Perception of speech, music and emotion by hearing-impaired listeners

Tim Metcalfe

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the Department of Music

The University of Sheffield

November, 2017

Contents

Abstract

Overview

The everyday tasks of perceiving speech, music and emotional expression via both of these media, are made much more difficult in the case of hearing impairment. Chiefly, this is because relevant acoustic cues are less clearly audible, owing to both hearing loss in itself, and the limitations of available hearing prostheses. This thesis focussed specifically on two such devices, the cochlear implant (CI) and the hearing aid (HA), and asks two overarching questions: how do users approach music and speech perception tasks, and how can performance be improved?

The first part of the thesis considered auditory perception of emotion by CI users. In particular, the underlying mechanisms by which this population perform such tasks are poorly understood. This topic was addressed by a series of emotion discrimination experiments, featuring both normal-hearing (CI-simulated) participants and real CI users, in which listeners heard stimuli with processing designed to systematically attenuate different acoustic features. Additionally, a computational modelling approach was utilised in order to estimate participants' listening strategies, and whether or not these were optimal. It was shown that the acoustic features attended to by participants were a compromise of those generally better-preserved by the CI, and those particularly salient for each stimulus.

In the latter half of the thesis, the nature of assessment of music perception by hearing-impaired listeners was considered. Speech perception has typically taken precedence in this domain which, it is argued, has left assessment of music perception relatively underdeveloped. This problem was addressed by the creation of a novel, psychoacoustical testing procedure, similar to those typically used with speech.

This paradigm was evaluated via listening experiments with both HA users and CI-simulated listeners. In general, the results indicated that the measure produced both valid and reliable results, suggesting the suitability of the procedure as both a clinical and experimental tool.

Lastly, the thesis considered the consequences of the various findings for both research and clinical practice, contextualising the results with reference to the primary research questions addressed, and thereby highlighting what there is left to discover.

Acknowledgements

Firstly, I would like to thank the White Rose College of Arts and Humanities (WRoCAH) for awarding me a doctoral studentship, without which I would have not been able to pursue a PhD. WRoCAH also provided invaluable additional funding, facilitating both the conduction and dissemination of my research.

Of course, I am grateful for the invaluable assistance and encouragement of my supervisors, Dr. Renee Timmers and Prof. Guy Brown, throughout the entire PhD process – for providing constant guidance on all aspects of my research, and for providing training and extra-curricular opportunities which have vastly improved my confidence and effectiveness as a researcher. I would also like to thank Dr. Harriet Crook for her assistance with the recruitment and scheduling of cochlear implant users, and for providing assistance in communicating with those users for whom speech was more difficult.

Thanks must also be extended to my manager during an internship at Starkey Hearing Technologies, Dr. Jason Galster, as well as my colleagues Dr. Kelly Fitz and Dr. Jingjing Xu, and the rest of the Audiology Research team. The experience had a profound impact upon my approach to research, offering me the opportunity to develop - both personally and professionally – within a welcoming, supportive environment.

For offering me my first experience of being a researcher, I would like to thank my undergraduate tutor, Dr. Beth Jefferies. The years I spent as a research assistant with her lab were what first made my impetus to pursue a PhD appear as a tangible reality.

Lastly, as is customary, thanks must be extended to my family, for encouragement and all-important financial support throughout my academic journey, and of course

to my partner, Flavia, for offering support so far-reaching that to document it here would prolong the thesis by its own length again.

