MUSIC PRE-PROCESSING FOR COCHLEAR IMPLANTS

Wim BUYENS

Jury:
Prof. dr. ir. Marc Moonen (supervisor)
Prof. dr. Jan Wouters (supervisor)
Dr. Bas van Dijk (co-supervisor)
Prof. dr. ir. Dirk Vandermeulen (chairman)
Prof. dr. ir. Hugo Van Hamme (secretary)
Prof. dr. Astrid van Wieringen
Prof. dr. ir. Toon van Waterschoot
Dr. Hamish Innes-Brown

Dissertation presented in partial fulfilment of the requirements for the degree of PhD in Engineering Science.

August 2015

© 2015 KU Leuven, Science, Engineering & Technology
Uitgegeven in eigen beheer, WIM BUYENS, KAPELLEN

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaandelijke schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Acknowledgement

Een eigen doctoraatsonderzoek, ik droomde er al jaren van. Na het afstuderen heb ik echter de kans om eraan te beginnen links laten liggen en ervoor gekozen om de academische wereld achter mij te laten. Na enkele omzwervingen in het bedrijfsleven ben ik tot inkeer gekomen en kon ik via een Baekelandmandaat alsnog een doctoraatsonderzoek aanvatten.

In de eerste plaats wil ik dan ook mijn promotoren Marc Moonen, Jan Wouters en Bas van Dijk bedanken om mij die kans te geven en om me bij het hele traject te begeleiden.

Het agentschap voor Innovatie door Wetenschap en Techniek (IWT) wens ik te danken voor de financiering, alsook Cochlear Technology Centre Belgium voor de bijkomende ondersteuning.

De voorzitter en de leden van de examencommissie zou ik willen bedanken voor de tijd die ze vrijmaakten om het proefschrift na te kijken en voor de waardevolle feedback die ze gaven om het manuscript nog te verbeteren.

Verder eveneens een woordje van dank aan mijn collega's voor alle bemoedigende woorden en voor de fijne momenten samen, zowel op de werkvloer als daarbuiten. In het bijzonder een woordje van dank aan mijn PhD maatje, Obaid Qazi. We zaten initieel in hetzelfde schuitje, maar zijn bootje is door omstandigheden iets eerder aangekomen dan het mijne.

Voor de muzikale selectie, de opnames en de muziekafmixwijsheden kon ik rekenen op Hans Buyens, Anthony Claeys, Bart Delacourt, Bart Dirckx, Ellen Peeters, Gunter Peeters, Michel Verkempinck en de Pop en Jazz afdeling van de stedelijke muziekacademie in Ekeren – Antwerpen. Dankjewel daarvoor!

Thomas Stainsby en Valerie Looi wens ik te bedanken voor de interessante discussies bij het opzetten van de experimenten en Tony Van den Eynde voor het helpen programmeren van de iPhones. Voor statistische vragen kon ik steeds terecht bij Kristof Buytaert.

Verder wens ik alle CI-gebruikers die hebben deelgenomen aan één of meerdere muziekexperimenten te bedanken voor hun medewerking en waardevolle feedback. Het is maar door naar hen te luisteren dat we een beter beeld kunnen vormen van hoe muziek voor een CI-gebruiker klinkt en hoe we het verder kunnen verbeteren.

Uiteraard is er ook een leven naast het doctoraatsleven. Aan vrienden en familie, dankjewel om mij op tijd en stond van achter de boeken te halen en voor de nodige ontspanning te zorgen, zowel bij het musiceren als bij een

feestje of een gezellig samenzijn. In het bijzonder heel veel dank aan mijn ouders om me te blijven steunen in alles wat ik onderneem.

Tot slot, wil ik nog één iemand speciaal bedanken: mijn vrouw, Inge. Bedankt dat je me bij dit doctoraatsproject bent blijven steunen, voor al het geduld dat je hiervoor hebt moeten opbrengen, en dankjewel om van ons huis een leuke thuis te maken voor ons en onze drie grote schatten. Evelien, Lotte en Kato, ook jullie bedankt voor het geduld terwijl papa 'zijn boek aan het schrijven was'. Het is altijd leuk om ’s avonds thuis te komen en verwelkomd te worden met een dikke knuffel. Binnenkort heb ik terug wat meer tijd en kunnen we eindelijk naar het zwembad met de veertien glijbanen...

Abstract

A Cochlear Implant (CI) is a medical device that enables profoundly hearing impaired people to perceive sounds by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea. The focus of most research on signal processing for CIs has been on strategies to improve speech understanding in quiet and in background noise, since the main aim for implanting a CI was (and still is) to restore the ability to communicate. Most CI users perform quite well in terms of speech understanding. On the other hand, music perception and appreciation are generally very poor.

The main goal of this PhD project was to investigate and to improve the poor music enjoyment in CI users. An initial experiment with multi-track recordings was carried out to examine the music mixing preferences for different instruments in polyphonic or complex music. In general, a preference for clear vocals and attenuated instruments was observed with preservation of bass and drums. Based on this knowledge, a music pre-processing scheme for mono and stereo recordings was developed which is

capable of balancing vocals/bass/drums against the other instruments. The scheme is based on the representation of harmonic and percussive components in the spectrogram and on the spatial information of the instruments in typical stereo recordings. Subsequently, the music pre-processing scheme was evaluated in a take-home experiment with post-lingually deafened CI users and different genres of music, providing encouraging results for building a tool for music training or rehabilitation programs.

Abstract (Dutch translation)

Een cochleair implantaat (CI) is een medisch apparaat dat het mogelijk maakt om mensen met zeer ernstig gehoorverlies geluiden te laten waarnemen door elektrische stimulatie van de gehoorzenuw met behulp van een reeks electrodes die geïmplanteerd zijn in het binnenoor. De focus van het onderzoek naar signaalverwerking voor CIs ligt op het verbeteren van spraakverstaan in stilte en bij achtergrondlawaai. Het voornaamste doel voor een cochleaire implantatie is dan ook het herstellen van de mogelijkheid om te communiceren. De meeste CI-gebruikers hebben een aanvaardbare spraakverstaanvaardigheid. Het waarnemen en appreciëren van muziek daarentegen wordt in het algemeen als matig ervaren.

Het voornaamste doel van dit doctoraatsproject is het onderzoeken en verbeteren van de muziekbeleving bij CI-gebruikers. Een eerste experiment met multi-track opnames werd uitgevoerd om de voorkeur voor de onderlinge verhouding tussen de verschillende instrumenten in een muziekmix te onderzoeken. In het algemeen ging de voorkeur uit naar een

muziekmix met zang op de voorgrond en met behoud van bas en drum. Gebaseerd op deze kennis werd een voorverwerkingsschema voor muziek ontwikkeld voor mono- en stereo-opnames dat in staat is de balans tussen zang/bas/drum en de andere instrumenten te wijzigen. Dit schema is gebaseerd op de voorstelling van harmonische en percussieve componenten in het spectrogram en op de ruimtelijke informatie van de instrumenten in typische stereo-opnames. Vervolgens werd het muziekvoorverwerkings-schema geëvalueerd in een take-home experiment met postlinguaal dove CI-gebruikers en verschillende genres van muziek. Deze evaluatie heeft bemoedigende resultaten opgeleverd om over te gaan tot het bouwen van een tool voor muziek-training of rehabilitatie.

List of acronyms and abbreviations

ACE	*Advanced Combination Encoders*
AGC	*Automatic Gain Control*
AMICI	*Appreciation of Music in Cochlear Implantees*
ANOVA	*Analysis of variance*
ASA	*American Standards Association*
ASC	*Automatic Sensitivity Control*
B	*Bass guitar*
BM	*Binary Mask*
CA	*Contour Advance*
CI	*Cochlear Implant*
C-level	*Comfort level*
CS	*Contour Straight*

C-SPL	*SPL needed to reach comfort level stimulation*
D	*Drums*
dB	*Decibel*
dB HL	*Decibel Hearing Level*
DR	*Dynamic Range*
DSP	*Digital Signal Processor*
F0	*Fundamental Frequency*
FFT	*Fast Fourier Transform*
FMI	*Familiar Melody Identification*
G	*Guitar*
H	*Harmonic*
HPSS	*Harmonic/Percussive Sound Separation*
Hz	*Hertz*
ICA	*Independent Component Analysis*
ICC	*Intraclass Correlation Coefficient*
IDR	*Input Dynamic Range*
IIR	*Infinite Impulse Response*
IWT	*Institute for the Promotion of Innovation through Science and Technology in Flanders*
LIST	*Leuven Intelligibility Sentence Test*
LSD	*Least Significant Difference*
MAC	*Minimal Auditory Capability*

MBEA	*Montreal Battery for the Evaluation of Amusia*
MCI	*Melodic Contour Identification*
MIDI	*Musical Instrument Digital Interface*
ms	*Milliseconds*
N	*Number of subjects*
N	*Noise*
NH	*Normal Hearing*
NMF	*Non-Negative Matrix Factorization*
P	*Piano*
P	*Percussive*
pps	*Pulses per second*
RMS	*Root Mean Square*
S	*Signal*
SD	*Standard Deviation*
SDR	*Signal-to-Distortion Ratio*
SIR	*Signal-to-Interference Ratio*
SNR	*Signal-to-Noise Ratio*
SPL	*Sound Pressure Level*
SRT	*Speech Reception Threshold*
STFT	*Short-Time Fourier Transform*
SVM	*Support Vector Machine*
TFS	*Temporal Fine Structure*

T-level	*Threshold level*
T-SPL	*SPL needed for threshold level stimulation*
USA	*United States of America*
UW-CAMP	*University of Washington Clinical Assessment of Music Perception*
V	*Vocals*

Table of Contents

Acknowledgement....3
Abstract....6
Abstract (Dutch translation)....8
List of acronyms and abbreviations....10
Table of Contents....15
List of figures....19
List of tables....23
Chapter 1 : Introduction....25
1.1 The human auditory system....25
1.1.1 Anatomy of the human ear....25
1.1.2 Hearing loss....27
1.2 Cochlear implants....28
1.2.1 The CI system....28
1.2.2 Sound processing in CI....30
1.2.3 Performance of CI users....35

1.3 Cochlear implants and music 36
1.3.1 Music perception with CI 37
1.3.2 Music appreciation with CI 41
1.4 The music signal 42
1.4.1 Music complexity 43
1.4.2 Sound source separation techniques 43
1.5 Motivation 45
1.6 Objectives and outline 46
Chapter 2 : Music mixing preferences of cochlear implant recipients: a pilot study 51
2.1 Abstract 51
2.2 Introduction 52
2.3 Methods 56
2.3.1 Sound material 56
2.3.2 Subjects 58
2.3.3 Experiment 1 60
2.3.4 Experiment 2 61
2.3.5 Statistics 64
2.4 Results 65
2.4.1 Experiment 1 65
2.4.2 Experiment 2 68
2.5 Discussion 74
2.6 Conclusion 78
Chapter 3 : A Harmonic/Percussive Sound Separation based Music Pre-Processing Scheme for Cochlear Implant Users 79
3.1 Abstract 79
3.2 Introduction 80
3.3 Music pre-processing scheme 82
3.4 Objective testing 86
3.5 Subjective testing 90
3.6 Conclusion 92
Chapter 4 : A stereo music pre-processing scheme for cochlear implant users 93
4.1 Abstract 93

4.2 Introduction 94
4.3 Stereo music pre-processing scheme 98
4.3.1 Vocals & Drums Extraction 99
4.3.2 Bass frequency extraction 107
4.3.3 Stereo binary mask 108
4.3.4 Stereo music pre-processing output 113
4.4 Methods 113
4.4.1 Sound material 114
4.4.2 Subjects 115
4.4.3 Perceptual evaluation 116
4.5 Results 117
4.6 Discussion 119
4.7 Conclusion 123
Chapter 5 : Evaluation of stereo music pre-processing for cochlear implant users 125
5.1 Abstract 125
5.2 Introduction 126
5.3 Methods 129
5.3.1 Sound material 129
5.3.2 Subjects 130
5.3.3 Take-home device 131
5.3.4 Test procedure 133
5.4 Results 135
5.5 Discussion 140
5.6 Conclusion 143
Chapter 6 : Conclusions and future work 145
6.1 Music mixing preference experiment 146
6.2 Music pre-processing scheme 147
6.3 Take-home evaluation 148
6.4 Future work 149
6.4.1 Music feature preference 150
6.4.2 Advanced music pre-processing 151
6.4.3 Fitting a music map 152

Appendix.....153
Bibliography.....171
Curriculum Vitae.....184
List of publications.....185

List of figures

Figure 1-1: A normal functioning human ear with a sound wave traveling through the outer ear towards the middle ear and the inner ear. (Cochlear Ltd.)..... 26

Figure 1-2: A CI system with on the left the sound processor and the coil, and on the right the implant connected to the electrode array and the reference electrode (Cochlear Ltd.). 29

Figure 1-3: The external part consists of a sound processor placed behind the ear connected with the transmitter coil; the internal part consists of a receiver coil on the implant connected with the reference electrode and the electrode array. 30

Figure 1-4: Signal path of the sound processor converting acoustic waves picked up by the microphone into electrical stimuli sent to the implant. 31

Figure 1-5: Default filterbank with 22 channels in the CI system of Cochlear Ltd. 33

Figure 1-6: Loudness growth function for the conversion of acoustic levels into electrical levels as used in the Nucleus devices from Cochlear Ltd. 34

Figure 2-1: Electrodograms of song 2 from Table 2-1 for condition “Standard” (top left), “-6dBMusic” (top right), and “-12dBMusic” (bottom left). The mixture with vocals and guitar is visualized in red; the separate vocals are presented in black (bottom right). 62

Figure 2-2: Results for experiment 1 represented as boxplot indicating the preferred level settings for the different instruments (P=Piano, G=Guitar, B=Bass, D=Drums) relative to the vocal level (0 dB) for song 6 (Table 2-1). Individual level settings for CI subjects S1-S4 are shown in three configurations: configuration with two mix channels (all instruments together), configuration with three mix channels (piano/guitar together and bass/drum together) and configuration with five mix channels (all instruments separately available). 67

Figure 2-3: Results for experiment 2 indicating the pairwise comparison of condition “Standard”, “-6dBMusic” and “-12dBMusic” with 10 NH and 10 CI subjects (S2-S11). Each bar represents the percentage scores for the preference of the first condition over the second condition for the 180 comparisons (10 subjects, 6 songs and 3 repetitions). The dashed line represents chance level. Error bars indicate 95% confidence interval. 69

Figure 2-4: Ordinal rating scores for experiment 2 indicating the pairwise comparison of condition “Standard”, “-6dBMusic” and “-12dBMusic” with 10 NH and 10 CI subjects (S2-S11). Positive scores represent preference for first condition. Imperceptible (0), Slightly better (1), Better (2), Largely better (3) and Hugely better (4). Circles and stars represent outliers and extreme outliers. 70

Figure 2-5: Negative correlation between pairwise preference (in %) of “-6dBMusic” versus ‘Standard” and CI experience (in years) (Pearson’s $r(10) = -.70$, $p = .025$ and Spearman’s $\rho(10) = .72$, $p = .02$) 72

Figure 2-6: Preference rating scores for experiment 2 indicating the pairwise comparison of condition “Standard” versus “-12dBMusic” and “VocalBassDrum” versus “-12dBMusic” with 10 CI subjects (S2-S11) for all songs from Table 2-1 with a drum track available (songs 5 and 6). CI subjects are grouped based on their preference in the comparison “Standard” versus “-12dBMusic” (T1 (with 4 subjects) preferred “Standard”, T2 (with 3 subjects) had no preference, T3 (with 3 subjects) preferred “-12dBMusic”). Negative scores indicate the preference for condition “-12dBMusic”. Circles and stars represent outliers and extreme outliers. 73

Figure 3-1: Music pre-processing scheme for CI using harmonic/percussive sound separation (HPSS) 82

Figure 3-2: Visualization of the sliding analysis block for real-time processing of the harmonic/percussive sound separation (adopted from Ono et al., 2009) 86

Figure 3-3: Energy ratio of output signal for vocal, piano, guitar, bass and drum track of "The dock of the Bay" by Otis Redding. Left: for different values of the attenuation parameter 'S' in the music pre-processing scheme for CI. Right: for different Piano/Guitar/Bass attenuation relative to vocals and drums (0 dB). 89

Figure 3-4: Energy ratio of output signal for the percussive and harmonic components from the songs studied in Buyens et al. (2014) for different values of 'S'. Percussive components (straight line) include vocal and drum tracks; harmonic components (dotted line) include piano, guitar and bass tracks. Error bars indicate 95% confidence interval. 89

Figure 3-5: Graphical User Interface for pairwise comparison analysis with processed and unprocessed music excerpts. 91

Figure 4-1: Schematic of the stereo music pre-processing scheme for CI users which is enhancing vocals/drums/bass while attenuating the 'Other' instruments with parameter 'Attenuation'. It is based on "Vocals & Drums Extraction" of the input spectrogram (Input_L + Input_R with frequency > cut-off frequency) and a "Stereo Binary Mask" to exploit the spatial information in stereo recordings (based on Input_L, Input_R, Input_L-Input_R)..... 99

Figure 4-2: Boxplot with SNR improvement (dB) of the P-components with vocals/drums versus the other instruments for the multi-track recordings used in Buyens et al. (2014) as a function of the number of iterations (J) with and without applying the binary mask (BM) from (4.9). The window length of the STFT used in this graph is 185 ms. 103

Figure 4-3: Energy ratio of the P-components (j=15) for the different tracks of a typical pop song as a function of the STFT window length. The vocals (solid line) are separated as P-components with high STFT window length and as H-components with low STFT window length..... 105

Figure 4-4: Energy ratio of the P-components for different instrument samples from the instrument recognition tests MACarena and UW-CAMP. (j=15, STFT 185 ms) 106

Figure 4-5: Mean energy ratio of the P-components for eight solo bass guitar tracks processed with the music pre-processing scheme as a function of the cut-off frequency. Error bars represent 95% confidence interval. 108

Figure 4-6: SNR improvement for vocals/drums versus other instruments as a function of the stereo parameter θ from (4.18) for different stereo mixes with panning χ ranging from 0 to 100 for piano (panned to the left) and guitar (panned to the right). 111

Figure 4-7: Vocals/drums distortion indicated as the energy ratio of the P-components for the vocals/drums track as a function of the stereo parameter θ from (4.18) for different stereo mixes with panning χ ranging from 0 to 100 for piano (panned to the left) and guitar (panned to the right). 111

Figure 4-8: Removal of the off-centre vocals track (from 0 to 100) from the P-components of the music pre-processing scheme visualized as the energy ratio of the P-components for the vocals track as a function of the stereo parameter θ. 112

Figure 4-9: Individual results for 7 CI subjects with their preferred setting for the attenuation of the H-components for 24 song excerpts with low, mid and high complexity. The average preferred setting from the seven subjects for low, mid and high complexity songs are in the rightmost column. Error bars represent 95% confidence interval. 118

Figure 4-10: Mean preferred attenuation of the H-components for the 24 song excerpts with 7 CI subjects as a function of the complexity rating given by 12 NH subjects. Error bars represent 95% confidence interval. Straight line is the linear regression (R^2 = .43). 119

Figure 5-1: Screenshot of the music application on the take-home device including the music library access and navigation buttons, the turning wheel to adjust the balance and the “Vote”-button to store the preferred setting. 132

Figure 5-2: Mean preferred attenuation parameter setting (dB) over all songs for all subjects in the take-home test. Positive values represent an attenuation of ‘other instruments’, negative values (not shown) represent an amplification of ‘other instruments’, and the value zero represents the original balance. Error bars indicate 95% confidence interval. 138

List of tables

Table 2-1: Multi-track recordings for experiment 1 (song 6) and experiment 2 (song 1-6). The recordings include vocals and background music as indicated. Only the recordings of song 2 were provided by the original artist (V=Vocals, P=Piano, G=Guitar, B=Bass guitar, D=Drums). 57

Table 2-2: Overview with demographic and etiological information about the eleven post-lingually deafened CI subjects (all Cochlear™ Nucleus®) who participated in experiment 1 (S1-S4) and experiment 2 (S2-S11). 59

Table 2-3: Overview of the predefined relative level settings for the different tracks (in dB) for the conditions in the pairwise comparison of experiment 2. Condition “Standard”, “-6dBMusic” and “-12dBMusic” differ in vocals-to-instruments ratio only. Condition “VocalBassDrum” represents an audio mix with attenuated instruments in which the bass/drum is attenuated less than the piano/guitar..... 63

Table 3-1: Results of preference rating experiment, including preference for the processed songs, 95% confidence interval and median rating for the preferred condition. 92

Table 4-1: Demographic and etiological information of seven post-lingually deafened CI test subjects. 115

Table 5-1: Sound material for take-home test including the six music genres with their total duration together with the mean and standard deviation (SD) of the dynamic range (DR) over all songs. 130

Table 5-2: Demographic and etiological information of the twelve post-lingually deafened CI subjects participating in the study. 131

Table 5-3: Overview of subjects with SRT score for speech-in-noise (in dB), pitch discrimination (in semitones), preferred attenuation (in dB), median range of 3 repetitions, familiarity with the songs (in %), singing activity before CI, singing activity with CI, and Music Enjoyment (0-10) 137

Chapter 1: Introduction

1.1 The human auditory system

The auditory system is the sensory system for the sense of hearing. A brief introduction to the basic characteristics of hearing is provided in this section. In paragraph 1.1.1 the anatomy of the human ear is described and in paragraph 1.1.2 the different types of hearing loss are explained.

1.1.1 Anatomy of the human ear

In Figure 1-1 the functioning of the human ear is visualized. The outer ear which consists of the pinna and the ear canal picks up the sound and transfers it to the tympanic membrane, the separation between outer ear and middle ear. The sound causes the tympanic membrane to vibrate and these vibrations are transferred to the oval window of the inner ear by the

ossicular chain. The movement of the oval window causes pressure waves in the fluid-filled cochlea or inner ear, and makes a small membrane in the cochlea called the basilar membrane move. The cochlea is tonotopically organized. Due to the mechanical properties of the basilar membrane, maximum displacement is reached near the base of the cochlea for high frequencies and near the apex for low frequencies. The organ of Corti, which is located on the basilar membrane, contains inner and outer hair cells. The movement of the basilar membrane causes the bending of the hairs on the hair cells (the stereocilia), which results in depolarization of the cell and the release of neurotransmitter. The action potential, which is hereby evoked on the auditory nerve, is traveling via the central auditory system up to the auditory cortex, where the stimulus is further processed.

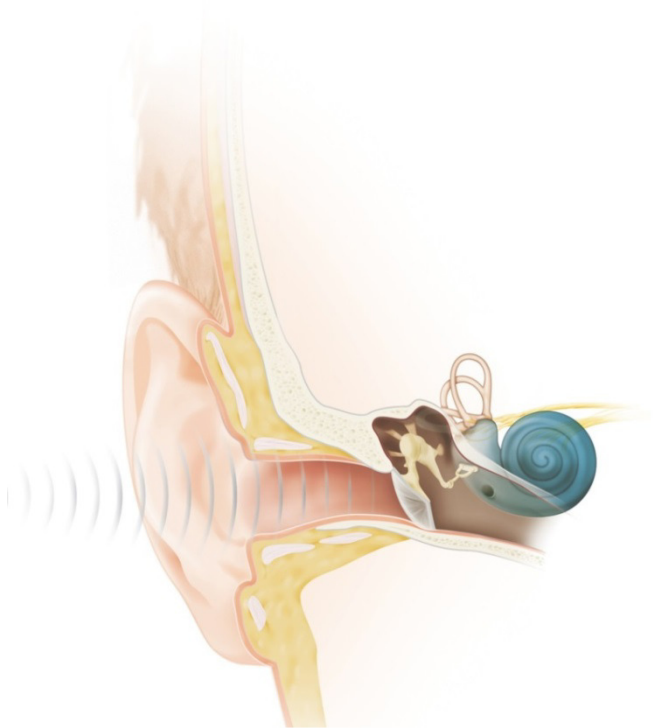


Figure 1-1: A normal functioning human ear with a sound wave traveling through the outer ear towards the middle ear and the inner ear. (Cochlear Ltd.)

1.1.2 Hearing loss

If some parts of the auditory system are not functioning well, the sensitivity to sounds drops, which is called hearing loss. The most common types of hearing loss are conductive, sensorineural or mixed.

With *conductive hearing loss* the transmission of the sound from the outer ear to the inner ear is damaged. This can be congenital or acquired at a later age and can be treated with medication, surgery or hearing aids. Some causes for this type of hearing loss are middle ear infection, fluid in the middle ear or malformations of outer ear, ear canal or middle ear. *Sensorineural hearing loss* can also be acquired or congenital and is caused by damage in the inner ear, mostly the loss of hair cells. This hearing loss is typically permanent and cannot be treated with medication or surgery. Causes of this type of hearing loss include: viral and bacterial diseases such as toxoplasmosis, rubella, bacterial meningitis, mumps; ototoxic drugs; exposure to loud sounds; presbycusis; meniere's disease; etc. Finally, *mixed hearing loss* is hearing loss with a conductive and a sensorineural component.

The degree of hearing loss, expressed in units of dB HL, refers to the severity of the hearing loss. Normal hearing is assumed when having a dB HL between -10 and 25. Other categories are defined as mild (25-40 dB HL), moderate (40-55 dB HL), moderately severe (55-70 dB HL), severe (70-90 dB HL) and profound (> 90 dB HL) (Clark, 1981). People with mild and moderate hearing loss can mostly be fitted with traditional hearing aids, but these devices are definitely not adequate for people with severe-to-profound hearing loss. The latter may be candidates for a surgically implanted medical

device, called a cochlear implant, if they meet the necessary implant criteria. Cochlear implants are described in more detail in section 1.2.

1.2 Cochlear implants

A cochlear implant (CI) is a medical device that provides auditory sensations to subjects with severe-to-profound hearing loss by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea. Paragraph 1.2.1 describes the CI system. In paragraph 1.2.2 the processing of sound into electrical stimuli is explained and in paragraph 1.2.3 the performance of CI users is discussed.

1.2.1 The CI system

A CI system consists of two parts (Figure 1-2): an external part, including the sound processor and the transmitter coil, and an internal part, including the implant, the electrode array and the reference electrode. The sound processor is worn behind the ear of the CI user (Figure 1-3) and consists of two microphones, a custom-made electronic chip with digital signal processor (DSP) and a battery.



Figure 1-2: A CI system with on the left the sound processor and the coil, and on the right the implant connected to the electrode array and the reference electrode (Cochlear Ltd.).

The sound is picked up by the microphones, or can be inserted directly through an accessory connected to the sound processor such as a Personal Audio Cable or TV/HiFi Cable. The incoming sound is converted into electrical stimuli with a so-called sound processing strategy (paragraph 1.2.2). The electrical stimuli and the necessary power are transferred to the implant with the transmitter coil, which remains on the skin on top of the implant with a magnet. Under the skin, the receiver coil picks up the information and the power, and electrical pulses are transmitted to the electrodes on the electrode array implanted in the cochlea. These electrical pulses cause action potentials in the auditory nerve fibres, which are sent on to the brain, giving the sensation of sound. In current multi-channel CI systems the electrode array contains up to 22 electrodes distributed evenly on the array. When high frequencies are processed through a CI, electrodes close to the base will be stimulated, and when low frequencies are processed, electrodes close to the apex are stimulated. Due to the large current spread in the cochlea, a large overlap in excited populations of neurons occurs when stimulating

neighbouring electrodes. This large current spread reduces the effective (independent) number of channels in the CI system.

The work presented in this thesis has been performed with CI users with a CI system from Cochlear Ltd., which is an Australian company. Other major manufacturers of CI systems are Advanced Bionics Corporation (USA), MED-EL (Austria) and Oticon Medical/Neurelec (France).



Figure 1-3: The external part consists of a sound processor placed behind the ear connected with the transmitter coil; the internal part consists of a receiver coil on the implant connected with the reference electrode and the electrode array.

1.2.2 Sound processing in CI

The sound processor is responsible for converting acoustic waves into electrical stimuli, which are then transmitted to the implant and stimulated on the electrodes in the cochlea. A number of sound processing strategies are available in current CI systems. A comprehensive review of early sound processing strategies is provided by Loizou (1998). A more recent review

with the latest sound coding strategies from the different manufacturers is described by Wouters et al. (2015). The majority of the current CIs from Cochlear Ltd. are programmed with the default Advanced Combination Encoders (ACE) strategy (Vandali et al., 2000), which is described in more detail below. Adjustable parameters which are set during the fitting of a CI, such as the threshold level (T-level) and the comfortable level (C-level) per channel, are stored in the sound processor in a so-called map. Different maps can be adjusted for different listening environments. A schematic overview of the signal path is shown in Figure 1-4. A description of the different blocks is provided in the following paragraphs.

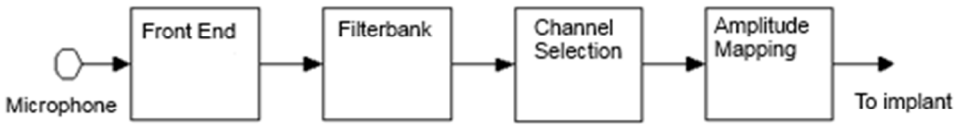


Figure 1-4: Signal path of the sound processor converting acoustic waves picked up by the microphone into electrical stimuli sent to the implant.

Front-end

The input signal picked up by the microphone (or directly inserted with an accessory cable) is first processed in the front-end processing block. The front-end processing includes pre-emphasis, automatic gain control and sensitivity control. A pre-emphasis of 6dB/octave in frequency is applied to the incoming signal, which increases the gain for high frequencies in order to improve the representation of speech sounds in the electrical pattern. The automatic gain control (AGC) applies a linear gain up to the kneepoint and infinite compression beyond that level to minimize distortions of loud sounds. The last part of the front-end processing is automatic sensitivity control (ASC), which automatically selects the adequate sensitivity setting in a certain environment. Sensitivity refers to the effective gain of the sound

processor and affects the minimum acoustic signal strength required to produce stimulation. At a very low sensitivity setting, higher sound pressure levels are needed to stimulate threshold or comfortable levels, whereas at a higher sensitivity setting less signal strength is needed. The sensitivity setting determines when the AGC will start acting and is aligned to stimulation at comfortable level (C-level).

Filterbank

After the front-end processing, the signal is fed through a filterbank which consists of a series of partially overlapping band pass filters. The audio signal is split into a number of frequency bands simulating the auditory filter mechanism in normal hearing. In the implant system from Cochlear Ltd. a 128-points Fast Fourier Transform (FFT) with sampling frequency 16 kHz is used as filterbank and FFT frequency bins are combined to determine 22 frequency bands. The filter bands are spaced linearly from 188 Hz to 1312 Hz and thereafter logarithmically up to 7938 Hz (Figure 1-5). Each filter band is allocated to one intra-cochlear electrode in the implant system according to the tonotopic organisation of the cochlea.

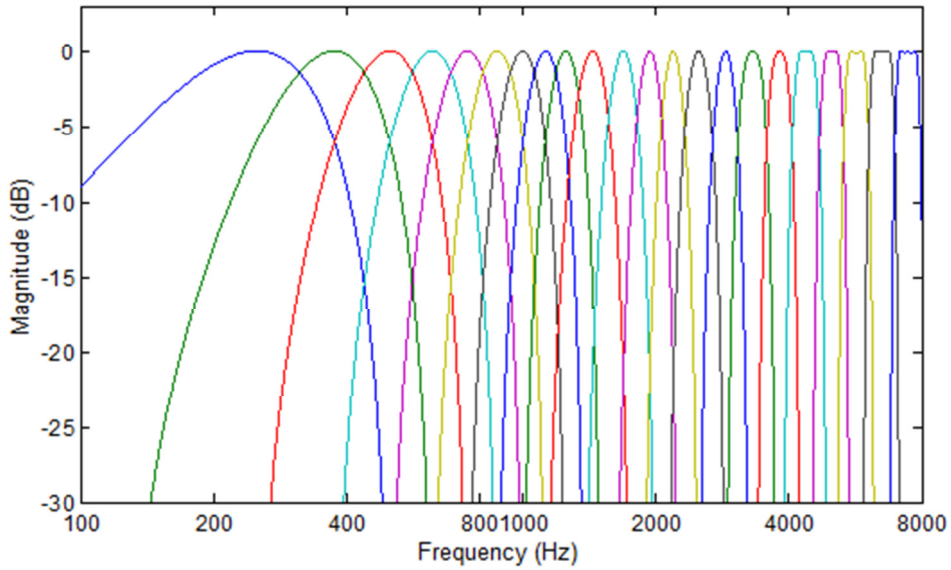


Figure 1-5: Default filterbank with 22 channels in the CI system of Cochlear Ltd.

Channel selection

Quadrature envelope detection is applied on the outputs of the band pass filters to extract the envelopes. In the ACE strategy not all channels are selected for stimulation. In each analysis window, only the *n* channels with the highest amplitudes are selected, the so-called maxima. The number of maxima selected for stimulation is programmed during fitting (default = 8).

Amplitude mapping

In amplitude mapping the amplitude conversion from acoustic level into electrical level is performed in each filter band. The conversion function, which is described in terms of T-SPL and C-SPL, is shown in Figure 1-6. T-SPL refers to the sound pressure level required to stimulate at threshold level (default = 25 dB SPL), and C-SPL refers to the sound pressure level required to reach comfort level (default = 65 dB SPL). Threshold level and comfort level are determined for each channel during the fitting session. The input

dynamic range (IDR) of typically 40 dB between T-SPL and C-SPL is mapped on to the electrical dynamic range between threshold level and comfort level, typically about 8-10 dB. The dynamic range is thus compressed with the loudness growth function displayed in Figure 1-6.

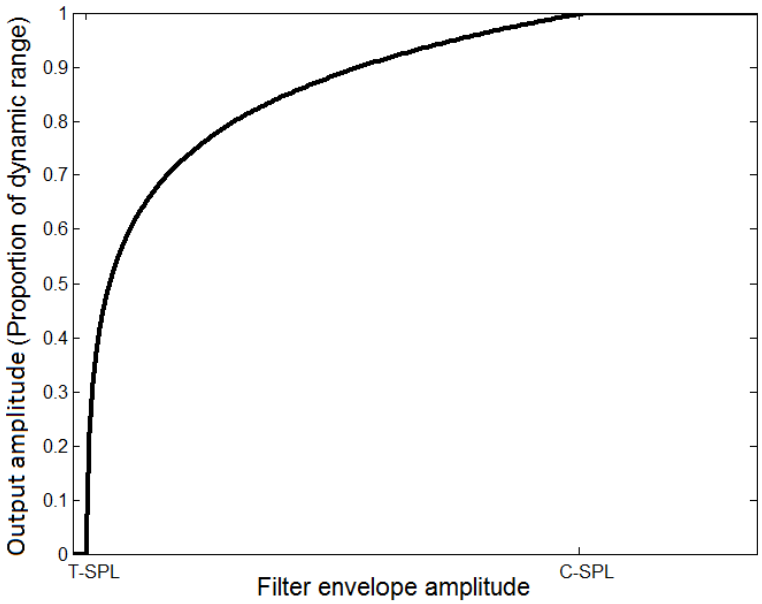


Figure 1-6: Loudness growth function for the conversion of acoustic levels into electrical levels as used in the Nucleus devices from Cochlear Ltd.

The compressed channel outputs are sent to the implant with an RF link between transmitter and receiver coil. Biphasic electrical current pulse trains with a fixed rate are modulated with the compressed channel outputs and are mapped to the electrodes in the cochlea. Low-frequency channels are mapped to apical electrodes, high-frequency channels to basal electrodes. The stimuli on the different electrodes are presented non-simultaneous (only one electrode is stimulated at a time) to avoid channel interaction (Wilson et al., 1991). The range of stimulation rates available in current devices range from 250 pps per channel to 3500 pps per channel (default = 900 pps).

1.2.3 Performance of CI users

CIs are the first successful man-made neural interfaces to restore the function of a sensory system. Nowadays, CI users reach good speech understanding in quiet environments (Zeng et al., 2008), although variability in performance across subjects is present. However, in adverse listening conditions with background noise CI users need 5-25 dB higher Signal-to-Noise Ratio (SNR) compared to normal-hearing (NH) listeners, and even more with fluctuating noise types (Fu and Nogaki, 2005).

The performance of CI users on pitch-related tasks, such as melody recognition and pitch discrimination, is also poor compared to NH listeners. Pitch is defined by the American National Standards Institute (1994) as “that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low”. It is a perceptual attribute, which is for periodic sounds the perceptual counterpart of the physical attribute fundamental frequency (F0). A complex-tone discrimination task was performed with 8 NH and 46 CI subjects by Gfeller et al. (2002a). NH subjects had a mean minimum threshold of 1.13 semitones (range 1-2 semitones). However, persons with a CI demonstrated considerable variability (across subjects) with a mean minimum threshold of 7.56 semitones (range 1-24 semitones). In normal hearing, pitch is determined by three different cues: (1) the place of stimulation in the tonotopically organized cochlea, (2) the variations in a wave shape within single periods of periodic sounds, the so-called temporal fine structure (TFS) (perceived from 500 Hz up to around 1500 Hz) and (3) the periodicity (perceived from 300 Hz - 500 Hz).

In CI, the pitch perception is inadequate due to the poor spectral resolution and the poor temporal coding. Spectral resolution is restricted by the

number of channels available in the CI. Due to the spread of excitation in the cochlea the number of effective (independent) channels is reduced (Chatterjee and Shannon, 1998). Also the poor neural survival and the limited dynamic range of electrical hearing have a negative impact on the spectral contrast (Loizou et al., 2000).

Increasing the stimulation rate provides an increasing pitch percept up to around 300 Hz. With higher stimulation rates, amplitude modulation can be used to convey a similar pitch percept, but also in this case the pitch cue breaks down above 300 Hz modulation. As TFS provides useful information for NH listeners in the perception of pitch and localisation, and in the segregation of different sound sources, many attempts to introduce TFS in commercial sound coding strategies have been made. Unfortunately, they do not show a significant benefit to date (Wouters et al., 2015). In the ACE sound coding strategy TFS cues are not present. Only the envelopes of the filterbank are used to modulate a pulse train with fixed carrier frequency.

