typically underestimated, especially for objects further away than approximately three meters [9, 14, 106]. Results of recent psychological experiments [14] support the assumption, that the presence of visual cues influences the auditory perception of sound source distance. Further, it was shown that participants could memorize visual properties of the environment during the experiment, which later improved the perception of distance. Additionally, very accurate responses could be obtained when candidates were allowed to inspect the test room visually before performing the localization tasks in the dark [14]. The authors explain that auditory room-size perception is dependent of the acoustic characteristics of the room (e.g. reverberation time), visual perception is instead based on geometric properties and is therefore more effective in getting information about size, shape and materials. While visual stimuli, which are placed within reachable distance to the observer, can influence the auditory distance perception [10], another study [106] on auditory distant judgements totally failed to observe any bias resulting from simultaneously present visual cues. According to Väjamäe et al [99], the reason for these inconsistencies among several studies may be due to "the different strategies for encoding distant or near, within-the-reach space".

4.1.2 Time Domain Effects

As described above, vision has a relatively high influence on the perception of spatial intermodal relations. In contrast, there is evidence that audition generally dominates when senses interact in time, meaning that the auditory system provides more accurate information about effects in time domain and temporal relations [4, 66, 89]. his phenomenon is often described as *temporal ventriloquism*, analogous to the *spatial ventriloquism* discussed above [66, 99].

Indeed, it was shown that visual perception can be altered by sound. For instance, it was discovered that a single visual flash accompanied by multiple auditory beeps induces the perception of multiple illusory flashes [87]. In another series of experiments [66] , a visual temporal order judgement task was performed, where two flashes were temporally framed by a leading and a trailing sound. Morein-Zamir et al [66] noticed that the sound trailing the second of the two consecutive visual flashes biases the perceived onset of the second light, thus, extending the perceived interval between the events and in further consequence increasing the performance. The effect could only be observed as long as the corresponding stimuli occurred in a time interval of approximately 200ms. This time interval is referred to as *intermodal temporal contiguity window*, during which intermodal events are perceived as being concurrent [99]. The intermodal temporal contiguity window is asymmetric, and studies on synchronization between natural video material and corresponding audio show an easier identification of audio-visual asynchrony when audio is leading up to the video than for delayed audio [87]. It is suggested that the asymmetry reflects an adaption to the physical laws and is caused by the rather low speed of sound and the corresponding physical delay of the audio compared to the visual stimulus [99].

Another example of auditory bias for audiovisual stimuli is the so-called *auditory driving effect*. It describes the mutual influence between the perceived rate of a repetitive sound ('flutter') and the perceived rate of a repetitive flashing light ('flicker') [66, 99]. Interestingly, it was observed that subjects were able to match the rates of two modulated stimuli more precisely for within-modality matches than for cross-sensory matches. Moreover, it was shown that a change in the physical rate of the auditory flutter did influence the perceived rate of the flickering flashing lights [31]. Surprisingly, this effect was not observed the other way round, which is another indication for the existence of temporal ventriloquism. Similar observations were also made with single audio-visual stimuli, where sound can alter the perceived temporal dimensions of the visual event [101].

Guttman et al [34] studied the auditory encoding of visual temporal sequences. Their results show that the "perceptual system automatically and obligatorily abstracts temporal structure from its visual form and represents this structure using an auditory code, resulting in the experience of hearing visual rhythms". In their experiment, subjects were asked to judge whether two successive visual stimuli with

temporally changing contrast followed the same or different rhythms. The visual sequences were presented in conjunction with congruent or incongruent sound and it was found that a mismatched auditory stimulus significantly disrupted the task and thus reduced the performance of the participants, whereas altering the nature of the visual pattern had almost no consequences. Additionally, congruent auditory information improved the performance in comparison to no-sound trials [34, 99].

4.1.3 Other Audio-Visual Effects

While most research on audio-visual interaction is related to spatial- and time domain effects, fewer studies have been performed in context with the perception of loudness, timbre and other musical characteristics.

Schutz and Lipscomb [85] considered the influence of visual information in the context of a musical performance, namely on the auditory perception of marimba stroke types. Identical recordings of a single marimba note were paired with video-recordings of a professional marimba player performing different stroke types (e.g. normal, staccato, legato). It was demonstrated that the subjects were biased by the visual stimulus in a predictable manner. Audio samples paired with a visually short stroke were perceived as 'more staccato' compared to samples paired with a legato stroke, and vice versa. In a similar study [86], the authors found that longer stroke gestures, while not actually producing longer notes on the marimba, resulted in perceptually longer sounding notes through the integration of cross-modal information. The authors suggest that "this finding contradicts previous research showing that audition dominates temporal tasks such as duration judgement".

In order to reveal relations between vision and audition in context of the subjective experience of music, Vines et al [100] investigated the cross-modal integration for perceiving a musical performance of two clarinet players. Participants made constant judgement on tension and phrasing, in both isolated (see or hear the performance) and combined (see and hear the performance) conditions. The visual information served to support the sense of phrasing and additionally conduced to both extend and reduce the sense of tension at different passages throughout the piece. Results of a similar experiment [96] show that the facial expressions of performers influenced

the judgement of emotional valence of sung intervals and that "the visual aspects of music performance are automatically and preattentively registered and integrated with auditory cues".

Studies on the evaluation of noise in an audio-visual context have shown, that the perceived annoyance or loudness of sounds can be reduced by a simultaneous presentation of an additional visual input [26,36]. For example, still pictures were found to induce reductions of perceived loudness on the average by about 2.5%. For moving pictures, the average reduction was about 5% and even larger loudness reductions could be obtained (on the average by about 8%) in case that also the recording equipment for audio and video was in a moving position (for instance a video camera in a car). Additionally, it was shown that the perceived loudness of red objects can be rated 15% higher than the loudness of green objects [36].

4.1.4 Cross-Modal Attention

Selective attention is an area which has been investigated quite extensively within the recent years. While a great part of research is concerned with the selection within a single sensory modality, this chapter summarizes some attention mechanisms related to audio-visual interaction.

In our everyday environment, our attention must be coordinated between different modalities [23]. In such situations it is often assumed that the integration of information obtained from more than one sense can be advantageous [99]. For instance, when following a conversation on a crowded party, the understanding of what is being said can be improved by selectively concentrating on the speaker's voice and integrating both the auditory content of the dialog and the visual information corresponding to the speaker's lip movement, while at the same time neglecting irrelevant information from both modalities [94]. Vice versa, the almost omnidirectional sensitivity of the auditory system can be beneficial in situations when alerting signals come from a direction outside the current view [99]. However, beside the advantages regarding the integration of congruent cross-modal information, costs can be involved when attention needs to be divided between more than one sensory modality and distractions occur [5, 51, 99].

Selective attention is a term that describes the act of focusing on task-relevant incoming sensory stimuli while irrelevant, competing information from other modalities is ignored [93]. Spence and Driver [93] argue, as location is an 'amodal' property which can be conceived by integrating information from different senses, spatial factors can have significant impact on dual-task performances. Generally, multiple senses are involved to produce a stable representation of space. For example, when manipulating an object, visual, aural and tactile information will be received approximately from the same spatial position and the object's position may be available in all senses simultaneously. In a series of experiments on endogenous spatial attention the authors showed, that even when visual and auditory tasks are entirely unrelated, auditory performance in active dual-task conditions is more efficient when the relevant visual and auditory stimuli are presented from the same external location, rather than from different spatial sources. Also, they observed worse performance when attending concurrently to separate task-relevant spaces in hearing and in vision. Thus, dividing attention across auditory and visual modalities seems to be more efficient for streams with a common spatial source. Conversely, they concluded that it is difficult to ignore distractive sounds coming from actively attended visual location. Further, results from spatial cueing experiments conducted by the same authors show that when people choose to deliberately direct their spatial attention in one primary sensory modality on a region of space, the same location is typically attended by other modalities as well, although in an attenuated form. However, it was shown that, although it is difficult, spatial attention in different modalities can to some extend still be directed in different locations [92]. These results illustrate that attentional resources for processing auditory and visual streams are not completely independent.

Unsurprisingly, studies on exogenous (involuntary) attentional orienting processes have found cross-modal links in spatial attention as well. Spatially non-predictive cues (meaning that their location does not provide any information about the location of the subsequent target stimulus) in one modality result in a rapid, unintended shift of spatial attention towards the same location for subsequently presented stimuli in a different modality, even when participants are explicitly instructed to ignore the spatially uninformative initial cue. This cross-modal effect of increased perceptual sensitivity at the cued location occurs even when the modality of the cue is not

related to the participant's task [92, 93].

In another study by Bonnel et al [8], it was shown that identification of a stimulus in one modality impaired the ability to identify a stimulus in a different modality. The authors suggest that these results support the assumption that the trade-off in sensitivity was due to cost of shared attention. Small, yet significant capacity limitations during visual and auditory perceptions have also been detected in an experiment where participants were asked to recognize test tones and test letters under the condition of selected or divided attention [61].

4.1.5 Mixing with Eyes Closed

It is a common opinion among mixing engineers that looking at computer-screen while mixing will affect critical listening skill negatively [17, 18, 41, 52, 56]. The reason behind is that mixing decisions are more likely to be based on the visual representation of music, the value of level- or gain reduction meters or the shape of equalizer-curves instead of what ears are taking in [17, 56]. When taking the visual stimulation away, the conscious mind will focus more on the audio [52]. Indeed, many reputable sound engineers claim that mixing music while not looking at the screen will improve hearing, both by reducing distractions and by allowing more of the brain's processing power to be used on that sense [75].

Lehmann [55] investigated the influence of eyes-closed versus eyes-open listening by collecting EEG data and found that listening to music with eyes closed increased brain activity in all cortical regions. He explains that this may support a more focused listening experience than the eyes-open situation [38].

In order to overcome this problem, it is suggested to relinquish visual cues from time to time during the mixing process and thus, listen to the audio in the same way as the audience will listen to it [17, 18]. Another technique used by some engineers is, while keeping their eyes closed, to bypass and re-activate the audio processing many times until they cannot remember whether the processing is active. Then, by slowly clicking on the bypass button during playback, this method allows the engineer to decide more objectively on whether the audio processing is appropriate or even necessary [56].

Chapter 5

Existing Studies

When mixing audio on a DAW, users need to share their attention between visual and auditory information. Furthermore, it is important to find an optimum balance between the two modalities [68]. A couple of studies exist that investigate the influences of graphical user interface designs in such situations.

Mycroft, Reiss and Stockman [67] found that an interface which requires scrolling has negative impact on the critical listening reaction time. Participants were asked to listen to specified instruments in an eight-channel mix on headphones and to determine whether there was a change in panning. The playback started with all files centred, but one of the specified instruments was continuously moved until it was panned hard left or hard right. Simultaneously to the critical listening task, the participants were asked to match the frequency curve of a target equalizer with a pre-equalized source equalizer, so that both curves were visually as close as possible. This was done using graphical interfaces of different complexity. It was shown that the interface with the least amount of controls had the quickest overall reaction time, while the interface with the highest visible-load (which also required scrolling to access both equalizers, as they did not fit on the screen at the same time) produced significantly slower reaction times. The authors argue that the reason for this may be due to ergonomic issues on the one hand, but on the other hand the visual task may also consume most of the working memory and attentional capacity, leaving less remaining for the other modalities. Furthermore, the fragility of the Short Time Memory may be conflicting with complex graphical user interfaces in creative

support software [67].

In a similar study [68], the same authors investigated whether different User Interface designs influence the coordination of critical listening and interface manipulation tasks. The participants were again required to perform a listening and a visual task at the same time. Firstly, they were asked to identify which of three specified instruments in an eight channel mix underwent a change in volume as soon as they were able to discern it. Again, all files started at full volume, but one instrument was attenuated continuously until it was inaudible, making it easier to detect the effect further into the excerpt. Secondly, in the visual task, 16 channels (represented by four different UI designs: numbers, dials, faders and colors) were presented to the participants. Each channel had four parameters with values from 1 to 16. It was then required to compare channels 2-16 against the first channel and to decide whether they are different or the same, while listening to the audio. Consequently, the time taken to hear the volume changes and the accuracy of the visual comparison were evaluated.

The average time needed in order to detect the attenuation of the audio was almost constant across all interfaces, suggesting that none of the designs distracted attention from the critical listening task. However, the analysis of the visual task revealed significantly less channel matching for UI designs using dials, while a lower error rate was found with faders, numbers and colors. The reason why the values of the latter controls could be compared more successfully may be imputable to visual perception. The human eye is rather capable of comparing differences in two dimensional locations and differences in line length, but is less precise in comparing angles [27]. Therefore, the authors suggest that dials may do not allow quick access to mix information. In this respect, faders and numbers are performing well but there is a high risk that they get less effective at lower zoom levels. Colors, on the other hand, work well at small views and can also convey quantitative information [68].

In a similar study [69], participants were asked to detect visual information on a 24-channel digital mixing interface with and without overviews, while identifying changes in the panning of different audio tracks in a mix. The visual task and the listening task had to be performed for three different interfaces. The designs consisted of a single-page channel strip mixer (no scrolling required to access all

channels), a stage metaphor (where the mix is broken down into the dimensions width, depth and height) also presented on one page, and a scrolling channel strip mixing interface. The analysis of the results revealed that overview designs did not only increase the amount of correctly identified panning positions in comparison to the scrolling interface, but also improved the efficiency of the visual search task.

Another thesis [105] investigates whether active VU-metering (averaging loudness meter [45]) while adjusting the gain reduction of a compressor plug-in influences mixing decisions or not. For this purpose, participants were asked to apply fixed compression to a lead vocal track in order to tame its dynamic range and thus integrate it into pre-mixed arrangements. More specifically, they were required to complete the task by manipulating only a single control, which was the compressor's gain reduction knob. Additionally, they were allowed to adjust the vocal-track fader in order to set the track's overall volume. Two different songs were used as stimuli (with their related lead vocals respectively). In one case (Song 1), vocals were recorded without any processing, thus, showing obvious dynamic problems [105].

For each song, test subjects were asked to complete the task twice, both with and without a visual VU-Meter. For the former case, all other parts of the display except the meter were covered with cardboard, and otherwise the screen was simply turned off (for the "VU-off" condition). Further, the gain reduction knob of the compressor was assigned to a programmable rotary encoder of a hardware controller [105].

The most prominent observation of this experiment was that in Song 1 subjects applied more gain reduction to the vocal track when the VU-meter was not visible. However, in case of the other song containing the pre-compressed vocals, no statistical significance was found. The author indicates that the lack of obvious dynamic problems may have made it harder to focus on the sound in this example, whereas in the first song the compression had a clear purpose and the effect was easier to hear. It is also worth mentioning that the settings for gain reduction were more consistent among the participants in the first example, which may also be related to the above consideration. Another interesting observation of this study is that the visual metering seemed to have diminished the variance of the gain reduction settings. Again, no significant statistical differences could be found regarding the time duration in order to complete the task between both conditions. Finally, the

author suggests to further investigate the influence of metering on mixing decisions in future studies, for example by using a purposely wrongly calibrated meter [105].

Lech and Kostek [54] engineered a novel gesture-based mixing interface and investigated if this novel approach of DAW controlling improved the ergonomics in comparison to traditional mouse and keyboard interfaces. Furthermore, with this new approach they wanted to assess the audio information visualization on mixing results. The mixing interface recognizes hand gestures in a video stream and can be operated with and without visual support. Again, the task for the participants (which consisted of ten experienced sound mixing engineers) was to mix eight audio channels to their own preference. They were asked to keep a fixed practice for all of the mixing methods, which included five different scenarios: Mixing via engineered interface and gesture control (both with and without visual support), mixing via engineered interface, but controlled with mouse and keyboard (both with and without visual support) and traditional mixing on a music production system with mouse, keyboard and MIDI controllers. Following this, the engineers were asked to fill in a questionnaire regarding the quality of precision, convenience and intuitiveness of the individual systems. They were also asked to rate their own mixes from best to worst [54].

The analysis of collected data revealed, that in six of ten cases the visualization of audio parameters resulted in a broader panorama and a more extensive use of shelving filters. Without support of displayed parameter values, participants seemed to concentrate more on the sound balance instead of what looked balanced in the visualizations. This was confirmed by nine out of ten engineers, who stated in the questionnaire that it was easier to focus on the sound when visual support was limited [54].

5.1 Summary of Literature Review

Mixing and mastering involves both technical and creative aspects, such as editing, restoration, equalizing, level adjustments or adding effects like reverb. As current Digital Audio Workstations provide the tools and flexibility to revise any setting at any point, mixing can be described as an iterative process which requires a lot of

concentration. Therefore, it is important to avoid any distractions in order to be able to maintain the creative flow.

However, a number of studies confirmed that complex graphical user interfaces, as found in modern music production systems, may distract attention from the sound itself. It was shown that certain interface designs, in particular interfaces with dials or interfaces requiring scrolling, had negative impact on critical listening. Generally, audio engineers seem to be able to concentrate more on the sound without displayed audio-parameter values. Furthermore, active VU-metering did affect mixing decisions in a task were participants were asked to apply compression to a vocal track.

All these findings indicate that the presence of visual stimuli can influence the perception of auditory information during sound mixing. Indeed, research has shown that information received simultaneously from different sensory modalities is usually not independent. Thus, the presence of visual information may give rise to cross-modal interaction between the auditory and visual modalities.

The literature review raises a number of questions. Firstly, as mixing and mastering usually involves a vast amount of different processes, it would be interesting to see if the results of former studies can be confirmed also for mixing tasks other than compression. Secondly, no attempts were made to quantify the degree of visual influence. At present - and to best of the author's knowledge - there is very limited information about whether the impact of visual stimuli is stronger for some tasks or test signals than for others. Further, this studies are the first to deliberately shift visual cues in order to provide evidence for and quantify the impact of visual information. Lastly, it seems questionable whether the observations are attributable solely to attention mechanisms or other cross-modal effects.

Chapter 6

Experiment 1: Loudness Matching

Digital audio mastering usually ensures consistent loudness, balance and sound across all tracks of an album [46]. Accordingly, the first task of the experiment was concerned with loudness matching between different sets of stereo audio mixes. For each pair of mixes, the subjects were asked to match the perceived loudness of the two tracks by adjusting the gain of one of the mixes.

6.1 Methodology

The whole experiment was divided into three conditions (performed in random order), and for each condition the participants were asked to perform the loudness matching task several times for three different stimuli pairs. For *Condition 1*, the computer screen was turned off and the task had to be completed without visual feedback. Conversely, for *Condition 2*, the headphones were turned off and the subjects were required to adjust the volume according to the visual display of two loudness meters. Finally, *Condition 3* provided both the aural and the visual feedback. However, one of the displayed loudness meters was purposely manipulated by a varying offset. This offset was either zero, $\pm 0.75\ dB$, $\pm 1.25\ dB$ or $\pm 1.75\ dB$, meaning that the meter was showing an incorrect loudness value in six out of seven cases. A detailed description of the individual conditions is provided below.

6.1.1 Setup

The experiment was performed on a digital audio workstation (*Toshiba Satellite* laptop computer, Intel Core i5, 8 GB RAM) in a lab at the *IEM* Graz. In order to provide a realistic, familiar working environment, the audio files were arranged on two different tracks within the digital audio workstation *Cockos Reaper*, running on a laptop computer. The tracks were named *Test* and *Reference* respectively.

Figure 6.1: Experiment Setup; Behringer X-Touch Mini, AKG K712 Headphones, computer screen showing the interface of two PPMulator Plus loudness meter plug-ins.

The subjects were asked to perform the loudness matching task by using a simple hardware rotary control, which output was connected to a VST gain plug-in (*Blue Cat Audio Gain Suite*) inserted on the *Test* track in *Reaper*. Consequently, when moving the rotary control on the midi hardware controller (*Behringer X-Touch Mini*), the transmitted midi data was converted into a corresponding increase or decrease of the plug-in's gain value. In order to avoid the restrictions of the 7- bit MIDI messages and - as a consequence thereof - the limited resolution of only 128 possible values, the controller was set to mode 'relative 2'. In that mode, the currently adjusted gain value of the plug-in is increased or decreased depending on direction and the velocity of the controller movement. The range of the gain plug-in was set to -30dB to +30dB. In order to prevent the audio from clipping, the faders of the Test track and the reference track were set to -10 dBFS.

By using the *OSC* protocol, it was possible to receive and process OSC messages sent from *Reaper* in *Pure Data*. Thus, important data, such as fader levels, time required, or the current value of the gain plug-in could be directly saved into a text

file for later evaluation. Similarly, OSC and MIDI messages could also be sent from *Pure Data* to *Reaper*.

Furthermore, the following commands were assigned to certain buttons and encoders of the Midi Controller:

- **Gain/Volume (Rotary Encoder):** Applies a gain change (range from -30 dB to +30 dB) to the *Test track*
- **Solo Test Track (Button):** Solos the *Test* track (and mutes the *Reference* track)
- **Solo Reference track (Button):** Solos the *Reference* track (and mutes the *Test* track)
- **Start (Button):** Starts the individual subtask/condition, and activates the stop watch implemented in *Pure Data*. It also reads parameters of the first sample from a list (this lists were generated with *Matlab* prior to the experiment). These lists are:
  - *Initial Gain*: Random numbers between [−6 dB, −2 dB] and [+2 dB and +6 dB], controlling the initial value of the gain plug-in inserted on the Test track. A different, random value was loaded for every repetition, and positive and negative gains were counterbalanced.
  - *Stimuli*: This random sequence of numbers determines which stimuli pair is played back next; 1 = Pink Noise, 2 = Pop A/Pop B, 3 = Pop A/Classic.
  - *Offset*: A random sequence of numbers in dB (Please refer to Section 6.2.3 for more information)
  - The first three repetitions for each condition were reserved for training (they are neglected in the evaluation of results). These were the first three values in the corresponding lists:
    - Initial gain: 5 dB, −3 dB, −6 dB.
    - Stimuli: 1, 2, 3 (Pink-Noise, PopA/PopB, Classic/PopA; in order to help subjects to become familiar with the different audio examples).

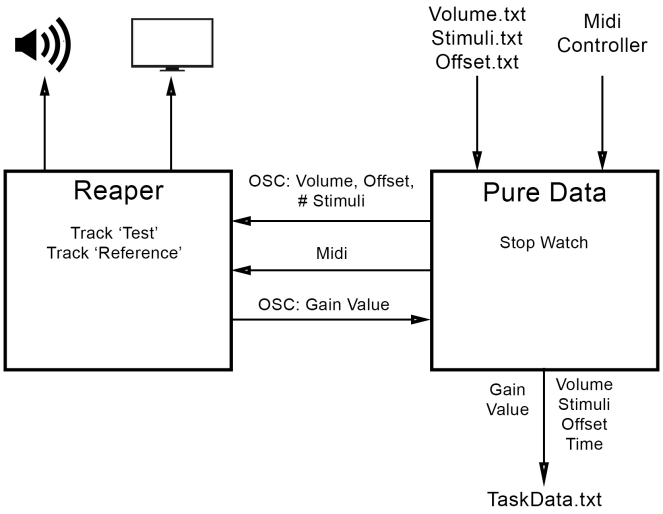


Figure 6.2: Routing and general setup in Experiment 1

  - Offset: 0,0,0 (no offset)

- **Pause (Button):** Stops the stop watch and the audio playback, in case the subject needs a break.
- **Resume (Button):** Resumes the stop watch
- **Play/Stop (Button):** Plays and stops the audio playback
- **Next (Button):** Saves the current parameters to a text file (initial gain, user gain, offset, stimuli pair, time required), reads the next parameters from the lists and jumps to the start of the next stimuli-pair (according to the entry in the Stimuli list) within the *Reaper* arrangement window.

The audio interface used for the experiment was a *RME Fireface UC* and the subjects were provided with *AKG K712* Headphones for monitoring.

The test subjects were instructed to apply a fixed gain value for the entire song, meaning that they were not allowed to make use of any automation. Moreover, they were notified that there was no time limit within which the task must be completed.

6.1.2 Stimuli

The following three different stimuli-pairs are used for the experiment:

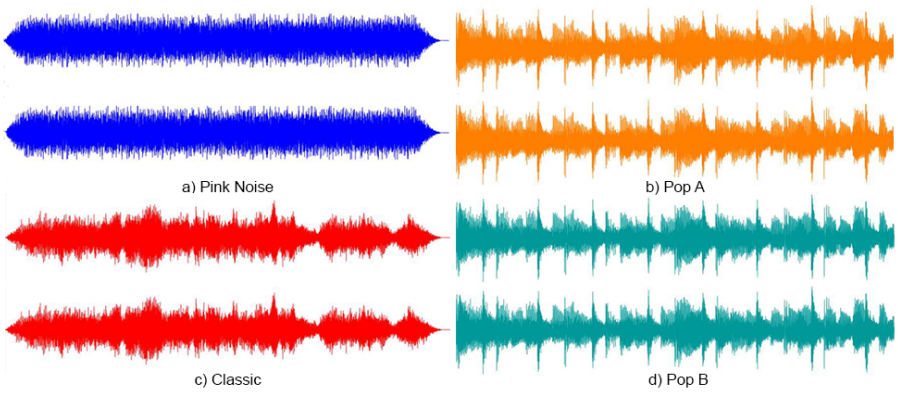


Figure 6.3: Test signals, waveforms

(a) **Pink Noise:** Two excerpts of stationary pink noise. The amplitude envelope of these files is very straight and the clips sound the same.

(b) **Pop A / Pop B:** Two different audio mixes created with commercial, pre-produced Pop/Pop-Rock audio loops. The excerpts feature bass guitar, acoustic guitar and piano, accompanied by a basic drum rhythm. While the overall sound and dynamics of the mixes are very similar, the instruments play slightly different patterns, making it more difficult to equal the loudness of the tracks than in the case of pink noise.