1 Literature review

1.1 The sense of hearing

1.1.1 Introduction

The sense of hearing, otherwise known as audition, is sensitivity to sound: fluctuations in pressure – usually in air – caused by vibrating objects (Plack, 2013). Sound plays a vital role in interpersonal communication, and in human-environment and human-object interaction. Audition facilitates the understanding of spoken language, appreciation of music, and awareness of myriad environmental sounds. Via fluctuations in air pressure, we are able to perceive and express information that is semantic, referential, spatial and emotional. The environment that we inhabit is most often filled with sound, and therefore information, that is potentially relevant, or was once relevant, for survival. The human ear has evolved, accordingly, to decode and to utilise this information. Of course, when access to this stream of information is disrupted or impoverished – as is the case with hearing loss – there are consequences for everyday life, relating to the aforementioned functions of audition. According to the oft-repeated aphorism, however, in order to understand hearing loss, one must first understand hearing itself (Wayner, 1990). Hence, to contextualise much of the upcoming discussion about the effects of hearing loss, there follows a detailed overview of how the sense of hearing functions in normally-hearing individuals. The processes involved are enormously complicated, and as such it is beyond the scope of this thesis to provide an exhaustive description of each intricacy. However, this description shall hopefully suffice to illustrate the sophisticated, sequential nature of audition, and thereby the challenges inherent in attempting to restore this sense, following hearing loss. Later in this section, focus will be placed more specifically on

aspects of hearing loss that are relevant for the perception of speech and music, and of emotion expressed via these media.

1.1.2 The mechanics of hearing

In normal-hearing (NH) listeners, sound waves are collected by the pinna (the visible part of the ear, on the outside of the head) and are filtered linearly according to the spatial location from which they originate, thus providing a sense of directionality dependent on the resulting frequency spectrum (Butler & Flannery, 1980). In fact, the head, pinna and auditory canal also have combined resonances, such that sound waves with frequencies important for speech perception (2-5 kHz) are enhanced in intensity by approximately 10-15 dB SPL (Pocock, Richards, and Richards, 2013; Welling, Ukstins, et al., 2013). From the pinna, sound waves are directed through the auditory canal – a tube around 2.5cm in length – until they hit the tympanic membrane, which vibrates according to the waveform of the sound (Bess & Humes, 2008). Sound is then transferred through the middle ear – an air-filled cavity containing three bones collectively referred to as the ossicles (see Figure 1). The tympanic membrane is attached to the malleus, which is connected to the incus which in turn is connected to the stapes. The purpose of these middle ear structures is to perform an efficient mechanical conversion between vibrations of the tympanic membrane and pressure waves in the fluid of the cochlear, maximising power transfer and reducing signal reflection (Andersson et al., 2006). This is achieved via leverage-based amplification by the malleus and incus, and pneumatic amplification – i.e. increased pressure due to the large discrepancy in surface area between the tympanic membrane and the stapes (Bear, Connors, & Paradiso, 2007). Without this process of impedance matching, much of the sound arriving at the inner ear would be reflected

back due to the relatively greater density of cochlear fluid, compared to air. Importantly, the middle ear also contains the stapedius and tensor tympani muscles, which underlie the acoustic reflex (Fox, 1996). In the event of a sudden, loud sound, these muscles contract, stiffening the ossicular chain and thereby reducing the transfer of vibrational energy to the cochlear, offering a degree of protection against auditory fatigue, or over-stimulation. The piston-like movement of the stapes sends vibrations into the bony labyrinth via the oval window, a membrane-covered entrance to the inner ear. As the oval window is contacted by the stapes, the round window – a second, flexible membrane located below the oval window – moves outward, thereby vibrating with approximately equal velocity but opposite phase (Gulya, Minor, Glasscock, & Poe, 2010). The movement of both the oval window and round window membranes is necessary for the movement of fluid within the cochlea, since this fluid is effectively incompressible, and the walls of the bony labyrinth are inflexible (Gulya et al., 2010).

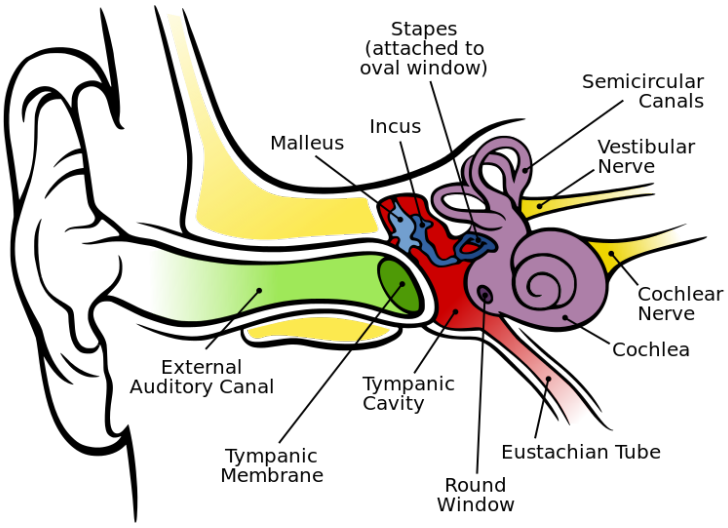


Figure 1: Illustrative diagram of the peripheral auditory system (Chittka & Brockmann, 2005). Used under Creative Commons Attribution 2.0 Generic License.