1.3 Cochlear implants and music

Although good speech understanding is reached for most CI users in a quiet environment, music perception and enjoyment is still rather poor. In the respective paragraphs 1.3.1 and 1.3.2 music perception and music appreciation with a CI are addressed in detail. Looi et al. (2012) demonstrated that next to the perceptual accuracy for pitch, rhythm and timbre also the appraisal of music is an important and valid consideration for the evaluation of music outcomes. Only a weak or no relationship was found between perceptual abilities and appraisal (Gfeller et al., 2008; Lassaletta et al., 2008; Wright and Uchanski, 2012; Drennan et al., 2015), indicating that

perception and appreciation are rather independent of each other. Therefore, these two aspects are described in separate paragraphs in this section.

1.3.1 Music perception with CI

Music can be defined as an organized sequence of sounds with a limited number of fundamental features, including rhythm, melody and timbre (McDermott, 2004). The perception of these features in CI users is described in this paragraph.

Rhythm perception

The temporal pattern in musical sounds in the frequency range of 0.2 Hz to 20 Hz is considered as rhythm. Lower frequencies relate to the overall dynamics of music and higher frequencies carry pitch information. On average, CI users perceive rhythm about as well as normal hearing listeners (McDermott, 2004). Several researchers investigated rhythm perception with CI users. A rhythm discrimination test was performed by Gfeller and Lansing (1991) with 18 adult CI users showing a mean score of 88% correct, which is comparable to the average score of the control group with 35 NH subjects. More recently, an average score of 93% correct was obtained by Looi et al. (2008) in a group of 15 CI users performing a similar rhythm discrimination task. Kong et al. (2004) performed a rhythm identification task with CI and NH test subjects in which they were asked to identify one of seven distinct rhythmic patterns. Scores for 4 NH subjects reached 100% and 3 CI subjects reached scores of 100%, 90% and 75%. Cooper, et al. (2009) assessed music perception in CI users with the Montreal Battery for the Evaluation of Amusia (MBEA). Higher performance was observed on temporal-based tests

(Rhythm and Meter) compared to pitch-based tests, which only show near-chance level performance. In Phillips-Silver et al. (2015), the capacity of CI users to feel the musical rhythm and to bounce in time to Latin Merengue music was revealed, especially when this music was presented in unpitched drum tones.

Melody perception

The recognition of melodies in CI users is poor if rhythm and verbal cues are removed. Tune identification was investigated by Kong et al. (2004) with and without rhythmic cues. In a melody recognition experiment with 12 familiar songs, NH subjects (N = 6) recognized near 100% correct in both conditions, whereas CI subjects (N = 6) only showed 63% correct with rhythm and near-chance level without rhythm. The influence of verbal cues on melody recognition was investigated amongst others by Fujita and Ito (1999) and Leal et al. (2003) and indicated higher scores on closed-set identification when words were present. Melody pattern recognition in which test subjects are asked to label two pitch sequences as the same or different (without rhythmic or verbal cues) was assessed by Gfeller and Lansing (1991) and relies on the ability to perceive changes in pitch. The average score over 18 CI users was 78% correct, which was lower than a similar task with rhythmic patterns. In more recent work, the melodic pitch perception in CI users was investigated with a melodic contour identification (MCI) task (Galvin et al., 2009a). Nine melodic contours were defined with 5 consecutive tones, which were either flat, falling, rising or any combination of two of them. While the performance with NH subjects was consistently high in these experiments (94.8%), the results with CI subjects showed high variability ranging from 14.1% to 90.7%. A significant effect of the instrument timbre and the presence of a competing instrument on the MCI task was observed in Galvin et al. (2009b). MCI and familiar melody identification (FMI) was assessed by

Milczynski et al. (2009) with 5 post-lingually deafened CI subjects with the default ACE strategy, showing average MCI results between 20% and 38% depending on the intervals used, and FMI results near-chance level.

The pitch direction discrimination, melody recognition, and timbre recognition tests from UW-CAMP (Nimmons et al., 2008) were assessed with CI users (N = 145) by Drennan et al. (2015) and were found feasible for routine clinical use, providing results consistent with previous thorough laboratory-based investigations. In Won et al. (2010) significant relationships were found between spectral-ripple discrimination and the pitch-direction discrimination and the melody recognition from UW-CAMP, showing that CI users who had better spectral-ripple discrimination ability had better music perception.

Timbre perception

According to ASA (1960) timbre is defined as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". In general, timbre perception is unsatisfactory in CI users and music sound quality is rated as less pleasant compared to normal hearing listeners. Gfeller et al. (2002b) performed timbre recognition and timbre appraisal experiments with CI subjects (N = 51), showing lower scores for discriminating the sound of 16 different musical instruments compared to NH subjects (CI 46.6% versus NH 90.9%). The quality rating of the musical instruments revealed lower scores for CI subjects compared to NH subjects, and particularly low ratings for stringed instruments. Instruments played in the high-frequency range were rated more scattered and more dull. Timbre is based on the physical characteristics of sound including the spectral energy distribution and temporal envelope, especially its attack and decay

characteristics. In Won et al. (2010) a significant correlation was found between the spectral-ripple thresholds and timbre scores, indicating that poor spectral resolution results in poor timbre perception.

A timbre recognition test for clinical use was constituted by Nimmons et al. (2008) and later validated by Kang et al. (2009). CI users recognized guitar most often (64% correct) and flute least often (31% correct). Whereas errors in normal hearing subjects tended to be within the same instrument family (Gfeller et al., 1998), errors in CI users were more diffuse. In Nimmons et al. (2008), flute was most often confused with cello, possibly due to the soft onset of both instruments and the flat articulation in the instrument recordings. This timbre recognition test is a subtest of the larger UW-CAMP test with pitch discrimination perception and melody recognition. Omran et al. (2010) presented the instrument recognition task from the MACarena test environment, which includes instrument samples that are different from the UW-CAMP instrument samples. The MACarena instrument samples consist of the beginning of a traditional Swedish folksong played by professional musicians on different instruments, whereas the UW-CAMP samples consist of a 5-note melodic sequence played with the same detached articulation. Due to the natural way of playing with vibrato, the flute was recognized better in MACarena compared to UW-CAMP. Another clinical music perception test including timbre recognition is the Appreciation of Music in Cochlear Implantees (AMICI) test developed by Spitzer et al. (2008) and validated by Cheng et al. (2013). It includes discrimination of music versus noise, musical instruments identification, musical style identification, and recognition of individual music pieces. Whereas the first task appeared to be an easy task for CI users, the results in the second and the third task showed high variability across subjects. The fourth task was the most challenging.

The effect of training on the recognition of musical instruments was investigated in CI users by Driscoll (2012), showing an improvement in musical instrument recognition after following a musical training program of three sessions a week for five weeks.

1.3.2 Music appreciation with CI

Musical backgrounds, listening habits and music enjoyment were investigated by Gfeller et al. (2000) using a questionnaire and post-lingually deafened adult CI users. A wide range of success with music was found, but in general, a drop in music enjoyment was indicated post implantation versus prior to hearing loss. Factors that were found to enhance music enjoyment include the ability to follow the musical score or words and to watch the performer. Furthermore, features in music that were found advantageous for music enjoyment contain simple musical structure, familiar music and clear rhythm/beat, whereas loud volume was described to impede music enjoyment. The listening environment was also found to have an influence on music enjoyment which is on the one hand enhanced by a quiet environment with good acoustics, and on the other hand deteriorated in a reverberant or noisy room. General music enjoyment correlated with time spent in listening to music. A negative correlation between subjective complexity and appraisal was described for CI users by Gfeller et al. (2003) with pop, country and classical music. The classical music excerpts were in general rated more complex, compared to pop and country, possibly due to the presence of a clear melody with lyrics and a strong simple beat in the pop and country music. When comparing quality ratings for music stimuli involving a single instrument, a solo instrument with background music and ensemble music, the single-instrument stimuli were preferred to music involving multiple instruments (Looi et al., 2007).

Music is an important aspect of well-being and social life for CI users (Gfeller et al., 2000). The poor music perception through the CI limits the access to and the enjoyment of this important social and cultural phenomenon. However, music training can improve music listening for many CI users (Looi et al., 2012) and has clinical relevance in relation to quality of life, enhanced participation within society, and subjective CI benefit. Music training can be offered to CI subjects using current technology. In Gfeller et al. (2002c) the effect of 12 weeks of training with different musical instrument excerpts was investigated with CI users (N = 24). Significant improvement in both timbre recognition and timbre appraisal was found after training, compared to a control group (without training). In Galvin et al. (2007) the effect of a computer-based training program of 30 minutes a day, 5 days a week with Melodic Contour Identification (MCI) was examined in CI users. An improvement in MCI performance was shown after training and – based on anecdotal feedback – also an improvement in music appreciation was reported.

1.4 The music signal

This section describes the music signal itself with in paragraph 1.4.1 a definition of music complexity and in paragraph 1.4.2 a short introduction on different sound source separation techniques to decompose a complex music signal in its different components.

1.4.1 Music complexity

The term ‘music complexity’ is difficult to capture in one clear definition. Two types of complexity are discussed in this section: objective complexity and subjective complexity.

Objective (or structural) complexity may be determined by calculating the amount of variability or uncertainty associated with a given song, which is - according to information theory - directly related to the amount of information and redundancy in the song (Gfeller et al., 2003). More information (i.e. more instruments playing different scores) increases objective complexity, whereas more redundancy (i.e. simple and repetitive melodic or rhythmic patterns having future musical events that can be predicted more quickly) decreases objective complexity.

Subjective complexity was defined by Price (1986) as “the perceived complexity level or information content, which is mutable and a function of the listener and past musical experience”. It is a result of the interaction of the objective complexity of a song and the listener’s musical knowledge, prior experience with the music style and/or familiarity with the song. A commonly used approach to identify the subjective complexity of music is a complexity rating experiment in which a test subject is asked to rate the (subjective) complexity of a song on a scale from 0 (not complex) to 100 (very complex) (Gfeller et al., 2003).

1.4.2 Sound source separation techniques

Complex music is a mixture of different instruments playing different scores. Sound source separation techniques make an attempt to separate this

mixture back into its different instruments, when multi-track recordings of the different instruments are not available. Depending on the number of available channels, sound source separation techniques can be divided in two main categories: single-channel methods and multi-channel methods (Virtanen, 2006).

In single-channel sound source separation techniques mostly a combination of several approaches is used, including model-based inference, unsupervised learning and psycho-acoustically motivated methods. In model-based inference a parametric model of the sound sources to be separated is employed in which the model parameters are estimated from the observed mixture signal. The sinusoids plus noise model is the most commonly used model (Smith and Serra, 1987; Serra, 1989, 1997). Unsupervised learning methods use a simple non-parametric model and estimate the sound source characteristics from the data based on Independent Component Analysis (ICA), Non-Negative Matrix Factorization (NMF) or Sparse Coding. In ICA, statistically independent latent sources are identified, which can be combined into harmonic and percussive components based on extracted features as in Uhle et al. (2003). The NMF approach was used by Helen and Virtanen (2005) for decomposing the spectrogram into elementary patterns. NMF decomposes a matrix in additive (not subtractive) components, resulting in a parts-based representation of the data. A pre-trained Support Vector Machine (SVM) was used to classify them into harmonic and percussive components. Sparsity constraints on the active elements were added by Abdallah and Plumbley (2004) and Virtanen (2003). In psycho-acoustically motivated methods the elementary time-frequency components of the incoming signal are categorized into their respective sound sources based on association cues such as spectral proximity, harmonic concordance, synchronous changes and spatial

proximity (Bregman, 1990). Ono et al. (2008) described a simple and fast algorithm to perform harmonic/percussive sound separation based on the “anisotropic smoothness” of the harmonic and percussive components in the spectrogram. Harmonic components appear smooth in the temporal direction, whereas percussive components appear smooth in the frequency direction in the spectrogram.

In the case when two or more channels are available, multi-channel methods can take advantage of the availability of spatial information, e.g., from recordings with multiple microphones placed at different positions, enabling acoustic beamforming or blind separation of convolutive mixtures to recover the sound sources. The typical karaoke problem to remove vocals from background music also exploits the spatial information of stereo recordings (York et al., 2004).

1.5 Motivation

Most music related research studies with CI users focus on the improvement of music perception. Interesting attempts were made to improve pitch perception, melody recognition and instrument discrimination. However, these (small) improvements in music perception not necessarily lead to an increase in music appreciation or enjoyment. As a matter of fact, music perception and appreciation are rather independent from each other. Since (the enjoyment of) music is an important aspect of well-being and social life for CI users, we focus in this research project on music appreciation and enjoyment rather than on music perception as such, and an attempt to improve music enjoyment is investigated.

A negative correlation was found for CI users between the appreciation of music and its (subjective) complexity (Gfeller et al. 2003), which means that less complex music is appreciated more, and more complex music is appreciated less. This negative correlation is investigated further in this research project with an attempt to reduce the complexity of music and an assessment of its impact on music enjoyment in CI users. A study on quality ratings for music with CI users showed higher ratings for music involving a single instrument, compared to solo instruments with background music and ensemble music (Looi et al. 2007). Based on these studies, the reduction of the number of instruments, or alternatively the attenuation of certain instruments is taken as approach for the reduction of music complexity in this research project. The attenuation of instruments is a way to reduce the structural complexity of music and is assumed to also reduce the subjective complexity. Music training or rehabilitation is another aspect in the reduction of subjective complexity and can also improve music listening for many CI users. A tool to facilitate music listening is constituted with the capability of adjusting the music according to the CI user's individual preferences. The tool is assessed in a take-home evaluation and may be used in future music training or rehabilitation programs.

1.6 Objectives and outline

The main objective of this PhD project is to investigate and to improve the poor music enjoyment in CI users. An initial experiment is carried out to gain more knowledge about music perception with CI users, in particular the preference for different instruments in polyphonic or complex music. Based on this knowledge a music pre-processing scheme is developed which is capable of modifying the relative instrument level settings in complex music

for CI. Subsequently, the music pre-processing scheme is evaluated in a take-home experiment with CI users.

In **Chapter 2** the music mixing preferences for normal hearing and CI subjects are investigated. In a first experiment, multi-track recordings and a mixing console are used to determine the preferred audio mix for CI users. In a follow-up experiment, a pairwise comparison is performed with predefined audio mixes with normal hearing and CI subjects. The content of this chapter is adopted from a publication in the *International Journal of Audiology*[1]. The following research questions are addressed in this chapter:

1) Is the original audio mix, composed for normal hearing subjects, also the ideal audio mix for CI users?

2) What are the preferred relative instrument level settings for the different instruments in typical pop music for CI users?

3) Is the music mixing preference for CI users dependent on the familiarity with the song?

In **Chapter 3** a music pre-processing scheme is described which is capable of modifying the instrument level settings of music while preserving vocals and drums. The scheme is evaluated objectively with the multi-track recordings used in Chapter 2 and subjectively using a preference rating experiment with NH listeners and CI-simulated pop/rock music excerpts. The content of this

[1] Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2014). Music mixing preferences of cochlear implant recipients: a pilot study. International Journal of Audiology, 53(5), 294-301.

chapter is adopted from a publication in the *Proceedings of EUSIPCO*[2]. The following research questions are addressed in this chapter:

1) Is it feasible to develop a signal processing scheme which is capable of enhancing vocals and drums, as investigated with CI users in Chapter 2?

2) Is the audio mix with enhanced vocals and drums preferred over the original audio mix for pop/rock music excerpts?

3) How strong is the preference for the selected audio mix?

In **Chapter 4** an extended stereo music pre-processing scheme is developed which is capable of modifying the relative instrument level settings on mono and stereo music recordings, similar to the level modifications with multi-track recordings in Chapter 2. The scheme is evaluated objectively with the multi-track recordings used in Chapter 2 and subjectively with CI users and pop/rock music excerpts from a popular hit list. The content of this chapter is adopted from a publication in *IEEE Transactions on Biomedical Engineering*[3]. The following research questions are addressed in this chapter:

1) Is it feasible to develop a signal processing scheme which is capable of modifying the relative instrument level settings in mono and stereo recordings as investigated with CI users in Chapter 2?

[2] Buyens, W., van Dijk, B., Wouters, J., and Moonen, M. (2013). A Harmonic/Percussive Sound Separation based Music Pre-Processing Scheme for Cochlear Implant Users. Proceedings of the 21st European Signal Processing Conference (EUSIPCO).

[3] Buyens, W., van Dijk, B., Wouters, J., and Moonen, M. (2015a). A stereo music pre-processing scheme for cochlear implant users. IEEE Transactions on Biomedical Engineering (in press).

2) How much modification (or attenuation) is preferred by CI users on original music excerpts with the proposed signal processing scheme?

3) Is the preferred modification dependent on the (subjective) complexity of the songs and on the subject's CI experience?

In **Chapter 5** the take-home evaluation is described in which the music pre-processing scheme is implemented as a custom-made application on an iPhone. Test subjects are asked to make their preferred balance between vocals/bass/drums and the other instruments with an adjustable parameter for different genres of music in a comfortable listening environment. Speech perception and pitch discrimination performance are measured, and a questionnaire about the subject's music listening habits is completed. The content of this chapter is submitted for publication in the *International Journal of Audiology*[4]. The following main research questions are addressed in this chapter:

1) Is the application with the music pre-processing scheme a good tool to improve music appraisal in CI users for different genres of music in a comfortable listening environment?

2) Is the preferred balance for CI users between vocals/bass/drums and other instruments dependent on the genre of music or on the familiarity with the songs?

[4] Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2015b). Evaluation of stereo music pre-processing for cochlear implant users. International Journal of Audiology (submitted for publication).

3) Is the preferred balance between vocals/bass/drums and other instruments correlated with speech performance, pitch detection abilities or CI experience?

Finally, in **Chapter 6** the main conclusions of the PhD thesis are drawn and some suggestions for future research are presented.

Chapter 2: Music mixing preferences of cochlear implant recipients: a pilot study[5]

2.1 Abstract

Objective: Music perception and appraisal are generally poor in cochlear implant recipients. Simple musical structures, lyrics that are easy to follow and clear rhythm/beat have been reported among the top factors to enhance music enjoyment. The present study investigated the preference for modified relative instrument levels in music with normal hearing and cochlear implant subjects.

[5] The content of this chapter is adopted from: Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2014). Music mixing preferences of cochlear implant recipients: a pilot study, International Journal of Audiology, 53(5), 294-301.

Design: In experiment 1, test subjects were given a mixing console and multi-track recordings to determine their most enjoyable audio mix. In experiment 2, a preference rating experiment based on the preferred relative level settings in experiment 1 was performed.

Study Sample: Experiment 1 was performed with four post-lingually deafened cochlear implant subjects, experiment 2 with ten normal hearing and ten cochlear implant subjects.

Results: A significant difference in preference rating was found between normal hearing and cochlear implant subjects. The latter preferred an audio mix with larger vocals-to-instruments ratio. In addition, given an audio mix with clear vocals and attenuated instruments, cochlear implant subjects preferred the bass/drum track to be louder than the other instrument tracks.

Conclusions: The original audio mix in real-world music might not be suitable for cochlear implant recipients. Modifying the relative instrument level settings potentially improves music enjoyment.

2.2 Introduction

A cochlear implant (CI) is a medical device enabling severe to profoundly deaf people to perceive sounds by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea (see Loizou (1998) for an introduction to cochlear implants). The CI's sound processor uses a sound coding strategy to determine an electrical stimulation pattern from the incoming sound. The main focus of signal processing research for CIs has been on strategies to improve speech understanding in both quiet and noisy

environments. Most CI recipients reach good speech understanding in quiet surroundings, however, music perception and appraisal generally remain poor. (see McDermott (2004) for review). Mirza et al. (2003) reported a significant degradation in music enjoyment after implantation by means of comparing a self-assessment scale in the period before deafness and after implantation. A similar decline in listening habits and music enjoyment after implantation was indicated by Tyler et al. (2000), Leal et al. (2003), Lassaletta et al. (2008) and Migirov et al. (2009).

Studies on music perception with CI subjects suggested a preference for simple monophonic melodies and rhythmic sounds, whereas complex polyphonic music pieces such as pop, rock or classical orchestral music were indicated as unpleasant, noisy or even annoying. Gfeller et al. (2000) reported simple musical structures and clear rhythm/beat amongst the top factors that enhance musical enjoyment for CI subjects. The effect of complexity on the appraisal of songs was studied by Gfeller et al. (2003) with CI subjects and NH subjects with pop, country and classical music. For CI subjects a negative correlation was found between complexity and appraisal, whereas for NH subjects this correlation was positive. The study also showed that CI subjects rated classical music as more complex than pop and country music. Several plausible explanations were indicated by Gfeller et al. (2003). On the one hand, from the standpoint of objective complexity, the structural characteristics of classical music tend to be more complex than pop or country music. Classical music tends to use more complex rhythmic structures and harmonic changes compared to pop and country music, which mainly consists of short simple melodies built over simple and repetitive rhythms and harmonic changes. On the other hand, the pop and country items studied by Gfeller et al. (2003) had lyrics while the classical items did not. The speech-type information from the lyrics was indicated as an

additional factor regarding complexity for CI subjects. The lyrics might have provided some CI subjects with enough useable acoustic information to more easily follow the sequence of events. In addition, both pop and country music are likely to contain a strong, simple beat which is more suitable for transmission through current-day implants than the structural features of instrumental classical songs.

Along with the enjoyment of rhythm and beat, the performance for CI subjects on rhythmic pattern perception tasks is nearly as good as for NH subjects (Kong et al., 2004; McDermott, 2004; Looi et al., 2008). Low frequency modulations, such as rhythm and beat, are well encoded in the electrical stimulation pattern with current CIs, whereas the accurate transmission of spectral information remains challenging (Loizou, 1998). Due to the limited access to fundamental frequency (F0) and timbre cues, the segregation of competing voices and/or music instruments is difficult. Galvin et al. (2009b) showed significantly worse performance in the melodic contour identification task in the presence of a competing instrument as masker. Zhu et al. (2011) reported on the inability of most CI subjects to use or combine pitch, timbre and/or timing cues in this melodic contour identification task with a competing masker, resulting in poor segregation. The masking effect of piano music on speech perception was also studied for NH and hearing impaired subjects by Ekström and Borg (2011). A piano composition played at high speed in a low octave had greater masking effect than when played at a low speed in a high octave. Quality ratings for music with a single instrument, a solo instrument with background music and ensemble music were studied by Looi et al. (2007), which indicated that CI subjects judged music with multiple instruments as less pleasant than music played by a single instrument. Donnelly et al. (2009) reported near-chance-

level performance for CI subjects in recognizing two- and three-pitch stimuli, demonstrating perceptual fusion of multiple pitches as single-pitch units.

The cited studies on music perception, in which music with simple musical structures and clear rhythm/beat were generally preferred over more "complex" music, all assessed existing music with an intrinsic level of complexity. No attempt was made to reduce the complexity of the "complex" songs themselves in order to increase music enjoyment. The present study investigated the effect of complexity reduction in "complex" songs on music enjoyment with CI subjects. From the many possible structural sources of complexity within music, the current study focused primarily on one source of complexity, which is the addition of background harmony or percussion to a solo voice. Other sources of complexity such as more complicated melody lines or more complex harmonies were not considered. Music heard on CD, radio, TV, and at concerts usually contains background harmony and rhythm, as opposed to being comprised of only a solo melody. The present study investigated the effects of modifying the relative amplitude of the lead vocal melody with respect to the harmonic and percussion accompaniments in real-world music, and the resulting listener sound-quality ratings. The universe of all possible forms/styles of music is enormous and cannot be fully addressed in one study. Consequently, the scope of this study was focused on one commonly heard type of music, pop music, which is constituted of a music band accompanying a vocalist singing the main melody. Such music bands typically include piano, guitar, bass guitar and drums.

The primary research question of the current study was to address the preference for modified relative instrument level settings of real-world pop music with post-lingually deafened CI subjects and NH subjects, including the

influence of song familiarity. In order to answer this question, two experiments were conducted. In experiment 1, CI test subjects (N = 4) were given a mixing console and multi-track recordings with which they were asked to make an audio mix that sounded most enjoyable to them. In experiment 2, a preference rating experiment was performed by NH (N = 10) and CI subjects (N = 10) based on the preferred relative level settings for the different tracks from experiment 1.

2.3 Methods

In this study it was hypothesized that CI subjects prefer a different audio mix than NH subjects. This hypothesis was investigated in a first experiment with a mixing console and multi-track recordings and in a second more controlled experiment with a preference rating analysis with predefined audio mixes.

2.3.1 Sound material

For this music experiment multi-track recordings from well-known as well as unfamiliar pop songs have been collected. Songs were chosen to consist of a vocal track with the main melody and background music with different complexity level, ranging from basic accompaniment with piano or guitar to a full music band typically consisting of piano, guitar, bass guitar and drums. In Table 2-1 the selected songs with the corresponding instrument tracks are shown. Songs 1 and 2 consist of sung lyrics with basic accompaniment of one instrument (piano or guitar), songs 3 and 4 include additional accompaniment with bass guitar, whereas songs 5 and 6 have additional drums.

The different tracks were collected either from artists and record companies, or by gathering professional backing tracks and by recording the vocals at the Pop and Jazz division of the Municipal Academy of Music in Antwerp (Belgium). Only the multi-track recordings of song 2 (Table 2-1) were provided by the original artist. The mean dynamic range of all tracks equals 10.1 dB (SD 2.7 dB), which is acceptable for the type of music used in this experiment (Pleasurize Music Foundation, 2013).

Table 2-1: Multi-track recordings for experiment 1 (song 6) and experiment 2 (song 1-6). The recordings include vocals and background music as indicated. Only the recordings of song 2 were provided by the original artist (V=Vocals, P=Piano, G=Guitar, B=Bass guitar, D=Drums).

	Song	V	P	G	B	D
1	Hallelujah (Leonard Cohen)	x	x			
2	Before I Go (Papermouth)	x		x		
3	Michel (Anouk)	x		x	x	
4	Hey Jude (excerpt A) (The Beatles)	x		x	x	
5	Hey Jude (excerpt B) (The Beatles)	x		x	x	x
6	Dock of the Bay (Otis Redding)	x	x	x	x	x

The separately recorded tracks were mixed down to MONO wav-files (44.1 kHz). From these tracks, small excerpts were selected based on the characteristics of the background music. These excerpts took on average 29 seconds and were played in a loop during the experiment. All tracks were normalized with the loudness measurement from the ReplayGain specification, which is a standard that describes a way of calculating and representing the ideal replay gain for a given track or album, and that provides guidance for players to make the required level adjustment during playback (ReplayGain, 2013). The loudness measurement in this ReplayGain specification includes three steps. In the first step, the audio signal is filtered

by an inverted approximation of the equal loudness curves to account for the frequency specific sensitivity of the human ear. A cascade of a 10^{th} order IIR filter designed by MATLAB's "yulewalk" function with a 2^{nd} order Butterworth high-pass filter provides a good approximation. The filter coefficients for both filters are listed in ReplayGain (2013). In the second step, the RMS level of the filtered signal is calculated every 50 ms. The 95^{th} percentile of all calculated RMS levels is finally chosen to represent the overall perceived loudness. The summation of the normalized vocal track and instrument tracks was assumed to represent the studio mix for normal hearing listeners. This assumption has been confirmed by a semi-professional studio mixer. In experiment 1, the multi-track recordings for song 6 (see Table 2-1) including vocal, piano, guitar, bass guitar and drums were combined in three configurations: a configuration with two mix channels containing vocals and background music respectively, a configuration with three mix channels containing vocals, piano/guitar and bass/drum respectively and a configuration with five mix channels containing vocals and all instruments separately. The mix channels for the different configurations were derived by adding the corresponding tracks and by normalizing them with ReplayGain. For comparing the results of the three configurations at a later stage all level settings were recalculated to individual instrument-to-vocal levels. In experiment 2 all songs listed in Table 2-1 were included. The sound samples can be made available by contacting the authors.

2.3.2 Subjects

Eleven post-lingually deafened CI subjects (all Cochlear™ Nucleus®) participated in the present study. A summary with demographic and etiological information can be found in Table 2-2. Post-lingually deafened CI

subjects were selected, because they might still have a memory of how music should sound. None of them was a professional musician before they deafened. All subjects used their everyday map with the default ACE sound coding strategy. The subjects signed a consent form and were paid for their travel expenses. Ethical committee approval was obtained. Subjects S1-S4 determined their preferred audio mix with a mixing console in experiment 1. The pairwise comparison experiments in experiment 2 were carried out with subjects S2-S11. Ten NH subjects with no self-reported hearing deficit also participated in experiment 2 and their age varied from 29 years to 65 years with an average of 45 years.

Table 2-2: Overview with demographic and etiological information about the eleven post-lingually deafened CI subjects (all Cochlear™ Nucleus®) who participated in experiment 1 (S1-S4) and experiment 2 (S2-S11).

Subject	Age (y)	Gender	CI exp. (y)	Etiology	Sound processor	Implant type
S1	73	Female	3	Progressive	Freedom	CI24RE(CA)
S2	59	Male	7	Unknown	Freedom	CI24R(CS)
S3	54	Female	2	Unknown	Freedom	CI24RE(CA)
S4	26	Male	8	Meningitis	Freedom	CI24R(CS)
S5	64	Male	8	Otosclerosis	CP810	CI24R(CS)
S6	60	Male	16	Otosclerosis	Freedom	CI24R(CS)
S7	42	Male	13	Progressive	Freedom	CI24R(CS)
S8	74	Female	4	Otosclerosis	Freedom	CI24RE(CA)
S9	50	Male	1	Colitis Ulcerosa	CP810	CI24RE(CA)
S10	62	Male	2	Progressive	CP810	CI24RE(CA)
S11	64	Male	11	Noise	CP810	CI24R(CS)

2.3.3 Experiment 1

The initial experiment was based on the multi-track recordings of song 6 from Table 2-1, which includes the typical instruments in a pop song: vocals, piano, guitar, bass guitar and drums. Three different configurations were assembled from the multi-track recordings of the selected song: a configuration with two mix channels (vocals and background music), a configuration with three mix channels (vocals, piano/guitar and bass/drums) and a configuration with five mix channels (vocals and all instruments separately). Four CI subjects (S1-S4) were asked to determine the most enjoyable audio mix by using a mixing console and the mix channels from the different configurations. During the experiment, the multi-track excerpt continuously played in a loop while the test subject could adjust the level for each mix channel separately. One trial run was given in advance with the mixing console in order to learn the system. No specific strategy was imposed on the test subjects to perform the task. However, it was suggested to listen to each mix channel separately before starting the experiment. Initially, all mix channel sliders on the mixing console were placed in the down position in order not to prime the test subject with a specific audio mix. The music was played in free-field in a sound-treated room and was limited to a maximum level of 65 dB(A) with all mix channels at maximum level. The experiment was performed on a laptop with multi-track editing software (Cubase Essential 5) connected to a MIDI mixing console (Behringer BCF2000) and a loudspeaker (Genelec 8020A). The experiment was repeated ten times in each configuration with the selected song (song 6 from Table 2-1). In order to prevent the test subject from using visual cues on the mixing console, a different random level offset was applied to each individual mix channel for each configuration and each run (0, -2 dB, -4 dB, -6 dB, -8 dB). Consequently, the determined audio mix resulted from listening to the music

rather than from looking at the mixing console. The test subject had no time constraint for the task and could take a break whenever necessary. Time to complete the mixing experiment varied between 1 hour and 2 hours. The preferred audio mix was determined by the slider levels for the different mix channels measured in dB relative to the vocals and individual instrument-to-vocals ratios were computed from the mix channel slider levels for comparison across the different configurations.

2.3.4 Experiment 2

In subsequent experiments, the preferred relative level settings for CI subjects were investigated further using a pairwise comparison analysis. In this experiment all songs from Table 2-1 were used to include different background accompaniment. The pairwise comparison analysis included four conditions with different relative level settings for the separately recorded tracks (Table 2-3). Condition “Standard” was composed by adding the normalized vocal track and instrument tracks, which was assumed to represent the studio mix for normal hearing listeners. Condition “-6dBMusic” and “-12dBMusic” originated from the two mix channel settings of the different CI subjects in experiment 1 with, respectively, 6 dB and 12 dB attenuation for the instruments relative to the vocals. To illustrate the effect of the different conditions, Figure 2-1 shows the electrodograms of condition “Standard” (top left), “-6dBMusic” (top right), and “-12dBMusic” (bottom left) of an excerpt from song 2 in Table 2-1. The mixture with vocals and guitar is presented in red; the separate vocals are plotted in black (bottom right). Whereas in the “Standard” condition the vocals are almost engraved in the background music, vocals start to stand out in the “-6dBMusic” and “-12dBMusic” condition.

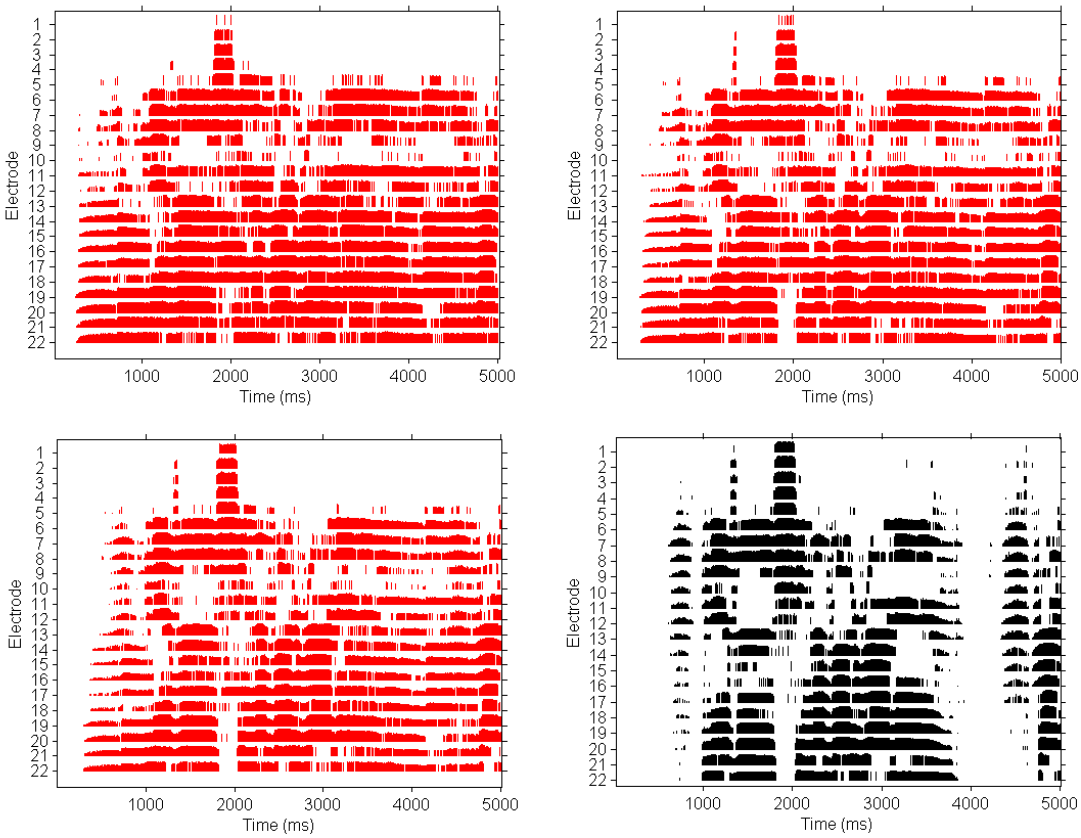


Figure 2-1: Electrodograms of song 2 from Table 2-1 for condition "Standard" (top left), "-6dBMusic" (top right), and "-12dBMusic" (bottom left). The mixture with vocals and guitar is visualized in red; the separate vocals are presented in black (bottom right).

Condition "VocalBassDrum" originated from the three mix channel settings in experiment 1, in which the bass/drum channel was less attenuated than the piano/guitar channel. In a first series of experiments, the conditions "Standard", "-6dBMusic" and "-12dBMusic" that differ only in vocals-to-instruments ratio as shown in Table 2-3 were pairwise compared. All possible pairs with these three conditions were presented in a random order and repeated three times for all 6 songs listed in Table 2-1, starting with the songs with basic accompaniment (one instrument) towards the songs with all instruments including drums. The total number of presented pairs was 54. The music was played in free-field through a loudspeaker (Genelec 8020A) at a fixed level of 65 dBSPL, which is below the AGC knee point in the CI signal

path, passing the signal with 0 dB gain. 10 NH and 10 CI subjects (S2-S11) were asked to select the condition that sounded most enjoyable to them and to quantify their preference with a rating score ranging from Imperceptible (0), Slightly better (1), Better (2), Largely better (3) to Hugely better (4). An additional experiment comparing the conditions "-12dBMusic" and "VocalBassDrum" with attenuated instruments and different bass/drum level was added for CI subjects only (S2-S11). Table 2-3 shows the corresponding relative level settings for the different instrument tracks. The preference rating between two conditions was indicated in percentages. The percentage scores represent the preference for the first condition over the second condition for all 180 comparisons (10 subjects, 6 songs and 3 repetitions). The percentage scores were thus calculated by dividing the frequency of preference for the first condition by the total number of comparisons (180), without taking into account the degree of enjoyment. The preference for the second condition is complementary (and is not shown). The additional preference rating score from Imperceptible to Hugely better indicated the strength of the preference.

Table 2-3: Overview of the predefined relative level settings for the different tracks (in dB) for the conditions in the pairwise comparison of experiment 2. Condition "Standard", "-6dBMusic" and "-12dBMusic" differ in vocals-to-instruments ratio only. Condition "VocalBassDrum" represents an audio mix with attenuated instruments in which the bass/drum is attenuated less than the piano/guitar.

	Vocals	Piano	Guitar	Bass	Drums
"Standard"	0	0	0	0	0
"-6dBMusic"	0	-6	-6	-6	-6
"-12dBMusic"	0	-12	-12	-12	-12
"VocalBassDrum"	0	-12	-12	-3	-3

2.3.5 Statistics

Given the small sample size in the first experiment, only individual results were shown. The non-parametric Friedman test with factor 'Run' was used to check the consistency over the repetitions. Differences in level settings between vocals and different instruments were also analyzed with the non-parametric Friedman test followed by a post-hoc Wilcoxon Signed-ranks test with Bonferroni correction. Intraclass Correlation Coefficient was used to check reliability of the data for every subject.