(c) **Pop A / Classic:** The third stimuli-pair consists of the Pop A file from the previous pair, but this time it is coupled with an extract of an orchestral performance. The instrumentation, tempo and musical intention of both tracks is completely different, thus, the loudness matching task is again exacerbated.

All files were 16 bit, 44.1 kHz stereo wav files and were normalized to an integrated program loudness of -18.5 LUFS (in accordance with the EBU-R128 standard).

6.2 Participants

Twelve participants with mixed age (19-35 years, 1 female, 11 male), all with prior experience in sound mixing and using a Digital Audio Workstation, were recruited from students at the Graz University of Technology. Eleven participants studied Audio Engineering and one studied Electrical Engineering. They were provided with

detailed instructions in written form. None of the subjects were paid for participating in the experiment and all of them participated voluntarily.

6.2.1 Procedure for Condition 1 - Auditory

For *Condition 1*, the computer screen was turned off and thus the subjects were required to adjust the gain control on the external midi-controller without any visual feedback. The LEDs on the midi controller were covered in order to prevent subjects from remembering their settings.

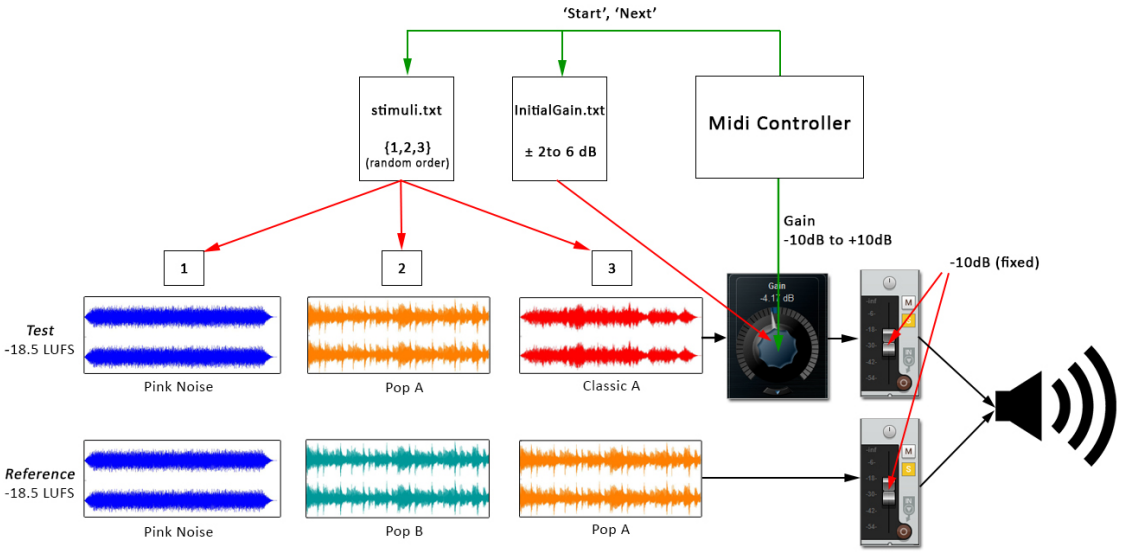


Figure 6.4: Schematic illustration of the experimental design for the unimodal aural condition

By pressing the *Start* button, the timer was activated and the first set of parameters was loaded from the prepared *.txt* files (a value between ±2 to ±6 dB for the initial gain of the gain plug-in on the *Test* track, and an integer between 1 and 3 which determined which stimuli-pair was played back next). The lists contained 12 values, meaning that each subject performed the loudness matching task 12 times in total (4 times per stimuli-pair) for this condition. The first three repetitions of the task were reserved for training and were not evaluated. The audio playback could be started or stopped by pressing the play/stop button. By using the *Test* and *Reference* buttons, the subjects could alternately solo the individual tracks. A rotary control labeled with gain was assigned the gain value (from -30 dB to +30 dB) of the VST plug-in inserted on the *Test* track. All track-faders were set to -10dB. The stimuli were

played back in a loop from start to finish until the *Next* button was pressed. Then, the current value of the gain plug-in was stored in a file, the next set of parameters was loaded and the locator was set to the beginning of the next readout stimuli-pair. The previous steps were repeated until the participant performed loudness matching for all of the 12 examples.

6.2.2 Procedure for Condition 2 - Visual

In *Condition 2*, the computer screen was turned on, but the subjects did not hear any sound. Therefore, they were asked to complete the task solely by means of visual feedback. The screen showed the graphical user interfaces of two loudness meter plug-ins with identical settings. One of them showed the (post-fader) loudness of the *Test* track, and the other one showed the (post-fader) loudness of the *Reference* track. The meters were arranged one below the other on the screen such that their current values were easily comparable.

The loudness meter plug-in used was the *PPMulator Plus* by *Zplane*. This plug-in offers several different modes and graphical representations. It was set to the preset 'EBU- R128 simple' with short term (time window of three seconds) metering. The subjects were not allowed to change these settings.

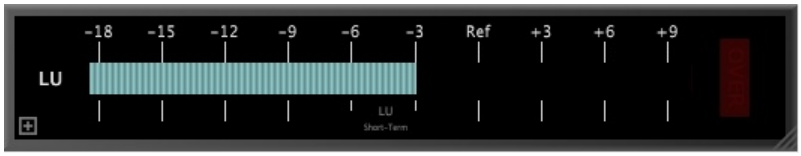


Figure 6.5: Graphical user interface of the PPMultor Plus plug-in by zplane in EBU-R128 mode with short time metering

It can be assumed that the difficulty of the visual loudness matching is dependent on the amount and speed of the fluctuations of the animated metering bars. For audio sources with flat envelope, like pink noise, the metering bar is almost static, while it moves moderately for more dynamic tracks, like it is the case for the Pop tracks or even more for the Classic track.

Except for the modified visual and aural conditions, the procedure and overall setup for this task was the same as in *Condition 1*. Again, the subjects were asked to repeat

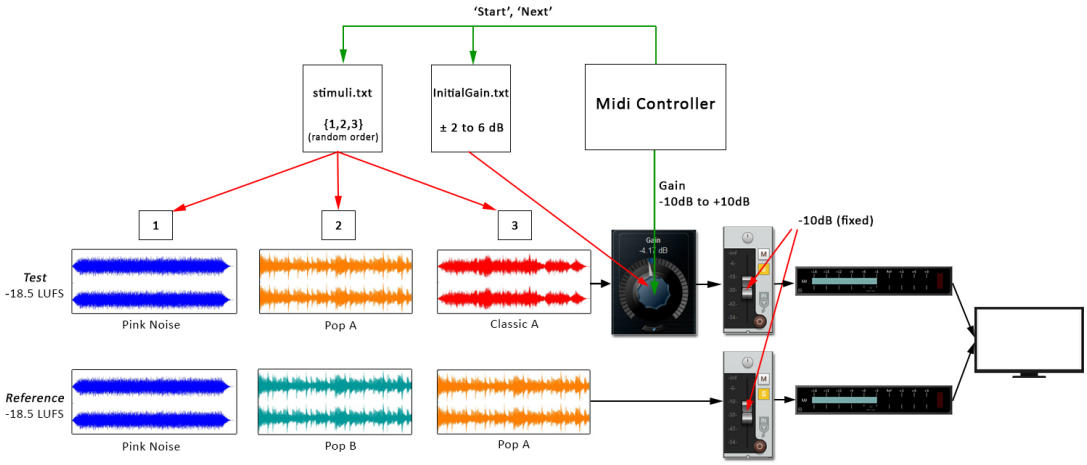


Figure 6.6: Schematic illustration of the experimental design for the unimodal visual condition

the loudness matching 12 times (three trials for training plus three repetitions for each of the three stimuli-pairs in random order).

6.2.3 Procedure for Condition 3 - Audio-Visual

Condition 3 combined the aural feedback from *Condition 1* and the visual feedback from *Condition 2*, meaning that the subjects could monitor their adjustments via headphones and via the loudness meters on the screen. However, in most cases, the readout of the meter on the *Test* track was purposely distorted by a certain degree of offset in the range of -1.75 dB to +1.75 dB. The exact possible values for the offset were -1.75 dB, -1.25 dB, -0.75 dB, 0 dB (no offset), 0.75 dB, 1.25 dB, and 1.75 dB, meaning that the meter displayed a false value in six out of seven cases.

Every value occurred three times for each stimuli-pair, leading to a total of 66 repetitions of the loudness matching task (including the three trials for training at the beginning). The sequence of values for the offset was randomized and different for every subject. It was stored in a text file and the offsets are sequentially loaded into *Pure Data* and added to the current value of the gain plug-in. The result was then sent via OSC to another gain plug-in inserted on a copy of the *Test* track. This track also held the manipulated metering plug-in (post fader), but its audio output was muted. Figure 6.7 shows the setup and track routing for *Condition 3*.

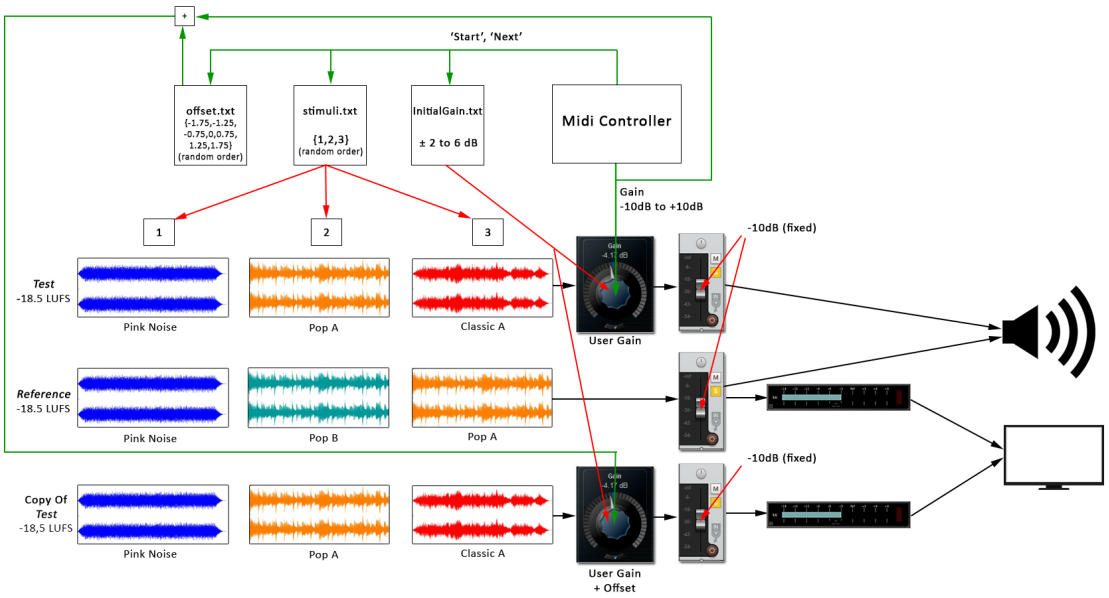


Figure 6.7: Schematic illustration of the experimental design for the bi-modal audiovisual condition.

The test subjects were not informed about the manipulated loudness meter as the goal of this task was to investigate and quantify the impact of the visual feedback on the mixing decisions of the participants. The values for the offset have been chosen carefully in order to be small enough to not engender mistrust [1], but large enough so that the differences were aurally and visually perceptible. The results of *Condition 3* could then be compared with the results from the unimodal conditions.

6.2.4 Summary: Loudness Matching Experimental Design

The experiment was structured into three conditions involving either sight or hearing, or both modalities at the same time. Twelve participating subjects were asked to match the loudness of different types test signals in a pairwise comparison by modifying the gain parameter of one of the tracks. Stimuli-pairs consisted of pink noise, two similar Pop mixes and orchestral music, which was again compared against one of the Pop mixes. In case participants were provided with visual feedback, two loudness meters were displayed on the screen. However, the meters were distorted by various degrees of offset. The task was repeated three times per subject for every

[1] During evaluation, it became apparent that the ±1.75 offsets seemed to have caused suspicion among subjects in case pink noise test signals were used

experimental condition and stimuli-pair. The key features of the loudness matching experiment are summarized in Table 6.1.

Condition	Controllable Gain	Test & Reference Track Fader Volume	Phones	Meter On Test Track	Meter on Reference Track	Repetitions
Aural	Controllable from -10 dB to 10 dB	Fader fixed at -10dB	On	Not visible	Not visible	$3 \times 3 + 3 = 12$ 3 stimuli a),b),c) á 3 repetitions + 3 repetitions for training
Visual	Controllable from -10 dB to 10 dB	Fader fixed at -10dB	Off	visible	visible	$3 \times 3 + 3 = 12$ 3 stimuli a),b),c) á 3 repetitions + 3 repetitions for training
Audio-Visual	Controllable from -10 dB to 10 dB	Fader fixed at -10dB	On	visible	visible	$3 \times 3 \times 7 + 3 = 66$ 3 stimuli a),b),c) á 3 repetitions for 7 different offsets + 3 repetitions for training

Table 6.1: Illustration of key parameters for the three conditions in Experiment 1

6.3 Results: Loudness Matching

In the following chapter, the results of the first experiment are discussed. In this section, data from the unbiased and biased conditions of the loudness experiment is evaluated. Later, in Section 7.3, the same procedure is applied to the results of the spectral matching task.

6.3.1 Hypotheses

Based on the previously provided literature review and personal observations during pilot experiments, it is assumed that the following main effects can be observed by evaluating data from the loudness matching experiment:

1. It is easier to match the loudness of two identical or similar signals than it is for signals which are completely different from each other.

2. The presence of visual cues affects the mixing decisions of the participants compared to situations in which no visual cues are provided. Thus, for the

same type of stimuli-pairs, gain settings related to unimodal auditory feedback are different compared to conditions in which visual cues are provided, even if there is no attention paid to the visual display.

3. Gain settings are significantly affected by visual cues. They are dependent on the degree of distortion applied to the displayed loudness meters and on which signals are being matched.

4. The more difficult it is to perform the task by ear, the more participants rely on the display.

   These hypotheses and other side effects will be further investigated in the following sections.

6.3.2 Data Preparation

Firstly, every three trials per subject, condition, stimuli-pair and offset were pooled to a single value, representing the mean of those three repetitions. In this way, the total 36 responses of all twelve participants per condition, stimulus and offset were reduced to twelve responses, whereby each value indicates the average setting per person. All of these pooled responses were separated into two main data groups, the unbiased data group and the biased data group. Each of these two data sets contain a number of subgroups á 12 data samples. The unbiased data group contains all subgroups where no offset was applied. In contrast, the biased data group is essentially composed of all subgroups in which subjects were potentially affected by the manipulated metering display. Therefore, it solely contains data from the audio-visual condition grouped by stimulus and offset. It also includes the audio-visual zero offset condition, allowing to compare biased responses against zero-offset responses. Table 6.2 illustrates the two data groups.

In order to be able to draw conclusions about significance when comparing results of the experiment, an appropriate statistical analysis technique had to be chosen. Some of the common statistical methods, such as the *Analysis of Variance* (ANOVA), require data to meet certain assumptions that relate to the distribution and variance of the data.

	Unbiased Group	Biased Group
Conditions	auditory, visual, audio-visual	audio-visual
Stimuli-Pairs	Noise/Noise, PopA/PopB, PopA/Classic	Noise/Noise, PopA/PopB, PopA/Classic
Offsets	Only zero offset	All 7 offsets
Number Of Subgroups	3 conditions x 3 stimuli-pairs x 1 Offset = **9**	1 condition x 3 stimuli-pairs x 7 Offsets = **21**

Table 6.2: Two main data groups, consisting of 9 and 21 subgroups á 12 responses respectively

The loudness matching experiment design required a statistical analysis method to investigate the effect of the two independent variables *Condition* and *Stimuli-Pair* on the dependent variable *Gain Setting*. As each subject was tested with all stimuli-pairs under all conditions, a repeated-measure two-way ANOVA (also referred to as within-within subjects ANOVA) could be used to examine whether gain settings were affected by test-condition or test signal and if there was any two-way interaction between these factors.

However, the repeated-measure ANOVA only produces valid results if the data is normally distributed and meets the assumption of sphericity (that is, the variance of difference between within-subject factor levels should be equal) [47].

6.3.3 Evaluation of Unbiased Data

The mean error and the standard error of the mean for all data collected in unbiased conditions, id est the unimodal auditory, unimodal visual and multimodal audio-visual no-offset conditions are illustrated in Figure 6.8. However, it is important to notice that the mean error describes the difference of the user settings compared to what the EBU-R128 loudness algorithm considered to be equal loudness. For example, it can be seen that the mean error is almost zero dB for pink noise judgements, regardless under which condition the task was performed. This illustrates that the subjective perceived equal loudness for pink noise is consistent with the EBU-R128 loudness model.

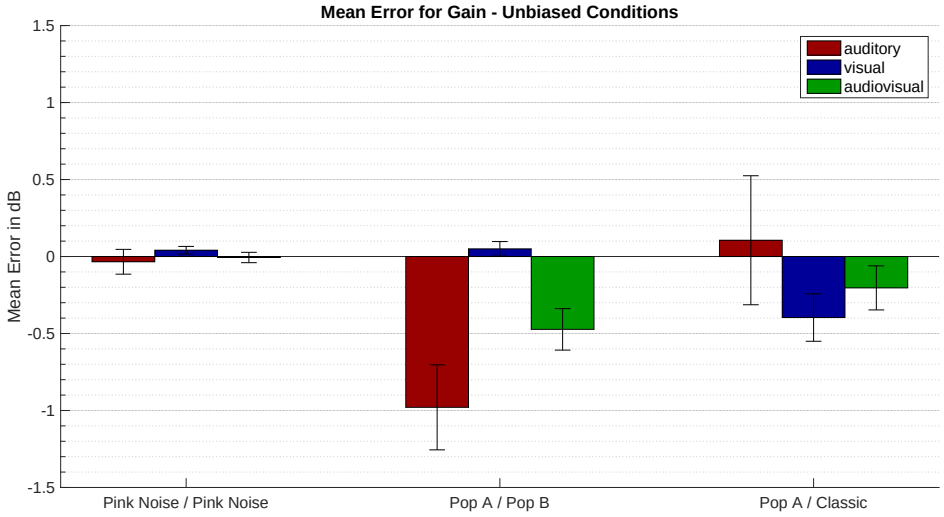


Figure 6.8: Evaluation of responses collected under unbiased conditions. Errorbars represent the standard error of them mean.

On the other hand, it appears that the most extreme deviations occurred for the Pop A - Pop B stimuli-pair, where participants judged the Pop A mix to be around 1 dB quieter than the Pop B excerpt, whereas the algorithm considered them to be of equal loudness. Referring to the results illustrated in Figure 6.8, it is evident that the responses for the pinknoise stimuli-pair were more consistent across all conditions than for the other stimuli-pairs. Moreover, it can be seen that the mean value derived from the audio-visual condition is always approximately centred in between the mean values of the unimodal conditions.

Another interesting observation is that, in contrast to the other conditions, it appears to be difficult to match the Pop mix and the classical mix visually. It seems that in those cases, participants have set the gain value slightly too low, even though the display of the meter was their only reference.

In order to examine the influence of the two independent variables *Condition* and *Stimuli-Pair*, a within-within subjects ANOVA on gain settings was performed. Shapiro-Wilk's test of normality confirmed the null- hypothesis of normally distributed data at a 5% significance level, which shows that in this case the repeated-measurement ANOVA is a valid statistical method. Also, no outliers were detected, as assessed by inspection of a boxplot for values greater than 1.5 box-lengths from the edge of the box. However, the Mauchly's test of sphericity [63]. indi-

cated that the assumption of sphericity was not met for the two-way interaction ($\chi^2(2) = 32,407, p < 0.001$), therefore a Greenhouse Geisser ($\varepsilon = 0.399$) correction [33] was used.

The results of the two-way repeated measures ANOVA are presented in the table below:

Main effects & interactions		**F-Value**	**Sig.**
Condition	$F(2, 22) =$	1.533	$p = 0.242$
Stimuli-pair	$F(2, 22) =$	3.493	$p = 0.063$
Condition x Stimulus	$F(1.59, 15.55) =$	4.612	$p = 0.031$

Table 6.3: Results of the 2-way repeated measures ANOVA

There was a statistically significant two-way interaction between condition and Stimuli-pair on the gain setting. Therefore, simple main effects were run.

Data are mean ± standard deviation, unless otherwise stated. The mean gain settings for the pink noise / pink noise and the Pop A / Classic stimuli pairs were not statistically significantly different across all conditions ($p = 1$, $p = 0.335$), as assessed by pairwise comparisons with Bonferroni adjustment. However, there was a statistically significant difference of -1.03 dB (95% CI, -1.81 to -0.24) between the auditory condition (-0.98 ± 0.96 dB) and the visual condition (0.05 ± 0.16 dB), $p = 0.011$, and the visual condition and the audio-visual condition (-0.47 ± 0.47 dB) for the Pop A / Pop B stimuli pair, $p = 0.008$.

It can further be observed, that the visual loudness matching of Pop A mix and the Classical mix was not as accurate compared to the other two stimuli pairs. According to the statistical analysis, this effect was marginally significant between the Pop A / Classic pair and the Pop A /Pop B pair ($p = 0.066$) and statistically significant between the Pop A / Classic pair and the pink noise pair ($p = 0.037$).

6.3.4 Evaluation of Biased Data

The mean gain settings and corresponding standard errors are shown in Figure 6.9 with respect to different offsets of the loudness meter, for all three stimuli pairs

respectively. The unbiased responses from the auditory and visual conditions are included in the graphs as well.

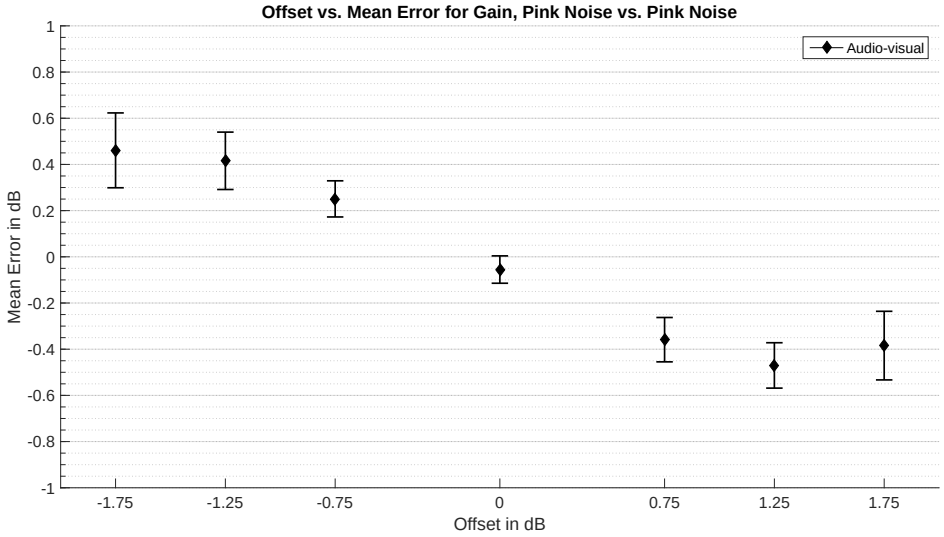
In general, it can be observed that positive offsets induced negative gain settings, and negative offsets induced positive gain settings. In fact, if subjects matched the loudness of the signals solely according to the display of the meters, the actual gain setting in order to make the visual outputs coincide as good as possible would be approximately the inverse of the offset. Thus, it is evident that the subjects were indeed influenced by the visual display, at least to a certain degree. If participants would have ignored the visual mismatch, all responses would have been approximately equal to the result of the audio-visual zero-offset condition. Three within-subjects ANOVAs on mean gain value, one per stimuli-pair, were performed on the basis of data from the *Unbiased Data* group.

Interestingly, all graphs appear to be symmetric with respect to the audio-visual zero-offset response, thus, the effect induced by the offset seems to be independent from sign.

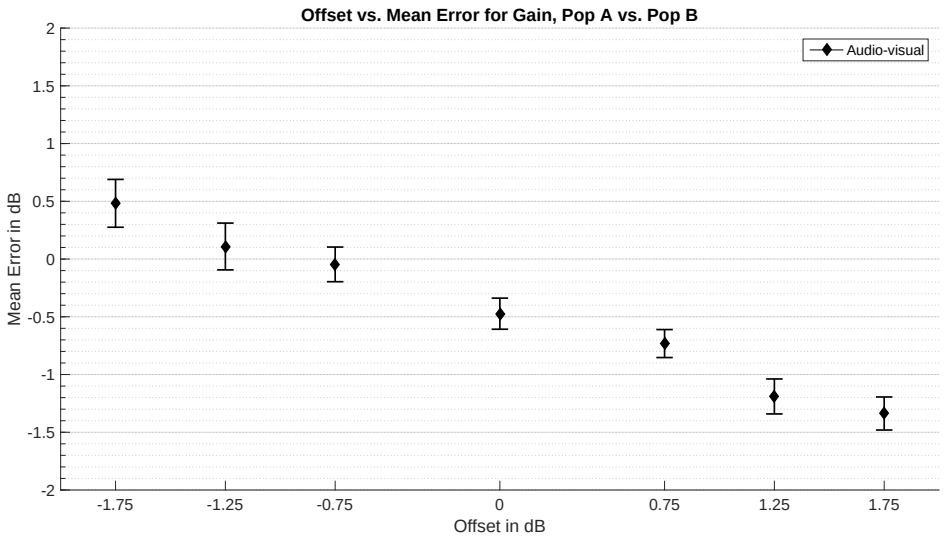
Comparisons of the three graphs reveal that the graph related to the Pink-Noise judgements shows a saturation effect for the extreme offsets of ± 1.75 dB. This effect cannot be observed for trials involving musical signals, as values are arranged with a more or less constant slope. The offsets provoked a maximum difference in gain settings of 0.93 dB for stimulus-pair 1 (Pink-Noise / Pink-Noise), 1.82 dB for stimulus-pair 2 (Pop A / Pop B) and 1.998 dB for stimulus-pair 2. It can be argued, that by dividing these values by the maximum offset range of 3.5 dB, the mean gain settings were spread out to a maximum of 26.5%, 52% and 57% of the maximum offset range, respectively. However, this source-signal dependent effect will be analysed in more detail in section 6.3.5, *A Characteristic Measure for Visual Influence*.