The cochlea itself is a spiral-shaped structure, containing two fluid-filled canals, sep-

arated by the cochlear duct (scala media), which is itself filled with fluid (Møller, 2000). Beginning at the oval window, vibrations produced by the stapes travel through an ascending canal (scala vestibuli), towards the cochlear apex, where the two canals merge at a passage known as the helicotrema (Møller, 2000). From here, the vibrations then travel through the descending canal (scala tympani), terminating at the round window. Separating the cochlear duct from the scala tympani is the basilar membrane, a pseudo-resonant structure which moves according to sound vibrations within the cochlea (Holmes & Cole, 1983). The basilar membrane has different characteristic frequencies along its length, owing to increased width and decreased stiffness at the apex and vice versa at the base (Von Békésy & Wever, 1960). This frequency-to-place mapping, whereby topographically neighbouring regions are sharply tuned to tones that are proximal in terms of frequency (and vice versa), is referred to as tonotopy.

Frequency may also be encoded by the rate of vibration of the basilar membrane, which is known as 'temporal coding'. As one portion of the basilar membrane vibrates according to its characteristic frequency, this movement causes some degree of synchronised movement in adjacent regions, resulting in an overall rate of vibration of the basilar membrane which is related to the frequency of the sound stimulus (Handel, 1993). Since auditory nerve firing tends to occur at a particular phase of a sound wave (a phenomenon referred to as 'phase-locking') the interval between firing is approximately an integer multiple of signal frequency (Oxenham, 2013). It remains to be determined exactly when, and to what extent and when either spatial or temporal encoding predominate, but modern theories of pitch perception largely acknowledge and incorporate both. In any case, the basilar membrane is central to the perception of frequency in sound.

Situated along the basilar membrane is the organ of Corti, a sensory epithelium (cell-layer) responsible for the mechanotransduction of sound pressure waves to electrical signals (Lim, 1986). As the basilar membrane vibrates (with a pattern of motion characterised as a travelling wave) (Von Békésy & Wever, 1960), a 'shearing' motion is induced between the organ of Corti and the tectorial membrane (a structure that runs parallel to the basilar membrane) (Lim, 1986). This motion causes the bundles of hair cells comprising the organ of Corti to be deflected. Specifically, the motion-sensing organelles (stereocilia) of these hair cells are deflected, which places tension on the filaments (tip links) connecting different rows of stereocilia (Hackney & Furness, 2013). Subsequently, these tip links pull open mechanically-gated ion channels located at their lower end (Beurg, Fettiplace, Nam, & Ricci, 2009), causing an influx of positively-charged potassium and sodium ions into the channel, thereby depolarising the respective hair cell. This leads to the release of neurotransmitter at the synapse between the hair cell and the auditory nerve fibre, producing an action potential that travels to the cochlear nucleus in the brainstem (Yost, 2003). From here, this electrical activity is propagated via a network of afferent nerve fibres, which synapse at the medial geniculate body, before travelling upwards to the primary auditory cortex, and beyond (Bess & Humes, 2008). Here, the basic acoustic attributes of auditory stimuli are decoded (frequency, intensity, etc.) and, in conjunction with a distributed network of brain structures specialised for processing higher-order properties of sound, significance is ascribed to the sounds perceived.