For the preference rating in experiment 2, the Chi-Square test was used to determine which of the two conditions was preferred and to analyze the effect of familiarity and singer's gender on audio mix preference. Chi Square statistics was applied on the absolute preference between two conditions (without taking into account the 5-step rating score). In the null hypothesis the frequency of preferred items were the same in both conditions. A significant Chi Square test resulted in rejecting the null hypothesis, which means a significant preference for one of the two conditions. The correlation between audio mix preference and CI experience was analyzed with Pearson's and Spearman's correlation coefficient. The One-Sample Wilcoxon Signed-ranks test was used to study the different audio mix preferences for the three defined groups in experiment 2.

2.4 Results

2.4.1 Experiment 1

Figure 2-2 shows the preferred level settings for the instrument tracks of song 6 (Table 2-1) for the configuration with two, three and five mix channels respectively. The different instruments of the song are shown on the horizontal axis, whereas the vertical axis represents the preferred level (in dB) relative to the vocal level (0 dB). Individual results are shown for the 4 CI subjects (S1-S4) with a boxplot including 10 repetitions per configuration per subject.

The reliability of the test results was checked with the intraclass correlation coefficient (ICC), showing high reliability for the results of subjects S2 (ICC = .89), S3 (ICC = .90) and S4 (ICC = .83) and low reliability for the results of subject S1 (ICC = .32).

To check the consistency over the 10 repetitions, a Friedman test with factor 'Run' was performed with N = 4 over the 10 measurements for each configuration. No significant effect of time was found for the configuration with two mix channels ($\chi^2(9) = 6.66$, $p = .67$), nor with the configuration with three mix channels ($\chi^2(9) = 14.48$, $p = .11$). However, in the configuration with five mix channels, a significant effect of time was found ($\chi^2(9) = 22.08$, $p = .009$). Consequently, further analysis was performed with the median of the 10 measurements in the two mix channel and three mix channel configuration.

In order to check the preferred relative instrument level differences, a Friedman test was performed with factor 'Instrument' for both configurations and revealed a significant effect of instrument for both configurations ($\chi^2(4) = 16$, $p = .003$). However, due to the low number of subjects a post-hoc analysis with Wilcoxon Signed-ranks Test showed no significant difference between vocal level and instrument level for the two mix channel configuration ($Z = -1.83$, $p = .068$), and no significant difference in the three mix channel configuration between vocal level and piano/guitar level, between vocal level and bass/drum level and between piano/guitar level and bass/drum level after Bonferroni correction ($Z = -1.83$, $p = .20$).

In the configuration with two mix channels (top row in Figure 2-2), a trend was found towards a preferred vocal level that is higher than the instruments level for all subjects with a median vocals-to-instruments ratio ranging from 4 dB (S2) to 16 dB (S3). In the configuration with three mix channels (middle row in Figure 2-2) a similar trend was found with a median preferred instrument level relative to the vocals in the range of -2 dB (S2) to -11 dB (S3) for piano/guitar and in the range of -1 dB (S2) to -9 dB (S3) for bass/drum. In addition, the median preferred level difference between piano/guitar and bass/drum in the three mix channel configuration was found in the range of 0 dB (S1) to 2 dB (S3). For completeness, the results for the configuration with five mix channels are shown on the bottom row in Figure 2-2, but revealed no clear preference for a particular instrument.

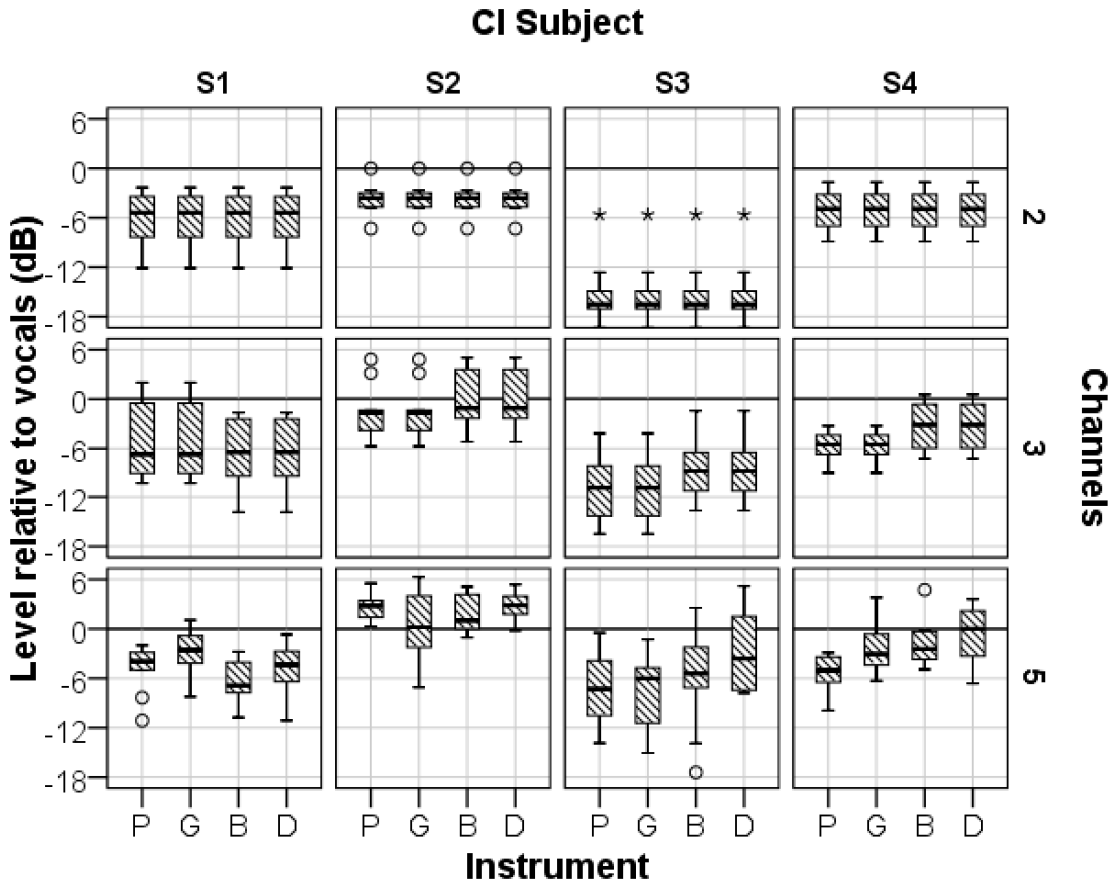


Figure 2-2: Results for experiment 1 represented as boxplot indicating the preferred level settings for the different instruments (P=Piano, G=Guitar, B=Bass, D=Drums) relative to the vocal level (0 dB) for song 6 (Table 2-1). Individual level settings for CI subjects S1-S4 are shown in three configurations: configuration with two mix channels (all instruments together), configuration with three mix channels (piano/guitar together and bass/drum together) and configuration with five mix channels (all instruments separately available).

The effect of the mix channel configuration on the vocal-to-instrument level was checked for the two mix channel and three mix channel configuration with a Wilcoxon Signed-ranks test, and revealed no significant difference between piano/guitar level in two and three mix channel configuration ($Z = -.73$, $p = .47$), and no significant difference between bass/drum level in two mix channel and three mix channel configuration ($Z = -1.46$, $p = .14$). However, for subject S3, a difference in piano/guitar and bass/drum level

between two mix channel and three mix channel configuration was found of 5.6 dB and 7.7 dB respectively.

2.4.2 Experiment 2

Figure 2-3 shows the absolute preference results for 10 NH and 10 CI subjects (S2-S11) for the pairwise comparison between condition “Standard”, “-6dBMusic” and “-12dBMusic” in which condition “Standard” represents the studio mix for normal hearing listeners and condition “-6dBMusic” and “-12dBMusic” correspond to an audio mix with the instrument tracks attenuated by 6 dB and 12 dB respectively. In Figure 2-4 the rating scores of the pairwise preferences are shown, representing the strength of the preference for the preferred condition ranging from Imperceptible (0), Slightly better (1), Better (2), Largely better (3) to Hugely better (4). Positive scores represent a preference for the first listed condition. The results for NH subjects indicated a significant preference for the condition with an equal balance between vocals and instruments (condition “Standard”) compared to condition “-6dBMusic” (80%), $\chi^2(1) = 64.8$, $p < .001$ and condition “-12dBMusic” (96%), $\chi^2(1) = 149$, $p < .001$. A significant preference was also found for condition “-6dBMusic” compared to condition “-12dBMusic” (97%), $\chi^2(1) = 161$, $p < .001$. On the contrary, the results with CI subjects indicated the opposite preference in the comparison between condition “Standard” and “-6dBMusic”. The condition with 6 dB attenuated instruments (condition “-6dBMusic”) was preferred over the condition with an equal balance between vocals and instruments (condition “Standard”) with 64%, which is significant ($\chi^2(1) = 13.9$, $p < .001$). Between condition “Standard” and “-12dBMusic”, no overall preference was found over all CI subjects ($\chi^2(1) = .36$, $p = .55$), whereas condition “-6dBMusic” was

significantly preferred over condition "-12dBMusic" (63%), similar to NH subjects ($\chi^2(1) = 11.8$, $p = .001$).

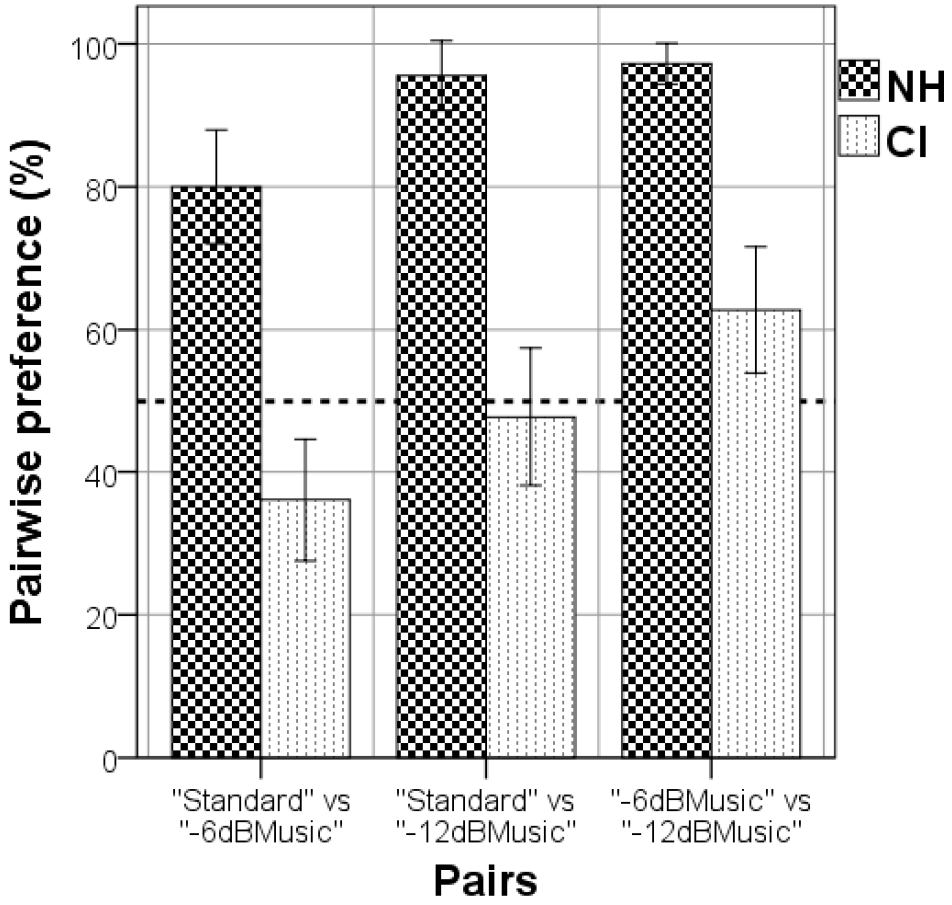


Figure 2-3: Results for experiment 2 indicating the pairwise comparison of condition "Standard", "-6dBMusic" and "-12dBMusic" with 10 NH and 10 CI subjects (S2-S11). Each bar represents the percentage scores for the preference of the first condition over the second condition for the 180 comparisons (10 subjects, 6 songs and 3 repetitions). The dashed line represents chance level. Error bars indicate 95% confidence interval.

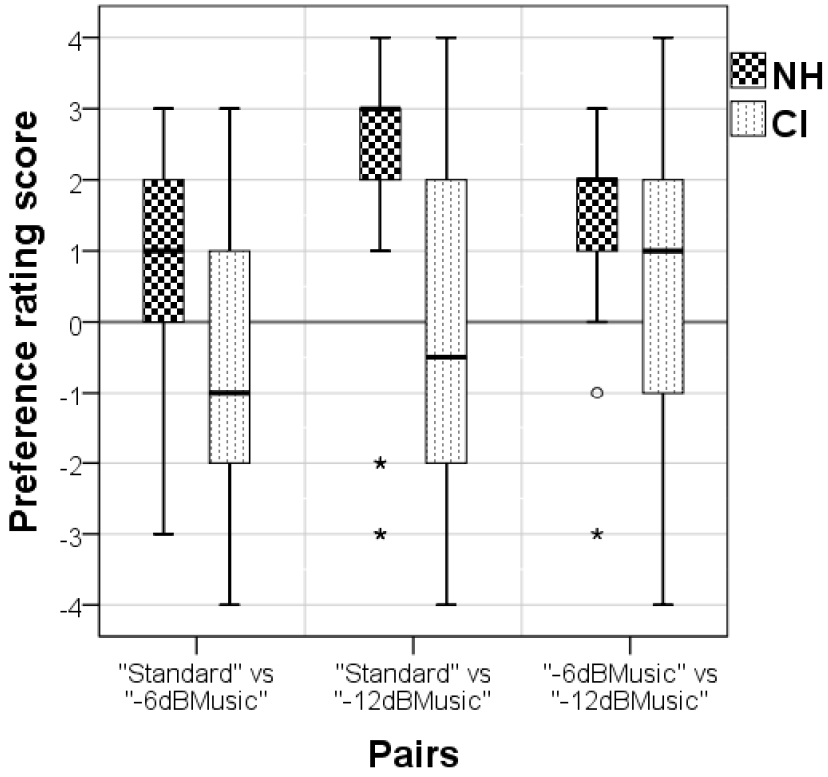


Figure 2-4: Ordinal rating scores for experiment 2 indicating the pairwise comparison of condition "Standard", "-6dBMusic" and "-12dBMusic" with 10 NH and 10 CI subjects (S2-S11). Positive scores represent preference for first condition. Imperceptible (0), Slightly better (1), Better (2), Largely better (3) and Hugely better (4). Circles and stars represent outliers and extreme outliers.

Following the preference rating experiment, the test subjects were asked about their familiarity with the songs provided, resulting in a classification of known (22/60) and unknown songs (38/60). No effect of familiarity of the songs on the preference rating was found (Pearson $\chi^2(1) = .49$, $p = .49$). However, in the pairwise comparison between condition "Standard" and "-6dBMusic", a preference of 66% was found for condition "-6dBMusic" in the group of unknown songs (95% confidence interval: 57%-75%) which is significant ($\chi^2(1) = 11.4$, $p = .001$), whereas in the group of known songs the preference (61% with 95% confidence interval: 48%-73%) was not significant, $\chi^2(1) = 2.97$, $p = .085$.

In the group of songs without drums (songs 1-4) which is equally divided between female (songs 1 and 3) and male singers (songs 2 and 4), a difference in preference rating was observed depending on the singer's gender. For the songs with a male singer a preference of 67% was found for condition “-12dBMusic” over condition “Standard” (95% confidence interval: 51%-83%), which is significant ($\chi^2(1) = 6.67$, $p = .010$), whereas for the songs with a female singer the preference (45% with 95% confidence interval: 27%-63%) is not significant ($\chi^2(1) = .60$, $p = .44$). The effect of the singer's gender in the pairwise comparison between condition “Standard” and “-12dBMusic” is significant (Pearson $\chi^2(1) = 5.71$, $p = .017$).

The preference for condition “-6dBMusic” with clear vocals and attenuated instruments versus condition “Standard” is negatively correlated with the CI experience of the test subject, meaning that less-experienced CI subjects showed stronger preference for the condition with attenuated instruments than experienced subjects. This correlation is significant with a Pearson’s $r(10) = -.70$, $p = .025$ and also with the non-parametric Spearman’s $\rho(10) = .72$, $p = .02$.

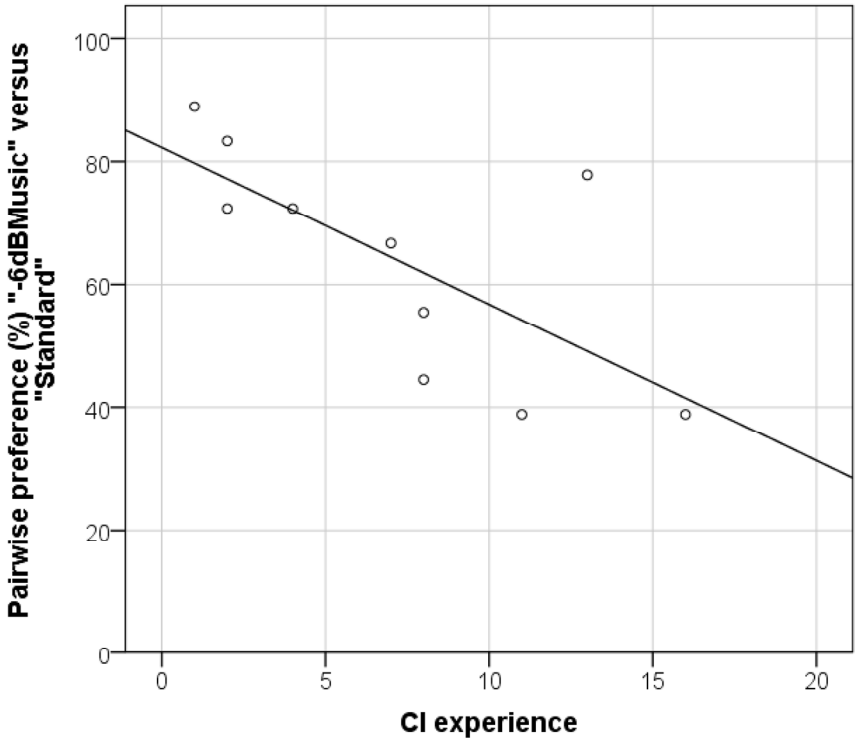


Figure 2-5: Negative correlation between pairwise preference (in %) of "-6dBMusic" versus 'Standard" and CI experience (in years) (Pearson's r(10) = -.70, p = .025 and Spearman's ρ(10) = .72, p = .02)

In an additional preference rating experiment with CI subjects (S2-S11) two conditions with attenuated instruments and different bass/drum level were compared (songs 5 and 6). The condition with prominent bass/drum track (condition "VocalBassDrum") obtained a significant preference of 87% (95% confidence interval: 82%-92%) over the condition with all instruments attenuated (condition "-12dBMusic") ($\chi^2(1) = 96.8$, $p < .001$). The preference rating results for the pairwise comparison between condition "VocalBassDrum" and "-12dBMusic" are compared to the preference rating results of the pairwise comparison between condition "Standard" and "-12dBMusic" for songs 5 and 6 in Figure 2-6. Because of the individual differences in preference rating between condition "Standard" and "-12dBMusic", the CI subjects were grouped based on these different preference ratings, resulting in a group with significant preference for condition "Standard" (N = 4) (One-Sample Wilcoxon Signed-ranks Test: p

< .001), a group with no preference between “Standard” and “-12dBMusic” (N = 3) (One-Sample Wilcoxon Signed-ranks Test: p = .46) and a group with significant preference for condition “-12dBMusic” (N = 3) (One-Sample Wilcoxon Signed-ranks Test: p = .003). In the comparison between condition “VocalBassDrum” and “-12dBMusic”, all groups had significant preference for condition “VocalBassDrum” (One-Sample Wilcoxon Signed-ranks Test: p < .001). It should be noted that even the group with subjects who significantly preferred condition “-12dBMusic” with attenuated instruments over condition “Standard”, also significantly preferred condition “VocalBassDrum” with prominent bass/drum over condition “-12dBMusic”.

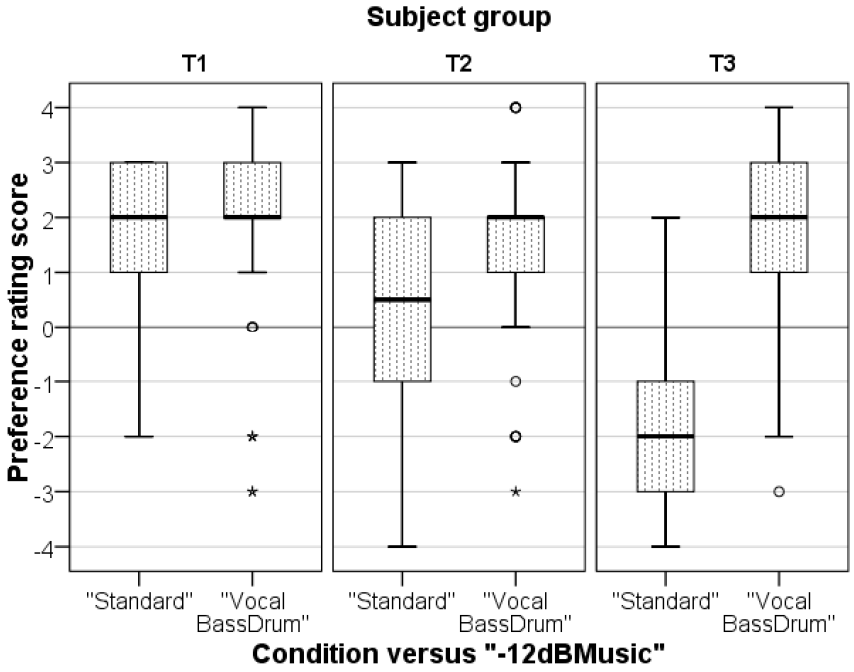


Figure 2-6: Preference rating scores for experiment 2 indicating the pairwise comparison of condition “Standard” versus “-12dBMusic” and “VocalBassDrum” versus “-12dBMusic” with 10 CI subjects (S2-S11) for all songs from Table 2-1 with a drum track available (songs 5 and 6). CI subjects are grouped based on their preference in the comparison “Standard” versus “-12dBMusic” (T1 (with 4 subjects) preferred “Standard”, T2 (with 3 subjects) had no preference, T3 (with 3 subjects) preferred “-12dBMusic”). Negative scores indicate the preference for condition “-12dBMusic”. Circles and stars represent outliers and extreme outliers.

2.5 Discussion

The primary research question of the current study was to address the preference for modified relative instrument level settings of real-world pop music with post-lingually deafened CI subjects and NH subjects. In order to answer this question, two experiments were conducted. In the initial experiment 1, CI subjects determined their preferred audio mix with a mixing console and multi-track recordings. Based on the preferred relative level settings of the different tracks in experiment 1, four conditions were defined for the pairwise comparison in experiment 2 with NH and CI subjects.

In experiment 1, CI subjects composed their preferred audio mix with all separately recorded tracks available. Due to the limited number of subjects only trends could be identified to be used later in experiment 2. A trend towards an audio mix with clear vocals and attenuated instruments was observed, especially in the configuration with two and three mix channels. In the configuration with five mix channels no consistent results were determined. The CI test subjects reported an increasing difficulty in making the audio mixes with increasing number of mix channels available. In subject S3, the overall preferred vocals-to-instrument ratio is decreasing in the configurations with more mix channels available. A reason for this could be that in the configuration with two mix channels, the focus for the test subject was only on the vocals-to-background music ratio, whereas in the other configurations the focus was on the relative level differences between all tracks. Because of the many repetitions of the song in the experiment, it might also be that the test subject became more familiar with the song in the more difficult five mix channel configuration towards the end, allowing more music in the audio mix.

Experiment 2 consisted of a pairwise comparison analysis with predefined relative level settings for the different tracks based on the results of experiment 1 in the configuration with two mix channels and three mix channels. The results obtained in experiment 1 with two mix channels are confirmed by experiment 2. CI subjects significantly preferred the condition with clear vocals and attenuated instruments compared to the condition with equally balanced vocals and instruments. For the limited songs provided, no significant effect of familiarity with the music was found on the preference rating. The condition with clear vocals and attenuated instruments was preferred for both known and unknown songs and might help CI subjects to follow songs with lyrics more easily.

The difference in preference rating between songs with male and female singers could be explained by the difference in difficulty in the segregation between vocals and background music. Gfeller et al. (2009) studied the impact of voice range on the recognition of song lyrics with CI subjects and found that the SNRs needed in song lyrics recognition for male vocalists were poorer than for female vocalists. Because segregating a male singer from background music might be more difficult than a female singer, the audio mix with attenuated instruments was preferred for songs with a male singer, whereas for songs with a female singer no significant preference was found in the pairwise comparison between condition “Standard” and “-12dBMusic”.

Informal feedback from CI subjects after the pairwise comparison between “Standard” and “-12dBMusic” revealed the dilemma between on the one hand the absence of clear vocals in the “Standard” condition and on the other hand the lack of bass/drums in the “-12dBMusic” condition. Therefore, an extra condition was added with clear vocals and attenuated instruments,

but with preservation of bass/drums (“VocalBassDrum”). This condition also showed up as a trend in the configuration with three mix channels in experiment 1. An additional pairwise comparison experiment between two conditions with attenuated instruments and different bass/drum level showed a significant preference for the condition with the louder bass/drum tracks, even for the CI subjects who initially preferred the attenuated instruments of the “-12dBMusic” condition over the “Standard” condition. The prominent presence of the bass/drum tracks together with clear vocals provided an audio mix that was more enjoyable for CI subjects. Gfeller et al. (2000) determined the ability to follow the musical score or words as a top factor for increasing musical enjoyment with CI subjects. According to the same study, features of music that enhance musical enjoyment include the presence of clear rhythm/beat and a simple musical structure. The preference results in the present study are in agreement with these findings.

The attenuation of background music is a possible way to reduce the objective complexity of a complex song. Moreover, clear vocals and clear drums might have reduced the perceived subjective complexity for CI subjects as well, because elements such as song lyrics or rhythm are more effectively transmitted and more accurately heard by CI subjects than pitch, timbre and harmony. Low frequency modulations, such as rhythm and beat, are well encoded in the electrical stimulation pattern with current CIs, whereas the accurate transmission of spectral and fine-structure information of sounds remains challenging due to channel interactions, limited number of stimulation channels and limited dynamic range. The described reduction in complexity resulted in more enjoyable music for CI subjects. This is in agreement with the findings of Gfeller et al. (2003), in which a negative correlation was determined for CI subjects between complexity and appraisal with different styles of music. Classical music was rated more

complex than pop and country music, which was suggested to be due to the lack of a clear melody and lyrics in classical music. These findings correspond with the results from the present study, in which music with a clear vocal melody and attenuated instruments was determined to be more enjoyable for CI subjects.

The large variability across CI subjects indicated that the modified relative instrument level settings as studied in the present study did not necessarily increase music enjoyment for all CI subjects. A negative correlation was found between CI experience and the preference for an audio mix with clear vocals and attenuated instruments, meaning that less-experienced CI recipients benefited more from the modified relative instrument level settings of the “original” audio mix than experienced recipients did.

The “original” audio mix, which was assumed to consist of an equal balance between vocals and instruments, was significantly preferred by NH test subjects. On the contrary, CI subjects preferred an audio mix with clear vocals and attenuated instruments. However, given the limited sample size and the variability among CI subjects, these findings might not be generalizable. Moreover, the results with this style of music should not be presumed to apply to other styles of music that have important structural differences. Consequently, similar tests with other musical styles would help to determine to what extent these modified settings are beneficial for other non-pop music.

2.6 Conclusion

Normal hearing subjects significantly preferred the “original” audio mix with an equal balance between the vocal melody and the different instruments. However, this preferred mix is not necessarily the ideal audio mix for CI subjects. For the pop songs provided in the present study, CI subjects significantly preferred an audio mix with clear vocals and attenuated instruments. In addition, given an audio mix with clear vocals and attenuated instruments, CI subjects significantly preferred the bass/drum track to be louder than the other instruments. The variability in the preference results across CI subjects indicated that the modified relative instrument level settings as studied in the current study did not necessarily increase music enjoyment for all CI subjects. In particular, the data indicated that less-experienced CI subjects benefited more from the modified relative instrument level settings in the original audio mix of complex songs than experienced recipients did.

Chapter 3: A Harmonic/Percussive Sound Separation based Music Pre-Processing Scheme for Cochlear Implant Users[6]

3.1 Abstract

Music perception and appreciation remain generally poor in cochlear implant (CI) users. Simple musical structures, a clear rhythm/beat and lyrics that are easy to follow are among the top factors to enhance music appreciation. A music pre-processing scheme for CI users is described in which vocals and percussion are elucidated using harmonic/percussive

[6] The content of this chapter is adopted from: Buyens, W., van Dijk, B., Wouters, J., and Moonen, M. (2013). A Harmonic/Percussive Sound Separation based Music Pre-Processing Scheme for Cochlear Implant Users. Proceedings of the 21st European Signal Processing Conference (EUSIPCO).

sound separation. The scheme is capable of modifying relative instrument level settings for improving music appreciation in CI users. The scheme is assessed with normal hearing subjects (N = 5) using a pairwise comparison analysis and CI-simulated music excerpts. All test subjects except one significantly preferred the music excerpts from the music pre-processing scheme.

3.2 Introduction

A cochlear implant (CI) is a medical device enabling severe to profoundly deaf people to perceive sounds by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea (Loizou, 1998). Although CI users reach good speech understanding in quiet surroundings, music perception and appreciation generally remain poor (McDermott, 2004). Simple musical structures and a clear rhythm/beat were reported amongst the top factors that enhance musical appreciation for CI users by Gfeller et al. (2000). A negative correlation was found between complexity and appreciation in CI users studied with pop, country and classical music (Gfeller et al., 2003). Classical music was rated as more complex than pop and country music. Several plausible explanations were provided including the presence of simple musical structures and lyrics in pop and country music. Since CIs were mainly developed for transmitting speech sounds, the presence of lyrics might facilitate CI users to follow more easily the sequence of events in complex music. In addition, both pop and country music oftentimes contain a strong, simple beat which is more suitable for transmission through current-day CIs than the structural features of instrumental classical music.

The preference for clear vocals and a strong rhythm/beat in CI users was studied by modifying relative instrument level settings in pop music by Buyens et al. (2014). A significant difference in preference rating scores was found between NH and CI subjects. For the pop songs provided, CI subjects preferred an audio mix with larger vocals-to-instruments ratio compared to normal hearing subjects, ranging from 4 dB to 16 dB. In addition, given an audio mix with clear vocals and attenuated instruments, CI subjects preferred the bass/drum track to be louder than the other instrument tracks. The relative instrument level settings were modified by altering the levels of the different, separately recorded, instrument tracks, which are, however, not widely available for most music. Therefore, a signal processing scheme is needed to modify the relative instrument level settings in complex music.

The paper is organized as follows. In section 3.3, the proposed music pre-processing scheme, which modifies the relative instrument level settings in complex music, is described in detail. In section 3.4 and 3.5, the music pre-processing scheme is assessed objectively and subjectively. Section 3.6 contains the conclusions.

3.3 Music pre-processing scheme

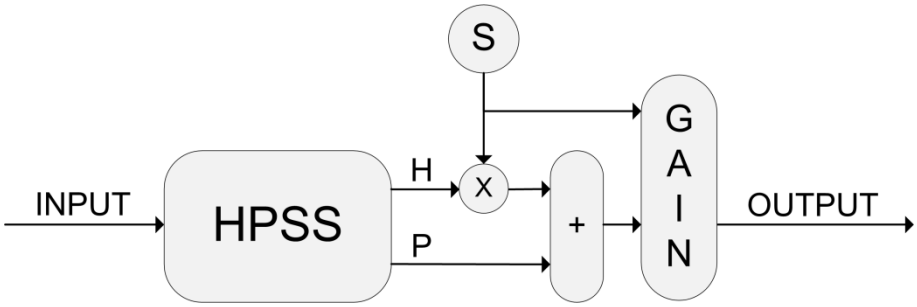


Figure 3-1: Music pre-processing scheme for CI using harmonic/percussive sound separation (HPSS)

The music pre-processing scheme for CI users, which performs vocal and drum enhancement on complex music, is shown in Figure 3-1. The harmonic/percussive sound separation (HPSS) separates harmonic (H) and percussive (P) components by exploiting the "anisotropic smoothness" of these components in the spectrogram (Ono et al., 2008). "Anisotropic smoothness" of sound is defined as partial differentials of the spectrogram in temporal or frequency direction: harmonic components are "smooth in temporal direction" because they are sustained and periodic; percussive components are "smooth in frequency direction" because they are instantaneous and aperiodic (Tachibana et al., 2010). The original power spectrogram $W_{\tau,\omega} = STFT[w(t)]$ from input signal $w(t)$, in which indices τ and ω represent time and frequency, respectively, is decomposed into the harmonic component $H_{\tau,\omega}$ and the percussive component $P_{\tau,\omega}$. In the original paper by Ono et al. (2008), the L_2 norm of the power spectrogram gradients was examined for evaluating the anisotropic smoothness, that is, $H_{\tau,\omega}$ and $P_{\tau,\omega}$ were found by minimizing:

$$J(\boldsymbol{H},\boldsymbol{P}) = \frac{1}{2\sigma_H^2}\sum_{\tau,\omega}\left(H_{\tau-1,\omega} - H_{\tau,\omega}\right)^2 + \frac{1}{2\sigma_P^2}\sum_{\tau,\omega}\left(P_{\tau,\omega-1} - P_{\tau,\omega}\right)^2 \qquad (3.1)$$

under the constraint of

$$H_{\tau,\omega}{}^2 + P_{\tau,\omega}{}^2 = W_{\tau,\omega}{}^2 \tag{3.2}$$

$$H_{\tau,\omega} \geq 0,\ P_{\tau,\omega} \geq 0 \tag{3.3}$$

where ***H*** and ***P*** are sets of $H_{\tau,\omega}$ and $P_{\tau,\omega}$, respectively, and σ_H and σ_P are parameters to control the weights of the horizontal and vertical smoothness.

Several algorithms were explored by Tachibana et al. (2012) to solve this optimization problem numerically, which were divided in two main approaches: power spectrogram estimation based on low-pass filtering and based on optimization. The parameters of the different algorithms were tuned using the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) as performance criteria. The "optimization under hard mixing constraint", referred to as "HM2", was found to outperform the other algorithms by Tachibana et al. (2012). After estimating the power spectrograms $H_{\tau,\omega}$ and $P_{\tau,\omega}$, a time-frequency mask was considered before applying inverse STFT to estimate the corresponding waveforms $h(t)$ and $p(t)$. From the different considered types of time-frequency mask, the binary mask was found most effective by Tachibana et al. (2012) to improve SIR. Therefore, in the music pre-processing scheme of Figure 3-1, the "HM2" approach is selected for the HPSS together with the binary mask, resulting in the following iteration formulae:

$$H_{\tau,\omega}{}^2 \leftarrow \frac{\alpha_{\tau,\omega} W_{\tau,\omega}{}^2}{(\alpha_{\tau,\omega}+\beta_{\tau,\omega})} \tag{3.4}$$

$$P_{\tau,\omega}{}^2 \leftarrow \frac{\beta_{\tau,\omega} W_{\tau,\omega}{}^2}{(\alpha_{\tau,\omega}+\beta_{\tau,\omega})} \tag{3.5}$$

where

$$\alpha_{\tau,\omega} = (H_{\tau+1,\omega} + H_{\tau-1,\omega})^2 \tag{3.6}$$

$$\beta_{\tau,\omega} = \kappa^2 (P_{\tau,\omega+1} + P_{\tau,\omega-1})^2 \tag{3.7}$$

in which κ is a tunable parameter that is optimized to maximize SIR of vocals and drums. The binary mask, which is applied on the input power spectrogram to eventually separate H and P, is defined as:

$$BM_{\tau,\omega} = 1 \qquad P_{\tau,\omega} > H_{\tau,\omega} \tag{3.8}$$
$$= 0 \qquad P_{\tau,\omega} \leq H_{\tau,\omega}$$

Vocal tones which contain 4-8 Hz quasi-periodic vibrations of the fundamental frequencies (F0s) and which do not sustain for a long time are contrasted to chord tones which contain very few fluctuations and which are temporarily maintained stationary. The nature of the respective tones is called “temporal-variability” and “temporal-stability” in Tachibana et al. (2010). Adjusting the time-frequency resolution of the short-time Fourier transform (STFT) in the HPSS calculations results in a different classification for the temporally-variable components. A STFT with long time window (100-500 ms) provides low temporal resolution and high frequency resolution. Consequently, sounds with long length and narrow bandwidth (temporally-stable sounds) appear “smoothly” in temporal direction (H-component) while sounds with moderately-or-very short length and moderately-or-very broad bandwidth (percussive and temporally-variable sounds) appear “smoothly” in frequency direction (P-component). Conversely, a STFT with short time window (30 ms) provides high temporal resolution and low frequency resolution resulting in the classification of temporally-variable sounds as well as temporally-stable sounds in the H-component.

In the music pre-processing scheme of Figure 3-1, the window length for the STFT in HPSS is 185 ms, which is a trade-off between good vocals/drums separation and latency. It corresponds to an 8192-points FFT with a sampling frequency of 44.1 kHz. As a result of this choice, the temporally-variable components (such as vocal tones) are classified as P-component. The

obtained P-component with vocals and drums is added to the harmonic component (H), which is attenuated with an adjustable parameter 'S' ranging from -∞ to 0 dB. The output spectrogram after addition becomes:

$$W_{\tau,\omega}^{out} = P_{\tau,\omega} + S * H_{\tau,\omega} \tag{3.9}$$

With the attenuation parameter 'S' at 0 dB, the output signal of the HPSS remains unaltered compared to the input signal since the binary mask from equation (3.8) is applied on the original input spectrogram to separate H and P. The final stage in the music pre-processing scheme applies a gain to the output signal as a function of the attenuation parameter 'S' to compensate for the decrease in output level due to the attenuated H-component.