Statistical Analysis - Pink Noise / Pink Noise pair

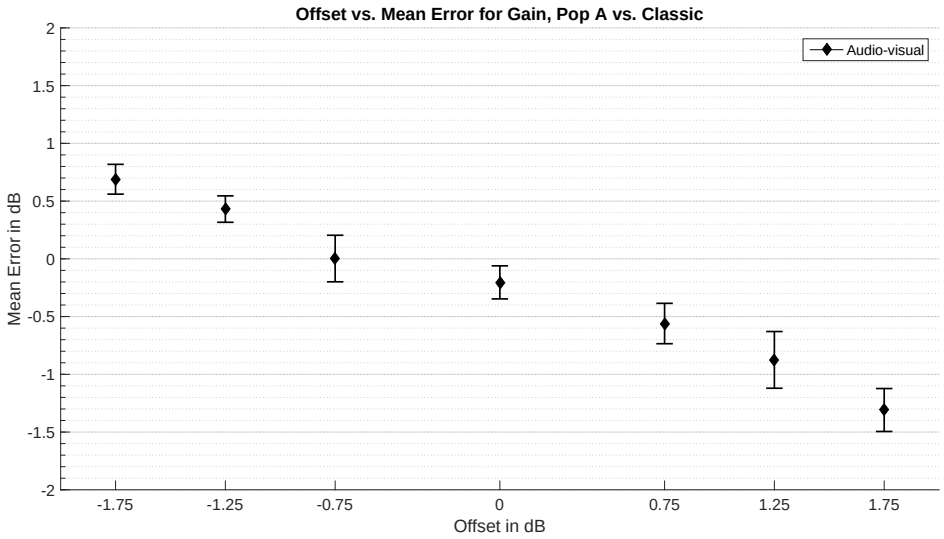
A Shapiro-Wilk test of normality indicated that 57% of *Biased Data* subgroups related to the pink noise stimuli-pair did not meet the assumption of normally distributed data. In addition, there were five outliers in this particular data set, as


Offset vs. Mean Error for Gain, Pink Noise vs. Pink Noise
Audio-visual
Mean Error in dB
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.75
-1.25
-0.75
0
0.75
1.25
1.75
Offset in dB

(a)


Offset vs. Mean Error for Gain, Pop A vs. Pop B
Audio-visual
Mean Error in dB
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1.75
-1.25
-0.75
0
0.75
1.25
1.75
Offset in dB

(b)



(c)

Figure 6.9: Mean gain settings related to the degree of (visual) offset for different stimuli pairs: [Pink Noise, Pink Noise] (a), [Pop A, Pop B] (b) and [Pop A, Classic] (c). Errorbars represent the standard error of the mean.

assessed by inspection of a boxplot. Therefore, a Friedman Test, which is robust against outliers, which does not assume normally distributed data and which thus is often used as an alternative method to the repeated-measurement ANOVA [95], was performed on gain settings. The results of the test for the pink-noise data are presented below.

Gain settings, applied to a pink-noise stimulus in order to match its loudness with that of a reference pink-noise signal in a pairwise comparison, were statistically significantly different for the different offsets, $\chi^2(2) = 47.143, p < .0005$. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. Gain settings were statistically significantly different between the following offsets:

Offsets	**-0.75 dB**	**-1.25 dB**	**-1.75 dB**
-0.75 dB	$p = 0.020$	$p = 0.003$	$p = 0.007$
-1.25 dB	$p = 0.001$	$p < 0.0005$	$p < 0.0005$
-1.75 dB	$p = 0.014$	$p = 0.002$	$p = 0.005$

Table 6.4: Pink-Noise vs. Pink-Noise loudness matching: Multiple comparisons revealed statistically significantly different gain settings between positive and negative offsets

Summarizing the results from the table above, it can be discerned, that gain settings

are statistically significantly different for all positive offsets compared to all negative offsets, and vice versa. Thus, for every degree of offset, there are significantly different results for 3 other levels of visual bias. As shown in Figure 6.9a, this behaviour can be traced back to the flattening of the curve at large offsets. Offset of ± 1.75 dB did not further increase the error.

Analysis - Pop A / Pop B pair

A normal distribution for datasets related to the other two stimuli-pairs was confirmed through a Shapiro-Wilk Test. There were no outliers in this datasets, as assessed by inspection of a boxplot. The evaluation of responses in conjunction with the Pop A / Pop B pair and the Pop A / Classic pair was therefore carried out with a within-subjects ANOVA, respectively. Mauchly's test of sphericity suggested that the assumption of sphericity had been violated, $\chi^2(2) = 32.582, p = 0.048$. Epsilon ($\varepsilon = 0.414$), as calculated according to Greenhouse & Geisser and was used to correct the results from the within-subjects ANOVA. In a pairwise comparison, where two different Pop Mixes stimuli had to be loudness matched, gain settings were statistically significantly different when various degrees of offsets were applied to one of the two displayed loudness meters, $F(2.482, 27.297) = 41.235, p < 0.0005$, partial $\eta^2 = 0.789$.

It can be observed, that for every distinct degree of meter distortion, there were on average 3.71 other offset conditions where subjects adjusted their gain levels statistically significantly differently. However, in most cases the gain settings were not statistically significantly higher or lower for neighbouring offsets. For instance, gain settings were not statistically significantly higher for the zero offset condition compared to the +0.75 dB offset condition, but they were statistically significantly higher compared to the +1.25 dB condition. Post hoc tests with Bonferroni adjustments elicited that gain settings resulting from a -1.75 dB and -0.75 dB offsets were statistically significantly higher compared to gain settings resulting from zero offset and positive offsets ($p < 0.002$). Further, a properly working meter (zero offset) also eventuated in significantly higher gain settings compared to offsets of +1.25 dB and +1.75 dB ($p < 0.001$). Likewise, participants rated signals louder (0.109 ± 0.2 dB) that the reference signal when the offset was -1.25 dB, and quieter than the

Gain settings, induced by a loudness meter distortion of ...	**... are statistically significantly different ($p < 0.05$) compared to gain settings induced by loudness meter distortions of ...**
-1.75 dB	≥ 0 dB
-1.25 dB	$\geq +1.25$ dB
-0.75 dB	≥ 0 dB
0 dB	≤ -1.25 dB, $\geq +1.25$ dB
$+0.75$ dB	-1.75 dB, $+1.75$ dB
$+1.25$ dB	≤ 0 dB
$+1.75$ dB	≤ 0.75 dB

Table 6.5: Pop A vs. Pop B loudness matching: Multiple comparisons revealed statistically significantly different gain settings between offsets of left and right column.

reference signal when the offset was +1.25 dB or more (-1.190 ± 0.151 dB). This difference was also statistically significant ($p < 0.002$). Finally, settings for offsets of +0.75 dB and +1.75 dB were statistically significantly different ($p = 0.015$).

The results, based on pairwise loudness matching of two different Pop mixes, are summarized in Table 6.5.

Statistical Analysis - Pop A / Classic pair

As shown in Figure 6.9c, the results from the Pop A / Classic pair look similar to the results from Pop A / Pop B comparisons.

Mauchly's test of sphericity revealed that the assumption of sphericity was violated, $\chi^2(2) = 20.971, p = .0431$, therefore, a Greenhouse-Geisser correction ($\varepsilon = 0.598$) was used. The various offsets applied to the loudness meter elicited statistically significant differences in gain settings, $F(3.589, 39.474) = 32.497, p < 0.0005, \eta^2 = 0.747$.

For the Pop A vs. Classic stimuli pair, the findings of the multiple comparisons with Bonferroni adjustment are comparable to the results in conjunction with the Pop mixes. They are presented in table 6.6.

Gain settings, induced by a loudness meter distortion of ...	**... are statistically significantly different ($p < 0.05$) compared to gain settings induced by loudness meter distortions of ...**
-1.75 dB	≥ 0 dB
-1.25 dB	≥ 0 dB
-0.75 dB	≥ 1.25 dB
0 dB	≤ -1.25 dB, $+1.75$ dB
$+0.75$ dB	≤ -1.25 dB, $+1.75$ dB
$+1.25$ dB	≤ -0.75 dB
$+1.75$ dB	≤ 0.75 dB

Table 6.6: Pop A vs. Classic loudness matching: Multiple comparisons revealed statistically significantly different gain settings between offsets of left and right columns.

On average, for every distinct degree of offset, there were 3.41 other levels of offsets inducing significantly different gain settings.

6.3.5 A Characteristic Measure for Visual Influence

Now, that it is evident that mixing decisions in conjunction with loudness matching are indeed significantly affected by visual cues, at least when both auditory and visual information is processed simultaneously (which is typically true when operating a Digital Audio Workstation), it seems appropriate to define a characteristic measure for the *Degree of Visual Influence* (DVI), which is probably dependent on the signals involved in the comparisons.

First, the gain errors illustrated in Figure 6.9 were related to the mean of the zero-offset conditions as opposed to zero. Accordingly, for all of the three graphs the respective mean of the zero offset condition was subtracted from all values. Therefore, the internal relation of the user responses was not affected. Next, all rectified gain settings were divided by the corresponding inverse offset. As explained earlier, if the test subject set the gain exactly to the inverse of the offset, the visual influence could be assumed to be 100%, because it can be argued that in this particular case the participant would have completely neglected the information retrieved from the

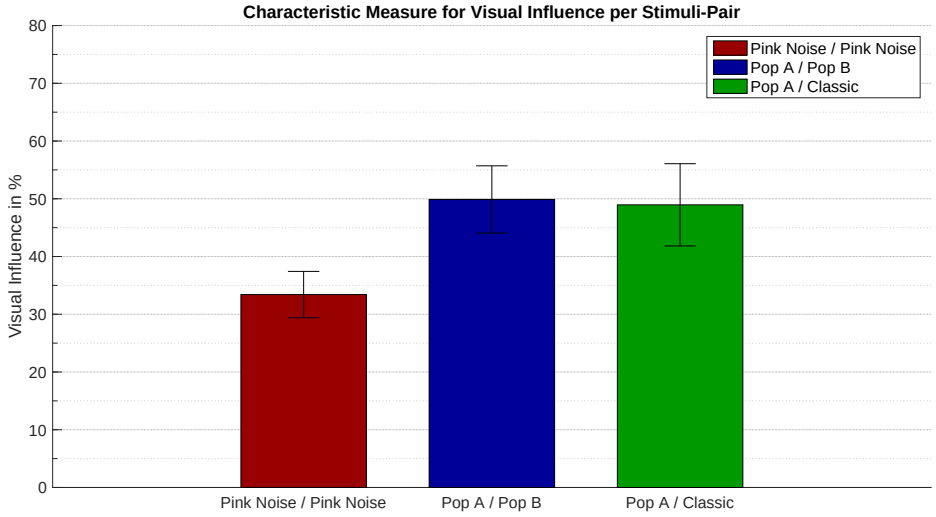


Figure 6.10: Characteristic measure for the degree of visual influence depending on test signals. A value of 100% may be interpreted as that the loudness was matched solely according to the display of the meters.

auditory modality. On the other hand, if the gain was set to zero despite of the offset, the visual influence was evaluated as zero percent. The DVI can also take values greater than 100% and smaller than 0%.

As described above, the characteristic measure for visual influence can be calculated for all trials with non-zero offsets. Means and standard errors of the mean for characteristic measure for visual influence are shown in Figure 6.10.

Overall, it can be observed that the visual influence was highest in combination with Pop A / Pop B test signals (49.88%), closely followed by the Pop A / Classic pair (48.95%). Participants seem to have been less affected by visual cues when matching noise signals (33.41%). Usually, it can be assumed that engineers are used to trust their tools in a real world mixing situation, however, in conjunction with pink noise signals, extreme offsets of ± 1.75 dB apparently have given rise to suspicion and probably caused the participants to question the display of the meter. Thus, results for the Pink-Noise stimuli-pair should be interpreted with caution.

A Shapiro-Wilk test of normality confirmed that data is normally distributed. No outliers have been detected, as assessed by inspection of boxplots. Mauchly's test of sphericity indicated that the assumption of sphericity had not been violated,

$\chi^2(2) = 2.934, p = 0.231$. Therefore, a one-way repeated measures ANOVA on DVI was performed, followed by Bonferroni corrected pairwise comparisons.

Test subjects were statistically significantly differently influenced by visual cues for the different test-signals, $F(2, 142) = 3.103, p = 0.048, \eta^2 = 0.042$. The amount of visual influence was marginally statistically significantly lower when pink-noise signals were matched (33.412 ± 34.04 %), compared to when two different Pop mixes were matched (49.88 ± 49.38 %), $p = 0.05$. Means, however, were not statistically significantly different compared to the Pop A / Classic pair.

6.3.6 Response time

Figure 6.11 illustrates the mean response time for the different conditions of the loudness matching experiment. On average, it took participants 25.5 seconds to complete a trial when adjusting the gain in a pink noise / pink noise comparison, 34 seconds in a Pop A / Pop B comparison and 36.5 seconds in a Pop A / Classic comparison. There is a tendency that the matching of musical test signals took more time than the matching of noise signals, and judgements were made faster when the aural modality was not involved.

The analysis of the reaction time also reveals that participants took noticeably more time to match the pink noise signals when the meter was distorted by the largest offsets of ± 1.75 dB, compared to smaller offsets or zero offsets.

A normal distribution of response times could be confirmed through a Shapiro-Wilk test, after a logarithmic transformation was applied to the data. No outliers have been detected after the transformation, as assessed by examination of boxplots. Therefore, a statistical analysis by means of a two-way repeated measures ANOVA on the transformed response time was performed. Mauchly's test of sphericity indicated that sphericity was not violated for the two-way interaction, $\chi^2(2) = 52.886, p = 0.727$.

The statistical analysis gives evidence that there is a statistically significant interaction between stimuli-pair and experimental condition, $F(10, 110) = 2.421, p = 0.012, \eta^2 = 0.18$. Thus, simple main effects with post hoc comparisons (with Bonfer-

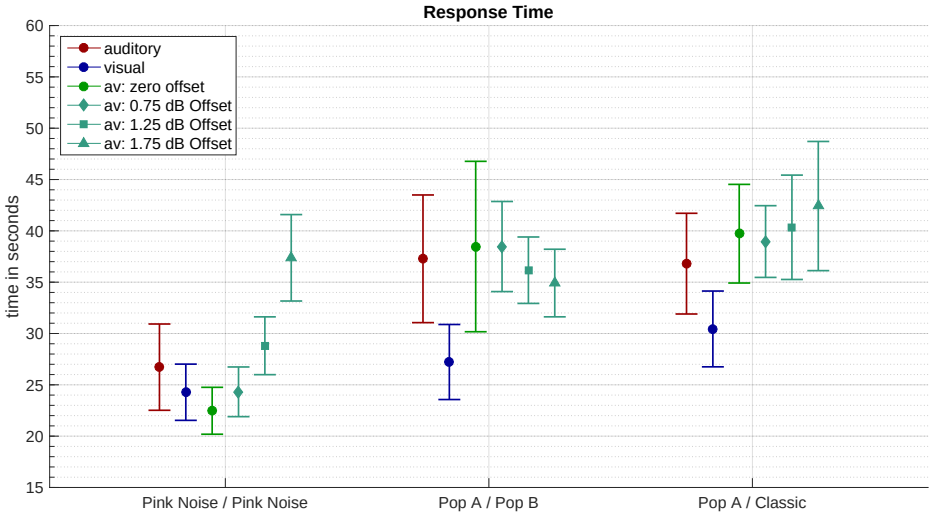


Figure 6.11: Response times per trial for the individual conditions of the experiment. Error bars represent the standard error of the mean.

roni adjustment) were run. Participants performed the task significantly faster when pink noise signals were matched, compared to when musical signals were matched. This was the case for the auditory condition (Pop A / Pop B: $p = 0.028$, Pop A / Classic: $p = 0.017$), the audio-visual zero offset condition ($p = 0.026, p = 0.004$), the audio-visual ± 0.75 dB condition ($p = 0.001, p < 0.0005$), the audio-visual ± 1.75 dB condition (Pop A / Classic: $p = 0.03$). Accordingly, it can be argued that the task was easier when noise signals had been matched than when musical signals had been matched. Surprisingly, there was not much difference in required time between the Pop A / Pop B and the Pop A / Classic stimuli-pairs, despite the fact that the Classic mix featured a larger dynamic range than the Pop mix. Further, the reaction time was only marginally significantly higher when the Pop A / Classic signals had to be matched visually, compared to when noise signals had to be matched visually ($p = 0.07$). This result gives evidence that the metering was probably equally useful for all three stimuli-pairs and somehow contradicts the assumption that the metering would be less helpful if the displayed loudness flickered due to dynamic material, as it was the case with the Classic mix. On the other hand, participants apparently had some difficulties to match the Pop A and the Classic mix visually as accurately as the other test signals, though this finding was only marginally statistically significant (see Figure 6.8).

The response times in conjunction with pink noise pairwise comparisons between the six different levels of condition were only marginally statistically significant ($p > 0.58$). Yet, this result gives evidence to the observation that large offsets caused mistrust towards the visual display. The tendency, that participants were faster when matching signals exclusively according to visual feedback, was not statistically significant.

6.3.7 Discussion

Based on results of previous studies [3, 105], it was expected that engineers do act differently depending on the display of the loudness meter. The analysis of gain settings in the context of a pairwise-comparison loudness matching experiment revealed that mixing and mastering decisions were indeed significantly affected by visual cues, which also coincides with the results of a study by Fastl [26], that the perceived loudness of sounds can be affected by the input of visual stimuli. In particular, the perception of equal loudness was influenced by the amount of convergence between the visual displays of the loudness meters. Gain settings were dependent on the amount of visual bias, and it was observed that at least responses associated with positive meter offsets were always significantly different from responses induced by negative meter offsets. The effect was even stronger when musical stimuli, such as Pop- or Classical mixes, were used as test signals. Therefore, it can be argued, that Hypothesis 3, which states that gain settings are significantly affected by visual cues and are dependent on the type of signals being matched, is true. Indeed, the derivation and analysis of a characteristic measure for visual influence indicated, that the visual bias was stronger in conjunction with musical pieces (around 50%) than in conjunction with noise signals (around 33%). However, it must be noted that in the case of the pink noise stimuli pair, the largest offsets of ± 1.75 dB probably provoked suspicion of the display of the meters, as it can be inferred from the non-linear behaviour at the extreme offsets in Figure 6.9a. In case of pink noise, the bias resulting from the largest offsets appeared to be around the same than for the other offsets, suggesting that the mismatch between vision and hearing was probably too obvious. Indeed, some participants mentioned an uncertainty in connection with the pink-noise stimuli pair. However, they rather questioned their auditory

perception or, that is to say, their ability to match the loudness correctly by ear than suspecting the display of the meters to be wrong. The facts that several outliers have been detected in connection therewith, and that the response time was a bit higher compared to smaller offsets, further support the assumption that some participants were concerned about the mismatch between vision and sound.

A very interesting conclusion can be drawn from the observed differences between the auditory, visual and audio-visual conditions. Apparently, when both auditory and visual information was provided, mean gain settings were approximately centred between mean gain settings of the unimodal conditions for all stimuli-pairs, respectively (see Figure 6.8). This finding suggests that information from both senses is equally weighted, and thus equally contributing to the perception of loudness, and augments the observations from several studies proving that input in one modality can alter activity in another modality [43, 53, 57, 64, 103]. Further, Baier et al [5] proposed that cross-modal suppression occurs when auditory and visual inputs are expected to be uncorrelated or distracting, whereas enhanced activity occurs when subjects are likely to benefit from reliably associated information in the two modalities. In a typical mixing task, visual information on the graphical user interface and auditory information is usually expected to complement each other, thus, information from both senses may be integrated in order to help making a reasonable decision. The assumption that both modalities are weighted equally is also supported by the fact that the degree of visual influence was around 50% (in case of musical test signals, see Figure 6.10).

According to the statistical analysis of unbiased data (unimodal conditions and the zero offset bimodal condition), gain settings were only statistically significantly different in case of the Pop A/ Pop B stimulus. Therefore, Hypothesis 2, which states that mixing decisions are affected by whether visual cues are provided, only holds for one of the three stimuli-pairs. This might be due to the observation that the perceived relative loudness of the Pop mix in the auditory condition did not correspond with the EBU-R128 loudness model. Participants judged the Pop A mix to be around 1 dB quieter than the Pop B excerpt, whereas the algorithm considered them to be of equal loudness. This finding coincides with the results of a study by Begnert et al [6]. Here, the EBU-R128 recommendation was investigated for its correspondence with perceived loudness, and differences up to ± 2.8 dB were

discovered when dissimilar program types were compared. In a similar study [91], different loudness models were tested, and differences between calculated loudness and perceived loudness were in the range of 0.5 dB to 2.5 dB.

The analysis of response time confirms Hypothesis 1. It was easier to perform loudness matching for noise signals than for more complex real-world signals, as participants took less time to complete the task for the pink noise stimuli-pair, especially when no visual feedback was provided. Gain settings were also more consistent across all conditions in case noise signals were adjusted. However, one would also expect that it is more difficult to match excerpts from different genres. Yet, when comparing the response times related to the loudness matching of a Classical mix and a Pop mix, no statistically significant differences have been found. Additionally, the degree of visual influence was higher than for noise signals, thus, it can be argued that the degree of visual influence is at least marginally dependent on how difficult it is to complete the task by ear, which confirms Hypothesis 4.

Chapter 7

Experiment 2: Spectral Matching

Experiment 2 refers to a common mixing and mastering situation where the timbre of two recordings has to be matched as closely as possible. A mixing engineer may face this problem when related instrumental or vocal performances within a song have been recorded on different days or with different recording equipment and therefore exhibit tonal differences which need to be corrected. Similarly, a mastering engineer has to make sure that all tracks on an album sound cohesive, as described in Chapter 2.3.

7.1 Equalizers

Shaping the spectral balance of a recorded sound is one of the most common tasks in mixing and mastering [75]. The most deliberate method of altering the spectral balance of a sound is achieved by boosting or cutting the amplitude of a certain frequency region with an equalizer [19, 76]. Therefore, equalizers can be used to either reduce or remove problematic frequencies of a sound, or they can be used to enhance certain frequency bands in order to emphasize important characteristics of a recording [19]. Different types of equalizers, such as shelving filters, low-, high- and band-pass filters, graphic equalizers and parametric and semi-parametric equalizers allow various level of control.

High-pass and low-pass filters are used to remove frequency content of a sound below

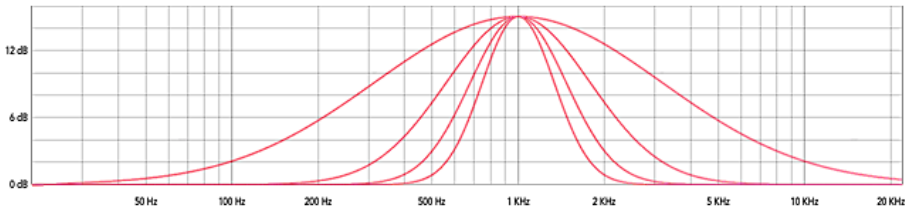


Figure 7.1: The frequency response of a parametric equalizer boosting 15 dB at 1 kHz with Q-Settings of (0.25/0.50/0.75/1.0). Reprinted from [1]

or above a certain cutoff-frequency. Typically, the only adjustable parameters of such filters are cutoff-frequency and filter slope (the amount of attenuation per octave in dB). A band-pass filter is essentially a combination of a high- and a low-pass filter. Graphical equalizers divide their frequency ranges in a number of bands, with adjustable gain (usually in form of vertical sliders) for each band. Parametric equalizers allow control of three independent parameters: center frequency, filter gain and Q factor. The filter gain determines the amount of boost or attenuation at the center frequency, and the Q factor controls the width of the frequency range being affected. It is inversely related to the 3 dB bandwidth, meaning that small Q factors affect a wider frequency range and high Q factors affect a narrower frequency range [19, 21]. On semi-parametric equalizers, one or more of these features are missing. For example, semi-parametric equalizers often provide a fixed Q factor [37].