It is important to note that higher-order interpretation or 'sense-making' of sound stimuli, is much more than a linear reflection of the physical properties of these sounds. Instead, the response of the auditory system is interpreted in accordance with top-down information about the nature and relevance of various attributes of

the stimulus being perceived. In describing this distinction, Bregman (1994) distinguishes between 'primitive' and 'schema-driven' processes in audition. The former refers to bottom-up, automatic processes of sensory pattern construction, which are argued to directly reflect "acoustical universals which humans at all times and in all places have encountered" (Hanley and Goolsby, 2002, p. 81). By contrast, schema-driven processes reflect learned associations between higher-order patterns of sounds and stored knowledge, and vary both cross-culturally and inter-individually (Bregman, 1994). According to most modern accounts of audition, primitive and schema-driven processes are almost always active concurrently, and act in a complementary fashion. Therefore, the overview of hearing mechanics provides useful context for the present discussion, is far from the end of the story, and in actuality the sense of hearing is vastly more complex.

1.1.3 Non-linearity in auditory perception

Functionally, there are various important nonlinearities between the mechanics of the human auditory system and the ways in which sounds are actually experienced. For example, frequency discrimination sensitivity varies according to frequency, with more fine-grained same-different judgements being possible, in general, for lower-frequency pairs of tones (Moore, 2003). Human perception of the frequency of a stimulus, usually referred to as pitch – i.e. "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale" (ASA, 1960) – is approximately logarithmic (with respect to base 2) in relation to stimulus fundamental frequency. That is, when sounds are doubled in frequency, they are usually interpreted as belonging to consecutive octaves. The most likely explanation for logarithmic perception is an evolutionary one: that it facilitates more finely tuned sensitivity

to the sounds that occur most often in one's environment. Since the majority of sounds experienced occur somewhere towards the lower end of the typical audible range (20Hz- 20,000 Hz), it makes sense to be able to distinguish with greater precision sounds occurring in this region. As Varshney and Sun (2013, p. 28) put it, this approach to perception allows us to "efficiently represent the statistical distributions of the natural world", improving perceptual resolution where it is most relevant to do so.

This nonlinearity also holds true for other aspects of the auditory signal, such as the perception of sound intensity, otherwise referred to as loudness. For example, a 1 kHz tone presented at 50 dB SPL is usually perceived as being approximately twice as loud as the same tone presented at 40 dB SPL (Moore, 2003). Again, this logarithmic scaling facilitates a better perceptual resolution for softer sounds, and facilitates efficient representation of an incredibly large range of sound intensities – painfully loud sounds are about one hundred trillion times more intense than sounds that are just audible (Smith, 2016). Varshney and Sun (2013) argue that, since humans are sensitive to relative, rather than absolute, differences between stimuli, a logarithmic mapping between external stimuli and perceptual experience is the statistically optimal way to reduce relative error in perception. In fact, this stimulus-percept mapping is pervasive for all kinds of stimuli – both within and beyond the auditory domain – and is enshrined in a basic principle of psychophysics known as the Weber-Fechner law (Varshney & Sun, 2013).

1.1.4 Summary

In summary, the sense of hearing is a highly complicated process consisting of many separate stages, each depending on the correct functioning of those preceding them.

In fact, the auditory system also incorporates complex feedback loops between various stages of processing, though a detailed discussion of such intricacies is neither necessary nor practical to include here.

When it works as intended, the auditory system is sophisticated and elegant in its function. However, the description of audition given above serves to highlight two major points, with respect to the loss of hearing:

A) There are numerous stages involved in the amplification and transduction of auditory input, and therefore there are many points during the process of audition at which it may be disrupted. Because of this, there exists a huge variety of potential hearing disorders, each characterised by heterogeneous symptomatologies and with substantial inter-individual differences. That is, individuals suffering from the same disorder of hearing may be relatively dissimilar.

B) Since the auditory system is hugely complex, restoring functionality following hearing loss is very difficult, and often fraught with complications. Although enormous progress has been made in recent years, sensory aids and prostheses that are intended to restore the sense of hearing are, at present, a relatively crude alternative when compared to the process outlined above. Although such devices are able to restore a functional sense of hearing, this sense is usually impoverished, to some extent, when compared to those with normal hearing.

Hereafter, this thesis shall consider what happens when normal hearing function is disrupted, examining the deficits that arise, how these can be circumvented, and in what ways these solutions affect the auditory signal. In particular, perception of speech and music by hearing-impaired listeners will be investigated, with a special focus on the specific challenges created by the use of cochlear implants and hearing

aids. To begin with, the causes and effects of different types of hearing impairment will be described in greater detail.