The scheme presented in Figure 3-1 is implemented in Simulink® (Mathworks®) with a sampling frequency of 44.1 kHz. To enable real-time processing of the harmonic/percussive sound separation, a sliding analysis block is defined which is visualized in Figure 3-2. It is assumed that the separation of a certain time-frequency bin is only weakly affected by the bins far away (Ono et al., 2009). The sliding block only contains *B* consecutive frames with sliding index *n* ($n \leq \tau \leq n + B - 1$) which are processed at a certain iteration. The following four steps are followed iteratively to update *H* and *P*:

1. Add a new frame to the sliding analysis block: $H_{n+B-1,\omega} = P_{n+B-1,\omega} = W_{n+B-1,\omega}/2$

2. Update *H* and *P* with equations (3.4), (3.5), (3.6), and (3.7) for the frames in sliding analysis block *n* ($n \leq \tau \leq n + B - 1$)

3. Extract the *n*th frame ($H_{n,\omega}$ and $P_{n,\omega}$) for further analysis with the binary mask from equation (3.8)

4. Increment *n* to slide the analysis block

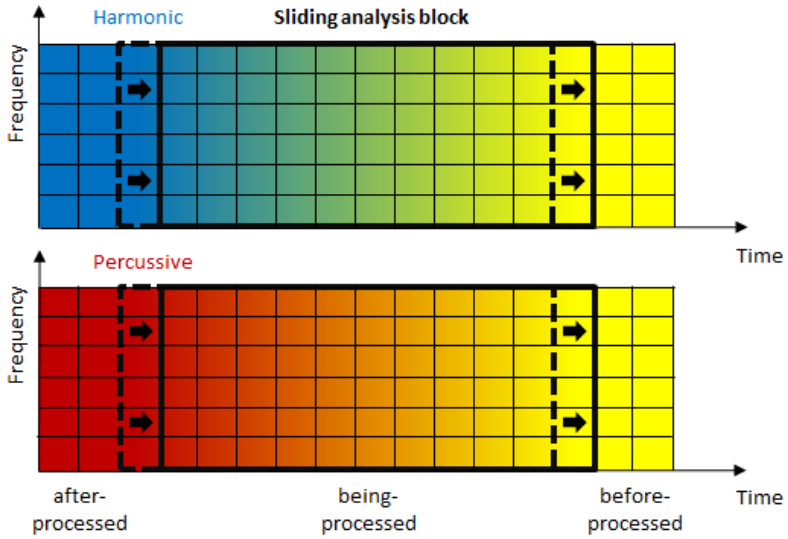


Figure 3-2: Visualization of the sliding analysis block for real-time processing of the harmonic/percussive sound separation (adopted from Ono et al., 2009)

3.4 Objective testing

The music pre-processing scheme in Figure 3-1 was evaluated objectively with multi-track recordings of pop music used in Buyens et al. (2014). A typical pop song consists of a vocal melody with piano, guitar, bass guitar and drums. The HPSS with a long time window separates vocals and drums (P) from the other instruments (H). The H-component attenuated with an adjustable parameter 'S' is added to the P-component, which results in the output signal. The remainder of the signal or "residual" signal is defined as the components of the original signal not present in the output signal.

Consequently, the sum of output and residual signal corresponds to the input signal. The performance of the separation of vocals/drums and the other instruments was analyzed with the energy ratio of the output signal and the residual signal for each track (Ono et al., 2008), and was calculated as

$$r_i^{output} = \frac{E_i^{output}}{E_i^{residual} + E_i^{output}} \tag{3.10}$$

where

$$E_i^{output} = < f_i(t), output(t) >^2 \tag{3.11}$$

$$E_i^{residual} = < f_i(t), residual(t) >^2 \tag{3.12}$$

in which <> represents the cross correlation operation and $f_i(t)$ the signal of the i-th track.

The energy ratio of the output signal is shown in Figure 3-3 (left) for the different tracks of the song 'The Dock of the Bay' by Otis Redding (Vocals, Piano, Guitar, Bass and Drums) from Buyens et al. (2014). The results are shown with attenuation parameter 'S' equal to 0 dB, -6 dB, -12 dB and -18 dB. With parameter 'S' at 0 dB, the output signal is equal to the input signal, resulting in a residual signal with zero amplitude and, consequently, an energy ratio of 'one' for each track. The vocal and drum track are preserved in the output signal regardless of the attenuation parameter, whereas the piano, guitar and bass guitar track shift towards the residual signal when decreasing the attenuation parameter. In Figure 3-3 (right), the energy ratio of the output signal is shown for the different tracks of the same song in different artificially constructed audio mixes with a fixed level for vocals and drums, and an attenuation of 0 dB, -6 dB, -12 dB and -18 dB for piano, bass and guitar relative to vocals/drums. The energy ratio is calculated with

formula (3.10) in which the residual signal is obtained from the subtraction of the output signal from the input signal. The shape of the curves in Figure 3-3 (left) and Figure 3-3 (right) are comparable, which indicates that the music pre-processing scheme visualized in Figure 3-3 (left) is capable of achieving the desired modifications in the relative instrument level settings as found in Buyens et al. (2014), visualized in Figure 3-3 (right). The curves of the piano, guitar and bass track, which are by definition equal in Figure 3-3 (right), show a certain spread in Figure 3-3 (left). Since in this song the piano is mainly playing sustained chords while the guitar is playing sharply, the guitar track appears to be more percussive than the piano track. Oftentimes, the opposite trend is observed since the piano is in fact a percussive instrument in which wired strings are struck by small hammers attached to the end of the keys.

Figure 3-4 shows the energy ratio of the output signal for the songs studied in Buyens et al. (2014) for different values of attenuation parameter 'S'. Energy ratios were calculated for each track in each song and were presented as harmonic and percussive components. The "harmonic" components (piano, guitar, bass) are attenuated in the output signal with decreasing attenuation parameter 'S', whereas the "percussive" components (vocals, drums) are mostly preserved in the output signal.

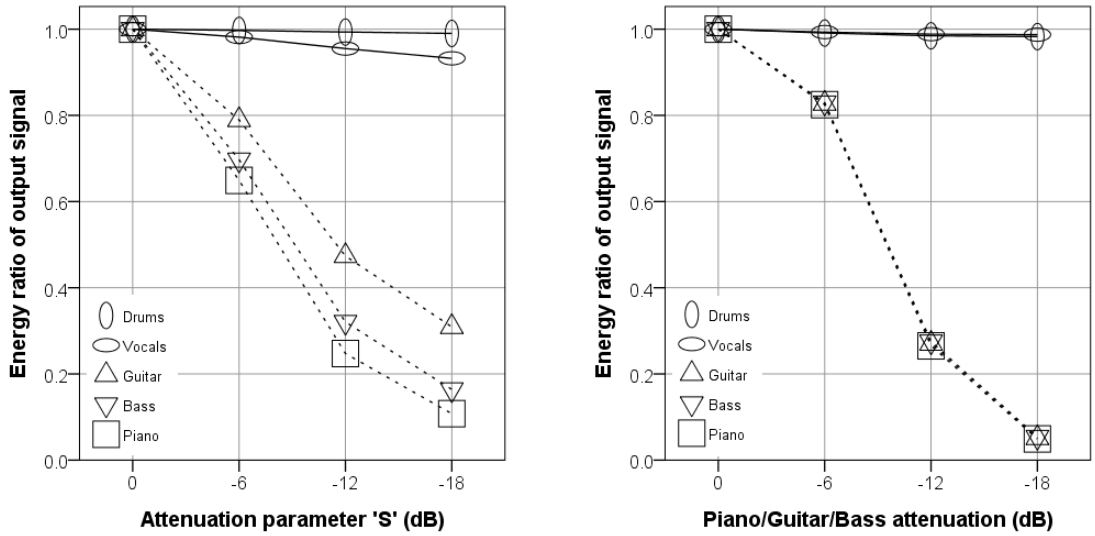


Figure 3-3: Energy ratio of output signal for vocal, piano, guitar, bass and drum track of "The dock of the Bay" by Otis Redding. Left: for different values of the attenuation parameter 'S' in the music pre-processing scheme for CI. Right: for different Piano/Guitar/Bass attenuation relative to vocals and drums (0 dB).

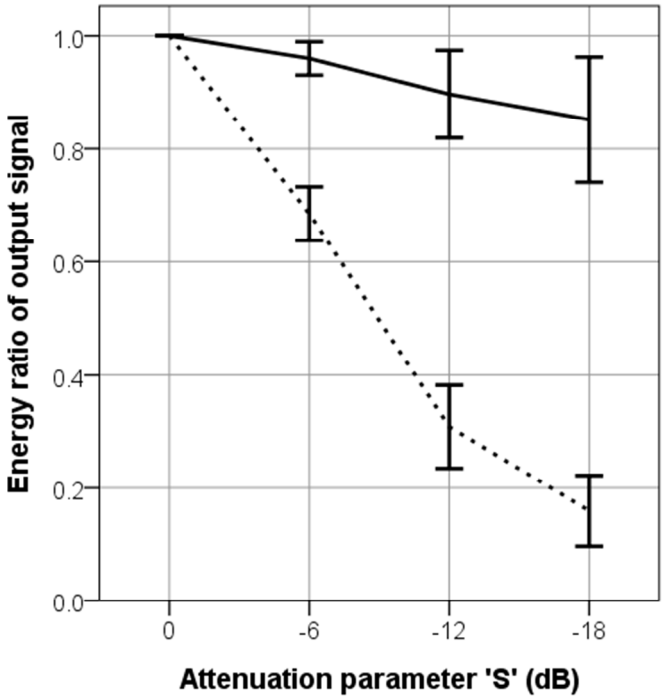


Figure 3-4: Energy ratio of output signal for the percussive and harmonic components from the songs studied in Buyens et al. (2014) for different values of 'S'. Percussive components (straight line) include vocal and drum tracks; harmonic components (dotted line) include piano, guitar and bass tracks. Error bars indicate 95% confidence interval.

3.5 Subjective testing

For the subjective evaluation of the music pre-processing scheme, a selection of pop/rock songs was used. The top twenty-five songs from the all-time greatest hits list of a popular radio station in Belgium (Joe FM) were gathered. Representative excerpts of the songs were selected with an average length of 24 seconds and were mixed down to MONO wave files with sampling rate of 44.1 kHz. The excerpts were processed with the music pre-processing scheme with an attenuation parameter 'S' equal to -18 dB.

A pairwise comparison was conducted with the processed and unprocessed (or original) condition for the 25 selected pop/rock songs. The pairs were randomly presented and repeated three times for each song. The music was presented to NH test subjects using a CI simulation with noise-band vocoder as used by Qazi et al. (2013). In each of the 22 frequency bands, the extracted temporal envelope after the maxima selection process was used to modulate a pink noise signal that was bandpass filtered corresponding to the analysis channel. All the modulated channels were then summed to produce the vocoded stimuli. Finally, all stimuli were equalized in rms level. The excerpts were played through headphones (Beyerdynamic DT-770 pro) in a silent room. The test subjects were asked to select the condition that was most enjoyable and to quantify their preference with a rating score ranging from *Imperceptible*, *Slightly better*, *Better*, *Largely better* to *Hugely better* (Figure 3-5).

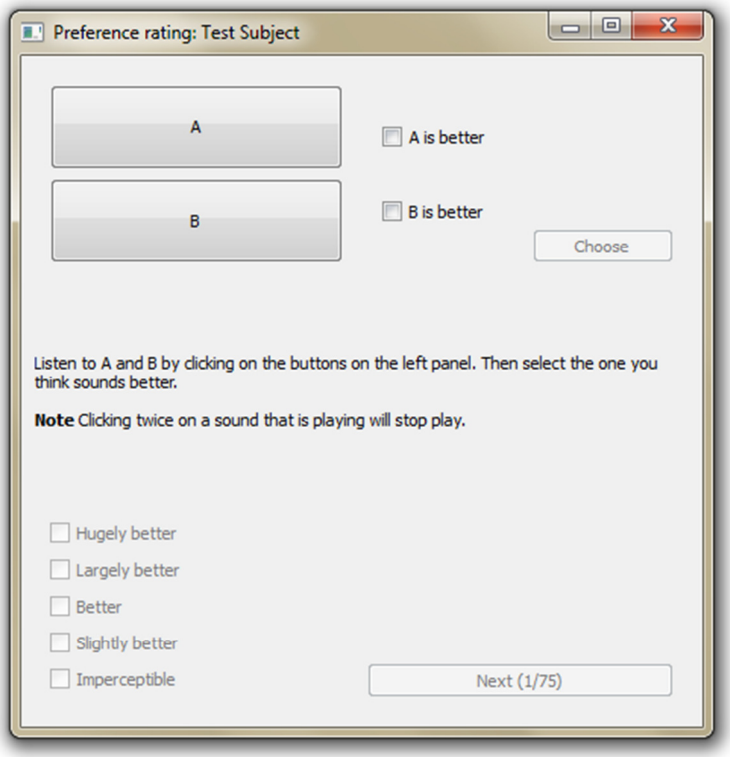


Figure 3-5: Graphical User Interface for pairwise comparison analysis with processed and unprocessed music excerpts.

The pairwise comparison is presented in percentage indicating the preference for the processed songs. The additional preference rating score from *Imperceptible* to *Hugely Better* is presented as median for the preferred condition and indicates the strength of the preference.

A pilot experiment with five NH test subjects was performed. The test subjects had no self-reported hearing deficit and their age ranged from 26 to 66 years old (average = 37 years). The results are shown in Table 3-1. Four test subjects significantly preferred the processed songs (Chi-square test: $p < .05$), with a median rating of *Slightly Better* or *Better*. Test subject S5, who is a professional DJ, significantly preferred the unprocessed songs (Chi-square test: $p < .05$) with a median rating of *Slightly Better*. In the processed songs the vocals are more understandable and the rhythm/beat is more noticeable. However, the background instruments, which appear very noisy

in the CI simulation, are reduced. This was preferred by most of the NH test subjects listening to CI simulation. The preference for the unprocessed songs in subject S5 could be explained by the professional activities of the test subject and the consequential familiarity with the songs. A preference rating experiment with CI subjects is necessary to confirm these results.

Table 3-1: Results of preference rating experiment, including preference for the processed songs, 95% confidence interval and median rating for the preferred condition.

Subject	Preference Processed (%)	95% Confidence Interval	Median Rating
S1	76	(66%-86%)	Better
S2	83	(74%-91%)	Slightly Better
S3	63	(51%-74%)	Better
S4	65	(54%-76%)	Slightly Better
S5	24	(14%-34%)	Slightly Better

3.6 Conclusion

A music pre-processing scheme aimed at improving music appreciation in CI users was described and evaluated. The scheme is capable of modifying the instrument level settings of music while preserving vocals and drums, which was found beneficial for the music appreciation in CI users. The scheme was evaluated subjectively using a preference rating experiment with NH test subjects and CI simulated music and showed significant preference for the processed music for all test subjects except one.

Chapter 4: A stereo music pre-processing scheme for cochlear implant users[7]

4.1 Abstract

Objective: Listening to music is still one of the more challenging aspects of using a cochlear implant (CI) for most users. Simple musical structures, a clear rhythm/beat and lyrics that are easy to follow are among the top factors contributing to music appreciation for CI users. Modifying the audio mix of complex music potentially improves music enjoyment in CI users.

Methods: A stereo music pre-processing scheme is described in which vocals, drums and bass are emphasized based on the representation of the

[7] The content of this chapter is adopted from: Buyens, W., van Dijk, B., Wouters, J., Moonen, M. (2015a). A stereo music pre-processing scheme for cochlear implant users, IEEE Transactions on Biomedical Engineering (in press)

harmonic and the percussive components in the input spectrogram, combined with the spatial allocation of instruments in typical stereo recordings. The scheme is assessed with post-lingually deafened CI subjects (N = 7) using pop/rock music excerpts with different complexity levels.

Results: The scheme is capable of modifying relative instrument level settings, with the aim of improving music appreciation in CI users, and allows individual preference adjustments. The assessment with CI subjects confirms the preference for more emphasis on vocals, drums and bass as offered by the pre-processing scheme, especially for songs with higher complexity.

Conclusion: The stereo music pre-processing scheme has the potential to improve music enjoyment in CI users by modifying the audio mix in widespread (stereo) music recordings.

Significance: Since music enjoyment in CI users is generally poor, this scheme can assist the music listening experience of CI users as a training or rehabilitation tool.

4.2 Introduction

A cochlear implant (CI) is a medical device enabling people with severe-to-profound sensorineural hearing loss to perceive sounds by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea (Loizou, 1998). This type of hearing loss is mostly caused by malfunctions in the hair cells of the cochlea and can be congenital or acquired after birth. Although CI users reach good speech understanding in quiet surroundings, music perception and appreciation generally remain

poor (McDermott, 2004). Simple musical structures, a clear rhythm/beat and lyrics that are easy to follow were reported amongst the top factors contributing to music appreciation in CI users by Gfeller et al. (2000). A negative correlation was found between (subjective) complexity and appreciation, studied with pop, country and classical music by Gfeller et al. (2003). CI users were asked to rate complexity and appraisal for different music excerpts on a scale from 0 to 100. Classical music was rated as more complex than pop and country music. Several plausible explanations were provided including the presence of simple musical structures and lyrics in pop and country music. Since CIs were mainly developed for transmitting speech sounds, the presence of lyrics may make it easier for CI users to follow the sequence of events in complex music. In addition, both pop and country music often contain a strong, simple beat which is well encoded in the electrical stimulation pattern of current CIs. Moreover, the performance for CI subjects on rhythmic pattern perception tasks is nearly as good as for NH subjects (McDermott, 2004; Kong et al., 2004; Looi et al., 2008). On the other hand, the accurate transmission of spectral and fine-structure information (as is often the case in instrumental classical music) remains challenging due to channel interactions, limited number of stimulation channels and limited dynamic range.

The preference for clear vocals and a strong rhythm/beat in CI users was demonstrated before by modifying relative instrument level settings in pop music by Buyens et al. (2014). A significant difference in preference rating scores was found between NH and CI subjects. For the pop songs provided, CI subjects preferred an audio mix with higher vocals-to-instruments ratio compared to NH subjects. In addition, given an audio mix with clear vocals and attenuated instruments, CI subjects preferred the drums/bass track to be louder than the other instrument tracks. Although individual differences

were observed across subjects, the potential for improving music appraisal by modifying the audio mix is apparent. The relative instrument level settings were modified by altering the levels of the different, separately recorded, instrument tracks, which are, however, not widely available for most music. To accomplish the same modification in relative instrument levels with mono or stereo music recordings, a specific signal processing scheme is needed. By using sound source separation techniques, vocals, drums and bass can be separated out, and then the residual signal is mixed in at a different level to provide the output of this specific signal processing scheme. Several approaches have been studied to tackle the sound source separation problem, and can be divided in two main categories: single-channel methods and multi-channel methods (Virtanen, 2006). The approaches used in single-channel sound source separation can be categorized roughly into model-based inference, unsupervised learning and psycho-acoustically motivated methods. Mostly, a combination of these approaches is used in practice. In model-based inference a parametric model of the sound sources to be separated is employed in which the model parameters are estimated from the observed mixture signal. The sinusoids plus noise model is the most commonly used model, introduced in music signal processing by Smith and Serra (1987), and Serra (1989, 1997). Unsupervised learning methods use a simple non-parametric model and estimate the sound source characteristics from the data based on Independent Component Analysis (ICA), Non-Negative Matrix Factorization (NMF) or Sparse Coding. Uhle et al. (2003) applied ICA to the spectrogram, and classified the extracted independent components into a harmonic and a percussive group based on features like e.g. percussiveness and noise-likeness. Helen and Virtanen (2005) utilized NMF for decomposing the spectrogram into elementary patterns and classified them by a pre-trained Support Vector Machine (SVM). Sparse coding methods represent a mixture

signal in terms of a small number of active elements chosen from a larger set (Olshausen and Field, 1997), and were used in the analysis of music signals by Abdallah and Plumbley (2004) and Virtanen (2003). Although good performance was demonstrated for many different signals, the aforementioned approaches often have their limitations. First, some sound sources are hard to model with a few fixed spectral templates and in practice require careful tuning of sophisticated models (Liutkus et al., 2014; Ozerov et al., 2012). Second, the typical assumption that different sources are characterized by different sets of spectra may not be realistic in many cases in music signals.

Instead of decomposing the sound sources as a combination of fixed patterns, in psycho-acoustically motivated methods the elementary time-frequency components of the incoming signal are categorized into their respective sound sources based on association cues such as spectral proximity, harmonic concordance, synchronous changes and spatial proximity (Bregman, 1990). A simple and fast algorithm to perform harmonic/percussive sound separation is based on the "anisotropic smoothness" of the harmonic and percussive components in the spectrogram (Ono et al., 2008). Harmonic components appear smooth in the temporal direction, whereas percussive components appear smooth in the frequency direction in the spectrogram. A (mono) music pre-processing scheme for CI users based on harmonic/percussive sound separation was described by Buyens et al. (2013) for a single-channel input. This scheme emphasized vocals and drums in complex pop music in order to improve music appreciation in CI users.

Multi-channel methods (in contrast to single-channel methods) can take advantage of the availability of spatial information, e.g., from recordings

with multiple microphones placed at different positions, enabling acoustic beamforming or blind separation of convolutive mixtures to recover the sound sources. The typical karaoke problem to remove vocals from background music also exploits the spatial information of stereo recordings (York et al., 2004).

In this paper, a stereo music pre-processing scheme for CI users is described, which is a stereo extension of the mono music pre-processing scheme from Buyens et al. (2013) with improved performance based on exploiting the spatial information in stereo recordings. The building blocks for this stereo music pre-processing scheme are described in detail in section 4.3. The evaluation of the scheme with CI users and pop music excerpts is presented in section 4.4 and 4.5.

4.3 Stereo music pre-processing scheme

The stereo music pre-processing scheme, which performs vocals, drums and bass enhancement on complex stereo music, is shown in Figure 4-1. The scheme contains a “Vocals & Drums Extraction” block applied on the input spectrogram and a “Stereo Binary Mask” block to promote components that are located in the centre of the stereo image. The output of the scheme contains the extracted signal (indicated as “Vocals, Drums, Bass”) mixed together with an attenuated version of the residual signal (indicated as “Other”).

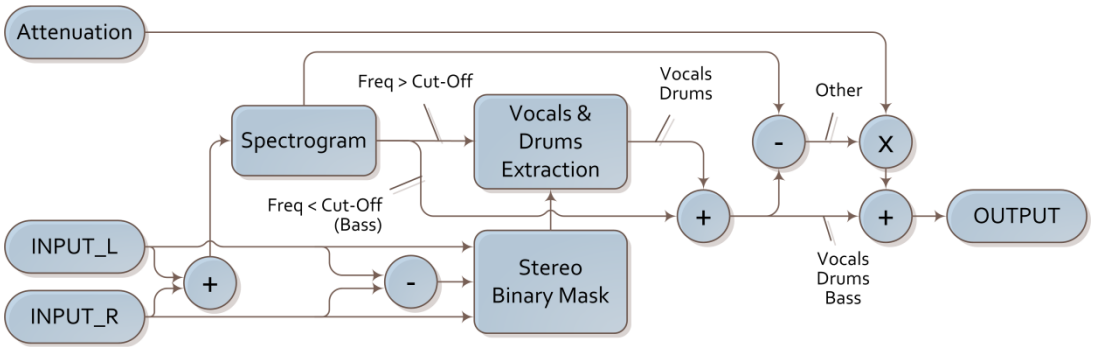


Figure 4-1: Schematic of the stereo music pre-processing scheme for CI users which is enhancing vocals/drums/bass while attenuating the 'Other' instruments with parameter 'Attenuation'. It is based on "Vocals & Drums Extraction" of the input spectrogram (Input_L + Input_R with frequency > cut-off frequency) and a "Stereo Binary Mask" to exploit the spatial information in stereo recordings (based on Input_L, Input_R, Input_L-Input_R).

In paragraph 4.3.1, the "Vocals & Drums Extraction" is described in detail, performing vocals and drums enhancement on a mono input signal which is – in the case of a stereo input – the sum of the left and the right channel. The processing of the bass frequencies is explained in paragraph 4.3.2. In paragraph 4.3.3 the typical mixing settings in stereo recordings are explained and the inclusion of the stereo mixing property is described as a constraint in the optimization problem with the "Stereo Binary Mask". Paragraph 4.3.4 concludes with the computation of the output of the stereo music pre-processing scheme.

4.3.1 Vocals & Drums Extraction

The "Vocals & Drums Extraction" block is adopted from the mono music pre-processing scheme for CI users in Buyens et al. (2013) and is based on the harmonic/percussive sound separation (HPSS) in Ono et al. (2008), which separates harmonic (H) and percussive (P) components by exploiting the "anisotropic smoothness" of these components in the spectrogram. "Anisotropic smoothness" is defined based on partial differentials of the

spectrogram in the temporal or the frequency direction: harmonic components are “smooth in the temporal direction” because they are sustained and periodic; percussive components are “smooth in the frequency direction” because they are instantaneous and aperiodic (Tachibana et al., 2010). The input spectrogram $W_{\tau,\omega} = STFT(w(t))$, which is calculated using the short-time Fourier transform (STFT) and a Hamming window, is decomposed into the harmonic components $H_{\tau,\omega}$ and the percussive components $P_{\tau,\omega}$. The indices τ and ω represent time and frequency, respectively. The L_2 norms of the spectrogram gradients are used as a metric for the anisotropic smoothness, that is, $H_{\tau,\omega}$ and $P_{\tau,\omega}$ are found by minimizing:

$$J(\boldsymbol{H}, \boldsymbol{P}) = \frac{1}{2\sigma_H^2}\sum_{\tau,\omega}\left(H_{\tau-1,\omega} - H_{\tau,\omega}\right)^2 + \frac{1}{2\sigma_P^2}\sum_{\tau,\omega}\left(P_{\tau,\omega-1} - P_{\tau,\omega}\right)^2 \quad (4.1)$$

under the constraint of

$$H_{\tau,\omega}{}^2 + P_{\tau,\omega}{}^2 = \left|W_{\tau,\omega}\right|^2 \quad (4.2)$$

$$H_{\tau,\omega} \geq 0,\ P_{\tau,\omega} \geq 0 \quad (4.3)$$

where ***H*** and ***P*** are sets of $H_{\tau,\omega}$ and $P_{\tau,\omega}$, respectively, and σ_H and σ_P are parameters to control the weights of the horizontal and vertical smoothness. This optimization problem can be solved numerically, which results in the following iteration formulae (Tachibana et al., 2012):

$$H_{\tau,\omega}^{2\ (j+1)} = \frac{\alpha_{\tau,\omega}{}^{(j)}\left|W_{\tau,\omega}\right|^2}{(\alpha_{\tau,\omega}{}^{(j)} + \beta_{\tau,\omega}{}^{(j)})} \quad (4.4)$$

$$P_{\tau,\omega}^{2\ (j+1)} = \frac{\beta_{\tau,\omega}{}^{(j)}\left|W_{\tau,\omega}\right|^2}{(\alpha_{\tau,\omega}{}^{(j)} + \beta_{\tau,\omega}{}^{(j)})} \quad (4.5)$$

where *j* is the iteration index, and

$$\alpha_{\tau,\omega}{}^{(j)} = (H_{\tau+1,\omega}{}^{(j)} + H_{\tau-1,\omega}{}^{(j)})^2 \tag{4.6}$$

$$\beta_{\tau,\omega}{}^{(j)} = \kappa^2(P_{\tau,\omega+1}{}^{(j)} + P_{\tau,\omega-1}{}^{(j)})^2 \tag{4.7}$$

$$\kappa = \frac{\sigma_H^2}{\sigma_P^2} \tag{4.8}$$

The tunable parameter κ is optimized to maximize the separation of vocals and drums from the other instruments by using (4.16) and the multi-track recordings (vocals, drums, bass, guitar, piano) used by Buyens et al. (2014). From the estimated spectrograms $H_{\tau,\omega}$ and $P_{\tau,\omega}$, a time-frequency mask is defined which is then used to estimate the corresponding waveforms h(t) and p(t). From the considered time-frequency masks, the binary mask was found more effective by Tachibana et al. (2012) to improve the separation performance compared to the Wiener mask or not applying any mask. The binary mask ($BM_{\tau,\omega}$) is defined as:

$$\begin{aligned} BM_{\tau,\omega} &= 1 \qquad P_{\tau,\omega} > H_{\tau,\omega} \\ &= 0 \qquad P_{\tau,\omega} \leq H_{\tau,\omega} \end{aligned} \tag{4.9}$$

Using this binary mask, the P- and H-components are computed from the input spectrogram as:

$$p(t) = ISTFT\left(BM_{\tau,\omega} \,.\, W_{\tau,\omega}\right) \tag{4.10}$$

$$h(t) = ISTFT\left(\left(1 - BM_{\tau,\omega}\right).\, W_{\tau,\omega}\right) \tag{4.11}$$

Figure 4-2 shows the separation performance versus the number of iterations with and without binary mask for the multi-track recordings used

by Buyens et al. (2014). The benefit of the binary mask on the separation performance is clearly seen. The separation performance is defined as the improvement in signal-to-noise ratio of the P-components $p(t)$, in which the signal (S) consists of vocals and drums and the noise (N) represents the other instruments. Since multi-track recordings are available from the songs used in Buyens et al. (2014), the signal-to-noise ratio can be calculated at both the input and the output as:

$$SNR = 20 * log_{10}\left(\frac{rms(S)}{rms(N)}\right) \tag{4.12}$$

To indicate the presence of the different instruments in the P-components, the energy ratio for the different tracks is used, which is calculated as:

$$r_p^i = \frac{E_p^i}{E_h^i + E_p^i} \tag{4.13}$$

where

$$E_p^i = < f_i(t), p(t) > \tag{4.14}$$

$$E_h^i = < f_i(t), h(t) > \tag{4.15}$$

in which <> represents the cross-correlation operation with time-lag zero and $f_i(t)$ the signal of track *i*. Therefore, the separation of vocals and drums from the other instruments can also be defined as the difference (δ) in energy ratio of the P-components for vocals/drums and bass/guitar/piano:

$$\delta = \left(r_p^{vocals/drums}\right) - \left(r_p^{bass/guitar/piano}\right) \tag{4.16}$$

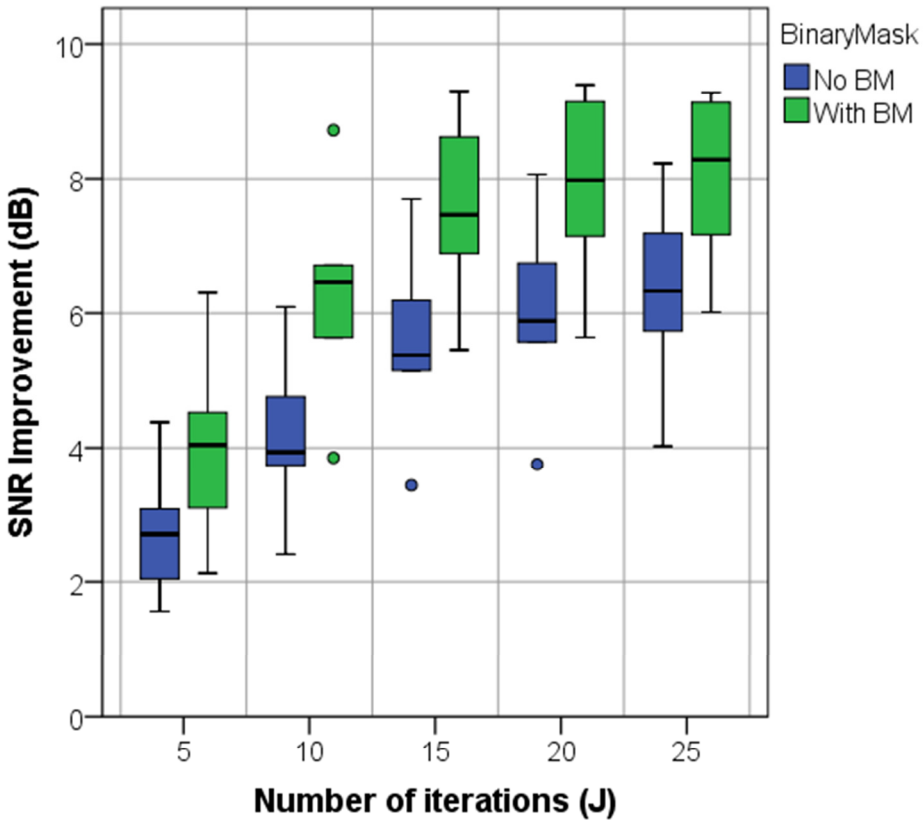


Figure 4-2: Boxplot with SNR improvement (dB) of the P-components with vocals/drums versus the other instruments for the multi-track recordings used in Buyens et al. (2014) as a function of the number of iterations (J) with and without applying the binary mask (BM) from (4.9). The window length of the STFT used in this graph is 185 ms.

The separation results from the "Vocals & Drums Extraction" block show that in addition to the drums, also the vocals can indeed be extracted as P-components based on their representation in the input spectrogram. "Temporally-variable" sounds are contrasted to "temporally-stable" sounds. The "temporally-variable" sounds (such as vocal tones) contain 4-8 Hz quasi-periodic vibrations of the fundamental frequencies (F0s) and do not sustain for a long time. On the other hand, the "temporally-stable" sounds (such as chord tones) contain very few fluctuations and are maintained stationary for a while (Tachibana et al., 2010). Adjusting the time-frequency resolution of the STFT results in a different classification for the temporally-variable components. For a STFT with long time window (100-500 ms) percussive and

temporally-variable sounds appear “smooth” in the frequency direction (P-components), whereas for a STFT with short time window (30 ms) temporally-stable sounds as well as temporally-variable sounds appear “smooth” in the temporal direction (H-components). This is illustrated in Figure 4-3 for a typical pop song with vocals, drums, bass, guitar and piano. The energy ratio of the P-components calculated with (4.13) is shown for every instrument as a function of the window length. Instruments with high energy ratio (close to one) appear as P-components, whereas instruments with low energy ratio (close to zero) appear as H-components. Figure 4-3 shows that the vocals appear in the H-components if a small STFT window length is adopted, whereas with a large STFT window length, the vocals appear in the P-components. A window length above 185 ms is good for vocals/drums separation, but the longer the window the more latency. Hence we choose in the “Vocals & Drums Extraction” block (Figure 4-1) a window length of 185 ms, resulting in the classification of temporally-variable components (such as vocals) as P-components. According to informal feedback, the small distortions on the vocals introduced by the algorithm are hardly noticeable for CI users. In speech, similar distortions introduced by a binary mask were studied by Qazi et al. (2012), showing no degradation in quality for CI users, as opposed to NH listeners, but a significant improvement in speech understanding in noise.

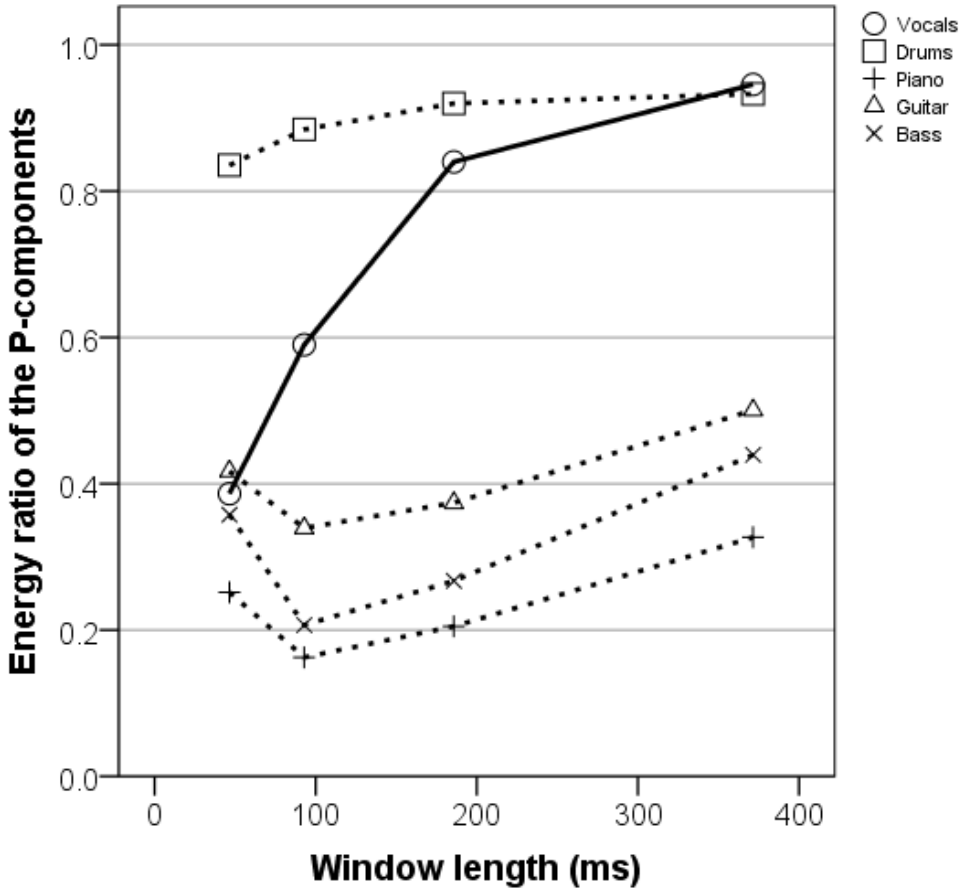


Figure 4-3: Energy ratio of the P-components (j=15) for the different tracks of a typical pop song as a function of the STFT window length. The vocals (solid line) are separated as P-components with high STFT window length and as H-components with low STFT window length.

The impact of the proposed settings of the "Vocals & Drums Extraction" on other instruments is shown in Figure 4-4. The energy ratio of the P-components is computed for different instruments. An energy ratio close to zero means the instrument is removed from the P-components, whereas an energy ratio close to one means the instrument remains in the P-components. Instrument samples are collected from the instrument recognition tests MACarena (Omran et al., 2010) and UW-CAMP (Nimmons et al., 2008; Kang et al., 2009). The MACarena samples consist of the first 8 bars of a traditional Swedish folksong played by professional musicians on different instruments. The UW-CAMP samples consist of a 5-note melodic sequence played on different instruments with the same detached

articulation. The majority of the instruments show low energy ratio and are thus removed from the P-components with the “Vocals & Drums Extraction”. However, the flute and the violin from the MACarena test show a higher energy ratio value due to the very strong *vibrato* in the music samples. This contrasts with the UW-CAMP flute and violin which have a low energy ratio value and no *vibrato*. The pitched percussion instruments (guitar and piano) are not completely removed in the P-components because of the clear percussive onset, except for the MACarena piano which plays the sequence of notes in *legato*.