7.2 Methodology

Whereas in Experiment 1 test subjects were asked to apply a gain change to the whole frequency range of a signal, the gain change was now applied only to restricted portions of the frequency spectrum by using one semi-parametric equalizer band with a fixed Q factor. Participants were now required to match the spectrum of two sound files as closely as possible by adjusting the center frequency and filter gain of this semi-parametric band. The stimuli-pairs to be matched consisted of two identical test files, whereby one of the files had been pre-equalized with the same type of semi-parametric equalizer. The test subjects were expected to compensate this pre-equalization by applying the inverse gain at the same center-frequency, thus, generating a flat frequency response at the output of the two serial equalizers. The main objective of this task was a) to increase the level of difficulty compared to

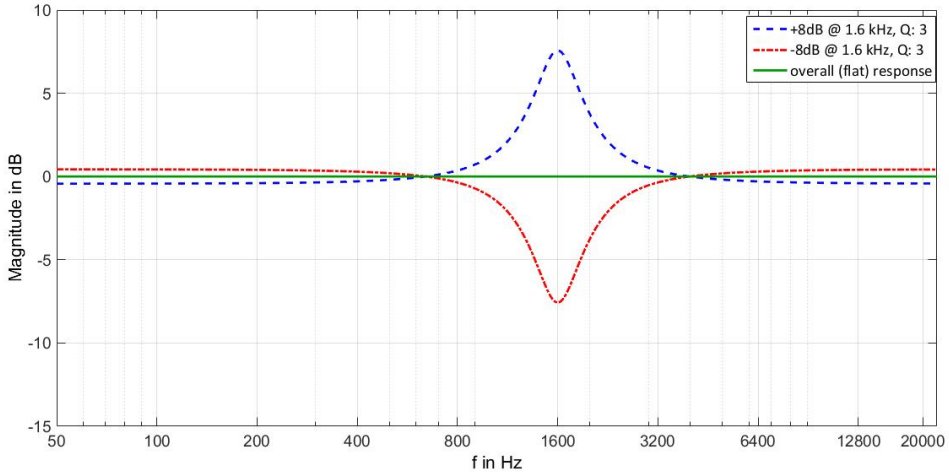


Figure 7.2: Measured frequency response of Fabfilter's Pro Q2 (single bell-band, slope 12 dB). Curves illustrate an 8 dB boost at 1.6 kHz and a Q factor of 3 (blue line), an inverse 8 dB cut at the same frequency (red line) and the overall flat frequency response of both bands in series (green line).

the first experiment by introducing a second parameter (frequency), b) to increase the complexity and amount of visual load and c) to still maintain a reasonable connection to a real-world mixing task.

As with the first experiment, the task had to be performed for two unimodal conditions (aural and visual) and one multimodal condition (audio-visual). To avoid confounding due to order effects, the order of the conditions was counterbalanced across the participants. As spectral matching is a very complex and even more time consuming task, the number of different stimuli-pairs was reduced. Thus, for each condition, subjects were asked to match the timbre of two different stimuli-pairs three times respectively.

While the computer screen was turned off for the single-modal aural condition, two equalizer plug-ins were arranged side by side on the screen for the visual and audio-visual conditions. The interfaces of the first equalizer showed both the EQ-curve of the active, user-adjustable, semi-parametric band as well as a real-time spectrum analyser lying underneath. Test subjects could then compare the spectrum of the test track to the spectrum of the target track, which was presented via a real-time spectrum analyser (with same look and settings) on the second plug-in. However, the displayed equalizer-curve was distorted by an offset for the audio-visual condition. Dependent on the type of test signal, different offsets were applied to the displayed

center frequency of the user-adjustable equalizer band, causing the EQ-plug-in to show a different EQ-curve and spectrum compared to what was audible. A more detailed description of the experimental setup is provided below.

7.2.1 Setup

Experiment 2 was performed in the same lab at the *IEM* Graz as the first experiment. Two different pairs of 16 bit, 44 kHz test files were arranged within the digital audio workstation *Reaper*, running on a laptop computer. Each pair of test files comprised a *Test* file and a *Reference* file.

The audio interface used was a *RME Fireface UC* and test subjects were provided with *AKG K712* mastering headphones, connected to the headphone output of the interface. The monitoring level was set to 83 dB SPL (referring to the pink noise sample at -10 dBFS) for all participants.

Figure 7.3: Experiment Setup: Behringer X-Touch Mini, AKG K712 Headphones, computer screen showing the interface of two Fabfilter Pro Q2 equalizer plug-ins.

Subjects were given control over two hardware rotary encoders in order to adjust the center frequency and the filter gain of the active, semi-parametric equalizer band. Both encoders were programmed velocity-dependent, meaning that rapid twists produced large parameter changes whereas slow rotations produced very small parameter changes (up to a minimum of around 1 Hz for frequency and 0.1 dB for filter gain). The hardware controller used was a Behringer X-Touch Mini and both rotary encoders were again set to 'relative mode 2'.

Midi messages coming from the controller were transmitted to *Pure Data*, translated to OSC messages and then sent to *Reaper*. The output of the first encoder was assigned to the center frequency parameter of a single band parametric equalizer plug-in (Fabfilter Pro Q2) with fixed Q-factor, the second encoder was assigned to the corresponding filter gain.

Multiple factors have been considered when choosing the bandwidth of the filter. Firstly, the resonances introduced by the pre-equalizer should be clearly audible and clearly visible in the spectrum analyzer. This requires a high Q-factor (small bandwidth), producing very narrow peaks or attenuations in the spectrum. This will help the subjects to identify the correct center frequency more precisely. Furthermore, the impact resulting from frequency misadjustments on the combined transfer function of the serial equalizers will be less obvious for wider bandwidths. Therefore, at higher Q-factors, a mismatch between the pre-equalizer center frequency and the user adjusted center frequency will be more audible. On the other hand, if the Q-factor is chosen too high, there is a risk that spectral masking may occur due to strong attenuations. Spectral masking describes the phenomenon that a sound of a certain frequency is made inaudible by the presence of a sound of different frequency with higher intensity [102]. Though all test signals were broadband signals with a rather flat spectrum, it could be assumed that the occurrence of masking effects is rather unlikely in case the bandwidth of the filter is chosen reasonably. However, preventively the filter-gain of the pre-equalizer was limited to positive gains only in order to completely avoid spectral masking effects which can possibly occur due to strong, narrow attenuations. Also, it is important that the band under consideration is actually unmasked in the original signal, otherwise participants can go infinitely low with the gain without any audible effect.

According to the considerations above, the Q-factor of the single active band was set to a value of 3. Analysis of the filter transfer-function revealed, that the plug-in specific Q-factor of 3 corresponded to a half-power bandwidth of about half an octave. This setting turned out to be a good compromise, because it still generates relatively steep peaks that are both easily visible and audible.

For each condition, the task was repeated three times per stimuli-pair. In order to be able to draw conclusions about the influence of different frequency ranges

on spectral matching, for each of the three repetitions the center frequency of the pre-equalizer was distributed across the following three different bark-bands:

Region 1		**Region 2**		**Region 3**	
Frequency Range	Center Frequency	Frequency Range	Center Frequency	Frequency Range	Center Frequency
770 Hz to 920 Hz	840 Hz	1480 Hz to 1720 Hz	1600 Hz	2700 Hz to 3150 Hz	2900 Hz

Table 7.1: Frequency-regions for center frequencies

In other words, the spectral matching task was performed once within every frequency region for both stimuli-pairs respectively. It was considered necessary to restrict the center frequencies of the filter to the regions listed above in order to be able to detect possible effects that may occur only in specific frequency ranges. In this way, while center frequencies were restricted from being totally random, participants were tested in a broad frequency range where the human ear is most sensitive. At the same time, this ensured that the variation of target center-frequencies is large enough in order to justify that the experiment is close enough to a real-world mixing task.

The same strategy has been utilized to reduce the influence of the pre-equalizer's filter gain, as results may be affected variously depending on the intensity of boosts. For that reason, the rage for the filter-gain value was limited to an interval from a minimum of 7 dB to a maximum of 10 dB. In this way, the pre-equalization was easily perceptible and extreme settings were avoided. As described earlier, the pre-equalization was restricted to boosts in order to avoid spectral masking effects.

For each of the stimuli-pairs, the center-frequency of the adjustable band was initialized at 450 Hz and 6 kHz on rotating basis in order to ensure that the target frequency is approached from both directions. Referring to Experiment 1, the following controls were accessible on the Midi-Controller:

- **Center Frequency (Encoder):** Controls the center-frequency (10 Hz to 30000 Hz) of the semi-parametric equalizer band
- **Filter Gain (Encoder):** Controls the filter gain (-30 dB to +30 dB) of the semi-parametric equalizer band

- **Reset Gain (Button):** Sets filter gain to zero. This function is intended for the unimodal aural condition, in case participants lose track of whether they are currently boosting or attenuating.

- **Solo Reference track (Button):** Solos the *Reference* track (and mutes the *Test* track)

- **Solo Test Track (Button):** Solos the *Test* track (and mutes the *Reference* track)

- **Start (Button):** Starts the individual subtask/condition, and activates the stop watch implemented in *Pure Data*. It also reads parameters of the first sample from a list (this lists were generated with *Matlab* prior to the experiment). These lists are:

  - *Center Frequency*: Random numbers within the frequency regions specified in Table 7.1, setting the center-frequency of the pre-equalizer on the *Test* track. For each of the three individual stimuli-pairs, the task is performed for all three frequency ranges in which target frequency was jittered (hence, there is a total of three repetitions per stimuli-pair)
  - *Stimuli*: This random sequence of numbers determines which stimuli pair is played back next; 1 = Pink Noise, 2 = Pop A/Pop A
  - *Offset*: A random number in percent (See Section 7.2.6).
  - *Filter Gain*: Random number on the interval [7,10] in dB, controlling the filter gain of the pre-equalizer on the *Test* track.
  - The first two trials for each condition are reserved for training and will not be evaluated. The first two values in the stimuli list are 1, 2. This should help the subjects to become familiar with the different audio examples. The first two values in the offset list are both zero.

- **Pause (Button):** Stops the stop watch and the audio playback, in case the subject needs a break.

- **Resume (Button):** Resumes the stop watch

- **Play/Stop (Button):** Plays and stops the audio playback. Test signals are played back in a loop.

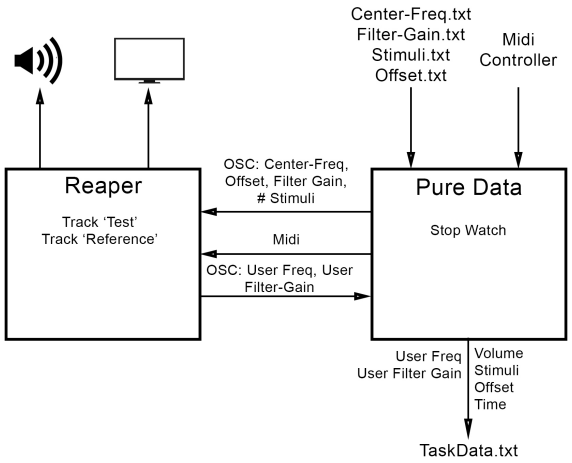


Figure 7.4: General experimental setup for Experiment 2

- **Next (Button):** Saves the current parameters to a text file (user adjustments for center-frequency and filter gain, pre-equalizer center-frequency and filter gain, stimuli pair, time required), reads the next parameters from the lists and jumps to the start of the next stimuli-pair (according to the entry in the *Stimuli* list) within the *Reaper* arrangement window. Every time this button is activated, the initial center-frequency of the user-adjustable filter band is set randomly between 400 Hz and 20 kHz for the next stimuli-pair.

Test subjects were instructed to apply a fixed center-frequency and a fixed filter gain for the entire duration of the test files, meaning that they were not allowed to make use of any automation. Moreover, they were notified that there was no time limit within which the task must be completed.

7.2.2 Stimuli

The spectral matching task was tested on basis of the following two pairs of audio examples:

(a) **Pink Noise:** Two 8 second excerpts of pink noise. Pink noise has equal power per octave.

(b) **Pop A / Pop B:** Two identical 30 second excerpts (chorus) of a commercial

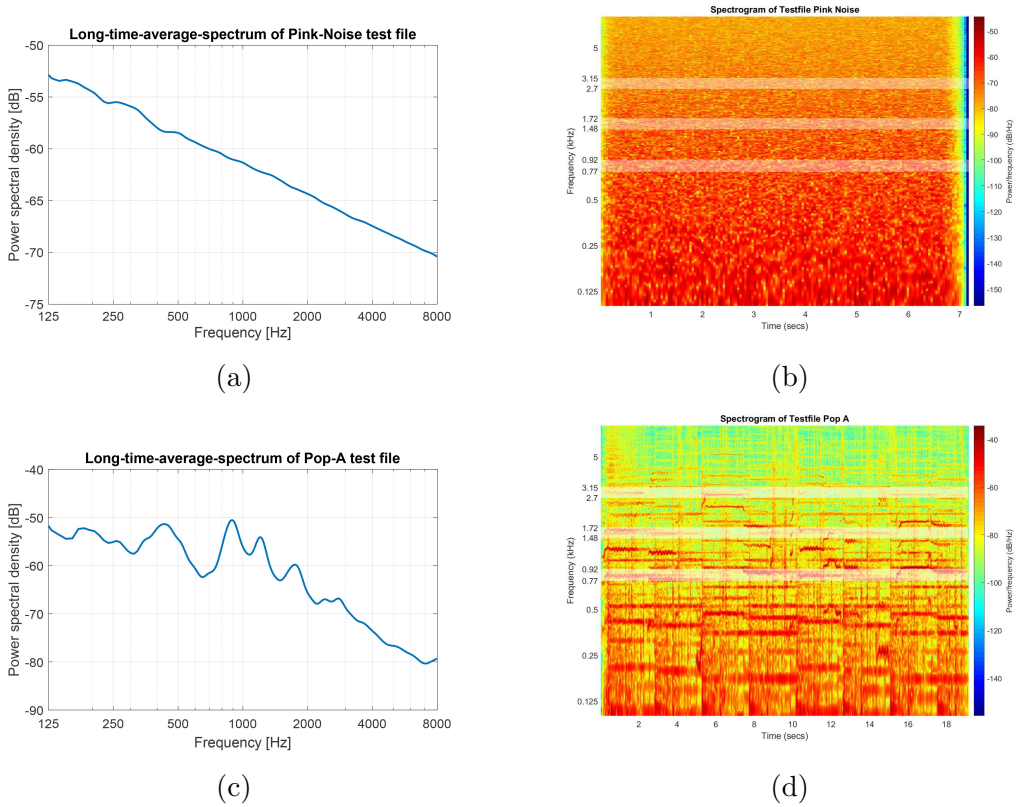


Figure 7.5: Long-time-average-spectrum (4096 point FFT with 2048 samples hop-size and 1/6 octave smoothing) for the Pink-Noise test-file (a) and the Pop A test-file (b). Spectrogram for the Pink-Noise test-file (c) and the Pop A test-file (d); Center-frequency regions are highlighted.

Pop-song. The excerpt features percussion, vocals, bass-guitar, guitars and orchestra.

All audio files were 16-bit, 44.1 kHz .wav stereo files, normalized to an integrated EBU-R128 measured loudness level of -18.5 LUFS.

7.2.3 Participants

There were twelve subjects of different ages (19-35 years, 2 female, 10 male) participating in the experiment. All of them were audio engineering students at the Graz University Of Technology and had prior experience in sound mixing. Five of the subjects also participated in the loudness matching experiment described in Chapter 6. None of the subjects were paid for participating in the experiment, and all of them participated voluntarily.

7.2.4 Procedure for Condition 1 - Auditory

The single-modal aural condition required the subject to perform the spectral matching task in absence of any visual cues. Thus, the computer screen was turned off, and the (meaningless) LED-rings around the hardware encoders were covered to prevent participants from being distracted or irritated by the status of the lamps. The task was started by pressing the *Start* button on the midi controller. A total of eight pairs of different test files (the task was performed four times per stimuli-pair) were loaded in random order. The participants could move on to the next example by pressing the *Continue* button, and their settings for the current files were saved to a text file. However, the first two examples were intended to familiarize the subject with the different test signals and therefore have not been evaluated.

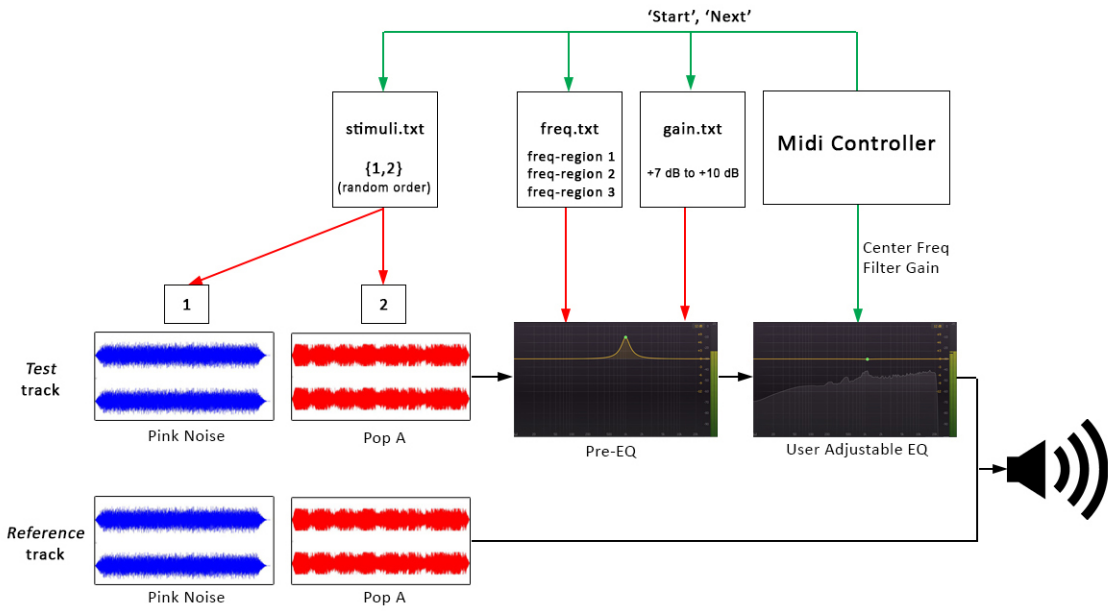


Figure 7.6: Schematic illustration of the experimental design for the unimodal aural condition

Subjects could use the *Solo Test Track* and *Solo Reference Track* buttons to switch playback between the tracks. While the *Reference* track was unaltered, the *Test* track had been pre-equalized with a single semi-parametric EQ-band in three different frequency bands per stimuli-pair and random gain (5-9 dB boosts). Test-takers were asked to match the sound of both tracks on their own discretion by adjusting the center-frequency and gain of a second, identical filter band in series to the first band. Theoretically, the *Test* track sounds closest to the *Reference* track if the sec-

ond band is adjusted in a way that it exactly inverts the first band (an inverse filter gain at the same filter frequency will produce an overall flat EQ-curve).

7.2.5 Procedure for Condition 2 - Visual

In contrast to *Condition 1*, the computer screen was now turned on, but the output of the headphones was turned off instead. Participants were required to match the test files by comparing a visual representation of the frequency spectrums. In order to do so, two EQ plug-ins were displayed side by side on the screen. The left plug-in, applied to the second insert slot on the *Test* track, was showing the EQ-curve of the user controllable semi-parametric band as well as a real-time spectrum analyser within the same user interface. The interface of the right plug-in, inserted on the Reference track, only showed a real-time spectrum analyser displaying the target spectrum.

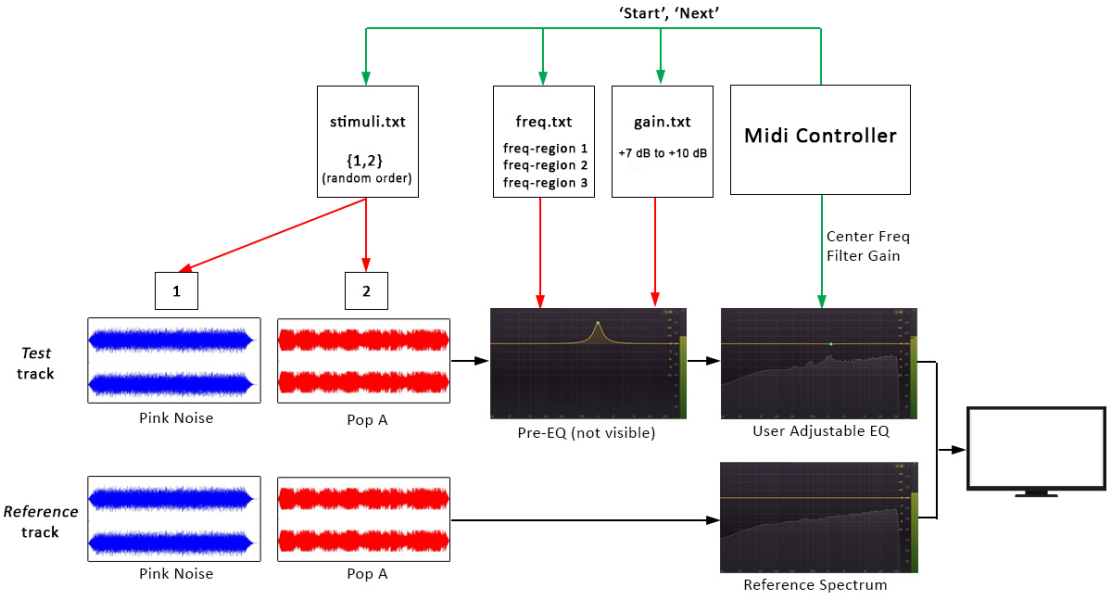


Figure 7.7: Schematic illustration of the experimental design for the unimodal visual condition

The following settings were used for all the equalizer plug-ins (*Fabfilter Pro Q2*):

- *Zoom-Level*: 12 dB
- *Mode*: Zero Latency

- *Channel Mode*: Left/Right
- *Spectrum-Analyzer*:
  - Post EQ
  - Response: very slow
  - Resolution: medium (2048 points)
  - Range: 90 dB
  - Tilt: 4.5 dB/octave
- Q-factor: 3.0

It is important to notice, that the EQ-curve and settings of the pre-equalizer were not visible to the subjects. Also, participants were not informed about the limited gain and frequency ranges in which the pre-equalization occurred. A real-world mixing task could only be constituted by allowing test-takers to make their adjustments according to their subjective perception, not due to artificially narrowed choices.

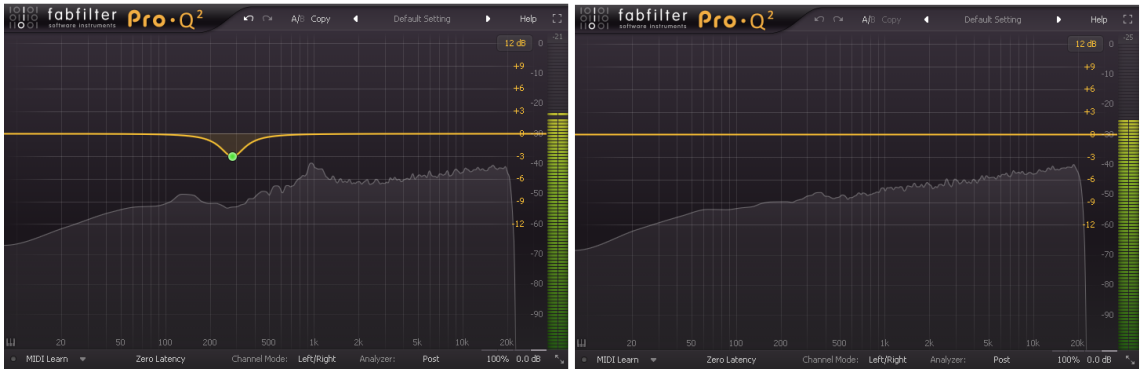


Figure 7.8: Example of displayed EQ-Plug-ins on the screen. Subjects were asked to match the frequency spectrum (light-grey area) shown on the left and the right window. *Left*: The interface shows the EQ-curve (yellow line) of the user-adjustable semi-parametric band (here it is set to a 3 dB cut at around 300 Hz). Additionally, it shows the spectrum of the Test track (in this case, it is showing the spectrum of the Pink-Noise stimulus). The bump in the spectrum results from a pre-equalization at 1 kHz with +8 dB. In order to remove the resonance, the user would have to dial in an 8 dB cut at 1 kHz. *Right*: The interface shows the unaltered reference spectrum of the Pink-Noise stimulus.

7.2.6 Procedure for Condition 3 - Audio-Visual

Condition 3 is essentially a combination of the previous two conditions. Participants were now supported with both aural and visual feedback via headphones and

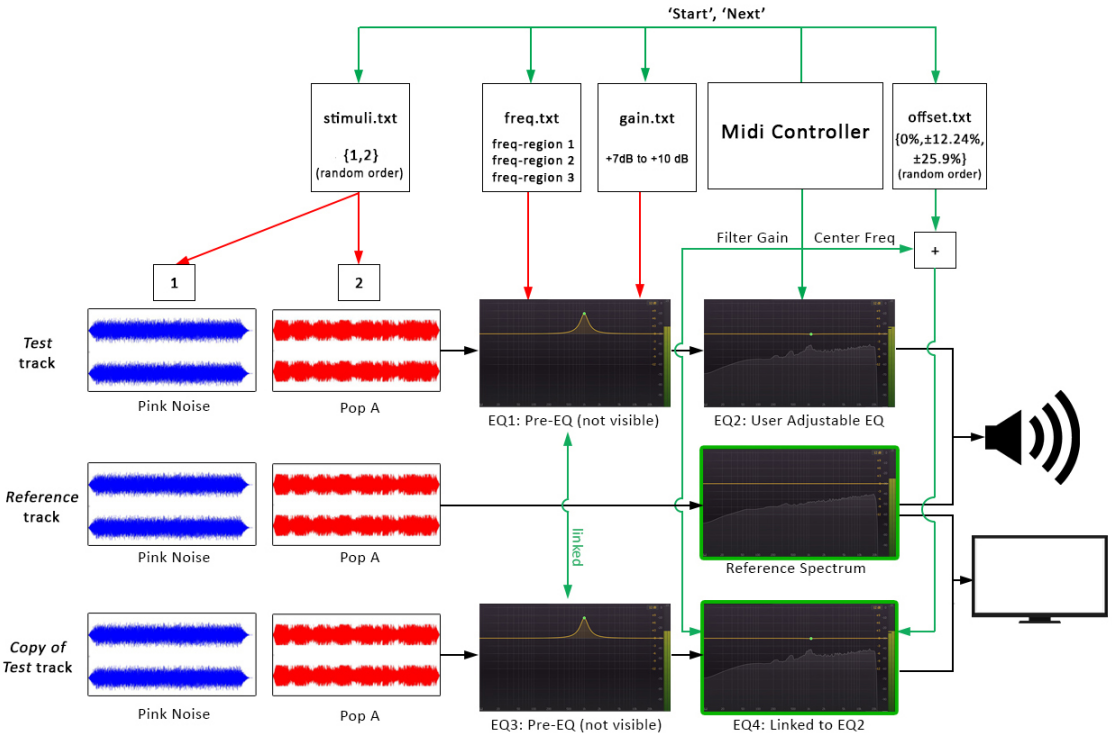


Figure 7.9: Schematic illustration of the experimental design for the multimodal audio-visual condition. Only green-bordered plug-in interfaces are visible on the screen (Reference Spectrum and EQ4).

computer screen. Based on the design of Experiment 1, the readout of the visual cue was again manipulated. This time, the center frequency of the user-adjustable EQ was distorted by an offset of either 0 %, ±6.12 %, ±12.24 % for the noise-signal or 0 %, ±12.24 % or ±25.9 % for the Pop-signal. It is important to notice that these offsets were only applied to the visual representation of the EQ-curve, meaning that the subjects were hearing the result of slightly different equalization compared to what was visible on the screen. However, the second user-adjustable parameter - the filter gain - was not distorted.