1.2 Hearing impairment

'Hearing loss' refers to a temporary or permanent, partial or total loss of the ability to hear in either one (monaural) or both ears (binaural) (Plack, 2013). It may be the most common disorder of communication in adults (Boone & Plante, 1993). According to a 2012 estimation by the World Health Organisation, approximately 360 million people worldwide suffer from disabling hearing loss (approximately 5% of the world population, as of 2017), denoting a loss greater than 40 dB for adults, or 30 dB for children (WHO, 2012). In persons aged 65 and over, the prevalence of disabling hearing loss is much higher – around 33% worldwide (WHO, 2012). Numerous risk factors for hearing loss have been identified by researchers, including: ageing, genetics, exposure to noise, perinatal developmental complications, traumatic injury, and a plethora of diseases and disorders (Lasak, Allen, McVay, & Lewis, 2014).

Broadly speaking, disruption to hearing function may be categorised according to one of three pathogeneses: conductive hearing loss, sensorineural hearing loss, and mixed hearing loss (i.e. a combination of the aforementioned) (Elzouki et al., 2011). 'Conductive hearing loss' refers to a mechanical failure of sound wave transmission that occurs prior to the sound reaching the inner ear – i.e. at the outer or middle ear (Elzouki et al., 2011). This is typically caused by inflammation of the ear canal or middle ear, and/ or accumulation of fluid – both of which are commonly caused by various ear infections – but may also be caused by perforation of the tympanic membrane, or by blockage of the ear canal by foreign objects (Sataloff & Sataloff, 2006). In any case, this type of hearing loss involves some type of physical 'barrier',

preventing the transference of sound waves to the inner ear, and further through the auditory system. Therefore, the treatment of conductive hearing loss typically targets the removal of this obstacle.

By contrast, sensorineural hearing loss refers to a disruption occurring at the inner ear or beyond, typically denoting either: tissue damage and/ or cell death at the cochlea, a failure of the cochlear nerve to transmit auditory information to the brain, or damage to the auditory cortex itself (Elzouki et al., 2011; Wong and Ryan, 2015). Of the above, the most apparent histological biomarker for sensorineural hearing loss is cochlear sensory cell damage and/ or loss (Schuknecht, Kimura, & Naufal, 1973). Unfortunately, these cells do not regenerate, and as such this loss is irreversible (Wong & Ryan, 2015). Accordingly, approaches to the treatment of hearing loss are premised on managing and circumventing its symptoms, as opposed to 'treating' the underlying causes. To this end, two of the most popular methods available for overcoming the effects of sensorineural hearing loss are hearing aids and cochlear implants (depending on the severity of the loss). Both of these solutions will be described in the next section.

1.3 How can hearing loss be circumvented?

1.3.1 Hearing aids

1.3.1.1 Candidacy Hearing aids are a widespread treatment option for individuals with sensorineural hearing loss (HL). Although anyone for whom interpersonal communication is impaired by HL may be a potential candidate, those with 'moderate' HL are typically expected to benefit the most from an HA (Bess & Humes, 2008). Moderate hearing loss is usually defined as a loss in the range of 41 to 55 dB

(Clark, 1981 – see Table 1). As the degree of HL moves away from this range, in either direction, the benefits of amplification are typically diminished. Individuals with only minimal HL do not experience great improvements, since they are usually able to cope adequately without amplification. For such individuals, 'assistive devices', providing for example amplification of the telephone, may be sufficient to meet their needs (Bess & Humes, 2008). Conversely, for those with severe to profound HL, amplification is often insufficient, since it is not feasible to provide the degree of amplification that would be required. In such cases, different treatment options (i.e. cochlear implants) might be considered.

Table 1: Summary table illustrating how different degrees of hearing loss may be classified, according to Clark (1981).