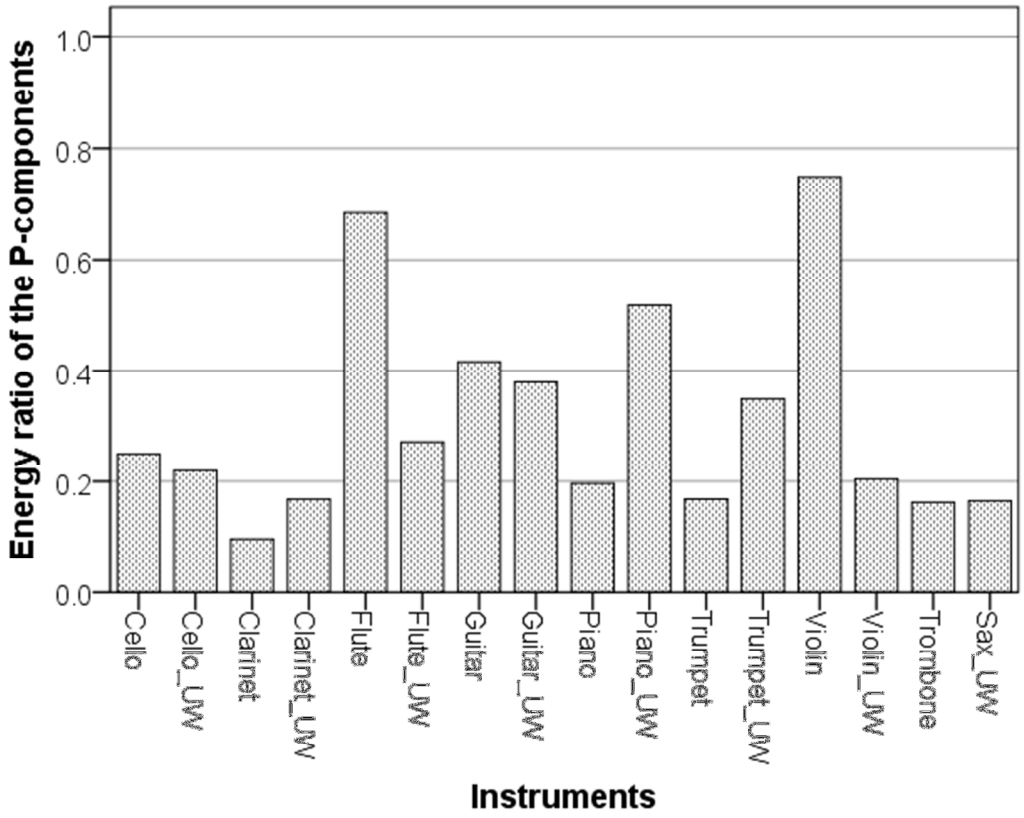


Figure 4-4: Energy ratio of the P-components for different instrument samples from the instrument recognition tests MACarena and UW-CAMP. (j=15, STFT 185 ms)

4.3.2 Bass frequency extraction

The bass guitar is classified with the H-components in the “Vocals & Drums Extraction” block, which means it is attenuated or removed from the P-components. In exploratory tests with CI subjects listening to the music pre-processing scheme from Buyens et al. (2013), which is based on the “Vocals & Drums extraction” block only, emptiness in the sound of the pre-processed music was reported, mainly because of the lack of bass frequencies. Bass guitar and the bass frequencies in general give more fullness to the music and the attenuation of bass frequencies may result in chopped (bass) drum sounds. Therefore, the bass frequencies should be preserved in the output signal of the music pre-processing scheme. Moreover, the preference for bass sounds was also found in the music mixing preference study with CI users by Buyens et al. (2014). To preserve the bass frequencies in the output of the music pre-processing scheme, the binary mask for the frequency bins below a chosen cut-off frequency is uniformly set to 1, whereas the binary mask for the higher frequency bins is calculated according to their percussiveness (Figure 4-1). To set the cut-off frequency, eight solo bass tracks from different pop songs were passed through the music pre-processing scheme and the energy ratio of the P-components for these bass tracks was evaluated for different cut-off frequencies (Figure 4-5). The cut-off frequency is incorporated in the binary mask formula (4.9) as follows:

$$\begin{aligned} BM_{\tau,\omega} = 1 \qquad & P_{\tau,\omega} > H_{\tau,\omega} \;\; or \;\;\; \omega \leq \omega_{cut-off} \\ = 0 \qquad & P_{\tau,\omega} \leq H_{\tau,\omega} \;\; and \;\; \omega > \omega_{cut-off} \end{aligned} \tag{4.17}$$

With a cut-off frequency of 400 Hz the bass guitar is mostly present in the P-components, therefore 400 Hz is chosen for the cut-off frequency in the music pre-processing scheme of Figure 4-1.

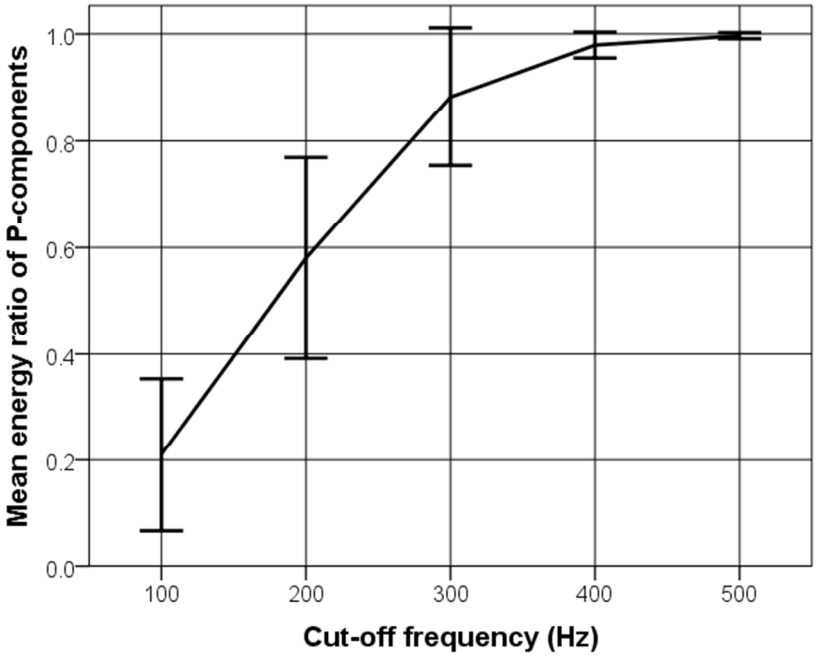


Figure 4-5: Mean energy ratio of the P-components for eight solo bass guitar tracks processed with the music pre-processing scheme as a function of the cut-off frequency. Error bars represent 95% confidence interval.

4.3.3 Stereo binary mask

Music is widely available in stereo recordings, which means that all instruments are mixed together in two channels (left and right). Stereo recordings aim to provide the listener with the experience of a live performance in which instrument sounds are coming from the direction of the corresponding musicians on stage (Kim et al., 2012). A typical stereo recording consists of vocals, (parts of the) drums and bass in the centre and other instruments, such as guitar and piano, panned to the left or the right. This property is also used in karaoke systems to remove the vocals from the instruments. A typical voice removal algorithm for stereo recordings was described by York et al. (2004), in which the high-pass filtered left and right channel were subtracted from each other and mixed together with the low-pass filtered left and right channel. Consequently, only vocals were removed

in the output, whereas the low-frequency content, such as drums and bass, was preserved. In contrast, the aim for the music pre-processing scheme is to emphasize vocals, drums and bass, which are typically mixed in the centre of the stereo image. Kim et al. (2012) identified the vocals in a stereo signal by comparing the spectrogram of the left channel, the right channel and the difference between left and right channel. The frequency bins for which the spectrogram of the difference is smaller than the spectrogram of the left and the right channel were classified as belonging to the vocals component. A binary mask to extract the centre part of the stereo image can thus be defined as:

$$\begin{aligned} BM_{stereo} = 1 \quad & \left(\theta * {W_{\tau,\omega}}^{diff}\right) < {W_{\tau,\omega}}^{L} \; and \; \left(\theta * {W_{\tau,\omega}}^{diff}\right) < {W_{\tau,\omega}}^{R} \\ = 0 \quad & \left(\theta * {W_{\tau,\omega}}^{diff}\right) \geq {W_{\tau,\omega}}^{L} \; or \; \left(\theta * {W_{\tau,\omega}}^{diff}\right) \geq {W_{\tau,\omega}}^{R} \end{aligned} \quad (4.18)$$

in which θ is a tunable parameter and ${W_{\tau,\omega}}^{L}$, ${W_{\tau,\omega}}^{R}$ and ${W_{\tau,\omega}}^{diff}$ the respective spectrogram of the left channel, the right channel and the difference between the left and the right channel. Applying the binary mask on the input spectrogram results in extracting the centre part of the stereo recording, which contains vocals, (part of the) drums and bass, but separation performance heavily depends on the broadness of the stereo image in the recording.

The “Vocals & Drums Extraction” block described in paragraph 4.3.1, separates vocals and drums from the other instruments based on a single-channel input signal. For stereo recordings, the performance of the “Vocals & Drums Extraction” can be significantly improved by exploiting the inherent spatial information of the instruments. To integrate this spatial information in the optimization problem (4.1) an additional stereo constraint (4.19) is

added next to constraints (4.2) and (4.3) which limits the P-components to the centre of the stereo image (cfr. formula (4.18)):

$$\left(\theta * {P_{\tau,\omega}}^{diff}\right) < {P_{\tau,\omega}}^{L} \; and \; \left(\theta * {P_{\tau,\omega}}^{diff}\right) < {P_{\tau,\omega}}^{R} \tag{4.19}$$

Consequently, the update formulae (4.4) and (4.5) are changed into:

$$P_{\tau,\omega}^{2\ (j+1)} = \frac{{\beta_{\tau,\omega}}^{(j)} \left|W_{\tau,\omega}\right|^2}{\left({\alpha_{\tau,\omega}}^{(j)} + {\beta_{\tau,\omega}}^{(j)}\right)} \tag{4.20}$$

$$P_{\tau,\omega}^{2\ (j+1)} = BM_{stereo} * P_{\tau,\omega}^{2\ (j+1)} \tag{4.21}$$

$$H_{\tau,\omega}^{2\ (j+1)} = \left|W_{\tau,\omega}\right|^2 - P_{\tau,\omega}^{2\ (j+1)} \tag{4.22}$$

with BM_{stereo} defined in (4.18) and $\alpha_{\tau,\omega}$ and $\beta_{\tau,\omega}$ defined in (4.6) and (4.7).

The SNR improvement for vocals/drums versus other instruments as a function of the stereo parameter θ from (4.18) is shown in Figure 4-6 for five different audio mixes of a typical pop song with vocals, drums, bass, guitar and piano. The audio mixes were artificially constructed with vocals, drums and bass in the centre of the stereo image and different stereo panning χ ranging from 0 to 100 for piano and guitar. Panning χ = 0 means no panning for piano and guitar, resulting in a mono signal, whereas panning χ = 100 means complete panning of piano and guitar respectively to the left and to the right of the stereo image. It is clear that there is no improvement in separation performance for the audio mix with panning χ = 0 (mono), nor for the other audio mixes with stereo parameter θ equal to zero. Increasing the stereo parameter θ results in an improved SNR, but as shown in Figure 4-7 also results in distorted vocals/drums.

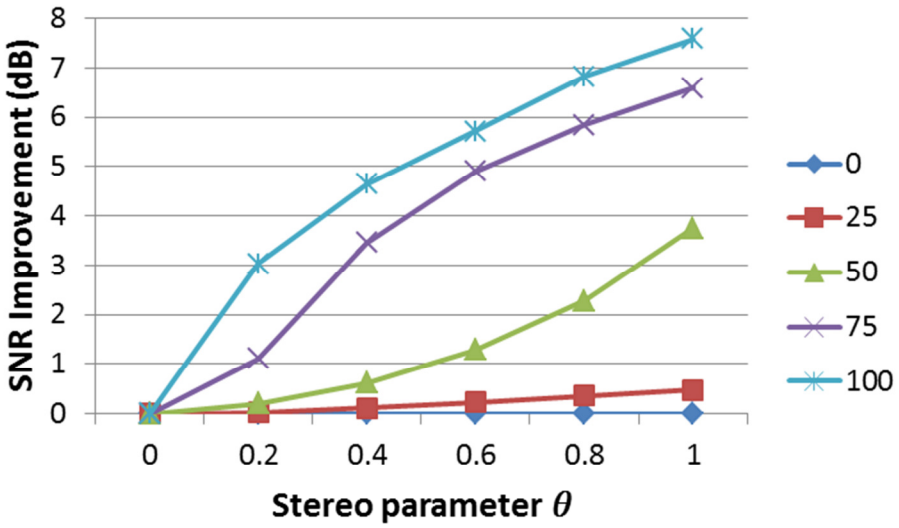


Figure 4-6: SNR improvement for vocals/drums versus other instruments as a function of the stereo parameter θ from (4.18) for different stereo mixes with panning χ ranging from 0 to 100 for piano (panned to the left) and guitar (panned to the right).

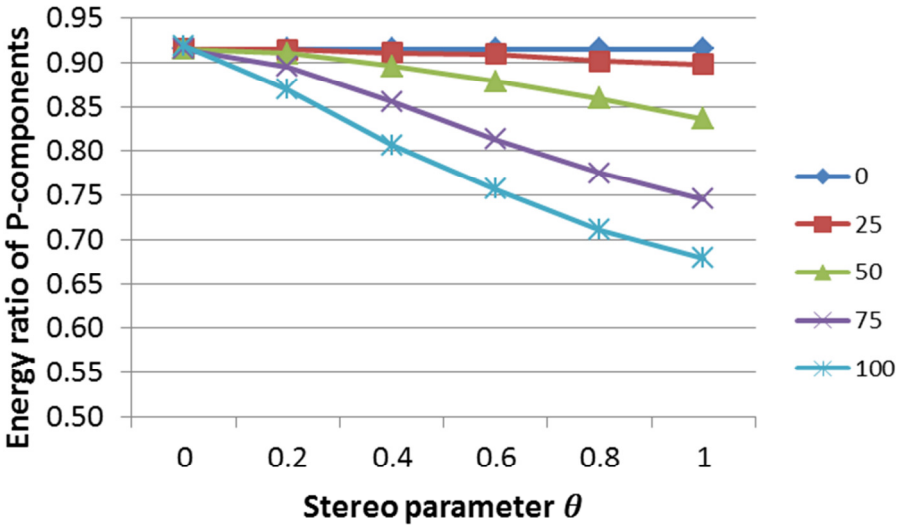


Figure 4-7: Vocals/drums distortion indicated as the energy ratio of the P-components for the vocals/drums track as a function of the stereo parameter θ from (4.18) for different stereo mixes with panning χ ranging from 0 to 100 for piano (panned to the left) and guitar (panned to the right).

In the case that the vocals are not mixed in the centre of the stereo image but panned to the left or the right, e.g. for background singers, the inclusion

of the stereo parameter θ in (4.18) and (4.21) influences the vocals separation. Figure 4-8 shows the influence of the stereo parameter θ on the separation of the vocals represented by the energy ratio of the P-components for the vocals track. Five different audio mixes of a typical pop song were artificially constructed atypically with bass, guitar and piano in the centre of the stereo image, and vocals and drums panned to left and right with panning level χ ranging from 0 to 100. Increasing the stereo parameter θ results in the removal of the vocals from the P-components depending on its panning level χ.

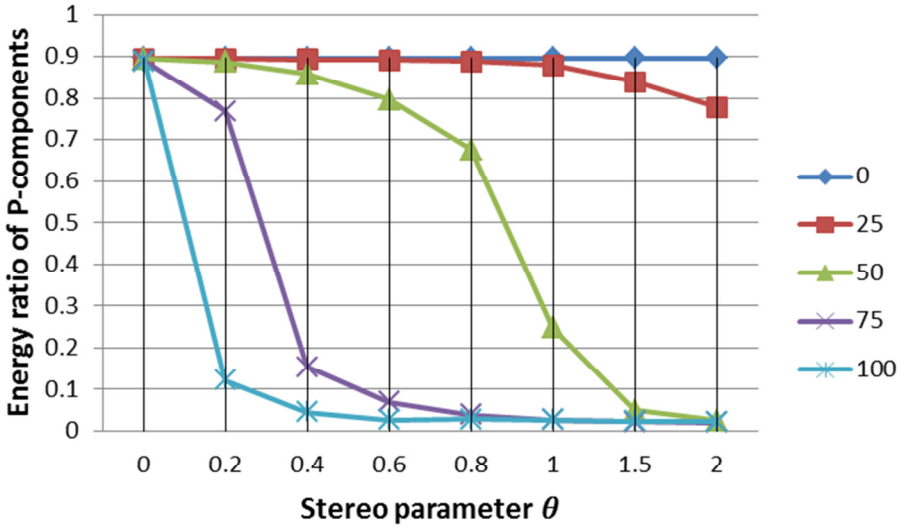


Figure 4-8: Removal of the off-centre vocals track (from 0 to 100) from the P-components of the music pre-processing scheme visualized as the energy ratio of the P-components for the vocals track as a function of the stereo parameter θ.

The parameter θ is determined to be 0.4, which is a good compromise between vocal distortion and instrument attenuation.

4.3.4 Stereo music pre-processing output

As shown in Figure 4-1, the obtained P-components with vocals, drums and bass are added to the H-components, which are attenuated with an adjustable parameter '*Attenuation*'. The output spectrogram after addition becomes:

$$W_{\tau,\omega}^{out} = P_{\tau,\omega} + Attenuation * H_{\tau,\omega} \quad (4.23)$$

The corresponding output waveform is obtained from the inverse STFT:

$$output(t) = ISTFT\left[\left(W_{\tau,\omega}^{out}\right) e^{j\angle W_{\tau,\omega}}\right] \quad (4.24)$$

in which the phase information from the input ($\angle W_{\tau,\omega}$) is reused. With the attenuation parameter equal to 0 dB, the output signal of the stereo music pre-processing scheme remains unaltered compared to the input signal. The final stage in the music pre-processing scheme applies a gain to the output signal as a function of the attenuation parameter to compensate for the decrease in output level due to the attenuated H-components.

4.4 Methods

The stereo music pre-processing scheme with an adjustable attenuation parameter was evaluated with CI subjects. This section includes a description of the sound material, the demographic and etiological information of the CI test subjects and the test procedure.

4.4.1 Sound material

For the perceptual evaluation of the stereo music pre-processing scheme, a selection of pop/rock songs was used from the top fifty songs in the all-time greatest hits list of a popular radio station in Belgium (Joe FM). Representative excerpts of the songs with an average length of 27 seconds and an average dynamic range of 10.0 dB (SD = 1.5 dB) were selected. The song excerpts were rms-equalized and stored as stereo wav-files with sampling rate of 44.1 kHz. A complexity rating experiment was performed with twelve NH test subjects with no self-reported hearing deficit. The subjects were recruited with an internal advertisement, had diverse musical background and were familiar with most of the music excerpts. The fifty song excerpts were played through headphones (Beyerdynamic DT-770 pro) in a silent room and the test subjects were asked to rate the musical complexity of the song on a scale from 1 to 100 with a slider in a graphical user interface on a laptop. No further definition or information was given to the test subjects in order not to prime them in the experiment. This resulted in a ranking of the fifty songs from least complex to most complex for every subject, which was quite consistent across subjects. The average ranking over all subjects was used to compose three groups of songs, containing the 8 least complex songs, the 8 most complex songs and 8 songs with medium complexity. These 24 excerpts were used in the experiment described in paragraph 4.4.3. The selected songs were categorized according to their complexity as low (1-8), mid (9-16) and high (17-24), and include Evergreens (2, 3, 5), Pop ballads (1, 4, 6, 7), Pop (13, 15, 16, 17, 21) and Rock excerpts (8, 9, 10, 11, 12, 14, 18, 19, 20, 22, 23, 24).

4.4.2 Subjects

Seven post-lingually deafened CI subjects (all Cochlear™ Nucleus®) participated in the present study. A summary with demographic and etiological information can be found in Table 4-1. The speech performance results - as provided by the clinic - were measured with an adaptive Speech Reception Threshold (SRT) test with LIST material and Speech Shaped Noise (Van Wieringen and Wouters, 2008). The familiarity with the songs was not explicitly assessed in the current study, but from informal feedback after the experiment, it was determined that in general the familiarity with the songs was rather low for the CI subjects. The subjects signed a consent form and were paid for their travel expenses. Ethical committee approval was obtained.

Table 4-1: Demographic and etiological information of seven post-lingually deafened CI test subjects.

Subject	Age (y)	Gender	CI exp (y)	SRT (dB)	Etiology	Sound processor	Implant type
CI1	51	Male	1	-2	Colitis Ulcerosa	CP810	CI24RE(CA)
CI2	63	Male	3	1	Progressive	CP810	CI24RE(CA)
CI3	60	Male	8	1	Unknown	Freedom	CI24R(CS)
CI4	75	Female	5	-1	Otosclerosis	Freedom	CI24RE(CA)
CI5	55	Female	3	-3	Unknown	Freedom	CI24RE(CA)
CI6	62	Female	1	-3	Progressive	CP810	CI24RE(CA)
CI7	28	Male	10	6	Meningitis	CP810	CI24R(CS)

4.4.3 Perceptual evaluation

For the evaluation of the stereo music pre-processing scheme, the twenty-four pop/rock songs, which were divided in three groups of eight songs with respectively low, mid and high complexity, were used. Test subjects were asked to select the preferred setting for the attenuation parameter in the range -6 dB to 30 dB, referring to the attenuation of the H-components in the output signal of the stereo music pre-processing scheme. A negative value means an amplification of the H-components. The settings were visualized on a graphical user interface with buttons numbered from 1 to 7, representing the range of the attenuation parameter from -6 dB to 30 dB (or reversed) in steps of 6 dB. The order was randomized in every trial to prevent the test subject from using the visual cue in the evaluation. The song excerpts were played continuously in a loop in free-field in a sound-treated room at a level of 60 dB(A). The experiment was performed on a laptop connected to a loudspeaker (Genelec 8020A). The songs were presented in random order and repeated three times. The actual instructions for the test subject were: (1) press the ‘play’ button to start listening to the first music excerpt, (2) listen to the seven modified music excerpts and identify the differences, and (3) select the music excerpt that is most enjoyable for you and press ‘next’ to store your preference and to move to the next music excerpt. The test subjects were allowed to take a break whenever necessary. The experiment took on average 2 to 2.5 hours per subject.

4.5 Results

Figure 4-9 shows the individual settings for the preferred attenuation of H-components for the three different groups of songs (low, mid and high complexity). In the rightmost column, the average scores from the seven subjects are indicated. All individual CI subjects preferred for every group of songs a setting with attenuated H-components which is significantly different from zero (One-Sample Wilcoxon Signed-ranks Test, $p < .05$). However, individual differences occur across subjects with on the one hand subject CI1 with low preferred attenuation between 0 dB and 6 dB and on the other hand subject CI4 with high preferred attenuation up to 24 dB for complex songs. Test subjects CI2, CI4, CI6 and CI7 preferred higher attenuation of H-components for songs with high complexity, whereas subjects CI1, CI3 and CI5 preferred the H-components to be attenuated irrespective of the complexity of the songs. The average preferred attenuation for each group (8 songs x 7 subjects x 3 repetitions) was analyzed with the Wilcoxon Signed-ranks Test with Bonferroni correction and is significantly higher for songs with high complexity compared to songs with low complexity ($Z = 4.71$, $p < .001$) or mid complexity ($Z = 2.64$, $p = .024$). The difference in preferred attenuation between low and mid complexity is not significant after Bonferroni correction ($Z = 2.08$, $p = .11$).

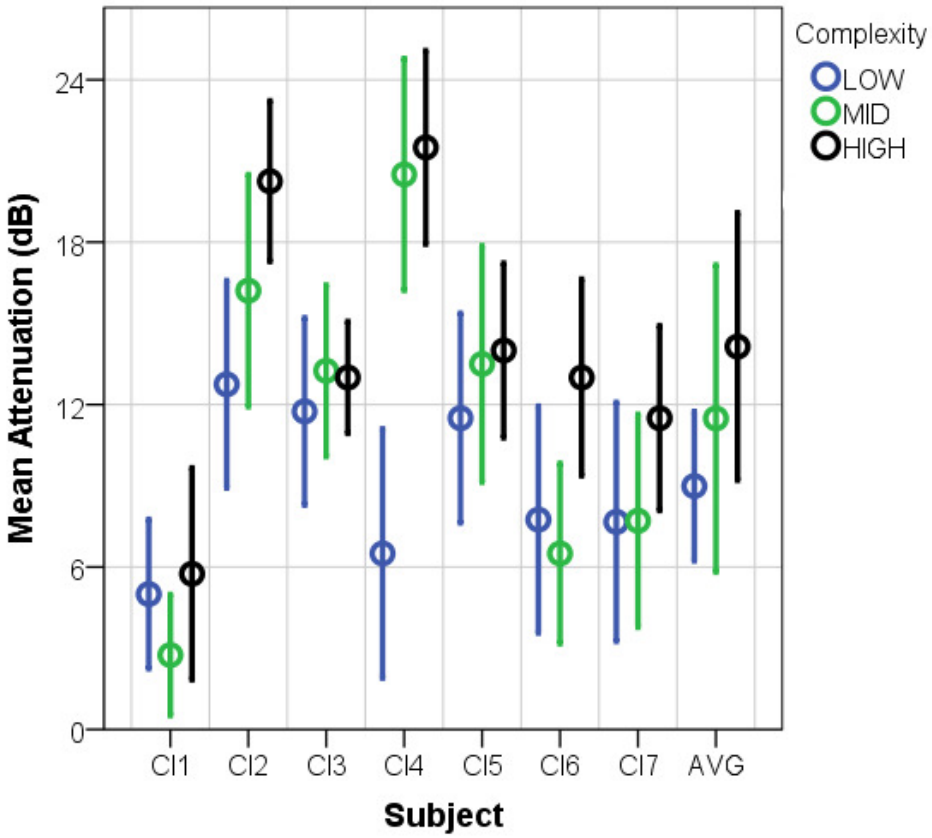


Figure 4-9: Individual results for 7 CI subjects with their preferred setting for the attenuation of the H-components for 24 song excerpts with low, mid and high complexity. The average preferred setting from the seven subjects for low, mid and high complexity songs are in the rightmost column. Error bars represent 95% confidence interval.

The preferred average setting of the attenuation parameter for the 24 song excerpts is positively correlated with the complexity of the songs as rated by the NH subjects (Pearson's $r(24) = .67$, $p < .001$) (Figure 4-10). For more complex songs, more modification of the song was preferred in terms of attenuation of the H-components; for less complex songs, the attenuation of the H-components was less needed or preferred to increase music enjoyment.

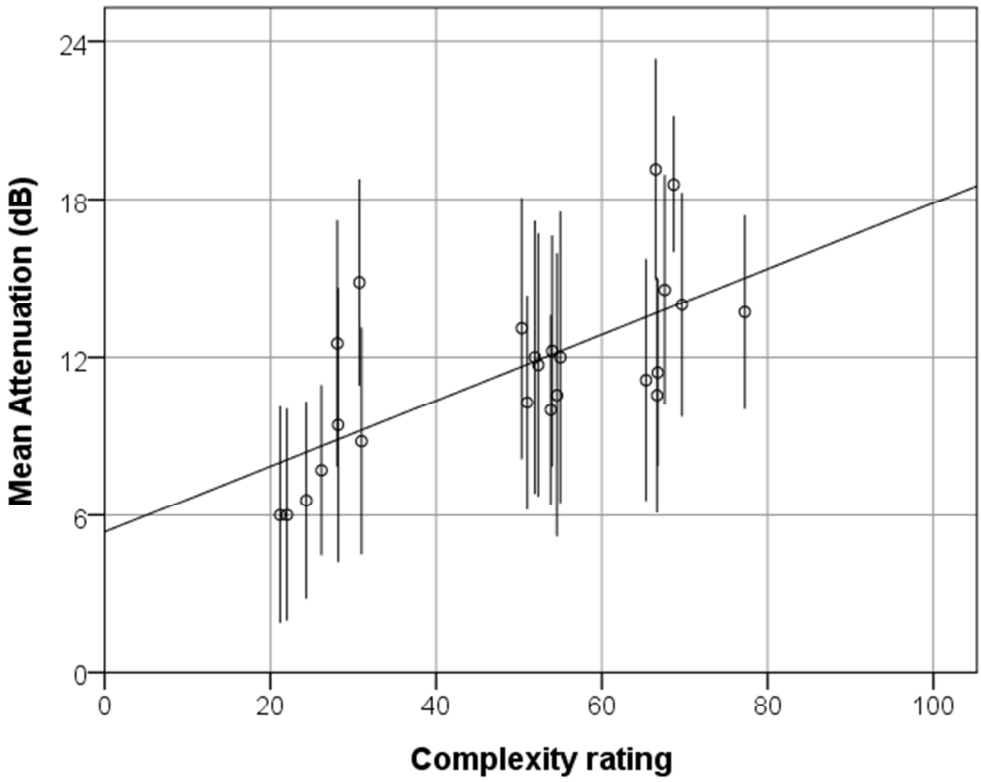


Figure 4-10: Mean preferred attenuation of the H-components for the 24 song excerpts with 7 CI subjects as a function of the complexity rating given by 12 NH subjects. Error bars represent 95% confidence interval. Straight line is the linear regression (R^2 = .43).

4.6 Discussion

A stereo music pre-processing scheme for CI users is described which has the potential to enhance music enjoyment by emphasizing vocals, drums and bass in complex music. The signal processing techniques used are applicable in real-time (with latency around 500 ms), which makes them interesting for music listening and/or music rehabilitation for CI users. When listening to songs from a music library, the latency is not an issue. However, for audiovisual music stimuli, the synchronization between audio and video is corrupted and should be resolved (if possible) by delaying the video signal with the (stable) latency. While attending a live performance, the latency

cannot be compensated and, moreover, the performance of the proposed signal processing scheme is likely to be suboptimal, since applying the stereo binary mask is only effective on stereo recordings or – in the case of a live performance – on two microphones capturing the sound in the sweet spot of the stereo image.

The “Vocals & Drums extraction” block from Figure 4-1 is based on the separation of H- and P-components, which has been developed and used to improve the performance of music information retrieval tasks such as chord detection and drum detection respectively. In this paper, it is used to enhance vocals and drums in an audio mix for CI users. Different approaches for drums extraction based on harmonic/percussive sound separation were described by Ono et al. (2008) and Fitzgerald (2010). A comparative evaluation of various harmonic/percussive sound separation algorithms based on the anisotropic continuity in the spectrogram was described by Tachibana et al. (2012). Five different methods were analyzed and the parameter sets for each method were tuned to maximize the signal-to-interference ratio and the signal-to-distortion ratio for H or P. Tachibana et al. (2010) used the same harmonic/percussive sound separation techniques for melody extraction by changing the STFT window length, which can be combined with the drums extraction as in Buyens et al. (2013) and Figure 4-3 to achieve the “Vocals & Drums Extraction” from Figure 4-1. The separation performance with mono signals, which is illustrated in Figure 4-2 for a selection of songs, can be significantly improved for stereo signals by exploiting the spatial properties of a typical stereo recording with vocals, drums and bass in the centre of the stereo image and the other instruments panned to left or right. By increasing the stereo parameter θ from (4.18), an increase in SNR for the vocals/drums track is achieved, which is visualized in Figure 4-6 for different artificially constructed audio mixes. The increase in

SNR is most apparent for audio mixes with instruments panned completely to left or right (panning χ = 100). However, increasing the stereo parameter θ might also introduce distortions in vocals/drums, which is shown in Figure 4-7 with the energy ratio for the vocals/drums track. Moreover, if the vocals are not panned in the centre of the audio mix, they may get distorted or completely removed in the output signal. Figure 4-8 shows that in an artificially constructed audio mix the vocals with e.g. panning χ = 75 or χ = 100 are eliminated almost entirely with stereo parameter θ = 0.4. Background vocals are typically panned off-centre and are thus likely to be attenuated or removed. If, on the other hand, the panning of the instruments in the stereo audio mix is not distinct or if the input signal is a mono signal, the stereo music pre-processing scheme relies only on the different representations of H- and P-components in the input spectrogram.

In general, two types of separation problems can arise. The vocals can be categorized as H-components and get attenuated, or harmonic instruments can be categorized as P-components and not attenuated.

The distortions of the vocals with the Vocals & Drums extraction mostly occur in vocal parts with long duration of vowels and flat articulation. In this particular case, parts of the vocals are categorized as H-components and are attenuated. These artefacts can be reduced by tuning the mono music pre-processing scheme less aggressive, and by relying more on the stereo binary mask, which enhances the centre part of the stereo image in which the vocals are mostly present. On the other hand, a violation of this typical stereo mixing property may lead to distortions in the vocals as mentioned before in Figure 4-7 and Figure 4-8.

Harmonic instruments or parts of them can be categorized as P-components, which limits the range of adjusting vocals/bass/drums versus the other instruments in the music pre-processing scheme. Tuned percussive instruments (such as the piano) have sharp onsets, which are categorized as P-components, whereas the remainder of the notes are categorized as H-components. This results in the reduction of piano notes to short percussive attacks with limited pitch information. Also here, the stereo binary mask can help with further reduction of the remaining instrument parts, if this instrument is panned off-centre in the stereo image.

Clear vocals and clear rhythm/beat were described to be top factors contributing to music appreciation in CI users by Gfeller et al. (2000). A study with multi-track recordings in which CI subjects (N = 10) were able to adjust the different instrument levels in pop music excerpts indicated a similar preference for vocals, drums and bass in complex music in Buyens et al. (2014). The modification of the relative instrument level settings in complex music was found to be beneficial for music appreciation in CI users. The music pre-processing scheme described in this paper is able to emphasize vocal, drum and bass tracks based on mono or stereo recorded music. Songs with different complexity rating were presented to CI subjects who were asked to determine their preferred setting for the attenuation of the H-components. By attenuating the H-components in complex music a reduction of structural complexity is achieved. The preference for a high attenuation of H-components for songs with high complexity, as found in the current study, is in agreement with the findings from Gfeller et al. (2003), in which a negative correlation was found between complexity and music appreciation in CI users. The songs that are rated more complex are appreciated less and are - in the current study - preferred with higher attenuation of H-components as opposed to songs with low complexity that

are already appreciated more, and thus require lower attenuation. The individual differences among subjects may relate to the experience of the CI user with listening to music. Subjects CI1 and CI7 are active music listeners, whereas the other subjects only report a more passive music listening involvement. The experienced music listener CI1, who is capable of distinguishing the different instruments in a complex song, is missing out certain instruments when attenuating the H-components, whereas a non-experienced music listener (such as subject CI4) benefits from the attenuated H-components to more easily follow the lyrics and the song in general. The sound material excerpts used in this study are pop/rock. In a future study, more different styles of music should be included together with a thorough investigation on the musical habits and music experience of the subjects and their speech and pitch performance to explain the individual differences among subjects.

4.7 Conclusion

A stereo music pre-processing scheme aimed at improving music appreciation in CI users was described and evaluated. The scheme is capable of modifying the instrument level settings of music while preserving vocals, drums and bass, which was found to be beneficial for music appreciation in CI users. The scheme was evaluated subjectively using pop/rock song excerpts with low, medium and high complexity. On average, the preferred setting for the attenuation parameter was found to be significantly different for the group of songs with low and high complexity. The music pre-processing scheme has the potential to improve music appreciation in CI users, in particular for complex songs. Individual differences have been observed across subjects.

Chapter 5: Evaluation of stereo music pre-processing for cochlear implant users[8]

5.1 Abstract

Objective: The improvement in music appraisal is evaluated in cochlear implant subjects by using a music pre-processing scheme on a take-home device in a comfortable listening environment. The pre-processing allows adjusting the balance between vocals/bass/drums and other instruments for different genres of music. The correlation between the preferred settings and speech and pitch detection performance is investigated.

[8] The content of this chapter has been submitted for publication as: Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2015b). Evaluation of stereo music pre-processing for cochlear implant users. International Journal of Audiology

Design: During the initial visit preceding the take-home test the speech-in-noise perception performance and pitch detection abilities are measured, and a questionnaire about the music listening habits is completed. The take-home device is provided including the music pre-processing scheme and seven playlists with six songs each. The subject is asked to adjust the balance with a turning wheel to make the music sound most enjoyable for all songs three times.

Study Sample: Twelve post-lingually deafened cochlear implant users have participated.

Results: All subjects prefer a balance significantly different from the original. Differences across subjects are observed which cannot be explained by perceptual abilities. An effect of Genre is found, showing significantly lower preferred deviation from the original balance for Evergreens compared to the other genres.

Conclusions: The music pre-processing scheme shows potential improvement in music appraisal with complex music and may be a good tool for music training or rehabilitation.

5.2 Introduction

A cochlear implant (CI) is a medical device that provides auditory sensations to subjects with severe-to-profound hearing loss by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea (Loizou, 1998). Most of the CI users reach good speech understanding in quiet environments, but speech performance in noise varies across subjects, and

the perception of music is generally unsatisfactory for CI users. Earlier reports demonstrated poor pitch perception and melody recognition, inadequate timbre perception and worse quality ratings for musical sounds compared to NH subjects (McDermott, 2004; Mirza et al., 2003; Kong et al., 2004). A sound coding strategy to improve music perception for CI users was proposed and evaluated by Laneau et al. (2006) and Milczynski et al. (2009). Interesting improvements were shown on pitch perception, melodic contour identification and familiar melody identification, but no improvement in music enjoyment with complex music was revealed. CI devices were developed and optimized primarily for transmitting speech sounds. The transmission of the features in musical sounds is limited in CI devices due to the poor pitch representation (Limb, 2006; Limb and Roy, 2013). At the level of electrode-auditory nerve interface further degradation of the musical sounds - and sound representation in general – is caused by the spread of electrical stimulation, the limited stimulation of apical cochlear regions, the potential mismatch between electrode place maps and ideal cochlear locations, and the limited phase-locking to auditory input with electric hearing. Furthermore, the dynamic range is considerably compressed in CI users with a reduced number of distinguishable steps in loudness, which also leads to an additional decrease in music sound quality. On the other hand, poor pitch and timbre perception does not necessarily result in poor music appreciation. Drennan et al. (2015) reported on a clinical evaluation of music perception, appraisal and experience in CI users and showed only a weak relationship between perceptual accuracy and music enjoyment, suggesting that perception and appraisal are relatively independent for CI users. Also by Gfeller et al. (2008) and Lassaletta et al. (2008) no or weak association was found between music enjoyment and music perception skills. Wright and Uchanski (2012) studied music appraisal and perception skills in CI users as well as NH subjects listening to CI-simulated music. The NH subject results

gave a reasonable model for many music perception skills of CI users, but not for the rating of music appraisal. Here again, only weak or non-existent correlations were found between appraisal ratings and music perception scores.