Again, the task was performed three times per stimuli-pair, whereas resonances were introduced in different frequency regions respectively (please refer to 7.1). It was further repeated for all possible offsets, leading to a total number of 32 repetitions of the audiovisual spectral-matching task (including two trials for training). The order of stimuli, offsets and frequency regions was randomized for every subject. As the goal of this condition was to investigate and quantify the influence of the visual cues on the participant's choice of center frequency, subjects have obviously

not been informed about the manipulated display.

7.2.7 Summary: Spectral Matching Experimental Design

The table below gives an overview about the core features in Experiment 2.

Condition	Controllable Parameters	Pre-EQ Settings	Phones	EQ on Test Track	EQ on Reference Track	Repetitions
Aural	*Center-Freq.:* From 350 Hz to 6 kHz *Filter-Gain:* From -30 dB to + 30 dB *Q:* 3.0 (fixed)	*Center-Freq.:* *Region 1*: 770-920 Hz *Region 2*: 1480-1720 Hz *Region 3*: 2700-3150 Hz *Filter-Gain*: +7dB to +10dB	On	Not visible	Not visible	$2 \times 3 + 2 = 8$ 2 stimuli a),b) in 3 freq. regions + 2 repetitions for training
Visual	*Center-Freq.:* From 350 Hz to 6 kHz *Filter-Gain:* From -30 dB to + 30 dB *Q:* 3.0 (fixed)	*Center-Freq.:* *Region 1*: 770-920 Hz *Region 2*: 1480-1720 Hz *Region 3*: 2700-3150 Hz *Filter-Gain*: +7dB to +10dB	On	*visible*: – User-adjustable EQ-curve – Spectrum of Test track	*visible*: Spectrum of Reference track	$2 \times 3 + 2 = 8$ 2 stimuli a),b) in 3 freq. regions + 2 repetitions for training
Audio-Visual	*Center-Freq.:* From 350 Hz to 6 kHz *Filter-Gain:* From -30 dB to + 30 dB *Q:* 3.0 (fixed)	*Center-Freq.:* *Region 1*: 770-920 Hz *Region 2*: 1480-1720 Hz *Region 3*: 2700-3150 Hz *Filter-Gain*: +7dB to +10dB	On	*visible*: – User-adjustable EQ-curve (distorted by 5 levels of offset per stimulus) – Spectrum of Test track	*visible*: Spectrum of Reference track	$2 \times 3 \times 5 + 2 = 32$ 2 stimuli a),b) in 3 freq. regions for 5 offsets + 2 repetitions for training

Table 7.2: Illustration of key parameters for the three conditions in Experiment 1

7.3 Results: Spectral Matching

7.3.1 Hypotheses

According to the literature review, personal observations during pilot experiments, and results from the loudness matching task, the hypothesis can be summarized as follows:

1. Resonances can be identified more easily and can be removed more accurately from stationary noise signals than from fluctuating program materials such as music.
2. Resonances can be removed with higher accuracy when the spectrum of the signal is visualized.
3. Engineers do act differently when equalizer parameters are represented visually, compared to situations where the EQ curve is not visualized.
4. Mixing decisions are significantly biased by visual cues in a way that the ability to completely remove a resonance is inversely commensurate to the degree of offset applied to the visual display.
5. The more difficult it is to complete the task by ear, the more users will be affected by visual cues.

Data from the spectral matching experiment is investigated by means of a statistical analysis in order to be able to verify or disprove the assumptions specified above.

7.3.2 Data Preparation

Results from the experiment can be categorized into two groups. The first group contains all responses from *unbiased conditions* including all conditions where either no graphical interface was displayed or the visual display was not distorted. This set of data can further be distinguished by dividing it by condition (auditory, visual or audio-visual), stimulus (pink noise or Pop mix) and frequency band. On

	Unbiased Group	Biased Group
Conditions	auditory, visual, audio-visual	audio-visual
Stimuli-Pairs	Noise, Pop	Noise, Pop
Offsets	Only zero offset	All 5 offsets
Frequency Regions	3	3
Number Of Subgroups	3 conditions x 2 stimuli-pairs x 1 Offset x 3 frequency bands = **18**	1 condition x 2 stimuli-pairs x 5 Offsets x 3 frequency bands = **30**

Table 7.3: Two main data groups, consisting of 18 and 45 subgroups á 12 responses respectively

the other hand, the second group contains only responses from the audio-visual condition. Subgroups of this *biased* data set can therefore be formed by separating responses associated with different offsets, stimuli and frequency bands. Each subgroup contains one value per subject, a total of 12 values.

To start with, unbiased data and biased data are discussed separately in the sections 7.3.3 to 7.3.6. After that, combined results are outlined in section 7.3.7.

7.3.3 Evaluation of Unbiased Data - Frequency Error

As with most sensual perceptions, the relationship between human pitch-perception and frequency is non-linear [102]. The ear rather responds approximately logarithmically to frequency, thus Frequency Error was computed as the relationship between center-frequency of the added resonance and center-frequency adjusted by the user in *Cent* [1] . The unit Cent can also be expressed in musical intervals, as 100 Cents correspond to one semitone. Figure 7.10 illustrates the mean *absolute* Frequency Error in conjunction with unbiased conditions. For an illustration of the mean Frequency Error, please refer to the Appendix.

One strong outlier was found in the data. It could be traced back to a trial were the participant accidentally hit the Next button on the Midi controller, thus, this value

[1]The relation between Cent and frequency ratio is given by: $Cent = 1200 \cdot \frac{\log_2 f_1}{\log_2 f_2}$

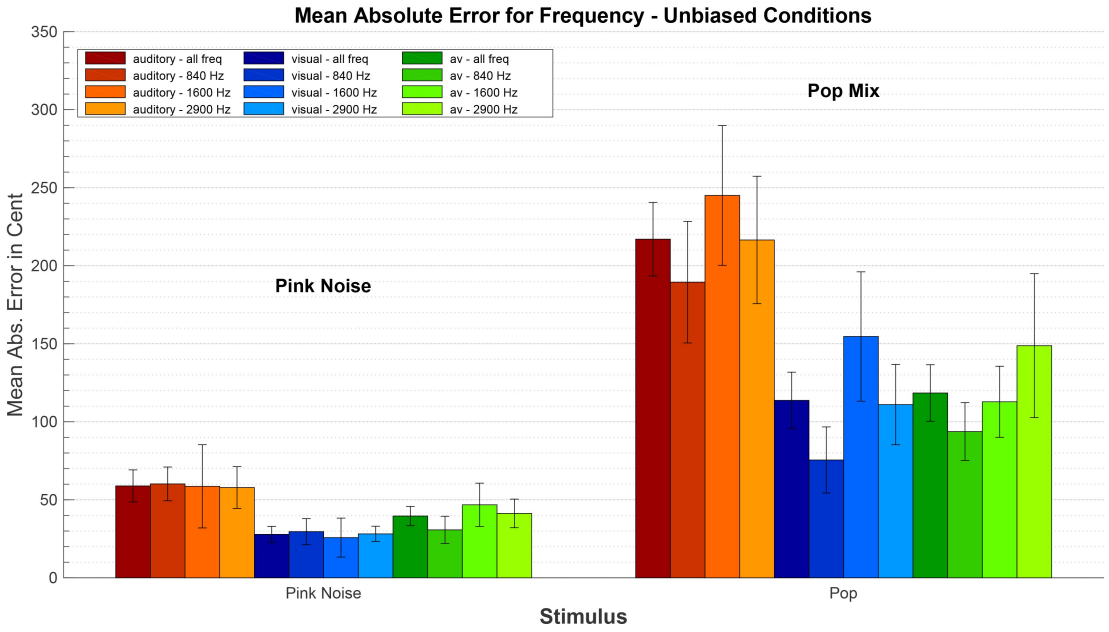


Figure 7.10: Mean Absolute Frequency Error for unbiased conditions. Error bars represent the standard error of the mean.

was replaced with the mean of the remaining data.

Regarding the differences between the different types of test signals, Figure 7.10 shows that participants achieved higher accuracy with regards to center-frequency when removing resonances from a noise signal than from a music mix. The mean absolute frequency error spanned a range of about 60 Cents in case of pink noise, and a range of 260 Cents in the case of the Pop mix. Another discovery with the data from this experiment is that the presence of visual feedback improved the ability to find the correct center-frequency by about 50%. In this context, an interesting observation is that the additional information from the auditory sensory modality did not improve precision compared to the unimodal visual condition. Further, different frequency regions generally seemed to have only marginal impact on the results.

The data turned out to not meet the assumptions for a 3-way repeated measures ANOVA. Seven outliers were detected, as assessed by inspection of box plots, and a Shapiro-Wilk test of normality revealed that only 39% of the data subgroups were normally distributed. Therefore, the ANOVA on absolute frequency error with independent variables condition, stimulus and frequency region was run both for the original data and for square root transformed data. Additionally, four remaining

Main effects & interactions		**F**	**Sig.**	ε
Condition	$F(2, 22) =$	18.854	$p < 0.0005$	
Stimulus	$F(1, 11) =$	95.412	$p < 0.0005$	
Frequency	$F(2, 22) =$	1.313	$p = 0.289$	
Condition × Stimulus	$F(2, 22) =$	4.194	$p = 0.029$	
Condition × Frequency	$F(4, 44) =$	0.260	$p = 0.902$	
Stimulus × Frequency	$F(1.26, 13.90) =$	2.143	$p = 0.164$	0.676
Condition × Stimulus × Frequency	$F(2.14, 23.53) =$	1.001	$p = 0.387$	0.535

Table 7.4: Results of the 3-way repeated measures ANOVA on square root transformed data. Epsilon indicates a Greenhouse-Geisser correction due to violated sphericity.

outliers in the transformed data were replaced by the nearest value not considered as an outlier. As a result thereof, 89% of the transformed data sets were normally distributed.

The results of the ANOVA on transformed data indicated, that neither the frequency region of the target center-frequency itself, nor any interaction between frequency region and the remaining within-subject factors had a statistically significant impact on the frequency error (see Table 7.4). Statistical analysis of the original non-transformed data by means of a 3-way repeated measures ANOVA led to the same conclusions. Consequently, and in accordance with the observations from Figure 7.10, frequency region was discarded as a significant within-subject factor for all following statistical analyses. As a result thereof, responses are not distinguished by their associated frequency band from now on, and data is grouped only by condition and stimulus and therefore averaged across the three trials per person in different frequency regions.

As shown in Table 7.4, there was a statistically significant 2-way interaction between the remaining within-subject factors stimulus and condition; hence, simple main effects were investigated. Removing resonances from pink noise test signals lead to significantly different absolute frequency errors depending on what sensory modalities were involved. Pairwise comparisons with a Bonferroni correction for multiple comparisons indicated, that providing just visual information (27.86 ± 26.39 Cent)

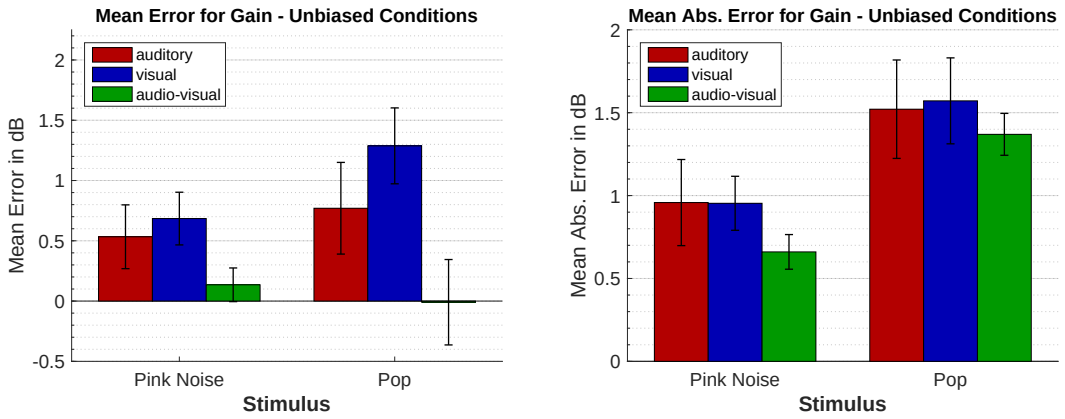
in the form of an equalizer curve on top of a real-time spectrum analyser significantly reduced the absolute frequency error, in comparison to equalizing solely by ear (58.9 ± 41.17 Cent), $p = 0.006$. Differences between the visual- and the audio-visual condition ($p = 0.313$) and between the auditory and audio-visual condition ($p = 0.193$) were not significant. Similar results were found for the Pop mix, however, the 2-way interaction can be attributed to the fact that both the visual and the audio-visual were significantly different from the unimodal auditory condition (whereas in case of pink noise, just the visual and the auditory condition were statistically significantly different). As indicated by the output of the ANOVA, the null hypothesis that the absolute frequency errors are the same for different sensory modalities involved could be rejected. Bonferroni corrected pairwise comparisons revealed that the frequency error was significantly higher in case the display was turned off. (auditory vs. visual: $p = 0.005$, auditory vs. audio-visual: $p = 0.001$). As expected, removing resonances from a real-world music mix also elicited statistically significantly higher absolute frequency errors. (Pop vs. noise, auditory: $p < 0.0005$; Pop vs. noise, visual: $p = 0.001$; Pop vs. Noise, audio-visual: $p = 0.004$).

7.3.4 Evaluation of Unbiased Data - Gain Error

The second factor that contributes to the overall error is misadjustments of filter gains. Ideally, if subjects found the correct center-frequency, they would need to adjust the gain inversely to the filter gain of the added resonance in order to make the spectrum of the test signals sound and look exactly the same. By means of Figure 7.11, both the mean gain errors and the mean absolute gain errors are illustrated.

Regarding the mean filter gain errors, there is a small tendency that cuts were undersized, except for the audio-visual condition. The mean absolute gain error appears to be quite moderate, within a range from 0.6 dB to about 1.5 dB. Again, mean absolute errors apparently were larger for the Pop mix than for the noise signal.

A two-way repeated measures ANOVA was performed to determine whether processing information from auditory or visual or combined audio-visual sensory modalities has an effect on the filter-gain error and absolute filter gain error in a spectral match-



(a) (b)

Figure 7.11: Mean gain error (a) and mean absolute gain error (b) for responses associated with conditions where the equalizer curves and spectrum analyser where either undistorted or not shown. Error bars indicate the standard error of the mean.

ing experiment. According to Shapiro-Wilk's test of normality, the assumption of normality was met, and no outliers were found, as assessed by inspection of boxplots with no outliers greater than ±3 box lengths. Mauchly's test of sphericity indicated, that there was sphericity for the interaction term ($p > 0.05$). In case the dependent variable was the non-absolute gain error, there was no statistically significant two-way interaction between Condition and Stimulus, $F(2, 22) = 1.317, p = 0.288$. However, the main effect of Condition showed a statistically significant difference between responses, $F(2, 22) = 1.317, p = 0.002$. Pairwise comparisons with Bonferroni corrections demonstrated that mean gain settings were statistically significantly different (the gain error was smaller) if information was received from both the auditory and visual sensory, compared to if only visual feedback was provided ($p = 0.001$). Secondly, the ANOVA was performed on absolute gain error. Again, no statistically significant interaction between Condition and Stimulus was found, $F(2, 22) = 0.091, p = 0.913$. Regarding main effects, absolute gain errors were statistically significantly smaller if noise signals were matched ($p = 0.002$).

7.3.5 Evaluation of Biased Data

A 3-way ANOVA on center-frequency error was run in order to determine if there was a 3-way interaction between *Stimulus*, *Offset* and *Frequency-Region*. Results are illustrated in Table 7.5. Only shared offsets were included in the analysis. There

were no outliers, as assessed by no values greater than 3 box-lengths from the edge of the box. Frequency-errors were normally distributed, as confirmed by Shapiro-Wilk's test of normality ($p > 0.05$). Mauchly's test showed that sphericity was met ($p > 0.05$), except for the main effect of *Offset*, $\chi^2 = 9.426, p = 0.01$. In this case, a Greenhouse-Geisser correction was used, $\varepsilon = 0.624$.

Main effects & interactions		**F**	**Sig.**	ε
Stimulus	$F(1, 11) =$	0.002	$p = 0.965$	
Offset	$F(1.25, 13.72) =$	77.597	$p < 0.0005$	0.624
Frequency	$F(2, 22) =$	1.696	$p = 0.206$	
Stimulus × Offset	$F(2, 22) =$	4.131	$p = 0.030$	
Stimulus × Frequency	$F(2, 22) =$	3.694	$p = 0.041$	
Offset × Frequency	$F(4, 44) =$	1.305	$p = 0.283$	
Offset × Stimulus × Frequency	$F(4, 44) =$	1.122	$p = 0.359$	

Table 7.5: Biased data; Results of the 3-way repeated measures ANOVA on center-frequency error. Epsilon indicates a Greenhouse-Geisser correction due to violated sphericity.

There was no statistically significant 3-way interaction between *Offset*, *Stimulus* and *Frequency Region*. On the other hand, the two-way interactions between *Stimulus* and *Offset*, as well as the interaction between *Stimulus* and *Frequency-Region* were statistically significant, as shown in Table 7.5. The latter interaction can be explained by the result of Bonferroni corrected pairwise comparisons, which indicated that center-frequency errors were statistically significantly different between pink noise (12.85 ± 7.74 Cent) and Pop (-46.121 ± 20.791 Cent) for medium frequencies ($p = 0.035$). However, other than that the main effect of *Frequency-Region* was not significant, thus, it can be concluded that different frequency regions were not accountable for inducing any substantial dependencies. Consequently, data is shown averaged over frequency regions of resonances.

Results displayed in Figure 7.12 enable to draw conclusions about the impact of various degrees of visual bias on the mean center-frequency error. Clearly, the relation between error and offset appears to be linear in a sense that larger offsets provoked equally larger errors. The results must be understood in such a way that negative offsets induced inverse frequency errors of equal magnitude, if subjects did act solely on basis of the display and neglected what their ears were taking in. The

main discovery with the data from this experiment is that the presentation of the plug-in's interface evidently affected the adjustments made by the subjects, even though the displayed EQ-curve was perhaps not in consensus with what they heard. This finding may imply that participants relied on what was displayed on the screen and at least partly took into account the visual information for their decision.

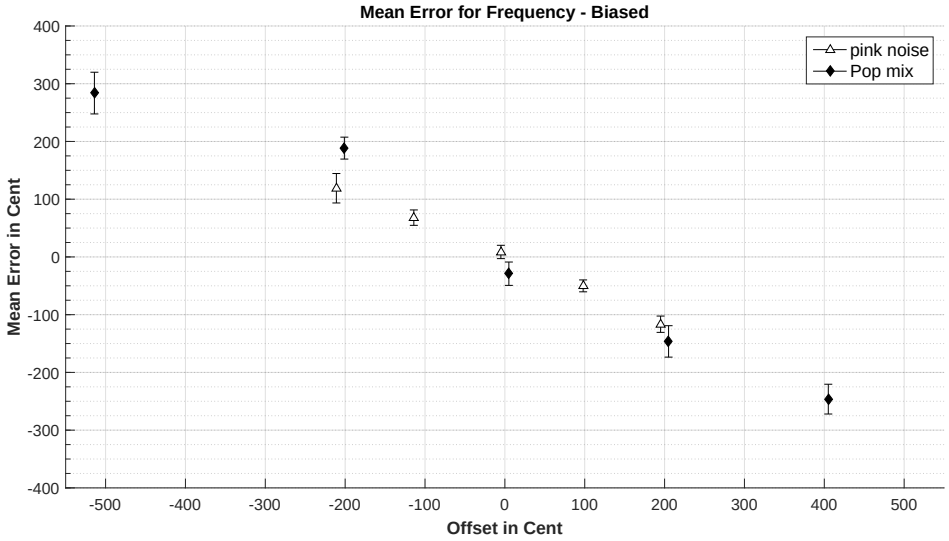


Figure 7.12: Mean frequency error as a function of visual bias. Subjects were provided with both auditory and visual information about the frequency spectrum of the test signals. The displayed center-frequency of the EQband was distorted by various degrees of positive and negative offsets. The musical asymmetry about the y-axis can be explained by the choice of symmetric offsets in percent. For instance, an offset of +25.9% corresponds to 4 semitones, whereas an offset of -25.9% corresponds roughly to 5 semitones. Error bars indicate the standard error of the mean.

Considering the results related to the two different test signals, the overall impression arises that the amount of visual influence was slightly affected by the nature of the stimuli being matched, as indicated by the two way interaction between *Offset* and *Stimulus*. For an offset of -200 Cent, pairwise comparisons revealed that the error was indeed marginally statistically higher for trials involving the Pop mix (188.477 ± 18.988 Cent) than pink noise (118.971 ± 25.503 Cent), $p = 0.054$.

Generally, it is hard to find a characteristic measure for the degree of visual influence comparable to the measure introduced in the course of the statistical analysis of the loudness matching experiment. It must be noted that in case of the spectral matching experiment, the impact on the overall error is much harder to predict due to the complex relationship of center-frequency error and filter-gain error. Therefore, their combined effect on the overall error will be investigated in section 7.3.7.

So far, only the common offsets were taken into account for the statistical analysis. Consequently, a set of one-way repeated measures ANOVAs on frequency error was performed in order to identify the differences elicited by all possible offsets per stimulus. An inspection of boxplots confirmed that there were no outliers in the data, and data was normally distributed for every combination of *stimulus* and *offset*, as assessed by Shapiro-Wilk's test of normality ($p > 0.05$).

Sphericity was not met for responses originating from pink noise test signals ($\chi^2(2) = 30.099, p = 0.001$, Mauchly's test), thus, a Greenhouse-Geisser correction ($\varepsilon = 0.47$) was used to correct the output of the ANOVA. Center-frequency errors were statistically significantly different depending on the various offsets applied to the EQ-curve, $F(1.882, 20.700) = 31.115, p < 0.0005, \eta^2 = 0.739$. Post hoc tests with Bonferroni corrections revealed that the errors were statistically significantly different between all offsets ($p < 0.05$), except for the difference between distortions of -200 Cent and -100 Cent ($p = 0.67$).

Similar results were found by analysing data originating from Pop mix test signals, though in this case sphericity was met ($\chi^2(2) = 11.722, p = 0.236$). Again, frequency error was statistically significantly different for varying amounts of visual bias, $F(2, 44) = 69.439, p < 0.0005, \chi^2 = 0.863$. The differences in frequency error for offsets were all statistically significant ($p < 0.05$), except for differences between neighbouring offset at the edges (-500 Cent vs. -200 Cent: $p = 0.20$, 200 Cent vs. 400 Cent: $p = 0.094$).

Figure 7.13 illustrates the mean absolute filter gain error subject to center-frequency offset of the displayed EQ-curve. The overall picture confirms that the varying amount of visual bias did not affect the filter gain adjustments in a meaningful way. This tendency was confirmed by the results of one-way ANOVAs on the absolute filter-gain error for pink noise ($p = 0.438$) and for the Pop mix ($p = 0.465$). Based on the findings in chapter6.3.3, it is not surprising that the gain error was generally higher for the Pop mix than for noise ($p = 0.011$).

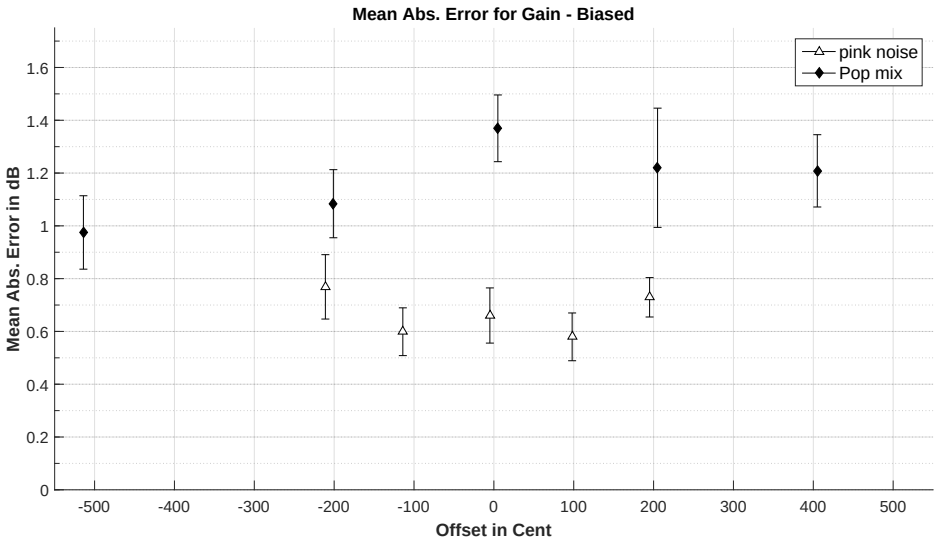


Figure 7.13: Mean absolute gain error as a function of visual bias (applied to center-frequency). Error bars indicate the standard error of the mean.