Degree of loss	**Range of loss (dB HL)**
Mild	26-40
Moderate	41-70
Severe	71-90
Profound	91+

1.3.1.2 Structure and function Modern digital hearing aids essentially consist of three components. Firstly, there is at least one microphone, which transduces sound waves present in one's environment. In order to provide directional information (used, for example, for noise cancellation), HAs may include two or more microphones, enabling differential processing of auditory signals, depending on the spatial location of their origin. Typically, this technology is used to preferentially amplify sounds in front of the listener, with the goal being to improve audibility of attended-to signals compared to background noise (Bess & Humes, 2008).

Once sound has been captured, it is passed to the next component of the hearing aid: a digital signal processor. Here, an analog-to-digital (A/D) conversion takes place, so that the signal may be easily manipulated. Next, various digital signal

processing (DSP) techniques are implemented, intended primarily to improve the signal-to-noise ratio (SNR) of incoming sounds. During this stage, sounds are typically low- and high-pass filtered, to remove unwanted frequencies, and amplified non-linearly. Specifically, wide dynamic range compression (WDRC) is used to accommodate the reduced range of sound intensities that hearing-impaired listeners are sensitive to, compared with the wide range of sounds naturally present in one's environment. Effectively, a multitude of different sound intensities captured by the microphone are 'squeezed' to fit into a narrower range. Particularly loud sounds ($\gtrsim$ 90 dB) are limited (i.e. compressed with a very high ratio), meaning that sounds exceeding some threshold receive no or negligible amplification (see Figure 2). The goal of this processing is to ensure that soft sounds are adequately amplified so as to render them audible, whilst intense sounds do not receive so much amplification that they would become uncomfortably loud. The exact thresholds at which input sounds are linearly amplified, compressed or limited, as well as the compression ratios employed across different frequencies and intensity levels can vary to a large degree from one HA user to another. In fact, the exact DSP employed can vary enormously across listeners, and across different HA manufacturers and models. However, most modern HAs incorporate to some extent: background noise reduction algorithms, feedback suppression and enhancement of speech intelligibility, for example via spectral subtraction (Murray & Hanson, 1992).

After amplification and any further processing stages have taken place, the signal is passed to a digital-to-analog (D/A) converter, before reaching the third component of the HA: a loudspeaker (or 'receiver'). The job of the loudspeaker is simply to deliver the analog output signal directly to the wearer's auditory canal (Bess & Humes, 2008). From this point onwards, the process of hearing can proceed as it would for

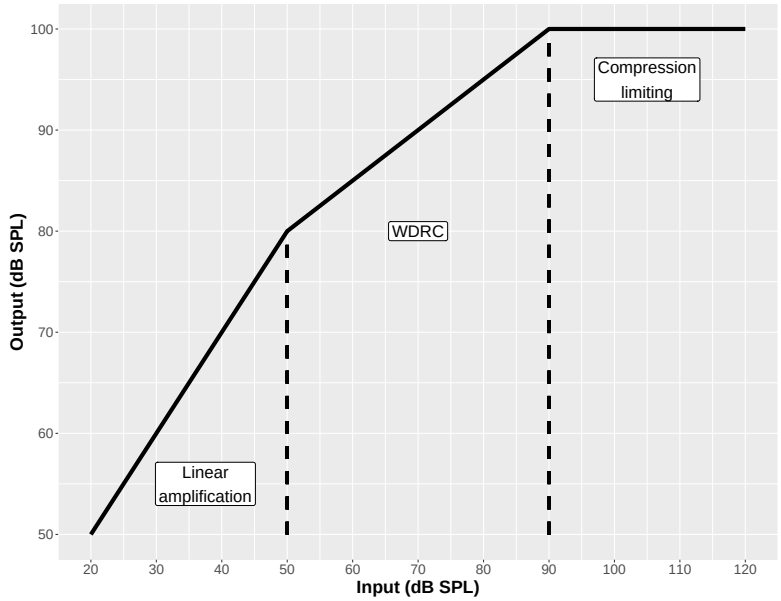


Figure 2: Example input-output function of a hearing aid, showing linear gain, WDRC and compression limiting. This processing ensures that soft sounds are amplified sufficiently, moderate sounds receive less amplification, and very intense sounds are either not amplified at all, or are attenuated.

an NH individual.