The effect of complexity on music appraisal was studied by Gfeller et al. (2003), which showed a strong negative correlation between complexity and appraisal for CI users. Music excerpts that are (subjectively) rated as less complex (such as country music) were appreciated more than music excerpts that are rated as more complex (such as classical music). Moreover, music that involves multiple instruments sounds, on average, less pleasant for CI users than music played by single instruments (Looi et al., 2007). The reduction of complexity by modifying the relative instrument levels in the audio mix of complex music was studied with CI users by Buyens et al. (2014). The subjects were asked to make their preferred audio mix by using multi-track recordings and a mixing console. Subsequently, a preference rating experiment was performed with predefined audio mixes, showing a significant preference for an audio mix with clear vocals and attenuated instruments, while preserving the bass/drum track. Clear rhythm/beat and lyrics that are easy to follow were also found among the top factors to enhance music enjoyment by Gfeller et al. (2000). Since multi-track recordings are not widely available for most commercial music, signal processing techniques are required to modify the relative instrument levels in complex music. A music pre-processing scheme that is able to perform these modifications in stereo music was studied in Buyens et al. (2015a) and was evaluated in CI users with pop/rock music excerpts. Results show that the music pre-processing scheme potentially improves music appraisal by adjusting the balance between vocals/bass/drums and other instruments. Individual differences across subjects were observed.

The current paper investigates the potential improvement in music appraisal with this music pre-processing scheme in more detail with different genres of music and in a home listening environment, which is more comfortable and arguably more realistic than a sound booth. In the following section, the sound material and the take-home device are described, together with the design of the take-home test. The subsequent sections provide the results, the discussion of the results, and the conclusions.

5.3 Methods

The stereo music pre-processing scheme, which was fully described in Buyens et al. (2015a), is evaluated with post-lingually deafened CI users in a take-home test. In paragraph 5.3.1 the sound material used for this evaluation is described. Paragraph 5.3.2 contains the demographic and etiological information about the subjects. In paragraph 5.3.3 the take-home device is presented and, finally, in paragraph 5.3.4 the test procedure is explained in detail.

5.3.1 Sound material

It is impossible to include all existing musical genres and subgenres in the present study. The former 'Head of music Belgium' of Mood Media and still a professional DJ was asked to provide a selection of different music genres with specific sound characteristics which include vocals and which are generally widespread. The six genres are: Disco/Funk/Soul, Evergreens, Latin,

Pop (Dance), Rock and songs with Dutch lyrics. For each genre seven songs are chosen by the DJ which represent the specific characteristics of that genre. The sound material consists of the commercially available stereo recordings with sampling frequency of 44.1 kHz, with a total duration of more than 150 minutes. Table 5-1 lists the different genres with total duration together with the average dynamic range for the genre (averaged over the 7 songs), which gives an indication of the compression used in the final mix of the music. For the take-home test, seven playlists are compiled containing one song from each genre of music, randomized per subject. To assess the familiarity with the songs, the list of songs is provided and subjects are asked to indicate whether they are familiar with the song or not.

Table 5-1: Sound material for take-home test including the six music genres with their total duration together with the mean and standard deviation (SD) of the dynamic range (DR) over all songs.

Music genre	Total duration (mm:ss)	Mean DR value (dB)	SD (dB)
Disco/Funk/Soul	24:36	10.0	1.9
Evergreens	22:57	11.0	1.7
Latin	29:48	9.3	2.9
Pop Dance	28:14	6.3	0.8
Rock	26:33	6.6	2.2
Dutch songs	25:32	8.6	2.1

5.3.2 Subjects

Twelve post-lingually deafened CI subjects (all Cochlear™ Nucleus®) participated in the study. They were recruited with an advertisement on social media and in newsletters and mailings from user groups. A summary

with demographic and etiological information can be found in Table 5-2. The subjects signed a consent form and were paid for their travel expenses. Ethical committee approval was obtained for this study.

Table 5-2: Demographic and etiological information of the twelve post-lingually deafened CI subjects participating in the study.

Subject	Age (y)	Gender	CI exp (y)	Etiology	Sound processor
S1	64	Male	3	Progressive	CP810
S2	29	Male	11	Meningitis	CP810
S3	56	Female	6	Progressive	CP910
S4	62	Female	7	Unknown	CP810
S5	62	Male	9	Otosclerosis	CP810
S6	29	Female	5	Unknown	CP810
S7	67	Male	9	Streptomizine	CP910
S8	70	Male	9	Congenital	CP810
S9	64	Male	17	Otosclerosis	CP910
S10	52	Female	3	Unknown	CP810 + CP910
S11	74	Male	10	Struma	CP810
S12	45	Female	14	Meningitis	CP910

5.3.3 Take-home device

The subjects received a take-home device in order to be able to listen to the sound material in a comfortable listening environment. This device is a commercially available iPhone5 with a custom-made application for the experiment containing all the sound material. A “Personal Audio Cable” is used to link the take-home device with the sound processor of the subject. The three important features of the application are (1) the music library access and navigation buttons, (2) the turning wheel for adjusting the balance between vocals/bass/drums and the other instruments and (3) the

"Vote"-button to store the preferred setting for each song on the device. Figure 5-1 shows a screenshot of the application on the take-home device.

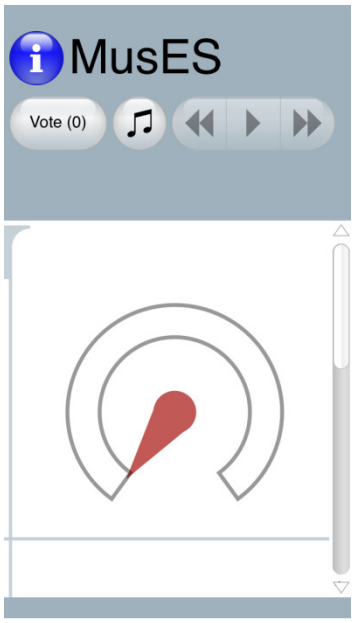


Figure 5-1: Screenshot of the music application on the take-home device including the music library access and navigation buttons, the turning wheel to adjust the balance and the "Vote"-button to store the preferred setting.

With the music library access button, the subject can select one of the seven playlists. With the navigation buttons the subject can navigate through the songs of the selected playlist. The song name and song progress are displayed on the screen while navigating and playing the songs. The turning wheel controls the attenuation parameter in the stereo music pre-processing scheme which was described in detail in Buyens et al. (2015a). The stereo music pre-processing scheme separates vocals, bass and drums from the other instruments by exploiting the representation of harmonic and percussive components in the input power spectrogram together with the spatial information contained in typical stereo recordings. The output signal with the extracted vocals, bass and drums is mixed together with the attenuated residual signal. The attenuation parameter ranges from -6 dB to +24 dB in which positive values represent an attenuation of the residual

signal, whereas negative values represent an amplification of the residual signal. An attenuation parameter of zero represents the original balance between vocals/bass/drums and other instruments. In order not to have an abrupt change when turning the wheel a mirrored version of the scale of the attenuation parameter is added to the original scale. A random rotation of this doubled scale is then applied, each time the navigation buttons are used to select a new song, in order not to prime the subject with a visual cue. Finally, the “Vote”-button stores the preferred attenuation parameter setting for each song. The counter on the “Vote”-button represents the number of times the preferred setting for a certain song has been stored. By pushing the “Vote”-button, the logging file on the take-home device is updated with one entry containing song name, song progress, preferred attenuation parameter setting and random rotation, together with a timestamp.

5.3.4 Test procedure

Before the start of the test with the take-home device, the subjects are asked to perform two initial experiments to determine their speech understanding performance in noise and pitch discrimination abilities, and to fill in a questionnaire about their music listening habits and music experience before and after the CI surgery. The speech understanding performance is measured with an adaptive Speech Reception Threshold (SRT) test with LIST material and Speech Shaped Noise (Van Wieringen & Wouters, 2008). The average of 2 SRT test runs (10 sentences) is used to indicate the speech-in-noise performance. If the difference between the two SRT test results is more than 2 dB, a third SRT test is performed, and the average of the last two SRT test results is used. The pitch discrimination abilities are measured with the pitch discrimination subtest from UW-CAMP, which is a 2-alternative forced-choice adaptive procedure to determine a

threshold interval for discrimination of complex pitch direction change (Nimmons et al., 2008; Kang et al., 2009). Synthesized complex tones with three different base frequencies are used (262 Hz, 330 Hz and 391 Hz). The tone with the base frequency and a higher-pitched tone determined by the random adaptive interval size are presented in random order and the subject has to indicate which of the two tones is higher in pitch. The amplitude of each tone was randomly roved +/- 4 dB to eliminate any loudness cues. The minimum tested interval is 1 semitone, and the maximum is 12 semitones or 1 octave. The test includes 3 randomly presented trials for each base frequency and ends when 8 reversals are reached. The threshold for each base frequency is calculated by averaging the last 6 reversals for each trial. Finally, the mean threshold over the three base frequencies is used to indicate the pitch discrimination ability. After the initial experiments the subjects are asked to complete a questionnaire which is based on the music questionnaire from Mirza (2003) translated into Dutch. It queries the subjects about their CI experience, music listening habits and appraisal, and it includes questions about playing an instrument and singing before deafness and after implantation.

After the initial experiments and the questionnaire, a take-home device is provided to the subjects together with a detailed manual. The subjects are trained to handle the take-home device. For the take-home test, the sound material is divided in seven playlists with six songs, one from each genre of music. The playlists are randomized for each subject with the only restriction of having one song from each genre of music in every playlist. The subject is asked to select one playlist at a time and to listen to the songs one by one while adjusting the balance between vocals/bass/drums and the other instruments with the turning wheel. If the preferred attenuation parameter setting is found, the subject can store the setting by pushing the “Vote”-

button. The subject is asked to listen to all the songs from the playlist three times in order to have three preferred attenuation parameter settings for each song. The same procedure is repeated for all playlists. Additional instructions are provided regarding the voting moment, especially for songs with a long intro. If in the beginning of a song only small or no changes are noticed, the voting moment should be delayed until the chorus, in which vocals and all instruments are present. After the first visit with initial experiments and training, the subject takes the device home for the listening test. The subjects are also asked to indicate their familiarity with the songs. The time foreseen for the take-home test is two weeks, with an optional intermediate visit if problems arise. During the final visit the logging file of the take-home device is read and feedback from the subject is discussed.

5.4 Results

The preferred attenuation parameter settings are stored in the take-home device three times for the seven songs from each of the six music genres. The average time spent before voting per song over all subjects is 92 seconds (SD 31 seconds). A first data check was performed before the analysis of the preferred settings. Three possible issues were considered. The first possible issue is an unintended voting, which appears as a preferred setting with song progress between 0 and 5 seconds. The second possible issue relates to an early voting during the intro of the song, whereas the instruction indicated that the preferred setting should be adjusted after the intro with vocals and instruments present. All early voting data were discarded before further data analysis. The third possible issue is missing or incomplete data, i.e. absent data in the returned take-home device after the music experiment. In total 3.7% of the data were discarded or missing after the first data check.

Three measurements with preferred attenuation parameter were registered for every song. A Friedman test with factor 'Run' was performed on the data (N = 12) to check the effect of time on the preferred attenuation setting. No significant effect was found ($\chi^2(2) = .41$, $p = .82$). Therefore, the median of the three measurements was taken for further analysis. However, when looking into individual results for subject S7, a median preferred attenuation of 13 dB was observed in the first run, whereas the median preferred attenuation for the second and the third run were only 5.5 dB and 8.5 dB respectively. For subject S9 the opposite trend was observed with in the first run a median preferred attenuation of 5.5 dB, and in the second and third run a median of 12.5 dB and 11 dB respectively.

A Friedman test with factor 'Genre' was performed on the median of the three measurements (N = 12) to check the effect of Genre on the preferred attenuation setting. A significant effect of Genre was found ($\chi^2(5) = 29.07$, $p < .001$). The post-hoc Wilcoxon Signed-ranks test (with Bonferroni correction) was performed to look into this significant effect. The songs with genre Evergreens had significantly lower preferred attenuation settings compared to Disco ($Z = -3.37$, $p = .015$), compared to Latin ($Z = -3.12$, $p = .03$), compared to Rock ($Z= -4.28$, $p < .001$), and compared to songs with Dutch lyrics ($Z = -4.41$, $p < .001$). The preferred attenuation for Evergreens was also lower compared to Pop, but this difference was not significant ($Z = -1.85$, $p = .96$).

The effect of familiarity of the songs on the preferred setting was investigated with a Mann-Whitney test. The test subjects had to indicate for each song, whether it was known or not. The group with known songs listed 36.5% of the items, the group with unknown songs 63.5%. No significant difference was found for the preferred attenuation setting between the two

groups (U = 26750, p = .30). However, for subject S8, a lower median preferred setting (11 dB) was found for familiar songs compared to unfamiliar songs (15 dB).

Table 5-3: Overview of subjects with SRT score for speech-in-noise (in dB), pitch discrimination (in semitones), preferred attenuation (in dB), median range of 3 repetitions, familiarity with the songs (in %), singing activity before CI, singing activity with CI, and Music Enjoyment (0-10)

Subject	SRT (dB)	Pitch (semitones)	Pref. Att. (dB)	Median Range (dB)	Familiarity (%)	Singing	Singing with CI	Music Enjoyment
S1	1	1	11	14	45	/	/	0
S2	8	3	4	8	31	X	X	5
S3	-4	1	7	11	45	/	/	9
S4	-2	2	11	10	48	/	/	7
S5	2	1	6	13	45	X	X	6
S6	4	3	8	14	64	/	/	10
S7	0	3	9	8	5	X	X	10
S8	0	2	13	10	33	/	/	0
S9	-1	3	10	7	38	X	X	0
S10	-2	2	16	7	41	X	/	9
S11	7	1	11	11	12	/	/	8
S12	9	2	10	10	31	X	X	9

The mean preferred attenuation parameter setting over all songs for all individual subjects is shown in Figure 5-2. Differences in mean preferred attenuation parameter settings were observed across subjects ranging from 4 dB (for subject S2) to 16 dB (for subject S10). An attenuation of 0 dB corresponds with the original balance of vocals/bass/drums and the other instruments. A One-Sample Wilcoxon Signed-ranks Test with Test Value = 0 was used to check the preferred attenuation setting against the original setting. This test was significant ($p < .001$) and thus revealed a preferred

attenuation setting significantly different from zero over all subjects. No correlation was found between the mean preferred setting and the SRT for speech-in-noise (Pearson's r(12) = -.35, p = .26, Spearman's ρ(12) = -.32, p .31), nor with the pitch detection performance (Pearson's r(12) = -.17, p = .59, Spearman's ρ(12) = -.21, p = .51), nor with the CI experience (Pearson's r(12) = -.29, p = .36, Spearman's ρ(12) = -.34, p = .29), nor with any item from the music listening habits questionnaire. An overview of the SRT scores for speech-in-noise, pitch discrimination scores and preferred attenuation settings is listed in Table 5-3.

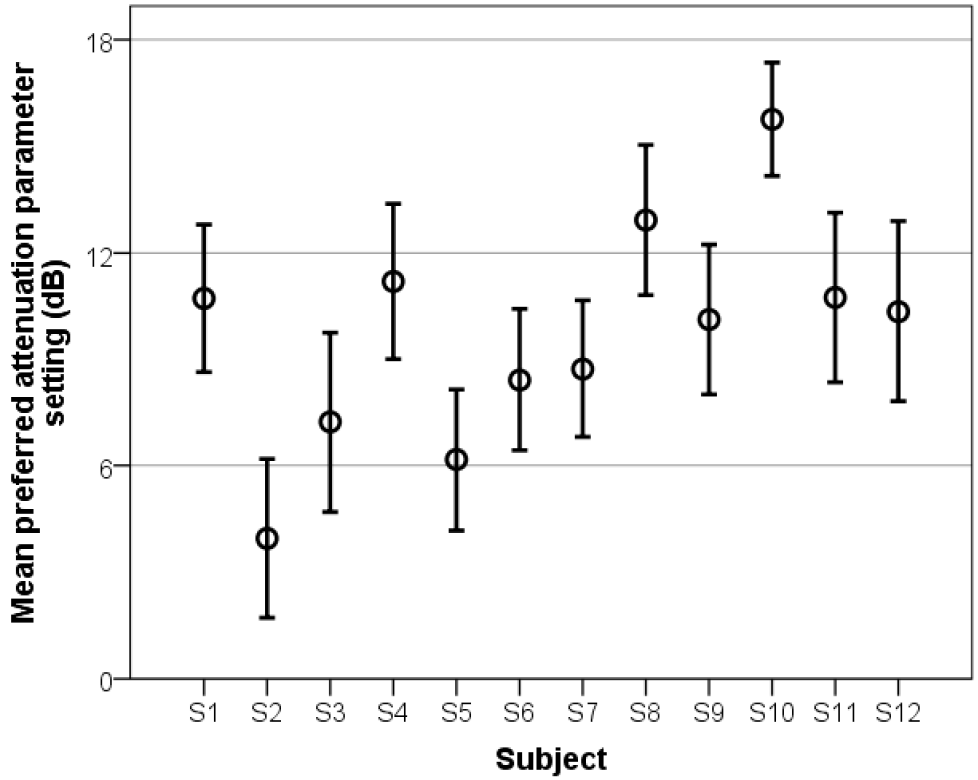


Figure 5-2: Mean preferred attenuation parameter setting (dB) over all songs for all subjects in the take-home test. Positive values represent an attenuation of 'other instruments', negative values (not shown) represent an amplification of 'other instruments', and the value zero represents the original balance. Error bars indicate 95% confidence interval.

The range of the three preferred settings collected for one song, which is defined as the subtraction of the maximum preferred attenuation and the

minimum preferred attenuation, can take a value from 0 dB to 30 dB (= 24 dB – (-6 dB)) and is a measure for the strength of the preference for the determined attenuation setting for that song. A small range indicates a strong preference for the determined attenuation setting and/or large audible effect from the music pre-processing, whereas a large range can be considered as a weak preference for the determined attenuation setting or small audible effect of the music pre-processing. Remember that the beginning value of the wheel was randomized for each voting of the song. The median range over all songs and all subjects was 11 dB. A significant effect of Genre was found on the computed range of the three preferred settings collected for every song (Friedman with factor 'Genre', $\chi^2(5) = 13.12$, $p = .022$). Post-hoc analysis with Wilcoxon Signed-ranks Test (with Bonferroni correction) revealed a significantly lower range for Pop compared to Disco ($Z = -3.14$, $p = .03$), and a significantly lower range for Rock compared to Disco ($Z = -3.54$, $p < .001$), or phrased in a different way the range of the three preferred settings collected for Disco songs was significantly higher compared to the range in Pop and Rock songs.

In the questionnaire, subjects were asked for their singing activities before and after implantation (see Table 5-3). The six subjects who reported singing activities before implantation had on average a significantly smaller range over the three preferred attenuation parameter settings (Mann-Whitney Test: $U = 4.5$, $p = .026$). The five subjects who indicated singing activities after implantation (see Table 5-3) were significantly more experienced CI users in terms of years of CI usage (Mann-Whitney Test: $U = 2$, $p = .01$) and showed a preference for lower attenuation parameter settings, although not significant (Mann-Whitney Test: $U = 6$, $p = .073$).

In the open question asking to share suggestions with other users for listening to music with a CI, some of the responses were as follows: “Ask for a dedicated music map”, “Don’t give up”, “Have patience and keep practicing”, “Start with songs that you know”, “Practice makes perfect”, “If you want to enjoy music, you need a lot of practice, it’s not about one day, one week or one month, it took me years”.

5.5 Discussion

The main objective of this study was to evaluate the improvement in music appraisal in CI subjects with a music pre-processing scheme for different genres of music and in a comfortable listening environment. The relation between the preferred attenuation parameter settings and the speech and pitch performance was investigated as well as the subject’s music listening habits. The music pre-processing scheme was described in Buyens et al. (2015a) and showed encouraging results in lab experiments with pop/rock song excerpts. Consequently, a take-home test was set up for the assessment of the music pre-processing scheme with different genres of music. After validating the registered input data, the consistency of the data over the three repetitions was checked. No effect of time was found, and thus the median of the three repetitions was used for further analysis. The median was preferred over the mean in order to exclude the influence from possible outliers.

During the feedback discussion when returning the take-home device, subject S7 indicated himself the difference in voting between the first run and the second/third run. For the first run, he focused on understanding the lyrics and consequently adjusted the attenuation parameter to emphasize

the vocals. For the second and the third run, however, he focused more on the song as a whole, and adjusted the attenuation parameter to a lower level than in the first run. In subject S9, a trend in the opposite direction was observed, showing a lower preferred setting for the first run and a higher preferred setting for the second/third run.

A significant effect of Genre was found for the preferred setting. The preferred setting was the lowest for the Evergreens. A possible explanation for this may be that these songs were less complex than the songs from the other genres, and consequently required less attenuation of the instruments for CI users. Gfeller et al. (2003) showed a negative correlation between complexity and appraisal for CI users and also Buyens et al. (2015a) found a preference for lower attenuation of instruments for songs with lower complexity compared to songs with higher complexity. Another possible explanation may be related to the different audio mixing habits in older recordings (such as Evergreens). Older recordings typically have a larger dynamic range where vocals stand out of the background music, whereas in contemporary music the vocals are more engraved in the music and the audio mix is heavily compressed (Vickers, 2010). Therefore, the vocals enhancement of the music pre-processing may be less needed in Evergreens compared to the other genres. On the contrary, during the feedback session subjects reported a strong advantage of emphasizing the vocals especially in the songs with Dutch lyrics. With the adjustment of the attenuation parameter, the subjects (all Dutch-speaking) were able to understand the lyrics better, and some of them indicated that this was less required for songs with English or Spanish lyrics, because they could not understand these lyrics anyway.

All subjects preferred a setting that was different from the original balance between vocals/bass/drums and the other instruments, but individual differences were observed across subjects. In Buyens et al. (2014), a significant negative correlation was found between the pairwise preference for an audio mix with attenuated instruments and CI experience, but this was not found in the current study. To understand the individual differences, the relationship with speech and pitch detection performance was investigated, but no correlation was found between the preferred settings and the speech perception performance or pitch detection abilities of the subject. The task in the experiment was to adjust the attenuation parameter in order to make the music most enjoyable, which seems to be unrelated with the perceptual abilities. Drennan et al. (2015) also showed only weak relationships between perceptual accuracy and music enjoyment, suggesting that perception and appraisal are relatively independent for CI users. Similar conclusions were found by Gfeller et al. (2008), Lassaletta et al. (2008), and Wright and Uchanski (2012).

The range over the three preferred settings collected for a song indicates the strength of the preference for the determined setting or the audibility of the music pre-processing effect. This was investigated for the different genres. The median range over all songs including all genres was 11 dB. For the Disco songs significantly weaker preference for the determined settings (i.e. higher range) was found compared to Pop and Rock, which showed a stronger preference for the determined setting. No significant difference in range was found for Evergreens, Latin and songs with Dutch lyrics. The explanation for this may be in the difference in instruments used in the different genres, with on the one hand Pop and Rock music with mostly vocals, guitar, piano, bass guitar and drums and on the other hand Disco music with additional hand clapping, extensive percussion, singers in harmony and prominent

brass sections, which may make the effect of the music pre-processing less noticeable. Subjects that indicated singing activities before implantation had on average smaller range, which indicates a stronger preference for the determined preferred settings or better ability to hear out the effect of the music pre-processing.

Informal feedback after the experiment was mostly positive, confirming the appreciation for clear vocals and reduction of the disturbing 'chaos' in the background. A small number of subjects reported a certain difficulty in finding the preferred setting, especially for unfamiliar songs or genres, but most of the subjects liked the experiment and enjoyed playing around with the balance between vocals/bass/drums and the other instruments. They reported that for most songs a difference between good and bad was noticeable by adjusting the attenuation parameter.

5.6 Conclusion

A music pre-processing scheme implemented on a take-home device was assessed with twelve post-lingually deafened CI subjects in a take-home test. The preferred setting for the adjustable attenuation parameter which balances vocals/bass/drums against other instruments was investigated for different genres of music. All subjects preferred an attenuation parameter setting to construct a balance with attenuated instruments significantly different from the original mix. Individual differences across subjects were observed. This could not be explained by perceptual abilities as speech perception performance or pitch detection. An effect of Genre was found, showing lower preferred attenuation settings for Evergreens compared to the other genres. Since the complexity of music was reduced with the music

pre-processing scheme, it may provide a good tool for music training or rehabilitation programs.

Chapter 6: Conclusions and future work

The main goal of this doctoral research project was to investigate and to improve the poor music enjoyment in CI users. An initial experiment was carried out to gain more knowledge about music perception with CI users, in particular the preference for different instruments in polyphonic or complex music. Based on this knowledge, a music pre-processing scheme was developed which is capable of modifying the relative instrument level settings in complex music on mono and stereo recordings. Subsequently, a take-home experiment was set up to evaluate the music pre-processing scheme with CI users.

The conclusions of this thesis are summarized in section 6.1, 6.2 and 6.3. Section 6.4 provides suggestions for future research.

6.1 Music mixing preference experiment

Although CI users reach good speech understanding in quiet surroundings, music perception and appraisal generally remain poor. Gfeller et al. (2000) described top factors to enhance music enjoyment for CI users including simple musical structures, clear rhythm/beat and lyrics that are easy to follow. Furthermore, Gfeller et al. (2003) found a negative correlation between complexity and appraisal for CI users, indicating a preference for less complex music. Also Looi et al. (2007) found a preference for music with a single instrument over music with multiple instruments. The preference for a reduction of structural complexity in music by attenuating harmonic instruments was investigated with multi-track recordings, including separate tracks for vocals, piano, guitar, bass and drums. After an initial experiment in which CI users were asked to make their preferred audio mix with a mixing console, a pairwise comparison was performed with predefined audio mixes with NH subjects and CI subjects. The results showed that NH subjects significantly preferred the “original” audio mix with an equal balance between the vocal melody and the different instruments. However, for the pop songs provided in the experiment, CI subjects significantly preferred an audio mix with clear vocals and attenuated instruments. In an additional experiment including audio mixes with different bass/drum level, a preference for the audio mix with preserved bass/drums on top of the clear vocals was found. Individual differences were observed across subjects. Nevertheless, the results indicated the potential improvement of music appraisal by modifying the relative instrument levels in complex music, especially by emphasizing vocals/bass/drums.

6.2 Music pre-processing scheme

The music mixing preference experiment showed a potential improvement in music appraisal for CI users by modifying the relative instrument levels in an audio mix. However, the experiment was performed with multi-track recordings, which are not widely available. A music pre-processing scheme was developed which is capable of modifying these relative instrument level settings on mono and stereo music recordings. It is able to emphasize vocals, drums and bass based on the representation of the harmonic and the percussive components in the input spectrogram, combined with the spatial allocation of instruments in typical stereo recordings. Vocals, bass and (part of the) drums are mostly mixed in the centre of the stereo image, whereas the other instruments are typically mixed to left or right as they would appear on stage. Objective evaluation of the pre-processing scheme revealed the capability of the scheme to modify the relative instrument levels in mono and stereo recordings which appeared beneficial in the music mixing experiment, i.e. emphasis on vocals/bass/drums. Subjective evaluation in the lab was performed first with NH subjects listening to CI-simulated music and later with post-lingually deafened CI subjects (N = 7) using pop/rock music excerpts with different level of complexity. Preferred settings for the balance between vocals/bass/drums and the other instruments were determined. On average, the preferred attenuation for the other instruments was significantly higher for the group of songs with high complexity versus the group of songs with low complexity. The reduction of the structural complexity of songs by attenuating the harmonic instruments may also lower the subjective complexity as perceived by CI users, which was found to be negatively correlated with music appraisal by Gfeller et al. (2003). Even though individual differences were observed, the subjective evaluation

demonstrated the potential to increase music appraisal for CI users with the music pre-processing scheme.

6.3 Take-home evaluation

The lab experiments with the music pre-processing scheme showed encouraging results with pop/rock song excerpts. Consequently, a take-home evaluation was performed in a comfortable listening environment with different genres of music (Disco, Evergreens, Latin, Pop, Rock and songs with Dutch lyrics). A take-home device (iPhone) was provided to the subject including a custom-made application with the music pre-processing scheme and seven playlists with six songs each. Test subjects were asked to make their preferred balance between vocals/bass/drums and the other instruments for all songs and to repeat this three times. The effect of speech-in-noise performance, pitch detection abilities and music listening habits on the results was investigated. Twelve post-lingually deafened CI subjects participated in the study. All subjects preferred an audio mix significantly different from the original. Individual differences were observed across subjects, which could not be explained by the perceptual abilities (speech-in-noise / pitch) of the subject. Also by Gfeller et al. (2008) and Drennan et al. (2015) only weak or no correlation was found between perceptual accuracy and music appraisal for CI users, which indicates that perception and appraisal are relatively independent. The preferred setting for Evergreens was significantly lower compared to the other genres of music. Since the music pre-processing scheme allows individual settings for music according to personal preferences, the custom-made application may provide a good tool for music listening, training or rehabilitation programs for CI users. The deliverable from Chapter 5 – a smartphone app that enables

CI users to adjust music according to their individual preferences, can become part of a commercially available Recipient App, running on a smartphone and connected with the sound processor. This can serve both recently implanted CI users as more experienced listeners. For recently implanted CI users, the step towards music may still be too big at start, and a tool that 'simplifies' music can be useful. The more experienced listeners may enjoy the music app, since personal modifications to music are possible that outperform simple equalizer adjustments. In clinics, the smartphone application can be used as training or rehabilitation tool for music.

6.4 Future work

The evaluation of the music pre-processing scheme in the take-home experiment showed encouraging results. A number of additional research questions for future research will be highlighted in this section, which can be addressed in a follow-up project. Individual differences were observed across subjects, meaning there is not one setting that suits all subjects. Next to these between-subject differences, there was also variability in the preference settings within-subjects. Effects of training, genre and familiarity were studied but they did not explain the variability entirely. Therefore, further investigation of the music and its features may be interesting to explain the results. Furthermore, the signal processing scheme was evaluated as a pre-processing stage, which means that the possible interaction between the music pre-processing scheme and the parameter settings in the CI signal path should be considered as well. Finally, the clinical parameters that can be adjusted to program the sound processor are mostly left at default values for music. The investigation of these clinically available

parameters and their impact on music appraisal could be an interesting topic for future research.

Three topics for future research are discussed in this section. In the first topic, specific music features and their importance for music appraisal in CI users are studied, resulting in an objective measure for music appraisal. In the second and the third topic this knowledge is exploited to develop an advanced version of the music pre-processing scheme and to investigate the fitting of a sound processor especially for music.

6.4.1 Music feature preference

To enhance the music appraisal in CI users, the features in music accountable for music appraisal can be further investigated as a first topic for future research. This can be achieved with a music appraisal experiment, in which CI subjects are rating music appraisal for a large number of music excerpts on a scale from 0 to 100 as by Gfeller et al. (2003). These excerpts should cover a wide range of music genres, complexity, instruments and vocals. On the other hand, different music features are extracted, which fully characterize the music excerpts and which may influence music appraisal in CI users. These features can include statistical measures describing the sound as well as features from the music information retrieval field of research, such as key, chords, harmonies, melody, dominant pitch, beats per minute or rhythm in the piece. By combining the rating scores and the inherent music features of the music excerpts, an objective measure or formula for music appraisal can be composed, which predicts the music appraisal for new music and which gives more information about the importance (or disturbance) of specific features for music appraisal in CI users. The music appraisal formula may take a weighted sum form of the

selected music features, in which the weights are estimated with machine learning techniques. In order to increase the music appraisal, the important features, which show up in the formula with high positive weights, should be enhanced and the disturbing features, which show up in the formula with high negative weights, should be reduced or removed.

6.4.2 Advanced music pre-processing

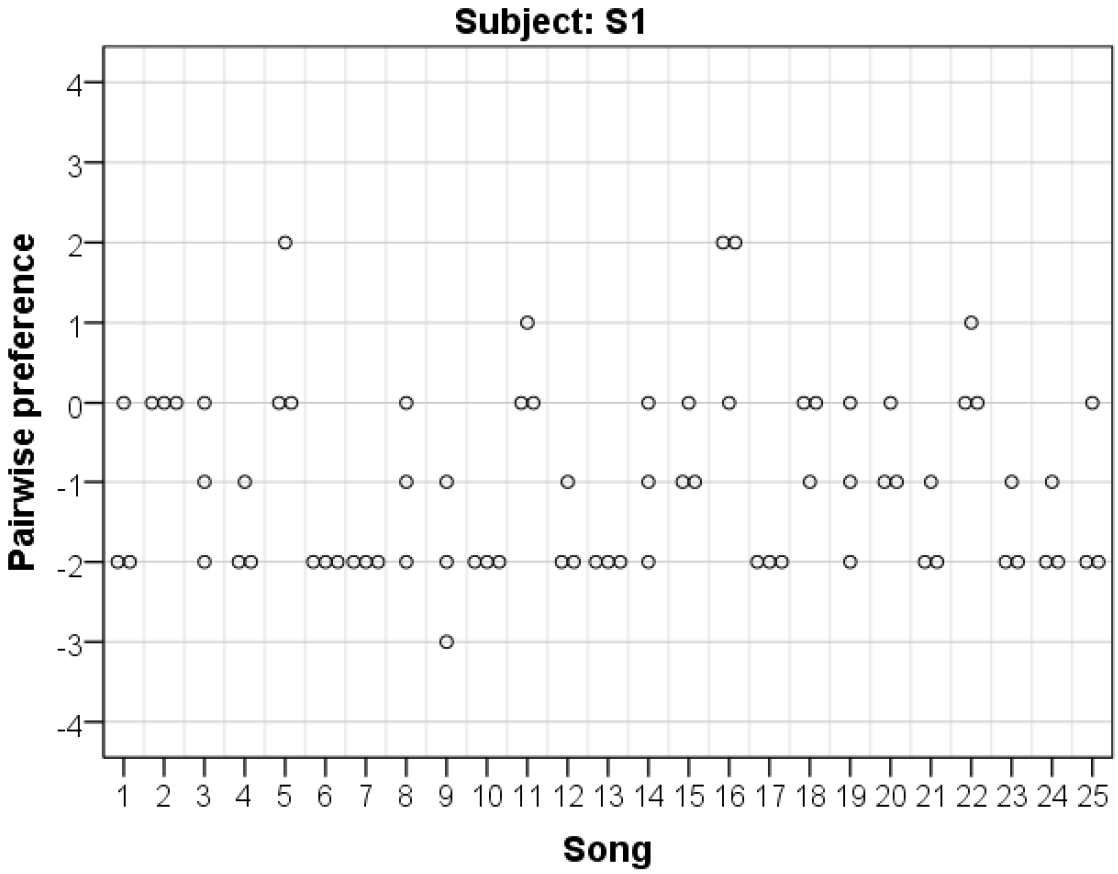
The music pre-processing scheme studied in this PhD thesis relied on sound separation techniques that are able to separate vocals/drums/bass from the other instruments, in order to enhance vocals/drums/bass in complex music. The attenuation of the other instruments in favour of vocals/drums/bass in the audio mix can be interpreted as a reduction of complexity and harmonicity, and an enhancement of rhythmicity, which are examples of music features to be included in the music appraisal formula from paragraph 6.4.1. A second topic for future research may investigate additional music features and their relation to music appraisal for CI users based on the music appraisal formula from paragraph 6.4.1. With this knowledge, the results of the clinical take-home experiment with the music pre-processing scheme and CI users from Chapter 5 can be further investigated. Moreover, the feasibility of additional music pre-processing to enhance the most important features and to reduce the most disturbing features from the music appraisal formula can be investigated. An advanced music pre-processing scheme can be developed which includes the adjustment of these new music features. The pre-processing stage is followed by the CI signal path, which is responsible for converting the audio signals into electrical stimuli. The interaction between music pre-processing and CI signal path should be investigated as well in this second topic.