7.3.6 A Characteristic Measure for the Degree of Visual Influence

As discussed in the previous section, one should bear in mind that the error related to center-frequency discrimination illustrated in Figure 7.12, cannot be interpreted as a total error. This is because it neglects the remaining error which may have occurred due to misadjustments of filter-gain. However, the inspection of Figure 7.13 gives evidence that filter-gain settings were not significantly different among the various offset conditions, which was confirmed by the results of an analysis of variances for both stimuli, $F(1.971, 21.681) = 0.085, p = 0.916, \varepsilon = 0.47$ (Greenhouse Geisser) and $F(4, 44) = 1.126, p = 0.356$, respectively. Thus, it can be argued that in this context an introduction of a characteristic measure for the degree of visual influence, by means of center-frequency error, is feasible as well.

Referring to the derivation described in more detail in section 6.3.5, the *Degree Of Visual Influence* for every trial was assessed by dividing the center-frequency error by the corresponding inverse visual offset. All values were corrected beforehand without affecting their internal relation by subtracting the mean of the zero-offset conditions, for both stimuli respectively. As such, the center-frequency errors were related to the corresponding mean of the zero-offset conditions as opposed to zero. As an example, if the center-frequency in the displayed EQ-curve was distorted by

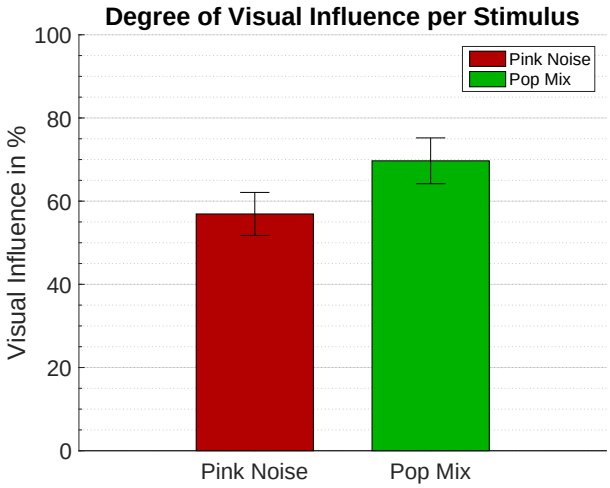


Figure 7.14: Characteristic measure for the degree of visual influence as a function of test signal. A value of 100% may indicate that information received from the auditory sensory was completely dominated by vision. Conversely, a value of 0% may be interpreted as that subjects were not affected by the presentation of visual cues at all.

-100 Cent, and the center-frequency error was +50 Cent, the visual influence would be $\frac{50}{-(-100)} \cdot 100\% = 50\%$.

The mean Degree of Visual Influence, as shown in Figure 7.14, was analysed statistically by means of a paired-sample t-test. No outliers were found in the data, as assessed by inspection of a boxplot. The differences between the degree of visual influence in the pink noise and Pop-mix trials were normally distributed, as assessed by Shapiro-Wilk's test ($p = 0.543$). According to the result of the paired-sample t-test, subjects were marginally statistically significantly stronger affected by visual cues for trials involving the Pop mix ($69.7 \pm 38.14\,\%$) compared to trials involving a pink noise test signal ($56.92 \pm 35.85\,\%$), $t(47) = 1.848, p = 0.071$.

Now, as the EQ-curve distortions were imbalanced across different stimuli, one could argue that larger average offsets might be the reason for the Visual Influence being higher for the Pop mix than for pink noise. However, when taking into account only the common offsets (-200 Cent and +200 Cent), the findings described above were even more pronounced. According to a paired-sample t-test, the *Degree Of Visual Influence* was statistically significantly higher when a Pop mix was used ($82.11 \pm 46.11\,\%$) as a test signal, compared to when noise signals were used ($58.07 \pm 34.45\,\%$), $t(23) = -2.128, p = 0.044$.

7.3.7 Evaluation of the Mean Cumulative Error

Heretofore, center-frequency errors and filter-gain errors were considered individually. In this section, the overall (or combined-) mean error is investigated whilst taking into account the interaction between the aforementioned parameters. Hence, the derivation of a characteristic measure for the *Mean Cumulative Error* is described below.

Firstly, the transfer functions for the added resonances and their associated user responses were calculated (per trial). Next, the addition of both transfer functions amounted to the difference spectrum. The cumulative error can be expressed as the area below the absolute difference spectrum, thus, the curve was integrated over a bark scale, whereby the range of the integral spanned 10 bark, placed around the center-frequency of the corresponding frequency region in which the resonance occurred, respectively (see Figure 7.15). As determined by the inspection of the absolute difference spectra, the curve was close to zero otherwise; as such the integral would not have contributed to the cumulative error outside the limits, anyway. Finally, the result of the integral was divided by 10, leading to a cumulative error in *dB per bark* for every single trial. The frequency-bark scale was necessary to guarantee that the error is weighted in accordance with the human frequency-perception. In the last step, the hereby computed cumulative errors were pooled by subject and grouped by test-signal, modality and offset (offsets of same magnitude were pooled as well).

In Figure 7.15, the derivation of the cumulative error is illustrated exemplary. Shown is the difference spectrum for all three trials of a single subject (Subject 1). No visual feedback was provided and in each trial the participant was required to remove a resonance from one of two (otherwise identical) Pop mixes, whereby the resonance was added in a different frequency region, respectively. The red curves (dashed: difference spectrum, solid: absolute difference spectrum) originate from a negative frequency error (-251 Cent) and a negative filter gain error (-1.79 dB), meaning that the cut was too big. Similarly, in case of the blue curves, the center-frequency adjusted by the user was too low (-91 Cent), but this time the cut was slightly too conservative (+0.42 dB). Lastly, referring to the green curves, the participant was able to find approximately the correct center-frequency (+28 Cent), but the cut was

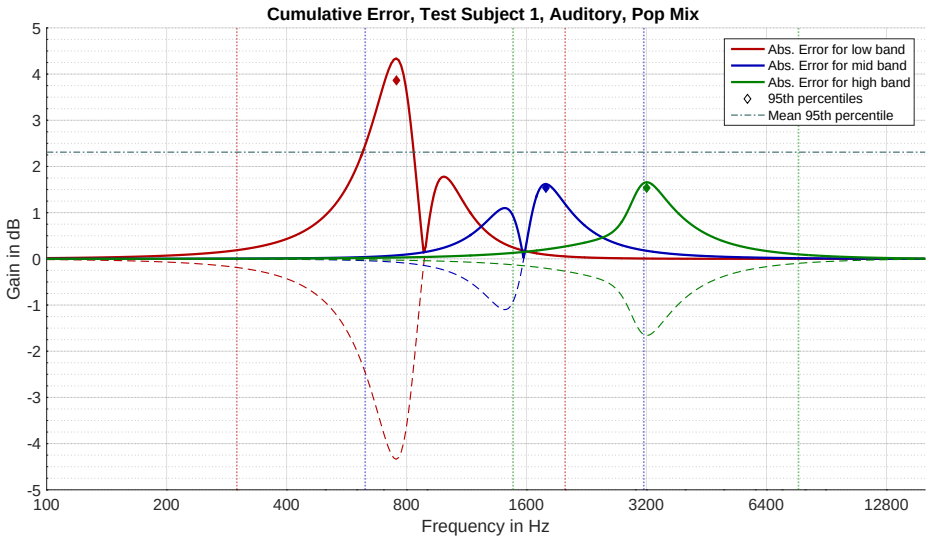


Figure 7.15: Cumulative Error with corresponding 95th percentiles. Difference spectra (dashed curves), absolute difference spectra (solid curves) and related integration limits (vertical dashed lines) are illustrated for three related trials of Subject 1 (auditory condition, Pop-mix).

again a little too drastic (-1.43 dB).

Now, the areas below the absolute difference spectra were computed by forming the integral over a bark scale within the integration bounds indicated by vertical dashed lines of the same colour, respectively. The frequency to bark transformation was performed according to [98]:

$$bark = \frac{26.81}{1 + \frac{1960}{f}} - 0.53 \tag{7.1}$$

Afterwards, for every curve the cumulative error was normalized per bark and averaged per subject, condition and test signal (for the example shown in Figure 7.15, the average cumulative error for Subject 1 was 0.76 dB per bark). Therefore, the *Mean Cumulative Error* can be described as the mean of the integrals over absolute differential spectra per bark.

The illustration of the *Mean Cumulative Error* in Figure 7.16 indicates a discernible pattern. Regardless of the nature of test signals being compared, the error was always smallest as long as participants were provided with unbiased visual presentations of the EQ-curves and real-time spectrum analysers. It is striking that ad-

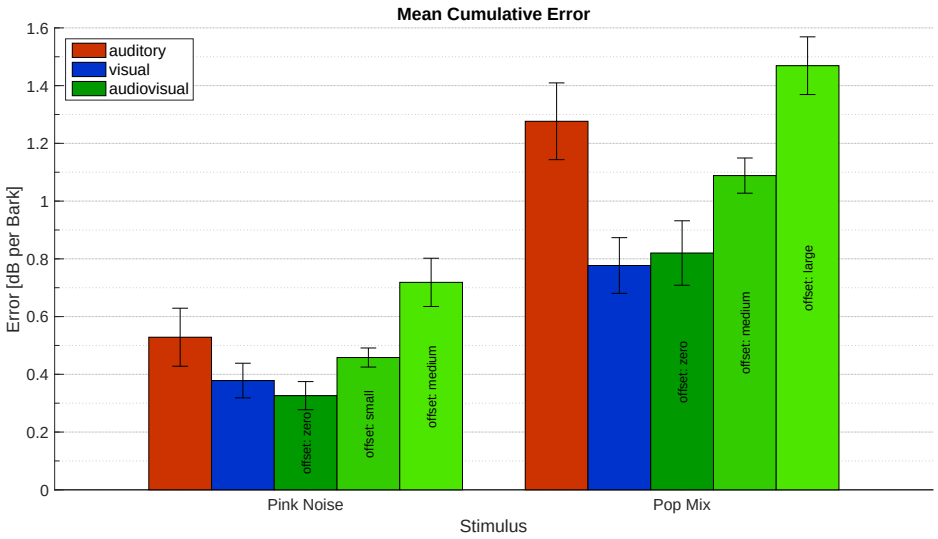


Figure 7.16: Mean Cumulative Error, taking into account the relationship between the center-frequency error and the filter-gain error. It can be interpreted as a measure for the difference between the combined transfer function of resonance filter and user filter, and an ideal flat transfer function. More precisely, the Mean Cumulative Error is defined as the mean area below the absolute differential spectra in dB per bark.

ditional auditory information did not further improve accuracy, as can be inferred from almost identical errors related to the visual- and the unbiased audio-visual condition.

Moreover, as anticipated, the cumulative error increased with increasing amount of visual distortion. Interestingly, it was in about the same range when the task had to be performed without any visual feedback as well. Figure 7.16 also enables comparisons between responses associated with different test-signals, and the overall picture confirms the findings from the previous chapters, that resonances could be removed more accurately from a noise signal than from a music mix. The Mean Cumulative error spanned ranges of about 0.3 to 0.7 dB/bark and 0.8 to 1.5 dB/bark, respectively.

The result of a two-way repeated measures ANOVA on cumulative error revealed a statistically significant higher-order interaction between *Stimulus* and *Condition*, $F(3, 33) = 3.022, p = 0.03, \eta^2 = 0.216$. Only shared offset groups (zero, medium) were included in this analysis for obvious reasons. The assumptions of normality, no outliers and sphericity were met, as assessed by the appropriate tests already utilized multiple times in the course of this thesis. As surmised by inspection of Figure 7.16,

participants were able to remove resonances more accurately from a noise source than from a full mix. This effect was also highly statistically significant, as assessed by pairwise comparison with Bonferroni correction between stimuli tested under same conditions ($p < 0.005$).

Afterwards, within-subjects ANOVAs were run for each test signal, including all of the five associated experimental conditions, respectively. *Mean Cumulative Error* was normally distributed if the Pop mix was used as test signals, and normally distributed after a square root transformation if pink noise was used, as confirmed by Shapiro-Wilk's test of normality ($p > 0.05$). No outliers were found and the assumption of sphericity was met, as assessed by Mauchly's test of sphericity (pink-noise: $\chi^2(2) = 13.614, p = 0.143$, Pop mix: $\chi^2(2) = 5.682, p = 0.775$). The results of the ANOVAs verified the assumption that there was statistically significant difference in *Mean Cumulative Error* between the five conditions illustrated in Figure 7.16, regardless of what stimuli had been used, $F(4, 44) = 8.005, p < 0.0005$ (pink noise), $F(4, 44) = 11.480, p < 0.0005$ (Pop Mix).

Next, post hoc tests with Bonferroni correction were performed separately for both types of stimuli. With reference to pink noise and deviations caused by different levels of visual bias, pairwise comparisons indicated that the Mean Cumulative Error was statistically significantly lower if the displayed user-adjusted center-frequency was undistorted, compared to medium offsets (audio-visual zero offset vs. audio-visual medium offset: $p = 0.002$, audio-visual medium offset vs. visual: $p = 0.026$). Further, differences between small offsets and medium offsets were at least marginally statistically significant ($p = 0.053$). No significant differences regarding the *Mean Cumulative Error* were found between unbiased conditions.

Similar results to those described above could be obtained in conjunction with the Pop mix source file. Bonferroni corrected pairwise comparisons among experimental conditions involving both sight and hearing revealed a statistically significantly decrease of the *Mean Cumulative Error* from $1.47 \pm 0.35\,\mathrm{dB/bark}$ in case of a large visual bias (4-5 semitones) to $1.09 \pm 0.21\,\mathrm{dB/bark}$ ($p = 0.022$) in case of a medium offset (2 semitones), and from to $1.09 \pm 0.21\,\mathrm{dB/bark}$ to $0.82 \pm 0.39\,\mathrm{dB/bark}$ between medium and zero offset ($p = 0.006$). Regarding differences among conditions involving different modalities, the overall error was significantly smaller in pres-

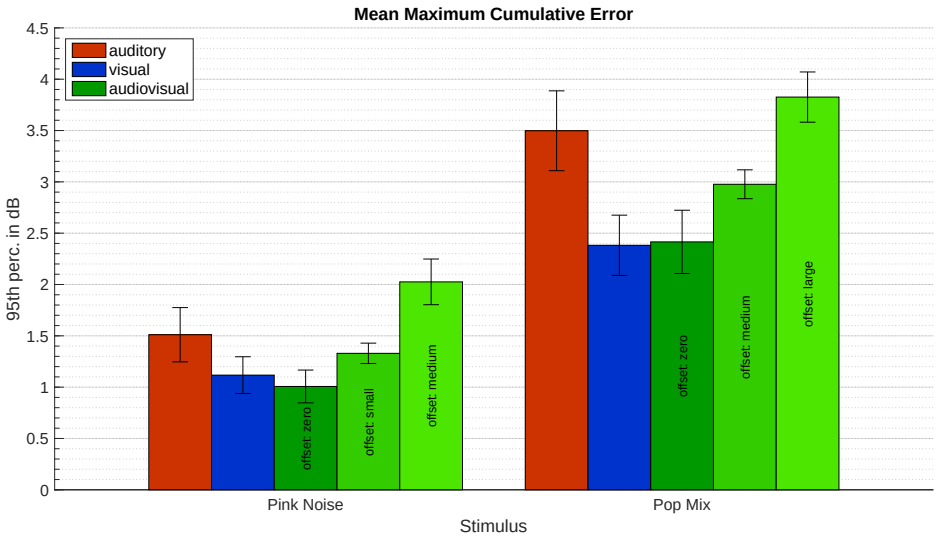


Figure 7.17: Mean Maximum Cumulative Error per condition and stimulus, as a characteristic measure for the audibility of the spectral matching error. Error bars represent the standard error of the mean.

ence of visual information ($p = 0.025$) and in presence of both visual and auditory information ($p = 0.036$), compared against the unimodal auditory condition.

In Figure 7.17, the *Mean Maximum Cumulative Errors*, or more specifically, the mean 95th percentiles of the absolute difference spectra, are shown. The derivation of this measure is based on the same principles as described in the context of the derivation of the *Mean Cumulative Error*, except that instead of the integral, the 95th percentiles are calculated between the limits indicated by the corresponding vertical dashed lines in Figure 7.15. For instance, the 95th percentile for the red curve in Figure 7.15 is calculated within the interval spanned by the vertical, red dashed lines, that is 3 Bark to 13 Bark or 300 Hz to 2000 Hz).

As discerned from Figure 7.17, a feature of great interest is the severity of the maximum errors themselves. In particular, the findings imply that on average, the strongest irregularities in the difference transfer function were in the range of 1 dB to 1.5 dB in case of pink noise and between about 2.5 dB and 3.5 dB for the Pop mix. Other than that, any decisive new insights can be gained from the evaluation of the *Mean Maximum Cumulative Error*, as the error proportions among different conditions and stimuli appear to be very similar to those in Figure 7.16.

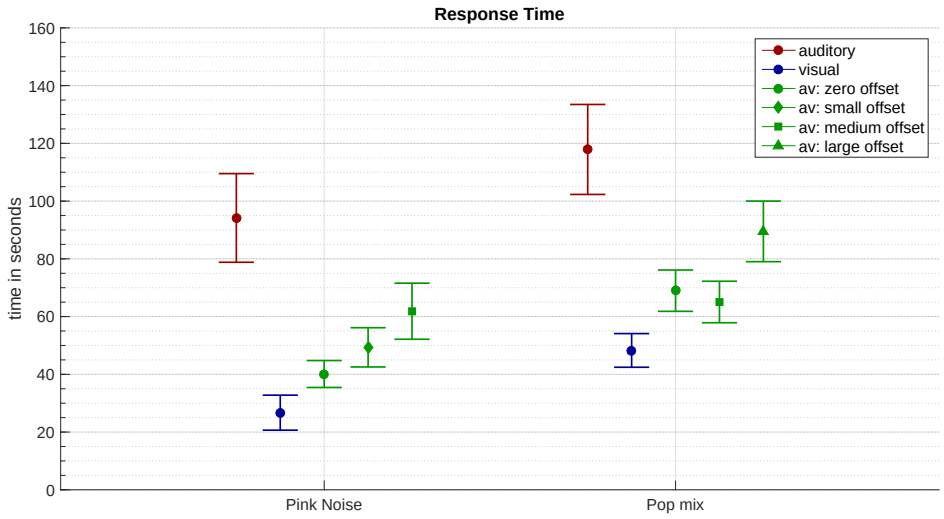


Figure 7.18: Response times per trial for the individual conditions of the experiment. Error bars represent the standard error of the mean.

7.3.8 Response Time

Test signals were played back in a loop until subjects approved their settings by pressing a button on the MIDI-controller. Further, there was no time limit to complete the task, leading to an average response time per trial of about 66 seconds. In Figure 7.18, the response times were grouped by test signal and experimental condition, whereby similar positive and negative offsets were pooled.

Firstly, the overall impression arises that various degrees of visual bias did not noteworthy affect the average time required to complete a trial, yet there was a slight tendency that time taken increased with increasing offsets. Secondly, response times were longest if sight was not involved and shortest if hearing was not involved. Regarding the condition demanding both modalities, the average time taken was in between the unimodal response times. Lastly, it appears that participants took slightly more time to equalize a full Pop mix compared to a noise signal.

A two-way repeated measures ANOVA was run in order to determine whether the dependent variable response time was statistically significantly affected by the independent variables stimulus and experimental condition (modalities involved). There were no outliers, as assessed by inspection of boxplots. Response times were normally distributed ($p > 0.05$) except for the [noise, visual] experimental condition ($p = 0.002$) and the [Pop, audio-visual zero-offset] condition ($p = 0.032$), as as-

sessed by Shapiro-Wilk's test. However, these violations of normality were considered to be not severe enough to warrant data transformation. Sphericity was not met for the interaction effect, as assessed by Mauchly's test, $\chi^2(2) = 18.848, p = 0.029$, thus, a Greenhouse-Geisser correction was used ($\varepsilon = 0.615$). No statistically significant two-way interaction was found between stimulus and condition, $F(2.460, 27.065) = 0.495, p = 0.653$. Both main effects showed a statistically significant difference in response time, $F(1, 11) = 19.084, p = 0.001$ (*Stimulus*), $F(1.853, 20.378) = 14.818, p < 0.0005$ (*Condition*). Bonferroni corrected post hoc test revealed, that participants took statistically significantly more time to remove resonances from a Pop mix than from a noise signal ($p = 0.001$). Further, response times were statistically significantly lower for trials involving vision exclusively than for trials involving hearing only ($p = 0.002$).

7.3.9 Discussion

Participants were asked to remove a single additive resonance from either pink noise or a full music mix including drums, bass, orchestra and vocals. They were allowed to compare the signals against an, apart from the added resonance, identical reference signal. During the experiment, subjects were required to perform this task under various experimental conditions employing visual or auditory modalities, or both modalities at the same time. In the context of multimodal perception, visual information, which was displayed in the form of a visualized EQ-curve and underlying real-time spectrum analysers, was deliberately distorted by various degrees of offset. The central research question was whether and to what extend the presence of visual stimuli reflecting audio parameter changes affects the process of sound mixing. In this section, the findings of the described experiment are summarized and discussed.

Firstly, Hypothesis 1 was confirmed, as subjects were able to remove resonances more accurately from a noise signal than from a musical piece, as indicated by the mean errors for center-frequency and filter-frequency, the mean cumulative error and response times. This observation confirms the findings of several authors [11, 50, 62, 97], that the perceptual resonance threshold level is lowest when dealing with noisy signals.

Secondly, results observed without visualization of audio parameters can be used to draw conclusions about the ability to aurally identify changes in the frequency spectrum of audio signals. Referring to Figure 7.10, the mean center-frequency error was in the region of 60 Cents for pink noise, which is slightly lower compared to the findings of a study on resonance-frequency discrimination by Gagné and Zurik [29]. According to their results, the just-noticeable difference for changes in resonance-frequency between 300 Hz and 2000 Hz induced by a second-order formant filter (producing peaks of about 7-8 dB) can be summarized by $\Delta f = 0.079 \cdot \frac{f_r}{\sqrt{Q}}$, which corresponds to a just-noticeable difference of about 90 Cents. Moreover, the nature of this formula coincides with the assessment that the comparative center-frequency error is not dependent on the center-frequency of the resonance. The authors further concluded that their formula holds for both periodic and white noise signals. However, the center-frequency error of about 220 Cent for a Pop mix, as illustrated in Figure 7.10, may signify that performance is significantly worse for fluctuating real-world signals.

With respect to filter-gain errors, the mean 95th percentiles of the *Maximum Cumulative Error* (see Figure 7.17) might be a good indicator for the amount of energy required for irregularities in the frequency spectrum to be detectable by the human ear. The results may imply that the just noticeable difference threshold for frequency peaks or valleys with varying bandwidths (resulting difference spectra can produce irregularities with bandwidths narrower than the bandwidth of the initial resonance, see Fig 8.10) is in the range of 1 dB to 1.5 dB for pink noise and in the range of 2.5 dB and 3.5 dB for a full Pop mix. However, this contradicts the widespread belief, especially related to audio mastering, that "slight EQ-adjustments can have a big impact on the overall sound" [7, 59], and that "with good monitoring, equalization changes of less than one half dB are audible" [46]. Further, in a study related to headphone equalization [79], the just-noticeable difference for one-third octave filter peaks was found to be around 1 dB for frequencies above 500 Hz. Likewise, peaks of 1 dB could be perceived for certain audio material, as confirmed by a study on the audibility of comb filter distortions [11]. On the other hand, it must be noted that for the experiment presented in this thesis, subjects were not only asked to detect resonances, but rather they were required to identify and remove them. As mentioned earlier, modifying the parameters of a semi-parametric EQ-band in or-

der to match the spectrum of two sound files is quite a challenging task, especially due to the fact that the impact on the cumulative error strongly depends on the relationship between center-frequency and filter-gain. Although subject might still have perceived a difference between the test signal and the reference, they probably stopped because they could not identify whether the difference was due to a mismatch of center-frequencies or filter-gains. Moreover, as shown in Figure 7.15, the 95th percentiles can be traced back to both peaks and valleys in the spectrum. According to research by Brücklein [12], "peaks are far more audible than valleys or dips". For these reasons, the observed mean 95th percentiles of the maximum error might be higher than the just-noticeable difference threshold observed in other studies cited above.

The readout of task-relevant information provided through the graphical user interface on the equalizer plug-in clearly improved the ability to find the correct center-frequency, but additional information from the auditory modality did not further improve precision in this regard. Furthermore, filter-gain errors tended to be positive, which can be explained by the fact that subjects progressively attenuated the gain in order to compensate the resonance, meaning that they generally approached their final setting in negative direction. Although this effect was observable only for the unimodal experimental conditions, it appears difficult to argue towards an improvement in accuracy as a consequence of bi-modal integration, as the filter gain error expressed in absolute values was about the same for different modalities involved (see Figure 7.11).