It should be noted that HA technology is constantly advancing – the above description constitutes a general overview of the core functionality of the HA, but this is rapidly being both improved and added to (Popelka & Moore, 2016). For example, increasingly sophisticated signal processing algorithms – incorporating techniques such as deep neural networks – are being developed (Wang, 2017), and the level of connectivity between HAs and other, relevant electronic devices (e.g. smartphones) is continuously expanding (Popelka & Moore, 2016). Therefore, even though the overall process outlined above is unlikely to change drastically in the near future, each processing stage is becoming progressively more intricate and efficient, with the associated hardware components becoming smaller and less obtrusive (Popelka & Moore, 2016).

1.3.1.3 Fitting

Patients are typically fitted with two hearing aids (i.e. 'binaural amplification'). However, monaural fittings are of course cheaper and may be offered for individuals with a milder HL, since this group have been shown to benefit less from binaural amplification. After an audiologist has selected an HA model that is suitable for the particular individual, its DSP characteristics can be 'tuned' in accordance with the patient's audiogram, and any other relevant characteristics (e.g. listening habits, etc.). Since the degree of HL usually varies across the frequency spectrum, it is beneficial to provide differential amplification across each frequency band, such that more amplification is provided where hearing is most deficient.

Once an HA has been fitted and programmed, the patient typically undergoes a period of rehabilitation, during which they become acclimated with their restored sensation of hearing, and the audiologist is able to make any necessary adjustments. Typically, the user also receives some guidance with respect to cognitive or behavioural strategies that can be used to maximise the effectiveness of the HA, including for example: wearing the HA as often as possible, lip-reading, avoiding very noisy environments, etc. (Kricos, 2017).

1.3.2 Cochlear implants

1.3.2.1 Candidacy

Cochlear implants are an alternative treatment to hearing aids, for individuals with a greater degree of hearing loss, who do not receive adequate benefit from amplification (Gifford, 2011). Patients must have already used HAs, and it must be demonstrated that they provide little or no benefit (for example, in cases where prescribed target gains are not obtainable). Typically, candidates for cochlear implantation will have bilateral severe (71-90 db HL) to profound (91+ dB HL) hearing loss, although patients with moderate low-frequency loss, sloping to severe/

profound loss may also be considered, depending on criteria set by both the implant manufacturer and the individual's insurance provider (Gifford, 2011). In addition to pure-tone audiometry, candidacy may also be decided based upon individuals' speech recognition performance. Again, the exact performance threshold signifying candidacy is not universally agreed upon. Depending on the manufacturer and/ or insurer, when listening to speech presented at 60 dB SPL, with amplification, this threshold varies between 40 to 60 percent correct (Gifford, 2008; Nilsson, Soli, and Sullivan, 1994).

Of course, there are also more qualitative considerations when establishing CI candidacy. Patients must be deemed medically fit to undergo the implantation surgery, and adult patients must usually demonstrate at least some ability to communicate using spoken language (often in conjunction with lip-reading) (BCIG, 2016). Even for eligible candidates, the decision to undergo cochlear implantation is not necessarily straightforward, and there are several potential drawbacks that might influence an individuals decision. Foremost, the CI has a much greater overall impact on the way that individuals perceive sound. At present, it is not possible for CIs to restore sound sensation in such a way that mimics normal hearing, and therefore adult patients may be discouraged by the need to readjust to this different 'mode' of hearing. Secondly, the cost of the CI is much higher than the HA. Although the costs may be state-funded, or paid for by medical insurers, there are nonetheless cases in which the cost presents a significant barrier. Other, more peripheral considerations include the typically shorter battery life of the CI relative to the HA, and the reduced flexibility, since the internal parts of the implant cannot be removed (without surgery). Lastly, and relevant for HA candidacy also, is the consideration that some individuals might be disinterested altogether in the restoration of hearing sensation, choosing instead

to communicate via sign language (Byrd, Shuman, Kileny, & Kileny, 2011). This is largely a cultural debate (and in the case of hearing-impaired children, potentially a legal or moral debate) which, in any case, falls well outside the scope of this thesis.