6.4.3 Fitting a music map

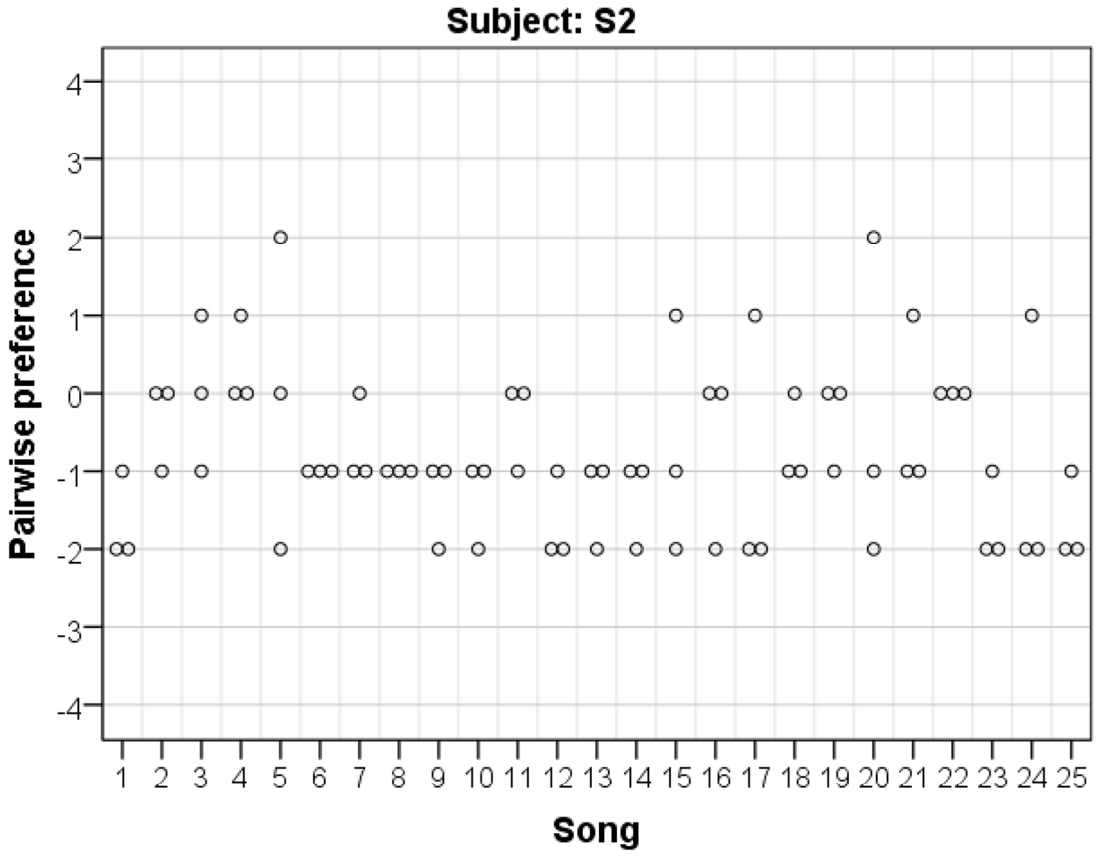
A large number of fitting parameters are available for the clinician to fine-tune the clinical map in the sound processor of the CI user. The vast majority of these parameters is set to their default value during the fitting of a CI, and is mostly changed for the improvement of speech understanding. The investigation of the clinically available fitting parameters and their relation to music appraisal may be a third topic for future research. The results from the music appraisal experiment (paragraph 6.4.1) can be reused, with additional features to describe the music excerpts related to the electrical representation of the music over the 22 electrodes. A new music appraisal formula can be composed based on these new features and the available music appraisal scores. The effect of the adjustment of clinical parameters on the electrical representation of music can be studied with this new music appraisal formula as an objective measure. The clinical parameters should be adjusted in order to increase the important electrical features and to reduce the disturbing electrical features as found in the new music appraisal formula. Recommendations for fitting a music map for increased music appraisal can then be formulated.

Appendix

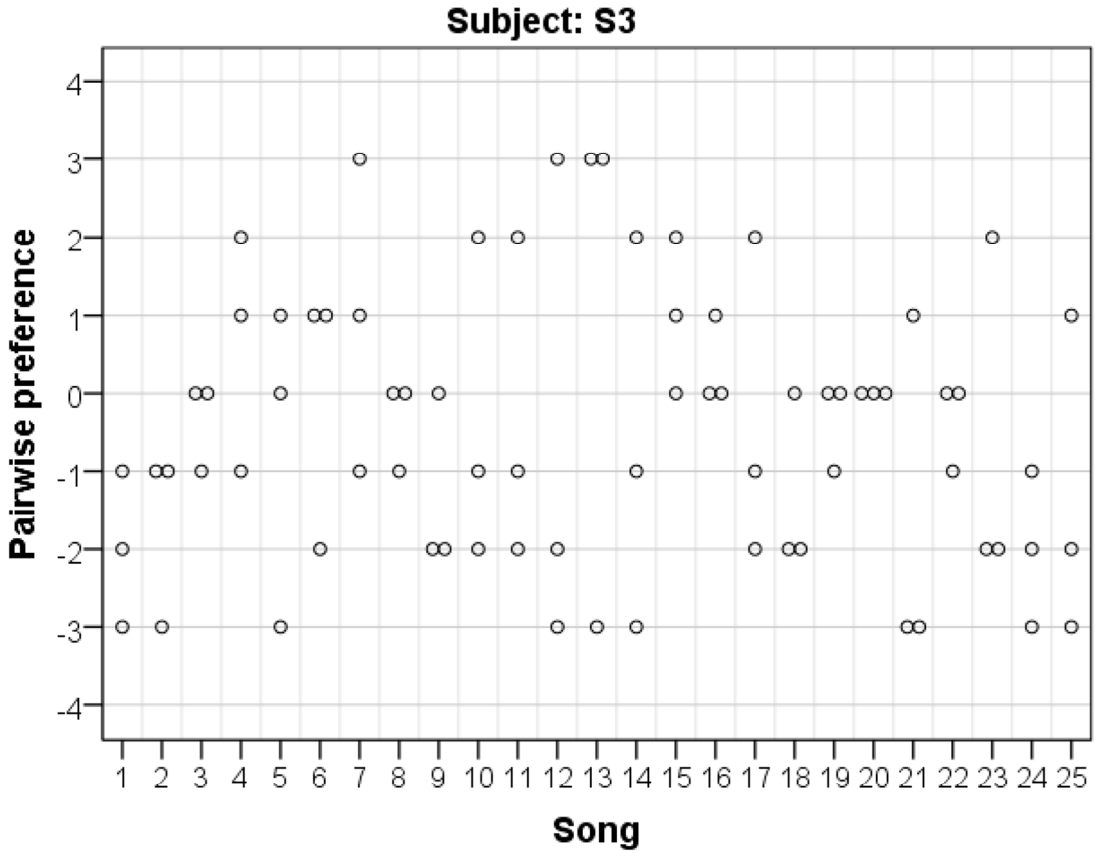
Subject S1 (Chapter 3): subjective evaluation with three rating scores per song (Imperceptible (0), Slightly better (1), Better (2), Largely better (3), Hugely better (4)). Positive scores indicate preference for unprocessed stimuli; negative scores indicate preference for processed stimuli.



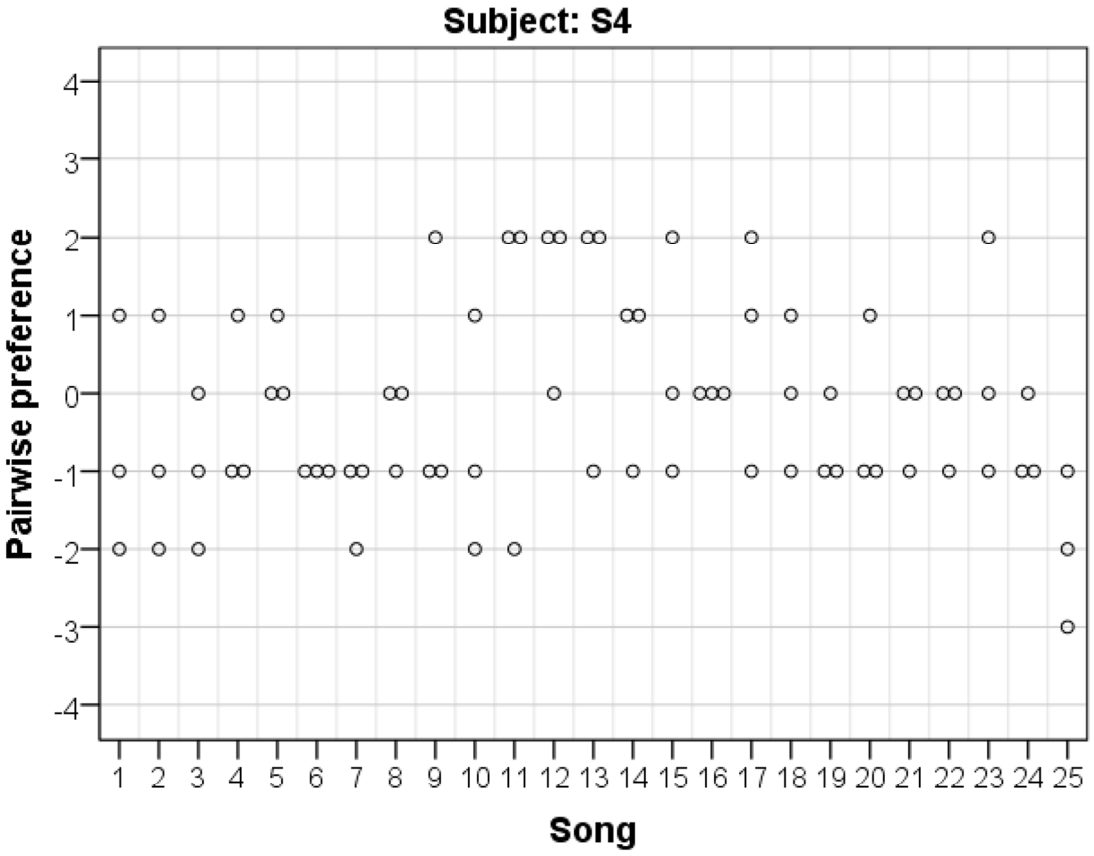
Subject S2 (Chapter 3): subjective evaluation with three rating scores per song (Imperceptible (0), Slightly better (1), Better (2), Largely better (3), Hugely better (4)). Positive scores indicate preference for unprocessed stimuli; negative scores indicate preference for processed stimuli.



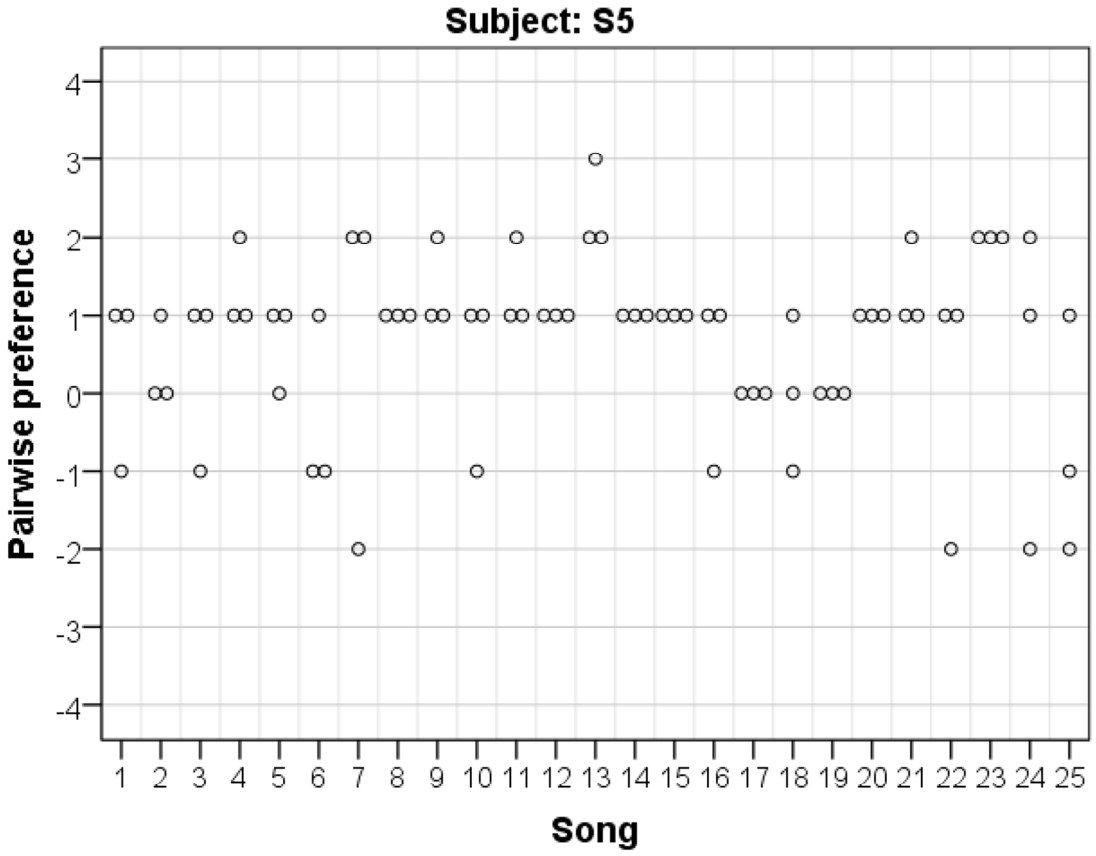
Subject S3 (Chapter 3): subjective evaluation with three rating scores per song (Imperceptible (0), Slightly better (1), Better (2), Largely better (3), Hugely better (4)). Positive scores indicate preference for unprocessed stimuli; negative scores indicate preference for processed stimuli.



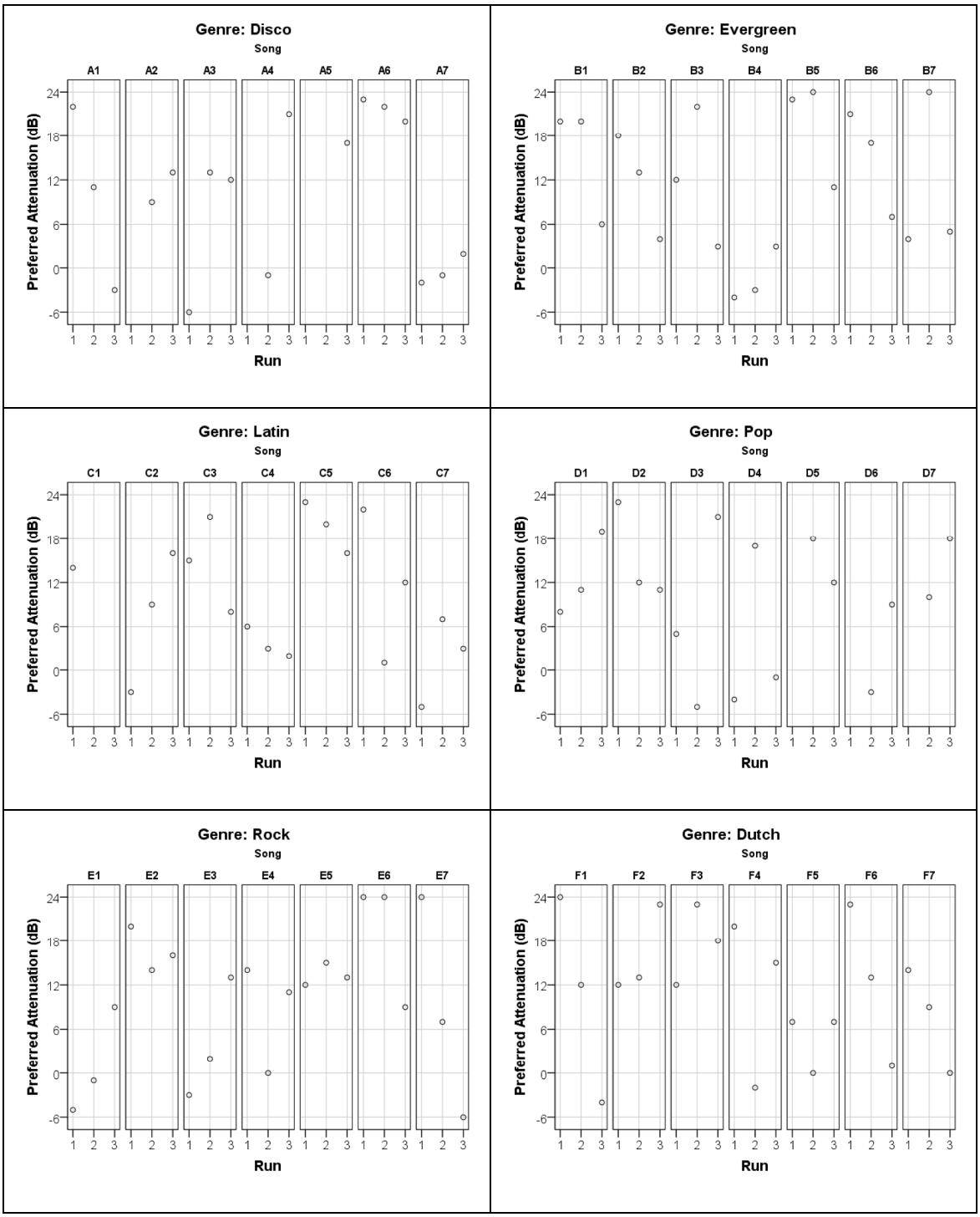
Subject S4 (Chapter 3): subjective evaluation with three rating scores per song (Imperceptible (0), Slightly better (1), Better (2), Largely better (3), Hugely better (4)). Positive scores indicate preference for unprocessed stimuli; negative scores indicate preference for processed stimuli.



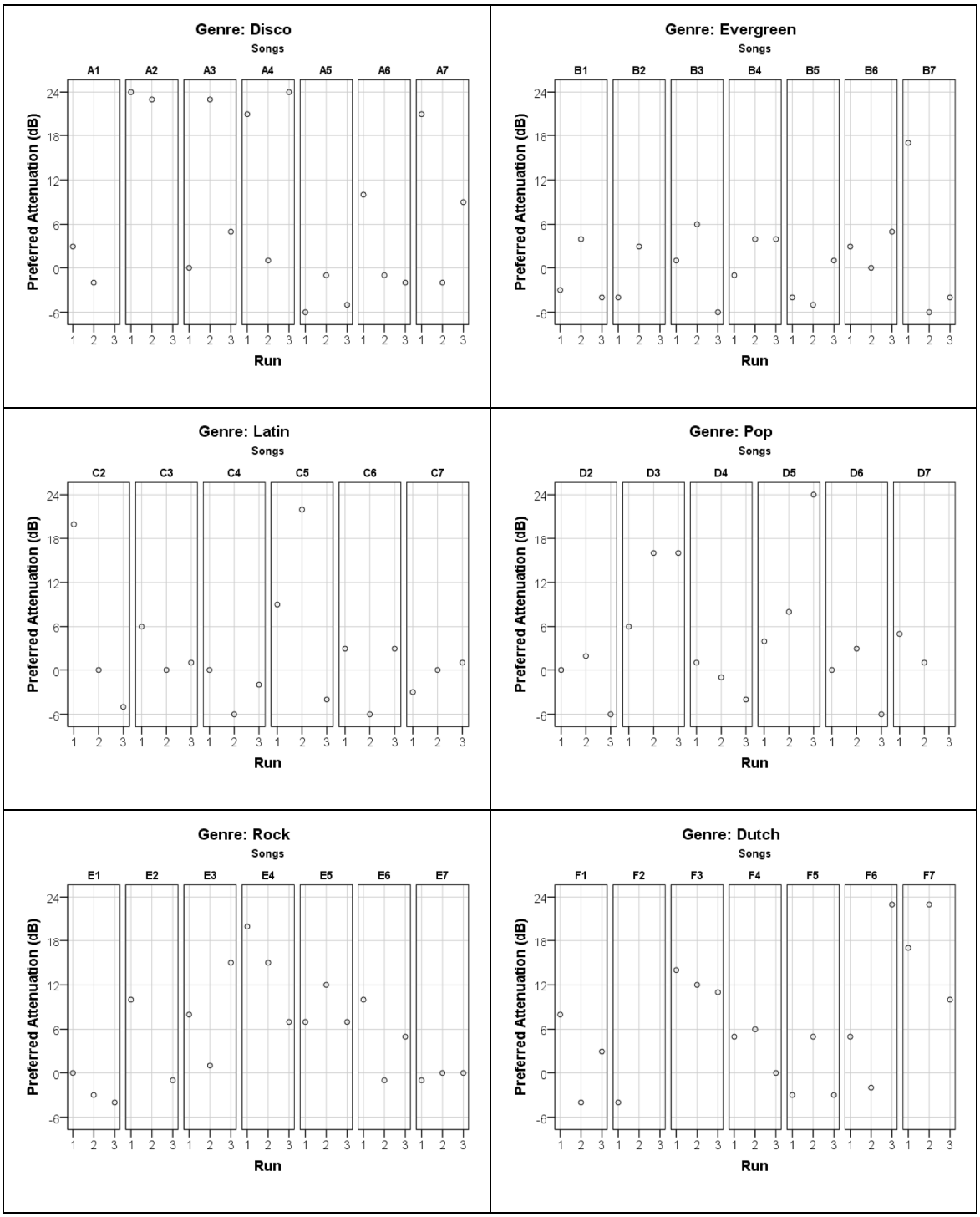
Subject S5 (Chapter 3): subjective evaluation with three rating scores per song (Imperceptible (0), Slightly better (1), Better (2), Largely better (3), Hugely better (4)). Positive scores indicate preference for unprocessed stimuli; negative scores indicate preference for processed stimuli.



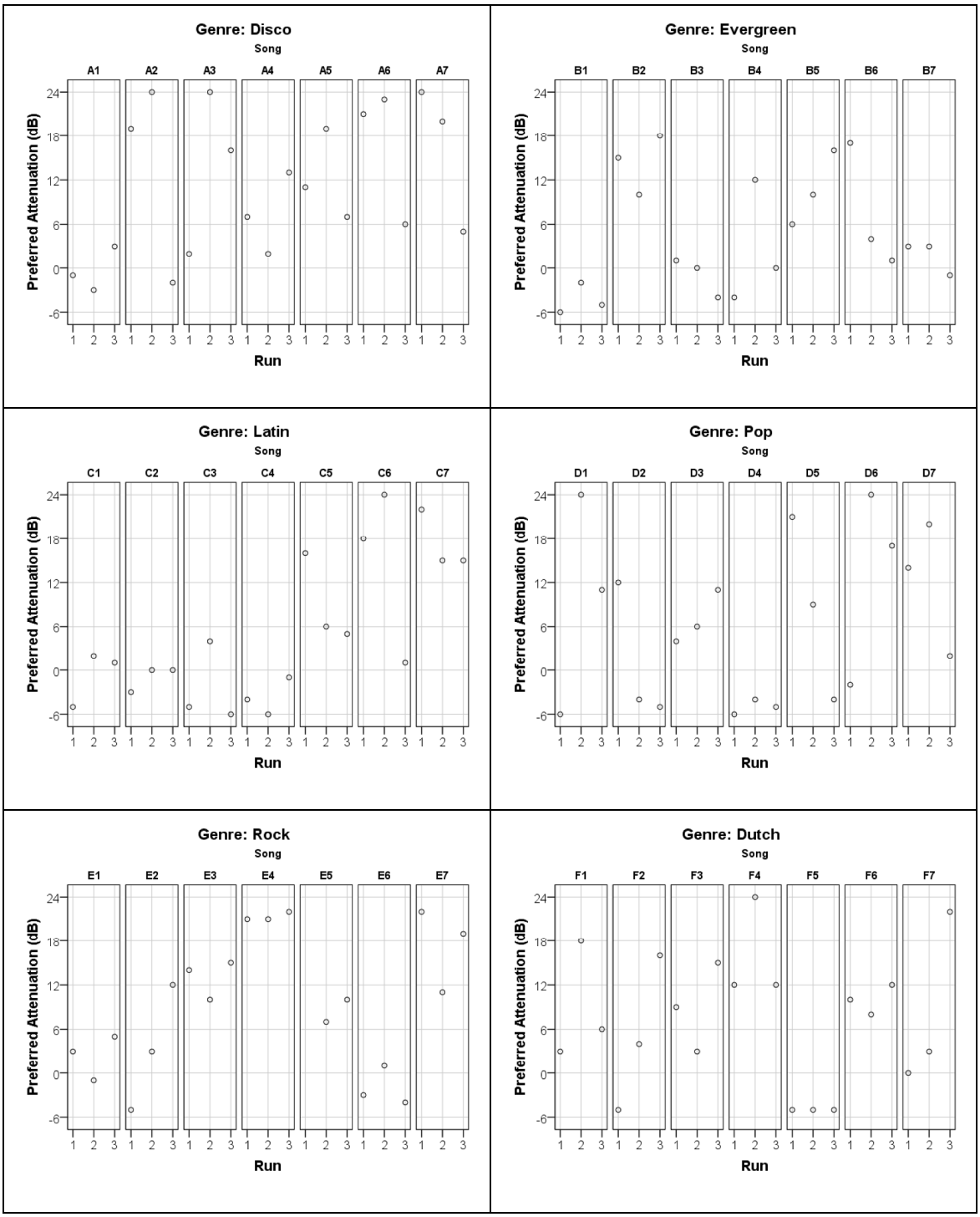
Subject S1 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



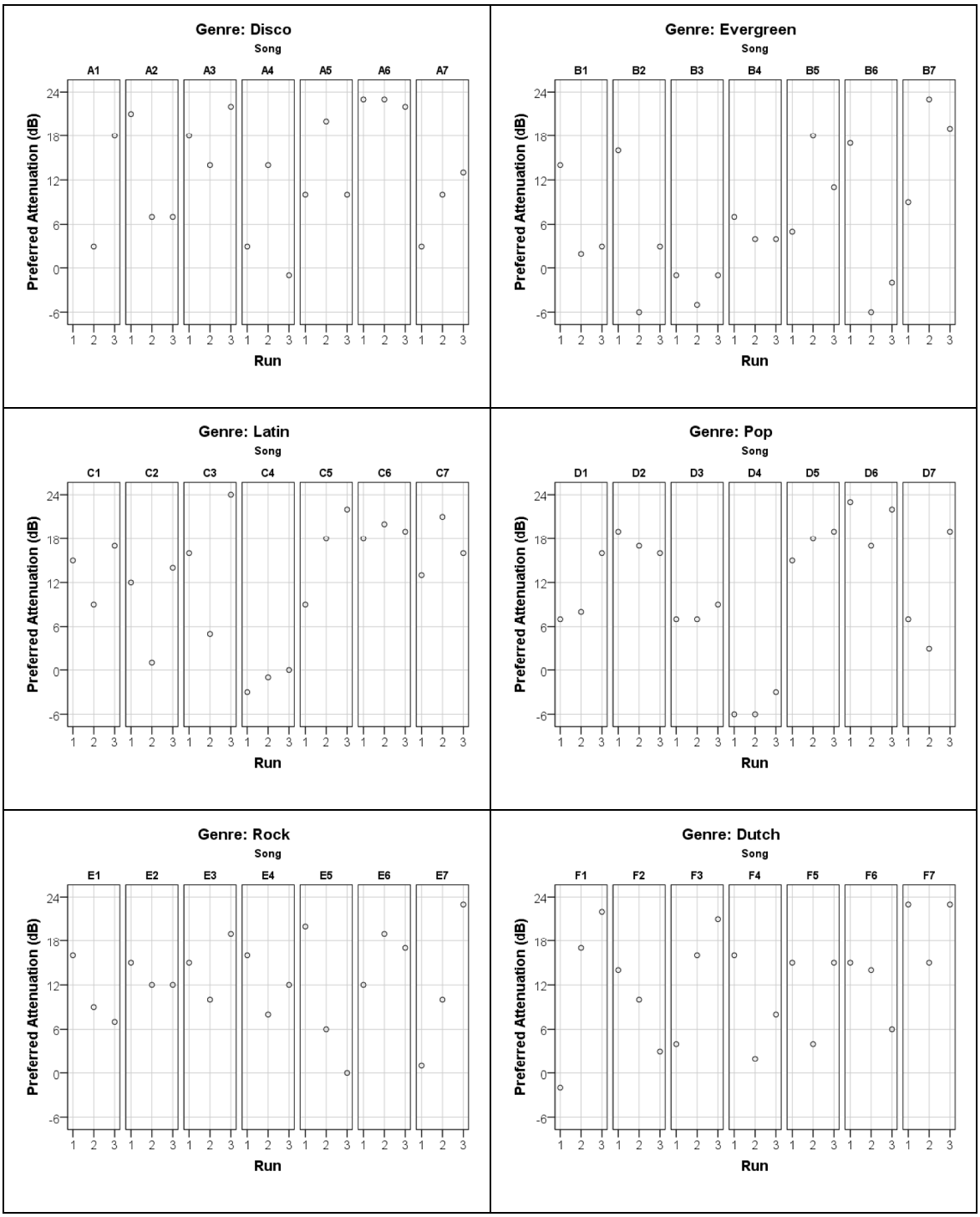
Subject S2 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



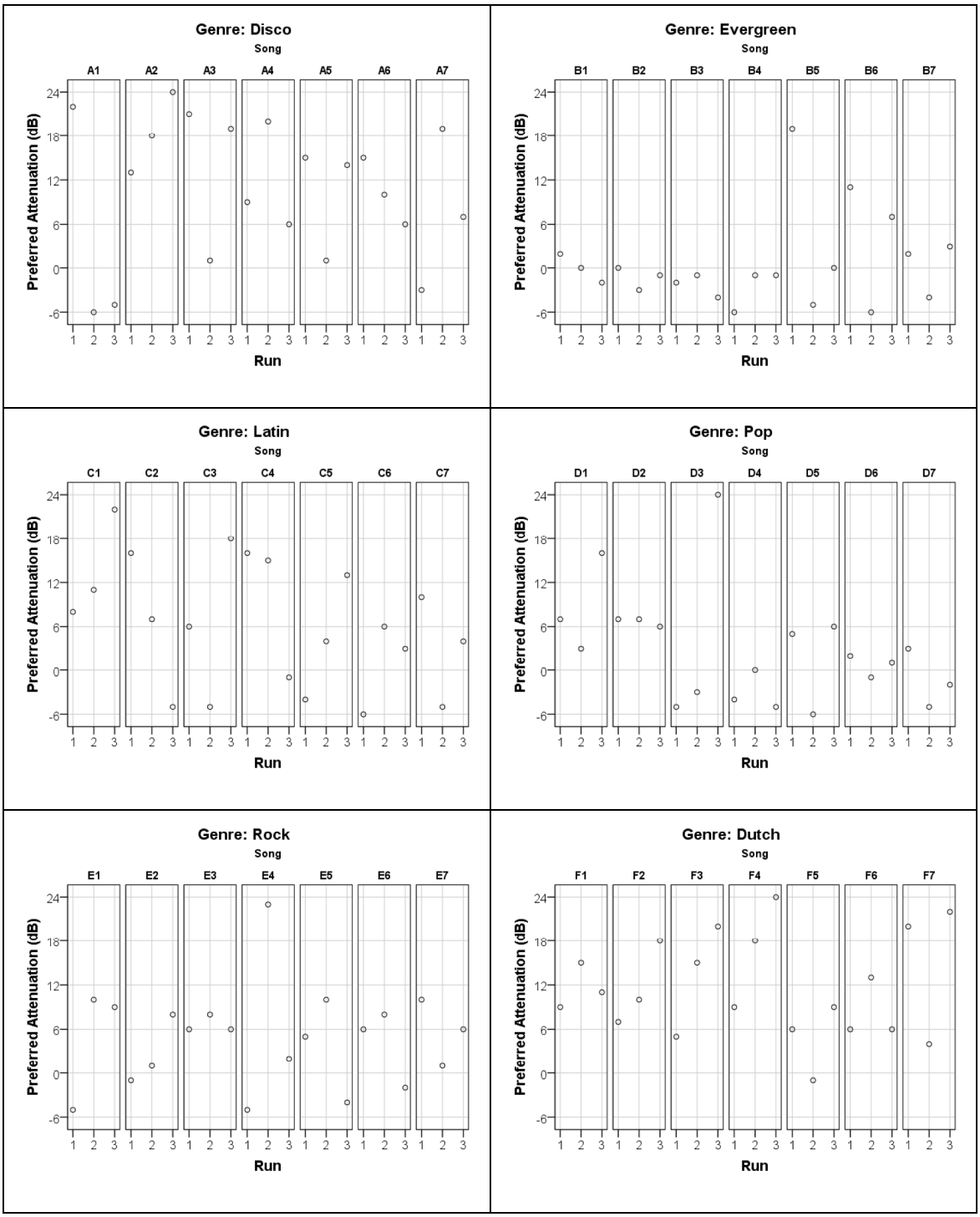
Subject S3 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



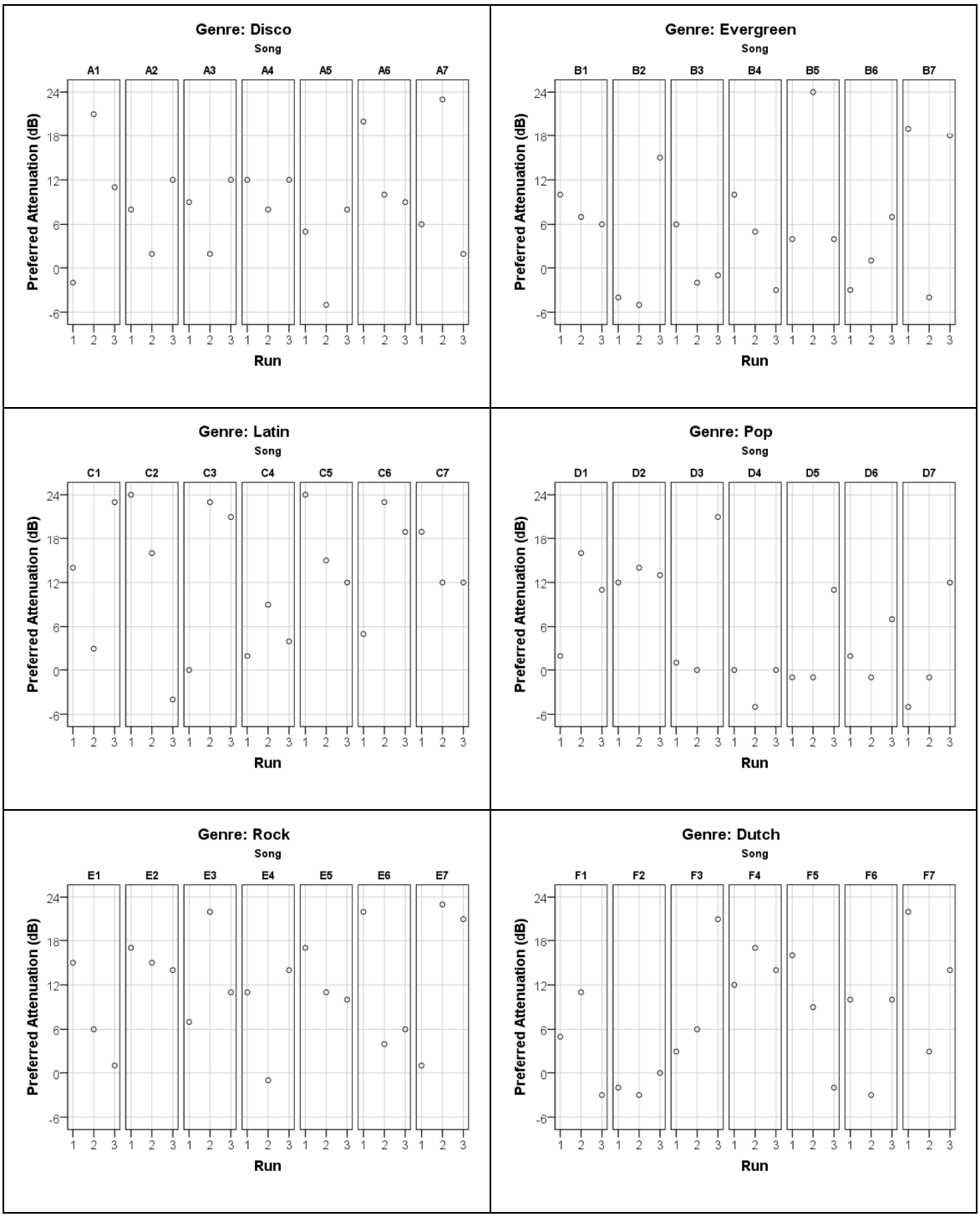
Subject S4 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



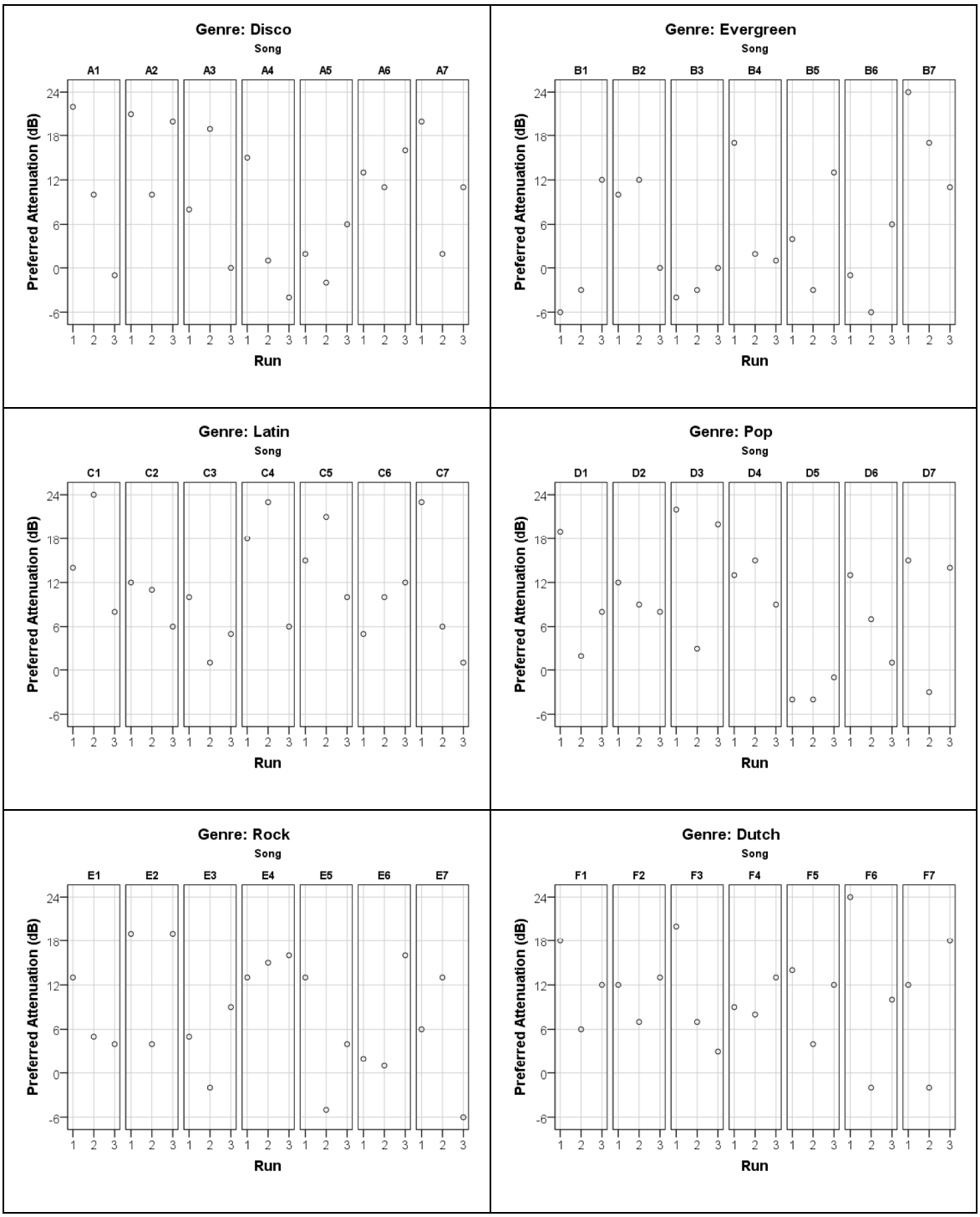
Subject S5 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