Results from the statistical analysis confirmed the suspicion that the ability to remove audible resonances from different test signals was reliant upon the accurateness of visual information reflecting audio parameter changes, confirming Hypothesis 4. Errors in center-frequency were linear-proportional to various levels of distortion applied to the center-frequency of the displayed equalizer curves. These offsets induced a mismatch between information received from vision and hearing, as changes in center-frequency manifested differently in the displayed real-time spectrum compared to how they were perceived aurally. Referring to the insignificant differences between response times obtained for audio-visual conditions and the linear behaviour of the center-frequency error as a function of offset, it can be assumed that the manipulated visual displays did not cause suspicion. As expected, negative

center-frequency offsets provoked positive center-frequency errors, and the error increased with increasing offsets. Further, it was shown that filter-gain settings were not attributable to the varying amount of visual bias. Therefore, it was possible to quantify the *degree of visual influence* solely by means of the center-frequency error. The high values obtained for this arguably fictional measure, around 55% for pink noise and as high as 70% for the Pop mix, clearly indicated a predominance of vision. Taking into account the finding that equalizing with (undistorted-) visual support has led to significantly improved performance, allows the presumption that subjects found the visual information particularly helpful to perform the task faster and more accurately (cf. Figures 7.16 to 7.18). Deliberately or not, based on this cognition and their previous experience, they might have given more weight to the visual stream than to the auditory stream. Further, visual influence was generally lower in the single-parametric dimensional loudness matching task, and higher for musical signals compared to noise signals. For all these reasons, the degree of visual influence might well be a function of the task's level of difficulty, which would confirm Hypothesis 5.

Finally, inspection of cumulative errors showed that performance was even improved in presence of moderately distorted visualizations, compared to scenarios that did not involve sight. This finding confirms Hypothesis 2 and implies that even a display with low precision, e.g. low frequency resolution, may still improve accuracy in an equalisation task. In practice, this suggests that whenever engineers run out of system resources during a session, they could reduce the frequency-resolution of real-time spectrum analysers without risking impairment of their mix.

Chapter 8

Conclusion

The purpose of this thesis was to identify and quantify the impact of displayed graphical user interfaces visualizing audio parameter changes and characteristics of the sound sources on an engineer's sound mixing decisions. After a short introduction to the topic in Chapter 1, the processes involved in mixing and mastering, their historical evolution and some important qualities of a typical mixing environment were reviewed in Chapter 2. Established user interface paradigms in audio processing were examined in Chapter 3, emphasizing on interface design strategies to avoid distractions and to reduce working memory load. Chapter 4 outlined the background necessary to understand the cross-modal interactions and summarized some attention mechanisms related to simultaneous presentation of vision and sound. Findings of existing studies related to the present thesis were presented in Chapter 5. Chapters 6 and 7 focused on methodology and data evaluation. Experimental designs of both the loudness matching and the spectral matching task were reported in great detail. Additionally, in Sections 6.3 and 7.3, results of both experiments were evaluated by means of a statistical analysis, respectively. In this section, the main findings of this thesis are revisited.

Data from both experiments gives clear evidence that mixing decisions were affected by the presence of task relevant visual cues. Mismatch between information streams received from vision and hearing, induced by purposely manipulated graphical displays, led to statistically significantly different responses for varying amounts of visual bias. The impact of manipulated visual stimuli on sound mixing was

stronger in the equalizing task than in the loudness matching task, and stronger for musical material than for noise signals. Hence, comparisons between results from both tasks suggest that streams coming from different sensory modalities may be weighted by their reliability, as it was shown that predominance of vision increased with increasing aural complexity of the task, which might well be explained by the modality appropriateness hypothesis [2, 99, 104] (cf. Section 4.1.1). Comparisons among unbiased experimental conditions involving different modalities confirm this observation. During the loudness experiment, information from both senses was apparently equally weighted, whereas vision prevailed hearing during the arguably more difficult equalization task, as inferred from almost identical errors between the audio-visual and the unimodal visual condition and significantly worse performance in case of the unimodal auditory condition. Therefore, evaluation of data obtained from two different mixing tasks performed with and without visual feedback could signify that the degree of visual influence might increase with the level of difficulty of the task.

Other than that, some task specific findings can be drawn on the results from the individual experiments. Regarding the accuracy of the EBU-R128 recommendation, a mismatch of about one dB was discovered between perceived and measured loudness of a full Pop mix. Further, results confirmed that subjects were more sensitive to differences in loudness when dealing with noisy signals compared to differences in commercial music material. However, the use of source files from different genres barely affected sensitivity. Moreover, results from the spectral matching experiment indicate that the perceptual resonance threshold was lowest for noise signals. As a consequence of center-frequency and filter-gain misadjustments, the remaining maximum absolute cumulative error in the difference spectrum was around 1 dB to 1.5 dB for pink noise, and around 2.5 dB to 3.5 dB for commercial music.

In conclusion, the results of this thesis should raise awareness among audio engineers that mixing with visual support can affect decisions during mixing. Engineers should keep in mind that attending both visual and auditory information may change the perception of sound. Therefore, involving sight to a smaller extend and modifying audio effect parameters with eyes closed from time to time may help to focus better on the sound.

Concerning future work, further research can be done to evaluate the impact of visual stimuli on the process of sound mixing with more focus on aesthetics. In the course of this thesis, audio-visual interaction effects have been investigated by means of two isolated, rather technical mixing tasks. In a subsequent step, it would be interesting to examine the influence of visual cues in a real-world mixing scenario, where multiple tracks need to be processed with different types of audio effects. Further, the question arises if the amount of visual impact might also be dependent on the nature and layout of the graphical user interface, for instance, if the effect will be the same on analogue mixing consoles. Hereof, this work can be seen as a basis for future research.

Bibliography

[1] 112dB Website. Redline equalizer - eq response curves. https://112db.com/redline/equalizer/?specifications. Assessed: 2016-04-09.

[2] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.

[3] Jon Allan and Jan Berg. Audio level alignment-evaluation method and performance of ebu r 128 by analyzing fader movements. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[4] Gisa Aschersleben, Talis Bachmann, and Jochen Müsseler. *Cognitive contributions to the perception of spatial and temporal events*. Elsevier, 1999.

[5] Bernhard Baier, Andreas Kleinschmidt, and Notger G Müller. Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information. *The Journal of Neuroscience*, 26(47):12260–12265, 2006.

[6] Fabian Begnert, Håkan Ekman, and Jan Berg. Difference between the ebu r-128 meter recommendation and human subjective loudness perception. In *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[7] B. Benediktsson. Eq - the most important part of mastering. http://music.tutsplus.com/tutorials/. Assessed: 2016-04-16.

[8] Anne-Marie Bonnel and Ervin R Haftser. Divided attention between simultaneous auditory and visual signals. *Perception & Psychophysics*, 60(2):179–190, 1998.

[9] Adelbert W Bronkhorst and Tammo Houtgast. Auditory distance perception in rooms. *Nature*, 397(6719):517–520, 1999.

[10] Julie M Brown, Krista L Anderson, Carol A Fowler, and Claudia Carello. Visual influences on auditory distance perception. *The Journal of the Acoustical Society of America*, 104(3):1798–1798, 1998.

[11] Stefan Brunner, Hans-Joachim Maempel, and Stefan Weinzierl. On the audibility of comb filter distortions. In *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

[12] Roland Bücklein. The audibility of frequency response irregularities. *Journal of the Audio Engineering Society*, 29(3):126–131, 1981.

[13] Marcio C Cabral, Carlos H Morimoto, and Marcelo K Zuffo. On the usability of gesture interfaces in virtual reality environments. In *Proceedings of the 2005 Latin American conference on Human-computer interaction*, pages 100–108. ACM, 2005.

[14] Esteban R Calcagno, Ezequiel L Abregu, Manuel C Eguía, and Ramiro Vergara. The role of vision in auditory distance perception. *Perception*, 41(2):175–192, 2012.

[15] Gemma Calvert, Charles Spence, and Barry E Stein. *The handbook of multisensory processes*. MIT press, 2004.

[16] Mark Cartwright, Bryan Pardo, and Josh Reiss. Mixploration: Rethinking the audio mixer interface. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 365–370. ACM, 2014.

[17] G. Cochrane. Mixing with your eyes closed. http://therecordingrevolution.com/2011/01/31/mixing-with-your-eyes-closed/, 2011. Assessed: 2015-07-30.

[18] G. Cochrane. The case for mixing with your eyes closed. http://therecordingrevolution.com/2013/06/28/the-case-for-mixing-with-your-eyes-closed/, 2013. Assessed: 2015-07-30.

[19] Jason Corey. *Audio production and critical listening: Technical ear training*. CRC Press, 2012.

[20] David Cronin. Feature into the groove: lessons from the desktop music revolution. *interactions*, 15(3):72–78, 2008.

[21] Gibson David. The art of mixing. *London, Continental*, 2009.

[22] Frederic Dehais, Catherine Tessier, and Laurent Chaudron. Ghost: experimenting countermeasures for conflicts in the pilot's activity. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2003.

[23] Jon Driver and Charles Spence. Crossmodal attention. *Current opinion in neurobiology*, 8(2):245–253, 1998.

[24] Konstantinos Drossos, Andreas Floros, and Konstantinos Koukoudis. Gestural user interface for audio multitrack real-time stereo mixing. In *Proceedings of the 8th Audio Mostly Conference*, page 2. ACM, 2013.

[25] Matthew Duignan. Computer mediated music production: A study of abstraction and activity. 2008.

[26] Hugo Fastl et al. Audio-visual interactions in loudness evaluation. In *Proceedings of the 18th International Congress on Acoustics*, volume 2, pages 1161–1166. Citeseer, 2004.

[27] Stephen Few. Save the pies for dessert. *Visual Business Intelligence Newsletter*, pages 1–14, 2007.

[28] Sebastian Fichtner. What music composition interfaces require. Master's thesis, University Of Konstanz, 2014.

[29] Jean-Pierre Gagné and PM Zurek. Resonance-frequency discrimination. *The Journal of the Acoustical Society of America*, 83(6):2293–2299, 1988.

[30] Gartner. Gartner says tablet sales continue to be slow in 2015. http://www.gartner.com/newsroom/id/2954317, 2015. Assessed: 2015-08-04.

[31] JW Gebhard and GH Mowbray. On discriminating the rate of visual flicker and auditory flutter. *The American journal of psychology*, 72(4):521–529, 1959.

[32] Kristian Gohlke, Michael Hlatky, Sebastian Heise, David Black, and Jörn Loviscach. Track displays in daw software: Beyond waveform views. In *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[33] Samuel W Greenhouse and Seymour Geisser. On methods in the analysis of profile data. *Psychometrika*, 24(2):95–112, 1959.

[34] Sharon E Guttman, Lee A Gilroy, and Randolph Blake. Hearing what the eyes see auditory encoding of visual temporal sequences. *Psychological science*, 16(3):228–235, 2005.

[35] Frank Halasz and Thomas P Moran. Analogy considered harmful. In *Proceedings of the 1982 conference on Human factors in computing systems*, pages 383–386. ACM, 1982.

[36] T Hashimoto and S Hatano. Effects of factors other than sound to the perception of sound quality. *17th ICA Rome, CD-ROM*, 2001.

[37] Cyrus J. Heiduska. Different eq types explained. http://www.ovnilab.com/articles/eqtypes.shtml. Assessed: 2016-02-16.

[38] Ruth Herbert. *Everyday music listening: Absorption, dissociation and trancing*. Ashgate Publishing, Ltd., 2011.

[39] Jay Hodgson. *Understanding Records: A Field Guide to Recording Practice*. Bloomsbury Publishing, 2010.

[40] David Miles Huber and Robert E Runstein. *Modern recording techniques*. CRC Press, 2013.

[41] C. Huff. Cannot get the mix right: Blame your eyes. http://www.behindthemixer.com/cant-get-mix-right-blame-eyes/. Assessed: 2015-07-30.

[42] Roey Izhaki. *Mixing audio: concepts, practices and tools*. Taylor & Francis, 2013.

[43] Jennifer A Johnson and Robert J Zatorre. Attention to simultaneous unrelated auditory and visual events: behavioral and neural correlates. *Cerebral Cortex*, 15(10):1609–1620, 2005.

[44] J. Kadis. Digital audio workstations. https://ccrma.stanford.edu/courses/192b/ProTools-Logic%20Lec.pdf, 2015. Assessed: 2015-07-13.

[45] Bob Katz. Integrated approach to metering, monitoring, and leveling practices, part 1: Two-channel metering. *Journal of the Audio Engineering Society*, 48(9):800–809, 2000.

[46] Bob Katz and Robert A Katz. *Mastering audio: the art and the science*. Butterworth-Heinemann, 2003.

[47] HJ Keselman, Joanne C Rogan, Jorge L Mendoza, and Lawrence J Breen. Testing the validity conditions of repeated measures f tests. *Psychological Bulletin*, 87(3):479, 1980.

[48] Gerald R Khoury and Simeon J Simoff. Elastic metaphors: expanding the philosophy of interface design. In *Selected papers from conference on Computers and philosophy-Volume 37*, pages 65–71. Australian Computer Society, Inc., 2003.

[49] Mario Köppen, Kaori Yoshida, and Pablo A Valle. Gestalt theory in image processing: A discussion paper. In *Proceedings of the 2007 IEEE Three-Rivers Workshop on Soft Computing in Industrial Applications*, pages 1–3, 2007.

[50] W Kuhl. Unterschiedliche bedingungen beim hören in einem raum und bei elektroakustischen übertragungen. *Rundfunktechnische Mitteilungen*, 13(5):205–208, 1969.

[51] Axel Larsen, William McIlhagga, Jeroen Baert, and Claus Bundesen. Seeing or hearing? perceptual independence, modality confusions, and crossmodal congruity effects with focused and divided attention. *Perception & psychophysics*, 65(4):568–574, 2003.

[52] H. Latham-Koenig. 20 audio mixing techniques you can experiment with to improve your mixes... http://www.clickmastering.com/audio-mixing.html. Assessed: 2015-07-30.

[53] Paul J Laurienti, Jonathan H Burdette, Mark T Wallace, Yi-Fen Yen, Aaron S Field, and Barry E Stein. Deactivation of sensory-specific cortex by crossmodal stimuli. *Journal of Cognitive Neuroscience*, 14(3):420–429, 2002.

[54] Michal Lech and Bozena Kostek. Testing a novel gesture-based mixing interface. *Journal of the Audio Engineering Society*, 61(5):301–313, 2013.

[55] AC Lehmann. Music listening proneness moderates the effects of eyes-open versus eyes-closed music listening on emotion-related subjectives and electrocortical responses. *Proceedings of the 5th Triennial ESCOM Conference*, 2003.

[56] K. Lewis. Mixing with your eyes closed. http://www.inmusik.co/news.php?news_id=110. Assessed: 2015-07-30.

[57] Emiliano Macaluso, Chris D Frith, and Jon Driver. Modulation of human visual cortex by crossmodal spatial attention. *Science*, 289(5482):1206–1208, 2000.

[58] James SP Macdonald and Nilli Lavie. Visual perceptual load induces inattentional deafness. *Attention, Perception, & Psychophysics*, 73(6):1780–1789, 2011.

[59] E. Manigio. Eq-settings for mastering: General tips. http://www.audiorecording.me/eq-settings-for-mastering-general-tips.html. Assessed: 2016-04-16.

[60] L Martinez and Félix Guerrero. Midi gestural control of a vst plugin using an ir sensor. page 11. IEEE, 2007.

[61] Dominic W Massaro and David S Warner. Dividing attention between auditory and visual perception. *Perception & Psychophysics*, 21(6):569–574, 1977.

[62] Ivo Mateljan, Heinrich Weber, and Ante Doric. Detection of audible resonances. In *Proceedings of the Third congress of Alps Adria Acoustics Association, Graz, Austria*, 2007.

[63] John W Mauchly. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11(2):204–209, 1940.

[64] John J McDonald, Wolfgang A Teder-Sälejärvi, and Steven A Hillyard. Involuntary orienting to sound improves visual perception. *Nature*, 407(6806):906–908, 2000.

[65] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[66] Sharon Morein-Zamir, Salvador Soto-Faraco, and Alan Kingstone. Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, 17(1):154–163, 2003.

[67] Josh Mycroft, Joshua D Reiss, and Tony Stockman. The influence of graphical user interface design on critical listening skills. *Sound and Music Computing (SMC), Stockholm*, 2013.

[68] Josh Mycroft, Joshua D Reiss, and Tony Stockman. The effect of differing user interface presentation styles on audio mixing. In *Proceedings of ICMEM 2015*. International Conference on the Multimodal Experience of Music, 2015.

[69] Josh Mycroft, Tony Stockman, and Joshua D Reiss. Audio mixing displays: The influence of overviews on information search and critical listening. . *Proc. of the 11th International Symposium on CMMR*, 2015.

[70] P. Myllys. User interface paradigms in digital audio workstations. Master's thesis, University of Arts Helsinki, 2014.

[71] Chris Nash. *Supporting virtuosity and flow in computer music*. PhD thesis, University of Cambridge, 2012.

[72] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[73] Donald A Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[74] BBC News Online. How the cd was developed. http://news.bbc.co.uk/2/hi/technology/6950933.stm, 2007. Assessed: 2015-07-29.

[75] Bobby Owsinski. *The mixing engineer's handbook*. Nelson Education, 2013.

[76] Bryan Pardo, David Little, and Darren Gergle. Building a personalized audio equalizer interface with transfer learning and active learning. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 13–18. ACM, 2012.

[77] Pekka Parhi, Amy K Karlson, and Benjamin B Bederson. Target size study for one-handed thumb use on small touchscreen devices. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 203–210. ACM, 2006.

[78] Antti Pirhonen, Stephen Brewster, and Christopher Holguin. Gestural and audio metaphors as a means of control for mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 291–298. ACM, 2002.

[79] A Ramsteiner and G Spikofski. Ermittlung von wahrnehmbarkeitsschwellen für klangfarbenunterschiede unter verwendung eines diffusfeldentzerrten kopfhörers. *Fortschritte der Akustik, DAGA Aachen*, page 581, 1987.

[80] Jef Raskin. *The humane interface: new directions for designing interactive systems*. Addison-Wesley Professional, 2000.

[81] Jarrod Ratcliffe. Hand motion-controlled audio mixing interface. In *Proceedings of New Interfaces for Musical Expression*, 2014.

[82] Joshua D Reiss and Andrew McPherson. *Audio Effects: Theory, Implementation and Application*. CRC Press, 2014.

[83] R Rensink. The management of human attention in visual displays. *Human Attention in Digital Environments*, 2012.

[84] Andrew T Sabin and Bryan Pardo. 2deq: an intuitive audio equalizer. In *Proceedings of the seventh ACM conference on Creativity and cognition*, pages 435–436. ACM, 2009.

[85] Michael Schutz and S Lipscomb. Influence of visual information on auditory perception of marimba stroke types. In *ANAIS do VIII International Conference of Music Perception and Cognition*, 2004.

[86] Michael Schutz and Scott Lipscomb. Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception*, 36(6):888–897, 2007.

[87] Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. Illusions: What you see is what you hear. *Nature*, 408(6814):788–788, 2000.

[88] Matt Shelvock. *Audio Mastering as Musical Practice*. PhD thesis, The University of Western Ontario, 2012.

[89] T Shipley. Auditory flutter-driving of visual flicker. *Science*, 145(3638):1328–1330, 1964.

[90] Ben Shneiderman and Benjamin B Bederson. Maintaining concentration to achieve task completion. In *Proceedings of the 2005 conference on Designing for User eXperience*, page 9. AIGA: American Institute of Graphic Arts, 2005.

[91] Esben Skovenborg and Søren H Nielsen. Evaluation of different loudness models with music and speech material. In *Proc. of the AES 117th Convention, San Francisco*, 2004.

[92] Charles Spence. Crossmodal attention. *Scholarpedia*, 5(5):6309, 2010.

[93] Charles Spence and Jon Driver. *Crossmodal space and crossmodal attention*. Oxford University Press, 2004.

[94] Charles Spence, Jane Ranson, and Jon Driver. Cross-modal selective attention: On the difficulty of ignoring sounds at the locus of visual attention. *Perception & psychophysics*, 62(2):410–424, 2000.

[95] Laerd Statistics. One-way repeated measures anova using spss statistics - statistical tutorials and software guides.
https://statistics.laerd.com/, 2015. Assessed: 2016-03-25.

[96] William Forde Thompson, Frank A Russo, and Lena Quinto. Audio-visual integration of emotional cues in song. *Cognition and Emotion*, 22(8):1457–1470, 2008.

[97] Floyd E Toole and Sean E Olive. The modification of timbre by resonances: Perception and measurement. *Journal of the Audio Engineering Society*, 36(3):122–142, 1988.

[98] Hartmut Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1):97–100, 1990.

[99] A Väljamäe, A Kohlrausch, S van de Par, D Västfjäll, and P Larsson. Kleiner m. audio-visual interaction and synergy effects: implications for cross-modal

optimization of virtual and mixed reality applications. *Handbook of presence*, 2005.

[100] Bradley W Vines, Carol L Krumhansl, Marcelo M Wanderley, and Daniel J Levitin. Cross-modal interactions in the perception of musical performance. *Cognition*, 101(1):80–113, 2006.

[101] Jean Vroomen and Beatrice de Gelder. Temporal ventriloquism: sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance*, 30(3):513, 2004.

[102] Johannes Webers. *Handbuch der Tonstudiotechnik für Film, Funk und Fernsehen.* Franzis Verlag, 2007.

[103] DH Weissman, LM Warner, and MG Woldorff. The neural mechanisms for minimizing cross-modal distraction. *The Journal of Neuroscience*, 24(48):10941–10949, 2004.

[104] Robert B Welch and David H Warren. Immediate perceptual response to intersensory discrepancy. *Psychological bulletin*, 88(3):638, 1980.

[105] M. Wessel. Mixing with compressors. Bachelor's thesis, Lulea University of Technology, 2015.

[106] Pavel Zahorik. Estimating sound source distance with and without vision. *Optometry & Vision Science*, 78(5):270–275, 2001.

Appendices

Additional Plots

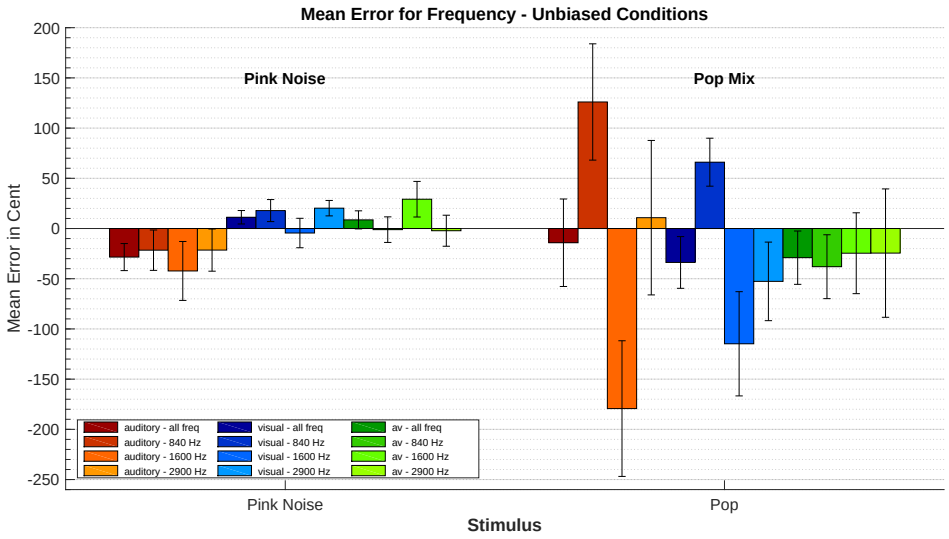


Figure 1: Mean Frequency Error for unbiased conditions (Spectral Matching experiment). Error bars represent the standard error of the mean.

Instructions

Experiment 1

In Anlehnung an einen wichtigen Arbeitsschritt beim Mastering einer Audio CD, bei dem üblicherweise alle Titel auf dieselbe Lautheit gebracht werden, soll in diesem Experiment die Lautheit zweier Audiomischungen angeglichen werden.

Zur Anpassung der Lautheit von Audiospur 1 an Audiospur 2 steht Ihnen am Midi-Controller ein einfacher Gain-Regler zur Verfügung (Endlosdrehregler ganz links am Gerät).

Die Anpassung der Lautheit soll für drei verschiedene Paarungen von Audiobeispielen erfolgen:

1) Rosa Rauschen & Rosa Rauschen
2) Poprock A & Poprock B (leicht unterschiedliche Ausschnitte mit ähnlicher Instrumentierung)
3) Poprock A & Klassik

Für jede der oben gelisteten Kategorien soll die Lautheitsanpassung mehrmals und unter verschiedenen Versuchsbedingungen durchgeführt werden. Mittels zweier Drucktaster („*Solo Spur 1*“ & „*Solo Spur 2*“) am Midi-Controller kann zwischen den beiden Audiospuren umgeschaltet werden, um deren Lautheit zu vergleichen. Die Audiobeispiele sind etwa 8 Sekunden lang und werden in einem Loop wiedergegeben, bis Sie Ihre Einstellung mit einem Druck auf die „*Weiter*“ Taste am Midi-Controller bestätigen und das nächste Beispiel abgespielt wird. Die Reihenfolge, in der die Stimulipaare auftreten, ist zufällig.

Für den Versuch gibt es kein Zeitlimit. Es ist wichtig, dass Sie konzentriert arbeiten und dass Sie die Lautheit der beiden Spuren so genau wie möglich und nach bestem Gewissen angleichen.

Insgesamt soll die Aufgabe unter drei verschiedenen Versuchsbedingungen durchgeführt werden. Die Beschreibung der individuellen Bedingungen, des Versuchsablaufes sowie der Bedienung des Controllers erhalten Sie auf separaten Beiblättern.

Versuchsbedingung A

Aufgabe:

Benutzen Sie den Gain-Regler am Midi-Controller, um die Lautheit zweier Audiobeispiele anzugleichen. Mit dem Gain-Regler steuern Sie die Lautstärke für Spur 1. Es soll jeweils nur eine Einstellung, passend für das gesamte Audiobeispiel, festgelegt werden (keine Automatisation).

Wiederholen Sie die Aufgabe 12 mal für verschiedene Paarungen von Audiobeispielen.

Hilfsmittel:

Gain Regler zur Steuerung der Lautstärke von Spur 1

Kopfhörer

Ablauf:

1. Drücken Sie den Taster „Start“ am Midi Controller, um der Versuch zu starten.
    a. Drücken Sie Play/Stop am Midi Controller, um jederzeit die Wiedergabe zu starten (oder bei Bedarf wieder anzuhalten).
2. Wählen Sie mittels der Taster „Solo Test“ und „Solo Referenz“ am Midi-Controller, welche Audiospur sie hören möchten. Mit diesen Tasten können Sie zwischen den beiden Audiobeispielen jederzeit hin- und herschalten.
3. Bewegen Sie den Endlosdrehregler „Gain“ am Midi-Controller, um die Lautstärke von „Spur 1“ zu steuern. Eine Drehung nach links bewirkt eine Pegelverringerung, eine Drehung nach rechts bewirkt eine Pegelerhöhung. Je schneller Sie den Regler drehen, desto größer die Änderungen. Ziel ist es, die Lautheit von „Spur 1“ der Lautheit von „Spur Referenz“ anzugleichen.
4. Wenn Sie mit Ihrer Einstellung zufrieden sind, drücken Sie die Taste „Weiter“ am Midi-Controller. Dies speichert Ihre Einstellung und das nächste Beispiel wird abgespielt.
5. Wiederholen Sie die Schritte 2-4, bis Sie die Lautheitsanpassung für alle 12 Audiobeispiele durchgeführt haben.

Hinweis: Sollten Sie während des Versuches eine Pause benötigen, drücken Sie bitte die Taste „Pause“, am Midi-Controller. Wenn Sie den Versuch anschließend wieder fortsetzen wollen, drücken Sie die Taste „Fortsetzen“.

Versuchsbedingung B

Aufgabe:

Benutzen Sie den Gain-Regler am Midi-Controller, um die Lautheit zweier Audiobeispiele anzugleichen. Mit dem Gain-Regler steuern Sie die Lautstärke für Spur 1. Es soll jeweils nur eine Einstellung, passend für das gesamte Audiobeispiel, festgelegt werden (keine Automatisation).

Wiederholen Sie die Aufgabe 12 mal für verschiedene Paarungen von Audiobeispielen.

Hilfsmittel:

Gain-Regler zur Steuerung der Lautstärke von Spur 1

Auf dem Bildschirm werden zwei EBU-R128 Lautheits Messgeräte angezeigt. Den aktuellen Lautheitswert in *LU* (*Loudness Units*, 1 *LU* entspricht 1 *dB*) der jeweiligen Audiospuren können Sie am jeweiligen Meter ablesen. Dieser kurzzeitig gemittelte Wert (3 Sekunden) wird nach dem EBU-R128 Standard ermittelt. Die Messeinstellungen der beiden Messgeräte sind ident.

Das obere Lautheits-Meter zeigt die aktuelle Lautheit von *Spur 1*. Das untere Lautheits-Meter zeigt die aktuelle Lautheit von *Spur Referenz.*

Hinweis: Die Lautheitsanpassung soll hier nur mit den visuellen Hilfsmitteln erfolgen, Kopfhörer sind für Kondition 2 demnach nicht erlaubt.

Ablauf:

1. Drücken Sie den Taster „Start“ am Midi Controller, um der Versuch zu starten.
    a. Drücken Sie Play/Stop am Midi Controller, um jederzeit die Wiedergabe zu starten (oder bei Bedarf wieder anzuhalten).
2. Bewegen Sie den Endlosdrehregler „Gain“ am Midi-Controller, um die Lautstärke von „Spur 1“ zu steuern. Eine Drehung nach links bewirkt eine Pegelverringerung, eine Drehung nach rechts bewirkt eine Pegelerhöhung. Je schneller Sie den Regler drehen, desto größer die Änderungen. Ziel ist es, die Lautheit von „Spur 1“ der Lautheit von „Spur Referenz“ anzugleichen.
3. Wenn Sie mit Ihrer Einstellung zufrieden sind, drücken Sie die Taste „Weiter“ am Midi-Controller. Dies speichert Ihre Einstellung und das nächste Beispiel wird abgespielt.
4. Wiederholen Sie die Schritte 2-3, bis Sie die Lautheitsanpassung für alle 12 Audiobeispiele durchgeführt haben.

Hinweis: Sollten Sie während des Versuches eine Pause benötigen, drücken Sie bitte die Taste „Pause“, am Midi-Controller. Wenn Sie anschließend den Versuch wieder fortsetzen wollen, drücken Sie die Taste „Fortsetzen“.

Versuchsbedingung C

Aufgabe:

Benutzen sie den Gain-Regler am Midi-Controller, um die Lautheit zweier Audiobeispiele anzugleichen. Mit dem Gain-Regler steuern Sie die Lautstärke für Spur 1. Es soll jeweils nur eine Einstellung, passend für das gesamte Audiobeispiel, festgelegt werden (keine Automatisation).

Wiederholen Sie die Aufgabe 66 mal für verschiedene Paarungen von Audiobeispielen.

Hilfsmittel:

Gain Regler zur Steuerung der Lautstärke von Spur 1

Kopfhörer

Auf dem Bildschirm werden zwei EBU-R128 Lautheits Messgeräte angezeigt. Den aktuellen Lautheitswert in *LU* (*Loudness Units*, 1 *LU* entspricht 1 *dB*) der jeweiligen Audiospuren können Sie am jeweiligen Meter ablesen. Dieser kurzzeitig gemittelte Wert (3 Sekunden) wird nach dem EBU-R128 Standard ermittelt. Die Messeinstellungen der Messgeräte sind ident.

Das obere Lautheits-Meter zeigt die aktuelle Lautheit von *Spur 1*. Das untere Lautheits-Meter zeigt die aktuelle Lautheit von *Spur Referenz.*

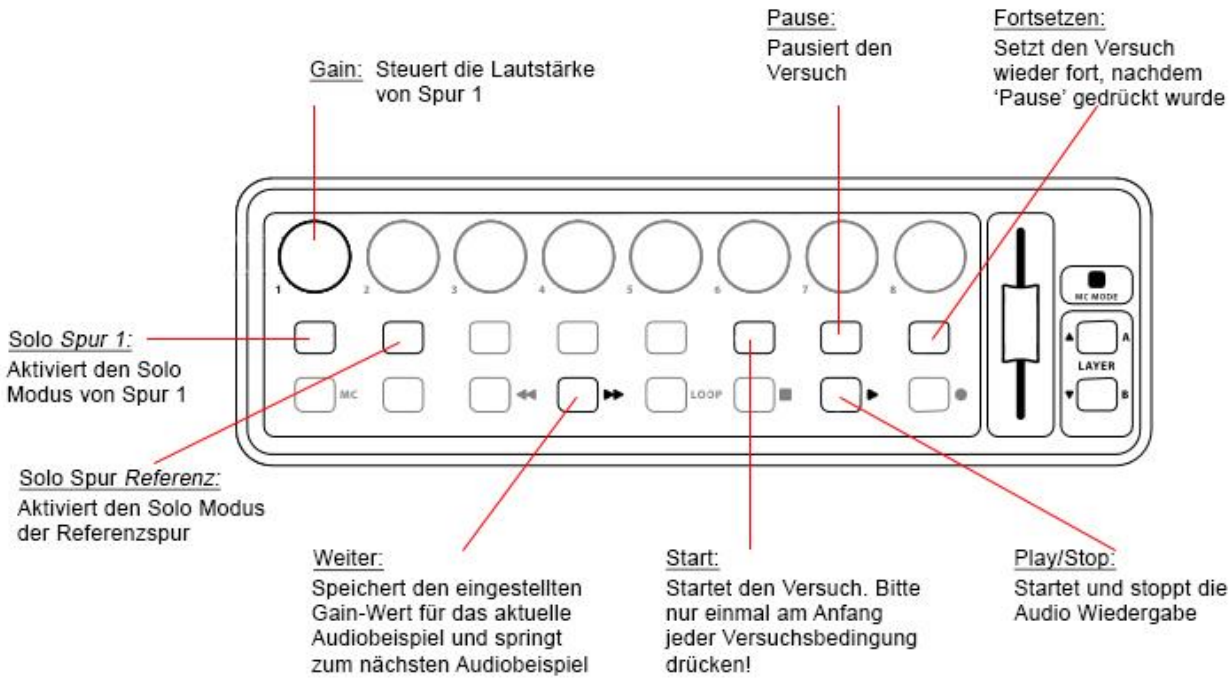
Ablauf:

1. Drücken Sie den Taster „Start“ am Midi Controller, um der Versuch zu starten.
    a. Drücken Sie Play/Stop am Midi Controller, um jederzeit die Wiedergabe zu starten (oder bei Bedarf wieder anzuhalten).
2. Wählen Sie mittels der Taster „Solo Test“ und „Solo Referenz“ am Midi-Controller, welche Audiospur sie hören möchten. Mit diesen Tasten können Sie zwischen den beiden Audiobeispielen jederzeit hin- und herschalten.
3. Bewegen Sie den Endlosdrehregler „Gain“ am Midi-Controller, um die Lautstärke von „Spur 1“ zu steuern. Eine Drehung nach links bewirkt eine Pegelverringerung, eine Drehung nach rechts bewirkt eine Pegelerhöhung. Je schneller Sie den Regler drehen, desto größer die Änderungen. Ziel ist es, die Lautheit von „Spur 1“ der Lautheit von „Spur Referenz“ anzugleichen.
4. Wenn Sie mit Ihrer Einstellung zufrieden sind, drücken Sie die Taste „Weiter“ am Midi-Controller. Dies speichert Ihre Einstellung und das nächste Beispiel wird abgespielt.
5. Wiederholen Sie die Schritte 2-4, bis Sie die Lautheitsanpassung für alle 66 Audiobeispiele durchgeführt haben.

Hinweis: Sollten Sie während des Versuches eine Pause benötigen, drücken Sie bitte die Taste „Pause“, am Midi-Controller. Wenn Sie anschließend den Versuch wieder fortsetzen wollen, drücken Sie die Taste „Fortsetzen“.

Bedienung

Behringer X-Touch Mini
Midi-Fernsteuerung von Reaper



Experiment 2

Der Mixing- oder Mastering Ingenieur steht häufig vor der Aufgabe, die Klangfarbe zweier Audiosignale aneinander anzupassen. Dies kann zum Beispiel der Fall sein, wenn zusammengehörige Instrumental- oder Gesangsaufnahmen an verschiedenen Tagen und/oder mit verschiedenem Equipment aufgenommen wurden, oder wenn Aufnahmen aus verschiedenen Studios auf einer Compilation zusammengeführt werden.

In Anlehnung daran ist es Ihre Aufgabe, zwei Testsignale mit Hilfe eines semiparametrischen EQ-Bandes klanglich anzugleichen. Im speziellen sollen Sie eine Resonanz im Frequenzspektrum des einen Signals entfernen, damit es exakt gleich klingt wie ein Referenzsignal.

Zur Anpassung des Frequenzspektrums des zu bearbeiteten Signals an das Spektrum des Referenzsignals stehen Ihnen am Midi-Controller zwei einfacher Endlosdrehregler zur Verfügung. Der mit *Freq* beschriftete Regler steuert die Centerfrequenz des Filters, der mit *F-Gain* beschriftete Regler steuert den Hub (Gain) des Filters.

Die Resonanz in Audiospur 1 kann bei unterschiedlichen Frequenzen und mit unterschiedlicher Stärke auftreten. Es handelt sich jedoch ausnahmslos um eine schmalbandige Anhebung des Spektrums, welche Sie durch eine entsprechende Absenkung bei derselben Frequenz exakt ausgleichen können.

Die Anpassung des Frequenzspektrums soll für zwei verschiedene Paarungen von Audiobeispielen erfolgen:

a) Rosa-Rauschen
b) Pop-Rock

Für jedes Stimulipaar soll die Aufgabe mehrmals und unter verschiedenen Versuchsbedingungen durchgeführt werden. Mittels zweier Drucktaster („*Solo Spur 1*“ & „*Solo Spur 2*“) am Midi-Controller kann zwischen den beiden Audiospuren umgeschaltet werden, um deren Frequenzspektrum zu vergleichen. Die Audiobeispiele werden in einem Loop wiedergegeben, bis Sie Ihre Einstellung mit einem Druck auf die „*Weiter*“ Taste am Midi-Controller bestätigen und das nächste Beispiel abgespielt wird. Die Reihenfolge, in der die Stimulipaare auftreten, ist zufällig.

Für den Versuch gibt es kein Zeitlimit. Es ist wichtig, dass Sie konzentriert arbeiten und dass Sie das Frequenzspektrum der beiden Spuren so genau wie möglich und nach bestem Gewissen angleichen.

Insgesamt soll die Aufgabe unter drei verschiedenen Versuchsbedingungen durchgeführt werden. Die Beschreibung der individuellen Bedingungen, des Versuchsablaufes sowie der Bedienung des Controllers erhalten Sie auf separaten Beiblättern.

Versuchsbedingung A

Audiospur 1 (der zu bearbeitenden Spur) wurde eine Resonanz hinzugefügt.

Entfernen Sie eine Resonanz mit Hilfe eine semi-parametrischen EQ Bandes, um das Frequenzspektrum zweier Audiobeispiele anzugleichen. Mit dem Frequenz-Regler steuern Sie die Center-Frequenz des Bandes, mit dem Gain Regler steuern Sie den Hub des Filterbandes. Es soll jeweils nur eine Einstellung, passend für das gesamte Audiobeispiel, festgelegt werden (keine Automatisation).

Wiederholen Sie die Aufgabe 8-mal für verschiedene Paarungen von Audiobeispielen.

Hinweis: Die Frequenz und Stärke der Resonanz variiert für jede Wiederholung.

Hilfsmittel:

- Regler zur Steuerung des Filter-Gains (Linksdrehung bewirkt frequenzselektive Pegelreduktion, Rechtsdrehung bewirkt frequenzselektive Pegelerhöhung)
- Regler zur Steuerung der Center-Frequenz (Linksdrehung verringert die Frequenz, Rechtsdrehung erhöht die Frequenz)

(Beide Regler reagieren auf die Drehgeschwindigkeit. Schnelle Drehungen bewirken drastische Änderungen, langsame Drehungen bewirken sehr feine Änderungen)

- Kopfhörer

Ablauf:

1. Drücken Sie den Taster „Start“ am Midi Controller, um der Versuch zu starten.
    a. Drücken Sie Play/Stop am Midi Controller, um jederzeit die Wiedergabe zu starten (oder bei Bedarf wieder anzuhalten).
2. Wählen Sie mittels der Taster „Solo Spur 1“ und „Solo Referenz“ am Midi-Controller, welche Audiospur sie hören möchten. Mit diesen Tasten können Sie zwischen den beiden Audiobeispielen jederzeit hin- und herschalten.
3. Bewegen Sie die Endlosdrehregler *Freq* und *F-Gain* am Midi-Controller, um das Filterband von „Spur 1“ zu steuern. Ziel ist es, die Resonanz auf „Spur 1“ zu entfernen und somit das Frequenzspektrum an das Spektrum der Referenzspur anzugleichen.
4. Wenn Sie mit Ihrer Einstellung zufrieden sind, drücken Sie die Taste „Weiter“ am Midi-Controller. Dies speichert Ihre Einstellung und das nächste Beispiel wird abgespielt.
5. Wiederholen Sie die Schritte 2-4, bis Sie die Lautheitsanpassung für alle 8 Audiobeispiele durchgeführt haben.

Hinweis: Sollten Sie während des Versuches eine Pause benötigen, drücken Sie bitte die Taste „Pause“, am Midi-Controller. Wenn Sie den Versuch anschließend wieder fortsetzen wollen, drücken Sie die Taste „Fortsetzen“.

Versuchsbedingung B

Audiospur 1 (der zu bearbeitenden Spur) wurde eine Resonanz hinzugefügt.

Entfernen Sie eine Resonanz mit Hilfe eine semi-parametrischen EQ Bandes, um das Frequenzspektrum zweier Audiobeispiele anzugleichen. Mit dem Frequenz-Regler steuern Sie die Center-Frequenz des Bandes, mit dem Gain Regler steuern Sie den Hub des Filterbandes. Es soll jeweils nur eine Einstellung, passend für das gesamte Audiobeispiel, festgelegt werden (keine Automatisation).

Wiederholen Sie die Aufgabe 8-mal für verschiedene Paarungen von Audiobeispielen.

Hinweis: Die Frequenz und Stärke der Resonanz variiert für jede Wiederholung.

Hilfsmittel:

- Regler zur Steuerung des Filter-Gains (Linksdrehung bewirkt frequenzselektive Pegelreduktion, Rechtsdrehung bewirkt frequenzselektive Pegelerhöhung)
- Regler zur Steuerung der Center-Frequenz (Linksdrehung verringert die Frequenz, Rechtsdrehung erhöht die Frequenz)

(Beide Regler reagieren auf die Drehgeschwindigkeit. Schnelle Drehungen bewirken drastische Änderungen, langsame Drehungen bewirken sehr feine Änderungen)

- Anzeige des Echtzeitspektrums beider Spuren auf dem Bildschirm
- Anzeige der eingestellten EQ Kurve

Das linkte Fenster zeigt die aktuell von Ihnen eingestellte EQ-Kurve, sowie das Frequenzspektrum des Signales welches Sie bearbeiten. Im rechten Fenster sehen Sie das Frequenzspektrum des Referenzsignals. Die Anpassung soll nur unter Verwendung dieser visuellen Hilfsmittel erfolgen.

Ablauf:

1. Drücken Sie den Taster „Start“ am Midi Controller, um der Versuch zu starten.
    a. Drücken Sie Play/Stop am Midi Controller, um jederzeit die Wiedergabe zu starten (oder bei Bedarf wieder anzuhalten).
2. Bewegen Sie die Endlosdrehregler *Freq* und *F-Gain* am Midi-Controller, um das Filterband von „Spur 1“ zu steuern. Ziel ist es, die Resonanz auf „Spur 1“ zu entfernen und somit das Frequenzspektrum an das Spektrum der Referenzspur anzugleichen.
3. Wenn Sie mit Ihrer Einstellung zufrieden sind, drücken Sie die Taste „Weiter“ am Midi-Controller. Dies speichert Ihre Einstellung und das nächste Beispiel wird abgespielt.
4. Wiederholen Sie die Schritte 2-3, bis Sie die Lautheitsanpassung für alle 8 Audiobeispiele durchgeführt haben.

Hinweis: Sollten Sie während des Versuches eine Pause benötigen, drücken Sie bitte die Taste „Pause“, am Midi-Controller. Wenn Sie den Versuch anschließend wieder fortsetzen wollen, drücken Sie die Taste „Fortsetzen“.

Versuchsbedingung C

Audiospur 1 (der zu bearbeitenden Spur) wurde eine Resonanz hinzugefügt.

Entfernen Sie eine Resonanz mit Hilfe eine semi-parametrischen EQ Bandes, um das Frequenzspektrum zweier Audiobeispiele anzugleichen. Mit dem Frequenz-Regler steuern Sie die Center-Frequenz des Bandes, mit dem Gain Regler steuern Sie den Hub des Filterbandes. Es soll jeweils nur eine Einstellung, passend für das gesamte Audiobeispiel, festgelegt werden (keine Automatisation).

Wiederholen Sie die Aufgabe 32-mal für verschiedene Paarungen von Audiobeispielen.

Hinweis: Die Frequenz und Stärke der Resonanz variiert für jede Wiederholung.

Hilfsmittel:

- Regler zur Steuerung des Filter-Gains (Linksdrehung bewirkt frequenzselektive Pegelreduktion, Rechtsdrehung bewirkt frequenzselektive Pegelerhöhung)
- Regler zur Steuerung der Center-Frequenz (Linksdrehung verringert die Frequenz, Rechtsdrehung erhöht die Frequenz)

(Beide Regler reagieren auf die Drehgeschwindigkeit. Schnelle Drehungen bewirken drastische Änderungen, langsame Drehungen bewirken sehr feine Änderungen)

- Anzeige des Echtzeitspektrums beider Spuren auf dem Bildschirm
- Anzeige der eingestellten EQ Kurve

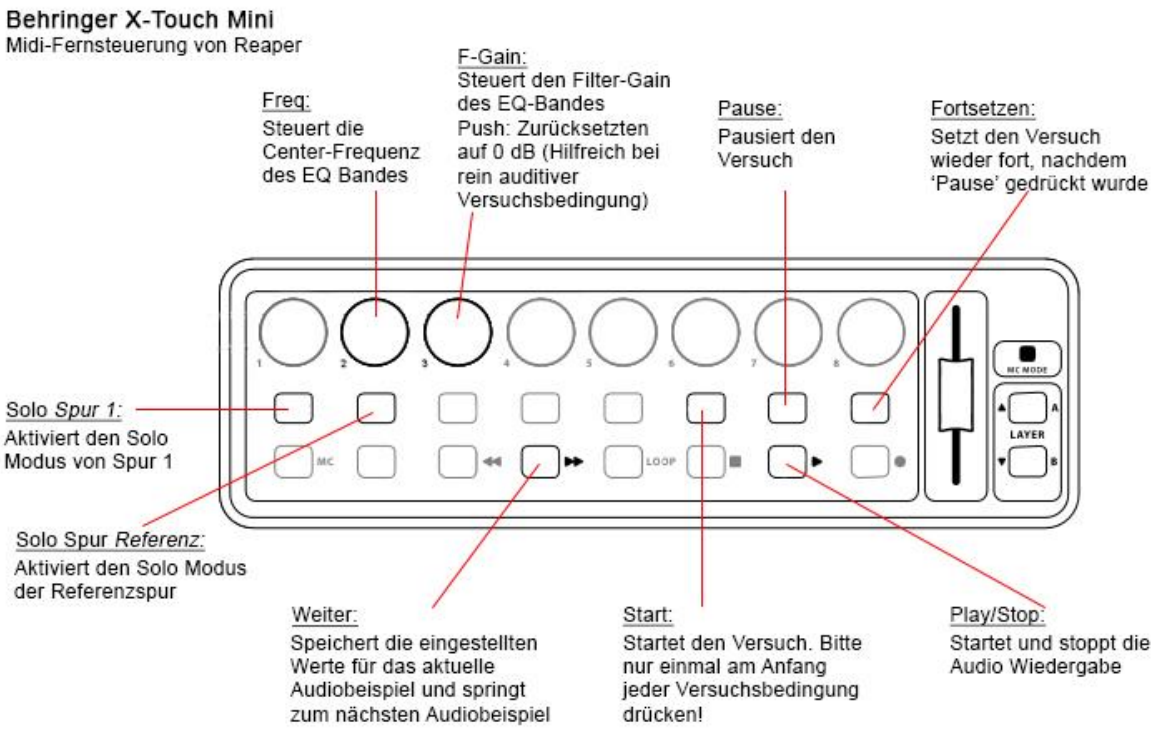
Das linkte Fenster zeigt die aktuell von Ihnen eingestellte EQ-Kurve, sowie das Frequenzspektrum des Signales welches Sie bearbeiten. Im rechten Fenster sehen Sie das Frequenzspektrum des Referenzsignals.

- Kopfhörer

Ablauf:

1. Drücken Sie den Taster „Start“ am Midi Controller, um der Versuch zu starten.
    a. Drücken Sie Play/Stop am Midi Controller, um jederzeit die Wiedergabe zu starten (oder bei Bedarf wieder anzuhalten).
2. Wählen Sie mittels der Taster „Solo Spur 1“ und „Solo Referenz“ am Midi-Controller, welche Audiospur sie hören möchten. Mit diesen Tasten können Sie zwischen den beiden Audiobeispielen jederzeit hin- und herschalten.
3. Bewegen Sie die Endlosdrehregler *Freq* und *F-Gain* am Midi-Controller, um das Filterband von „Spur 1“ zu steuern. Ziel ist es, die Resonanz auf „Spur 1“ zu entfernen und somit das Frequenzspektrum an das Spektrum der Referenzspur anzugleichen.
4. Wenn Sie mit Ihrer Einstellung zufrieden sind, drücken Sie die Taste „Weiter“ am Midi-Controller. Dies speichert Ihre Einstellung und das nächste Beispiel wird abgespielt.
5. Wiederholen Sie die Schritte 2-4, bis Sie die Lautheitsanpassung für alle 32 Audiobeispiele durchgeführt haben.

Bedienung:


Behringer X-Touch Mini
Midi-Fernsteuerung von Reaper
Freq:
Steuert die Center-Frequenz des EQ Bandes
F-Gain:
Steuert den Filter-Gain des EQ-Bandes
Push: Zurücksetzten auf 0 dB (Hilfreich bei rein auditiver Versuchsbedingung)
Pause:
Pausiert den Versuch
Fortsetzen:
Setzt den Versuch wieder fort, nachdem 'Pause' gedrückt wurde
Solo Spur 1:
Aktiviert den Solo Modus von Spur 1
Solo Spur Referenz:
Aktiviert den Solo Modus der Referenzspur
Weiter:
Speichert die eingestellten Werte für das aktuelle Audiobeispiel und springt zum nächsten Audiobeispiel
Start:
Startet den Versuch. Bitte nur einmal am Anfang jeder Versuchsbedingung drücken!
Play/Stop:
Startet und stoppt die Audio Wiedergabe
MC MODE
LAYER
MC
LOOP

Statement

Lukas KNOEBL
(Name in Blockbuchstaben)

0831387
(Matrikelnummer)

Erklärung

Hiermit bestätige ich, dass mir der *Leitfaden für schriftliche Arbeiten an der KUG* bekannt ist und ich diese Richtlinien eingehalten habe.

Graz, den 15.04.2016

.....
Unterschrift der Verfasserin/des Verfassers