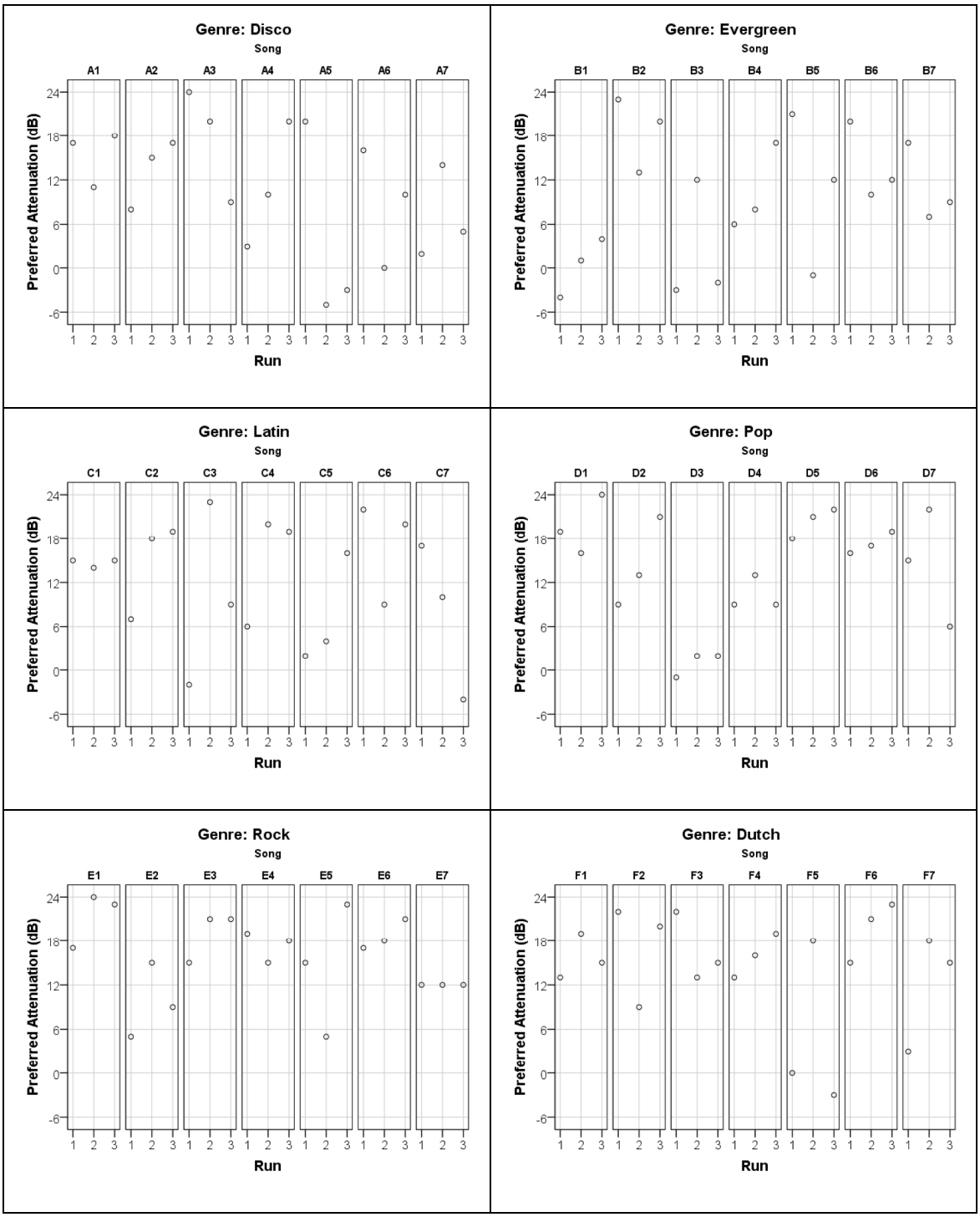
Subject S6 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



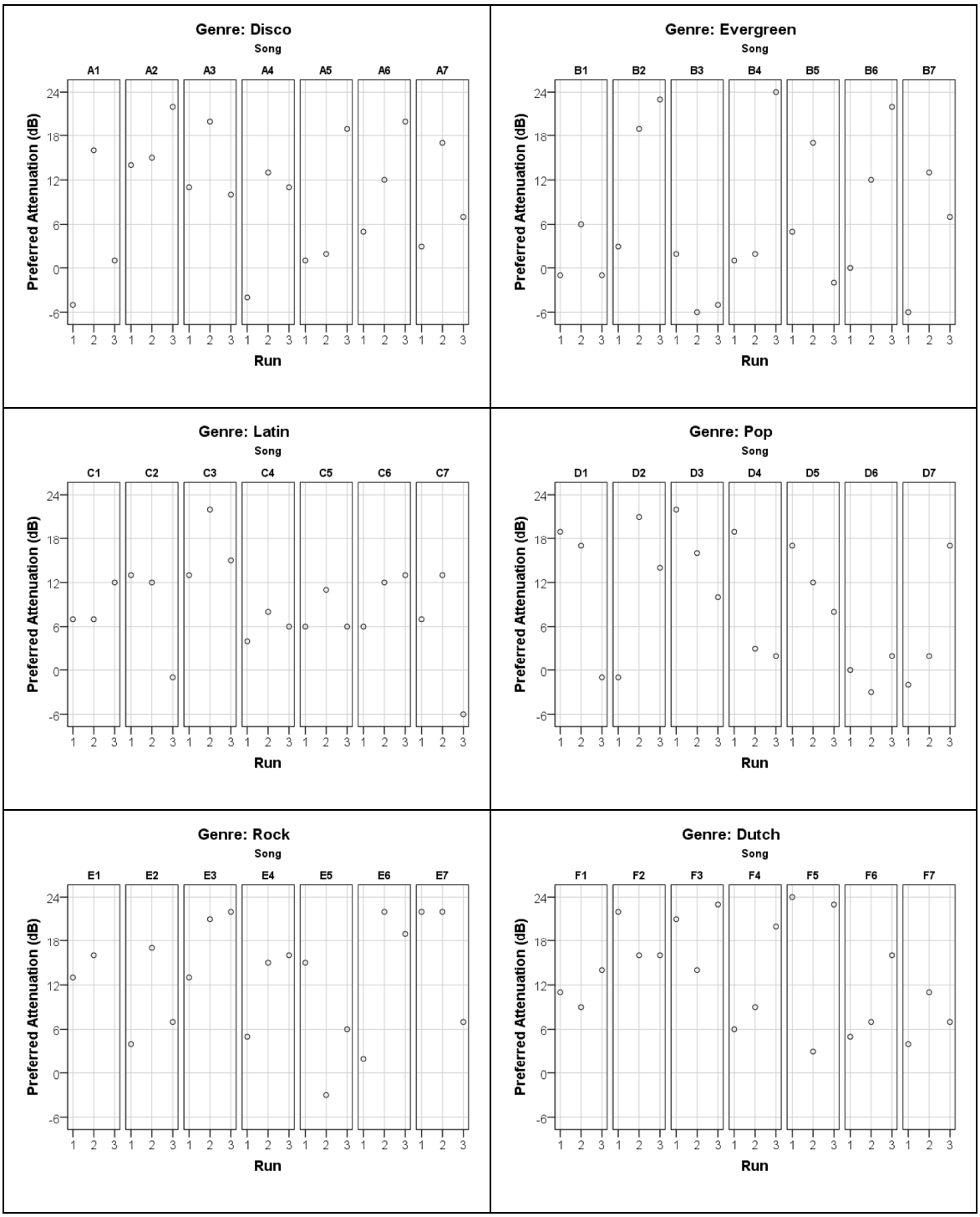
Subject S7 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



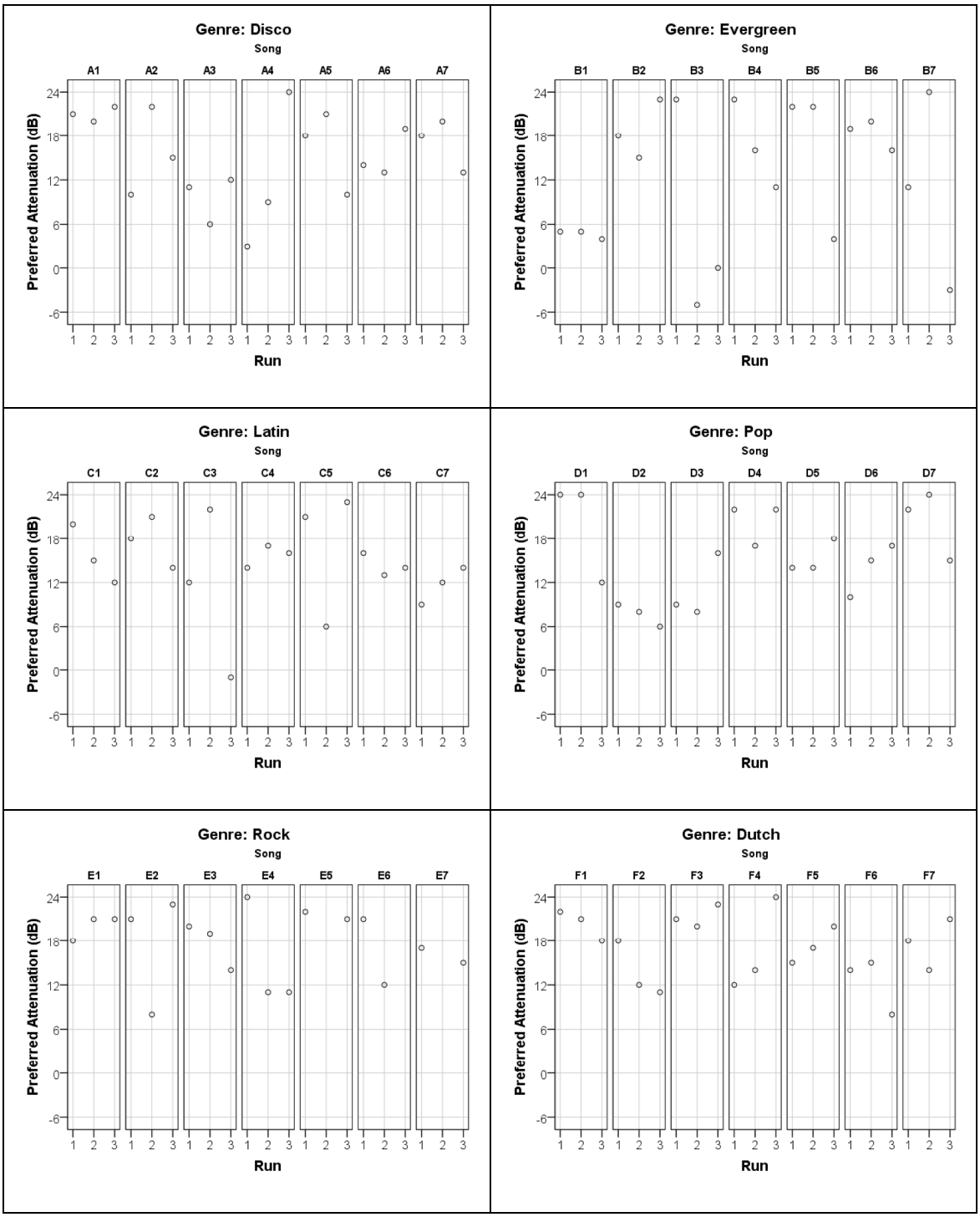
Subject S8 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



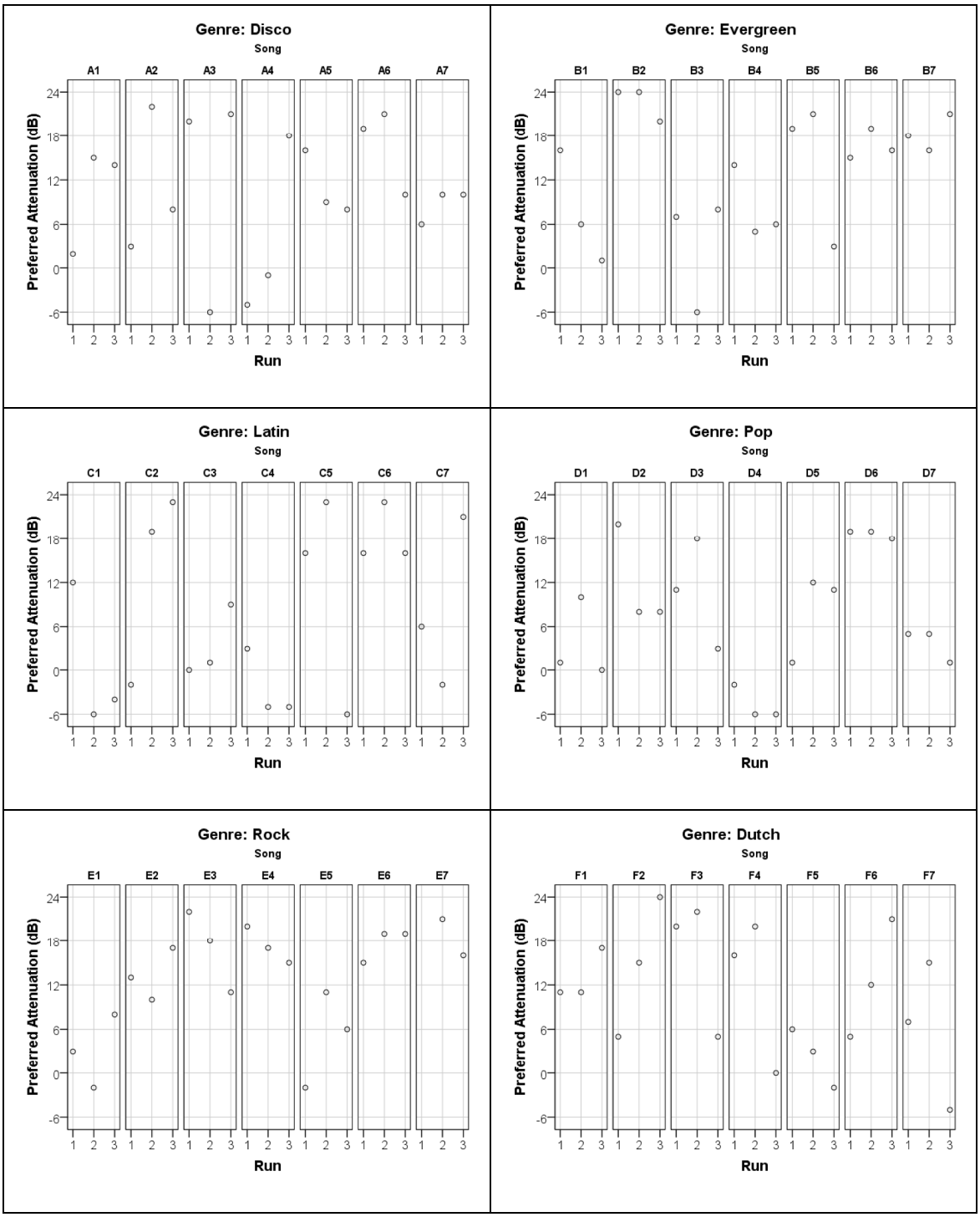
Subject S9 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



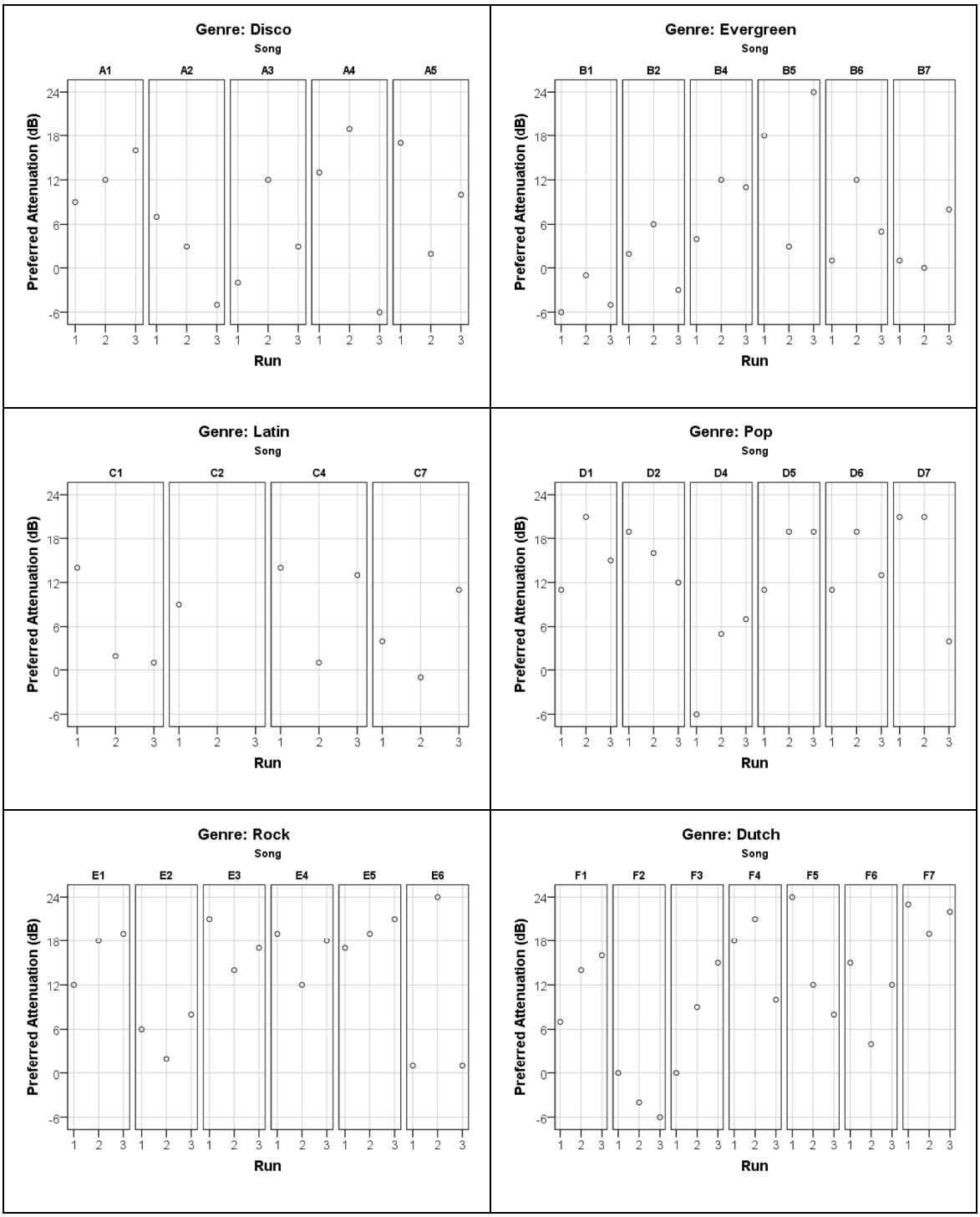
Subject S10 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



Subject S11 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



Subject S12 (Chapter 5): Take-home experiment with three preferred attenuation settings per song for different genres of music.



Bibliography

Abdallah, S.A. and Plumbley, M.D. (2004). Polyphonic music transcription by non-negative sparse coding of power spectra. *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)* (pp. 318-325).

ASA. (1960). American standards acoustic terminology, Acoustical Society of America, New York (USA).

Bregman, A.S. (1990). *Auditory scene analysis*. MIT Press.

Buyens, W., van Dijk, B., Wouters, J., and Moonen, M. (2013). A Harmonic/Percussive Sound Separation based Music Pre-Processing Scheme for Cochlear Implant Users. *Proceedings of the 21st European Signal Processing Conference (EUSIPCO).*

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2014). Music mixing preferences of cochlear implant recipients: a pilot study. *International Journal of Audiology,* 53(5), 294-301.

Buyens, W., van Dijk, B., Wouters, J., and Moonen, M. (2015a). A stereo music pre-processing scheme for cochlear implant users. *IEEE Transactions on Biomedical Engineering* (in press).

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2015b). Evaluation of stereo music pre-processing for cochlear implant users. *International Journal of Audiology* (submitted for publication).

Chatterjee, M., and Shannon, R.V. (1998). Forward masked excitation patterns in multielectrode electrical stimulation. *Journal of the Acoustical Society of America*, 103(5), 2565-72.

Cheng, M.Y., Spitzer, J.B., Shafiro, V., Sheft, S., and Mancuso, D. (2013). Reliability measure of a clinical test: Appreciation of Music in Cochlear Implantees (AMICI). *Journal of the American Academy of Audiology*, 24(10), 969-79.

Clark, J.G. (1981). Uses and abuses of hearing loss classification. *Journal of the American Speech-Language-Hearing Association*, 23(7), 493-500.

Cooper, W.B., Tobey, E., and Loizou, P.C. (2009). Music perception by cochlear implant and normal hearing listeners as measured by the Montreal Battery for Evaluation of Amusia. *Ear and Hearing*, 29(4), 618-26.

Donnelly, P.J., Guo, B.Z., and Limb, C.J. (2009). Perceptual fusion of polyphonic pitch in cochlear implant users. *Journal of the Acoustical Society of America*, 126(5), EL128-133.

Drennan, W.R., Oleson, J.J., Gfeller, K., Crosson, J., Driscoll, V.D., Won, J.H., Anderson, E.S., and Rubinstein, J.T. (2015). Clinical evaluation of music perception, appraisal and experience in cochlear implant users. *International Journal of Audiology*, 54(2), 114-23.

Driscoll, V.D. (2012). The Effects of Training on Recognition of Musical Instruments by Adults with Cochlear Implants. *Seminars in Hearing*, 33(4), 410-418.

Ekström, S.R. and Borg, E. (2011). Hearing speech in music. *Noise and Health*, 13(53), 277-285.

Fitzgerald, D. (2010). Harmonic/Percussive Separation using Median Filtering. *Proceedings of the 13th International Conference on Digital Audio Effects (DAFX10)*.

Fu, Q.-J. and Nogaki, G. (2005). Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology*, 6(1), 19-27.

Fujita, S. and Ito, J. (1999). Ability of Nucleus cochlear implantees to recognize music. *Annals of Otology, Rhinology and Laryngology*, 108(7), 634-640.

Galvin J.J. III, Fu, Q.-J., Nogaki, G. (2007). Melodic contour identification by cochlear implant listeners. *Ear and Hearing*, 28, 302–319.

Galvin, J.J. III, Fu, Q.-J., and Shannon, R.V. (2009a). Melodic contour identification and music perception by cochlear implant users. *Annals of the New York Academy of Sciences*, 1169(1), 518-33.

Galvin, J.J. III, Fu, Q.-J., and Oba, S.I. (2009b). Effect of a competing instrument on melodic contour identification by cochlear implant users. *Journal of the Acoustical Society of America*, 125(3), EL98-103.

Gfeller, K. and Lansing, C.R. (1991). Melodic, rhythmic, and timbral perception of adult cochlear implant users. *Journal of Speech, Language, and Hearing Research*, 34(4), 916-920.

Gfeller, K., Knutson, J.F., Woodworth, G., Witt, S., and DeBus, B. (1998). Timbral recognition and appraisal by adult cochlear implant users and normal-hearing adults. *Journal of the American Academy of Audiology*, 9, 1-19.

Gfeller, K., Christ, A., Knutson, J.F., Witt, S., Murray, K.T., and Tyler, R.S. (2000). Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients. *Journal of the American Academy of Audiology*, 11(7), 390-406.

Gfeller, K., Turner, C., Mehr, M.A., Woodworth, G., Fearn, R., Knutson, J.F., Witt, S., and Stordahl, J. (2002a). Recognition of familiar melodies by adult

cochlear implant recipients and normal-hearing adults. *Cochlear Implants International*, 3(1), 29-53.

Gfeller, K., Witt, S., Woodworth, G., Mehr, M.A., and Knutson, J.F. (2002b). Effects of frequency, instrumental family, and cochlear implant type on timbre recognition and appraisal. *Annals of Otology, Rhinology and Laryngology*, 111(4), 349-56.

Gfeller, K., Witt, S., Adamek, M., Mehr, M., Rogers, J., Stordahl, J., and Ringgenberg, S. (2002c). Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients. *Journal of the American Academy of Audiology*, 13(3), 132-145.

Gfeller, K., Christ, A., Knutson, J.F., Witt, S., and Mehr, M.A. (2003). The effects of familiarity and complexity on appraisal of complex songs by cochlear implant recipients and normal hearing adults. *Journal of Music Therapy*, 40(2), 78-112.

Gfeller, K., Oleson, J., Knutson, J.F., Breheny, P., Driscoll, V.D., and Olszewski, C. (2008). Multivariate predictors of music perception and appraisal by adult cochlear implant users. *Journal of the American Academy of Audiology*, 19(2), 120-34.

Gfeller, K., Buzzell, A., Driscoll, V.D., Kinnaird, B., and Oleson, J. (2009). The impact of voice range and instrumental background accompaniment on recognition of song lyrics. *Proceedings of the 7th Asia Pacific Symposium on Cochlear Implants and Related Sciences* (pp. 35-39).

Helen, M. and Virtanen, T. (2005). Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. *Proceedings of the 13th European Signal Processing Conference (EUSIPCO)* (pp. 1091-94).

Kang, R.S., Nimmons, G.L., Drennan, W.R., Longnion, J., Ruffin, C., Nie, K., Won, J.H., Worman, T., Yueh, B., and Rubinstein, J.T. (2009). Development and validation of the University of Washington Clinical Assessment of Music Perception test. *Ear and Hearing*, 30(4), 411-8.

Kim, H., Kim, T.H., and Park, J.S. (2012). A Real Time Singing Voice Removal System Using DSP and Multichannel Audio Interface. *International Journal of Multimedia and Ubiquitous Engineering*, 7(2).

Kong, Y.Y., Cruz, R., Jones, J.A., and Zang, F.G. (2004). Music perception with temporal cues in acoustic and electric hearing. *Ear and Hearing*, 25(2), 173-185.

Laneau, J., Wouters, J., and Moonen, M. (2006). Improved music perception with explicit pitch coding in cochlear implants. *Audiology and Neurotology*, 11(1), 38-52.

Lassaletta, L., Castro, A., Bastarrica, M., Pérez-Mora, R., Herrán, B., Sanz, L., de Sarriá, M.J., and Gavilán, J. (2008). Changes in listening habits and quality of musical sound after cochlear implantation. *Otolaryngology - Head and Neck Surgery*, 138(3), 363-67.

Leal, M.C., Shin, Y.J., Laborde, M.L., Calmels, M.N., Verges, S., Lugardon, S., Andrieu, S., Deguine, O., and Fraysse, B. (2003). Music perception in adult cochlear implant recipients. *Acta Otolaryngology*, 123(7), 826-35.

Limb, C.J. (2006). Cochlear implant-mediated perception of music. *Current Opinion in Otolaryngology - Head and Neck Surgery*, 14(5), 337-40.

Limb, C.J. and Roy, A.T. (2013). Technological, biological, and acoustical constraints to music perception in cochlear implant users. *Hearing Research*, 308, 13-26.

Liutkus, A., Rafii, Z., Pardo, B., Fitzgerald, D., and Daudet, L. (2014). Kernel spectrogram models for source separation. *Proceedings of the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)* (pp. 6-10).

Loizou, P.C. (1998). Mimicking the human ear. *IEEE Signal Processing Magazine*, 15(5), 101-130.

Loizou, P.C., Dorman, M., and Fitzke, J. (2000). The effect of reduced dynamic range on speech understanding: implications for patients with cochlear implants. *Ear and Hearing*, 21(1), 25-31.

Looi, V., Gfeller, K., and Driscoll, V.D. (2012). Music appreciation and training for cochlear implant recipients: a review. *Seminars in Hearing*, 33(4), 307-34.

Looi, V., McDermott, H., McKay, C., and Hickson, L. (2007). Comparisons of quality ratings for music by cochlear implant and hearing aid users. *Ear and Hearing*, 28(2), 59S-61S.

Looi, V., McDermott, H., McKay, C., and Hickson, L. (2008). Music perception of cochlear implant users compared to that of hearing aid users. *Ear and Hearing*, 29(3), 421-34.

McDermott, H.J. (2004). Music perception with cochlear implants: A review. *Trends in Amplification*, 8(2), 49-82.

Migirov, L., Kronenberg, J., and Henkin, Y. (2009). Self-reported listening habits and enjoyment of music among adult cochlear implant recipients. *Annals of Otology Rhinology and Laryngology*, 118(5), 350-55.

Milczynski, M., Wouters, J., and Van Wieringen, A. (2009). Improved fundamental frequency coding in cochlear implant signal processing. *Journal of the Acoustical Society of America*, 125(4), 2260-71.

Mirza, S., Douglas, S.A., Lindsey, P., Hildreth, T., and Hawthorne, M. (2003). Appreciation of music in adult patients with cochlear implants: a patient questionnaire. *Cochlear Implants International*, 4(2), 85-95.

Nimmons, G.L., Kang, R.S., Drennan, W.R., Longnion, J., Ruffin, C., Worman, T., Yueh, B., and Rubenstein, J.T. (2008). Clinical assessment of music perception in cochlear implant listeners. *Otology and Neurotology*, 29(2), 149-55.

Olshausen, B.A. and Field, D.F. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1?. *Vision Research*, 37(23), 3311–25.

Omran, S.A., Lai, W., and Dillier, N. (2010). Pitch Ranking, Melody Contour and Instrument Recognition Tests Using Two Semitone Frequency Maps for Nucleus Cochlear Implants. *EURASIP Journal on Audio, Speech, and Music Processing.*

Ono, N., Miyamoto, K., Le Roux, J., Kameoka, H., and Sagayama, S. (2008). Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. *Proceedings of the 16th European Signal Processing Conference (EUSIPCO).*

Ono, N., Miyamoto, K., Kameoka, H., and Sagayama, S. (2009). A real-time equalizer of harmonic and percussive components in music signals. *Proceedings of ISMIR*, 139–144.

Ozerov, A., Vincent, E., and Bimbot, F. (2012). A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4), 1118-33.

Phillips-Silver, J., Toiviainen, P., Gosselin, N., Turgeon, C., Lepore, F., and Peretz, I. (2015). Cochlear implant users move in time to the beat of drum music. *Hearing Research*, 321, 25-34.

PleasurizeMusicFoundation (2013). http://www.pleasurizemusic.com (date last viewed 30/01/13).

Price, H. (1986). A proposed glossary for use in affective response literature in music. *Journal of Research in Music Education*, 34, 151-159.

Qazi, O., van Dijk, B., Moonen, M., and Wouters, J. (2012). Speech Understanding Performance of Cochlear Implant Subjects using Time-Frequency Masking based Noise Reduction. *IEEE Transactions on Biomedical Engineering*, 59(5), 1364-73.

Qazi, O., van Dijk, B., Moonen, M., and Wouters, J. (2013). Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility. *Hearing research*, 299, 79-87.

ReplayGain (2013). http://www.replaygain.org (date last viewed 30/01/13).

Serra, X. (1989). A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition. PhD thesis, Department of Music, Stanford University.

Serra, X. (1997). Musical sound modeling with sinusoids plus noise. *Musical Signal Processing*, 91–122.

Smith, J.O. and Serra, X. (1987). PARSHL: An analysis/synthesis program for nonharmonic sounds based on a sinusoidal representation. *Proceedings of the International Computer Music Conference (ICMC).*

Spitzer, J.B., Mancuso, D., and Cheng, M.Y. (2008). Development of a clinical test of musical perception: appreciation of music in cochlear implantees (AMICI). *Journal of the American Academy of Audiology*, 19(1), 56-81.

Tachibana, H., Kameoka, H., Ono, N., and Sagayama, S. (2012). Comparative evaluations of various harmonic/percussive sound separation algorithms based on anisotropic continuity of spectrogram. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 465-468).

Tachibana, H., Ono, T., Ono, N., and Sagayama, S. (2010). Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 425-428).

Tyler, R.S., Gfeller, K., and Mehr, M.A. (2000). A preliminary investigation comparing one and eight channels at fast and slow rates on music appraisal in adults with cochlear implants. *Cochlear Implants International*, 1(2), 82-87.

Uhle, C., Dittmar, C., and Sporer, T. (2003). Extraction of drum tracks from polyphonic music using independent subspace analysis. *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)* (pp. 843–847).

Van Wieringen, A. and Wouters, J. (2008). LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired

listeners for Flanders and the Netherlands. *International Journal of Audiology*, 47(6), 348-355.

Vandali, A.E., Whitford, L.A., Plant, K.L., and Clark, G.M. (2000). Speech perception as a function of electrical stimulation rate: using the Nucleus 24 cochlear implant system. *Ear and Hearing*, 21(6), 608-24.

Vickers, E. (2010). The Loudness War: Background, Speculation, and Recommendations. *Audio Engineering Society Convention 129*

Virtanen, T. (2003). Sound source separation using sparse coding with temporal continuity objective. *Proceedings of the International Computer Music Conference (ICMC)* (pp. 231-234).

Virtanen, T. (2006). Sound Source Separation in Monaural Music Signals. PhD thesis, Tampere University of Technology.

Wilson, B.S., Finley, C.C., Lawson, D.T., Wolford, R.D., Eddington, D.K., and Rabinowitz, W.M. (1991). Better speech recognition with cochlear implants. *Nature*, 352(6332), 236-238.

Won, J.H., Drennan, W.R., Kang, R.S., and Rubinstein, J.T. (2010). Psychoacoustic abilities associated with music perception in cochlear implant users. *Ear and Hearing*, 31(6), 796-805.

Wouters, J., McDermott, H., and Francart, T. (2015). Sound coding in cochlear implants: from electric pulses to hearing. *IEEE Signal Processing Magazine*, 32(2), 67-80.

Wright, R. and Uchanski R.M. (2012). Music perception and appraisal: cochlear implant users and simulated cochlear implant listening. *Journal of the American Academy of Audiology*, 23(5), 350-65.

York, G.W., Rondeau, C.M., and Fuller, D.F. (2004). Teaching Real-time DSP applications (Voice Removal) with the C6711 DSK and MATLAB. *Proceedings of the American Society for Engineering Education Annual Conference & Exposition (AEEE).*

Zeng, F.G., Rebscher, S., Harrison, W., Sun, X., and Feng, H. (2008). Cochlear implants: system design, integration, and evaluation. *IEEE Reviews in Biomedical Engineering*, 1, 115-42.

Zhu, M., Chen, B., Galvin, J.J. III, and Fu, Q.-J. (2011). Influence of pitch, timbre and timing cues on melodic contour identification with a competing masker (L). *Journal of the Acoustical Society of America*, 130(6), 3562-65.

Curriculum Vitae

Wim Buyens was born in Brasschaat, Belgium, in 1976. He received the degree of Master of Science in Electrical Engineering from the KU Leuven in 2000, with a Master's thesis on multi-microphone noise reduction techniques for hearing aids and cochlear implants. He worked as Research Engineer Speech Enhancement in L&H, as DSP Design Engineer in Philips Technology Campus Leuven and as Senior Research Engineer at Cochlear Technology Centre Belgium. In September 2009, he started a PhD in the Arenberg Doctoral School at KU Leuven on music pre-processing for cochlear implants with supervisors Prof. Marc Moonen (Dept. of Electrical Engineering - STADIUS) and Prof. Jan Wouters (Dept. of Neurosciences - ExpORL) and with co-supervisor Dr. Bas van Dijk (Cochlear Technology Centre Belgium). He received a Baekeland PhD grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT090274). His research interests lay in the area of speech enhancement, cochlear implant sound coding and signal processing for sound applications. Wim Buyens is married to Inge Verheyen and has three children Evelien, Lotte and Kato.

List of publications

Patents

Buyens W., “Music pre-processing for hearing prostheses”, Cochlear Ltd., Patent WO 2010068997 A1, Dec. 2008.

Buyens W., “Pre-processing of a channelized music signal”, Patent Application No. 14/329,518, July 2013.

Publications in international journals

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2014). Music mixing preferences of cochlear implant recipients: a pilot study. *International Journal of Audiology*, 53(5), 294-301.

Buyens, W., van Dijk, B., Wouters, J., and Moonen, M. (2015a). A stereo music pre-processing scheme for cochlear implant users. *IEEE Transactions on Biomedical Engineering* (in press).

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2015b). Evaluation of stereo music pre-processing for cochlear implant users. *International Journal of Audiology* (submitted for publication).

Contributions to conferences

Buyens, W., Moonen, M., Wouters, J., and van Dijk, B. (2010). Music pre-processing for cochlear implants. Poster at USMIR, Utrecht, Netherlands.

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2011). Music preference study with cochlear implant recipients using multi-track recordings. Poster at CIAP, Asilomar, USA.

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2012). Music preference of cochlear implant recipients studied with multi-track recordings. Presentation at ENT congress (B-Audio session), Antwerp, Belgium.

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2013). Evaluation of Music Pre-Processing Scheme with Cochlear Implant Recipients. Poster at CIAP, Lake Tahoe, USA.

Buyens, W., van Dijk, B., Wouters, J., and Moonen, M. (2013). A Harmonic/Percussive Sound Separation based Music Pre-Processing Scheme for Cochlear Implant Users. *Proceedings of the 21st European Signal Processing Conference (EUSIPCO).*

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2013). Music pre-processing assessment with cochlear implant users. Presentation at ENT congress (B-Audio session), Brussels, Belgium.

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2014). Clinical take-home evaluation of music pre-processing scheme for cochlear implant users. Presentation at ENT congress (B-Audio session), Bruges, Belgium.

Buyens, W., van Dijk, B., Moonen, M., and Wouters, J. (2015). Take-home Evaluation of Music Pre-Processing Scheme with Cochlear Implant Users. Poster at CIAP, Lake Tahoe, USA.