1.3.2.2 Structure and function In essence, the CI consists of the same three components as the HA: a microphone (or multiple microphones), a digital signal processor, and a receiver. As with the hearing aid, the microphone is used for the collection and transduction of sound; and the digital signal processor is responsible for A/D conversion, amplification and any additional processing of the signal. Unlike the hearing aid however, this signal is not necessarily converted back to analog. Following digital signal processing (DSP), the sound signal is sent to an external receiver placed behind one's ear, and is then relayed to an internal receiver implanted just beneath the skin. Here, the digital signal is converted to an electrical one, which is directed to an electrode array inserted into the cochlea. Typically, up to twenty-two electrodes are used, although not all will necessarily be functional – in an analysis of over 300 in-use implants, Schow, Friedland, Jensen, Burg, and Runge (2012) found that at least one electrode had been turned off in more than 50% of cases.

In order to preserve the cochlea's characteristic tonotopic mapping (the organisation of the cochlea such that increasingly basal locations resonate preferentially to increasingly high frequencies), the input signal is band-pass filtered, and the resultant signal from each band is sent to the corresponding intra-cochlear electrode. That is, lower frequency components of the input signal are directed to electrodes situated in more apical regions of the cochlea, and vice versa – such that regions of the cochlear that would normally resonate with particular frequencies are stimulated with approximately corresponding frequencies via intra-cochlear electrodes. Thus, some degree of place coding for the perception of pitch is preserved by the CI.

However, the preservation of temporal coding is more problematic. The rate of stimulation (pulses per second,per electrode) is usually constant, as opposed to varying dynamically according to the frequency of a given sound. Therefore, while the timing of auditory nerve firing is usually related to the frequency of the perceived signal in normal hearing, this is not typically the case with the CI (although recent research has explored how this limitation might be overcome (e.g. Landsberger, Vermeire, Claes, Van Rompaey, and Van de Heyning, 2016)). In fact, the independence of location and rate of electrical stimulation with the CI has been exploited by researchers seeking to disentangle place and temporal coding in pitch perception (Fearn, Carter, and Wolfe, 1999; Zeng, 2002).

Depending on the stimulation strategy, which varies according to implant manufacturer and/ or model, the implanted electrodes may be stimulated directly with the filtered sound signals in analog form (i.e. electrically current varying continuously over time). Alternatively, these signals may first be converted to patterns of biphasic square-wave pulses (Sandlin, 2000). Using the former approach, electrodes are usually stimulated simultaneously, whilst the latter strategy allows electrodes to instead be stimulated sequentially (at rates of up to and exceeding 2,000 pulses per second), minimising unwanted interaction between the different channels (Sandlin, 2000). In any case, the CI effectively bypasses damaged inner ear circuitry, caused by the deterioration and/ or loss of cochlear hair cells, and instead stimulates the cochlea directly via electrical current. Therefore, compared to the HA, which aims to amplify and process the acoustic signal *before* it reaches the human auditory system, the goal of the CI is to *bypass* a large part of the auditory system – replacing much of the functionality of the outer, middle and inner ear with what has been termed a 'bionic ear' (Pinyon et al., 2014).

Compared to the HA, the invention of the CI occurred relatively recently, and therefore it is more likely that the core structure and function of this device could see substantial changes in the not too distant future. Along with ongoing improvements being made to signal processing algorithms to maximise the performance of the modern CI, other more substantive developments have been explored in recent years. For example, Yip, Jin, Nakajima, Stankovic, and Chandrakasan (2015) proposed and demonstrated the feasibility of an entirely implantable device, in which the external microphone and transmitter are replaced with a piezoelectric sensor located at the middle ear. This is sensitive to the alternating current (AC) voltage generated by the mechanical movement of the ossicles, in response to sound. Additionally, researchers have begun to explore the possibility of utilising optical cochlear stimulation, rather than electrical stimulation, in order to achieve greater frequency resolution, and alleviate problems such as current spread (Johannsmeier et al., 2017). Lastly, as alluded to previously, recent research has explored the implementation of temporal coding of frequency by the CI, at least for lower frequencies (Landsberger et al., 2016). Therefore, in the long term, there is scope for major innovation, with respect to the structure and function of CIs. More specifically, the core processes outlined above will likely remain somewhat similar, but the hardware involved in implementing them could potentially change drastically.

1.3.2.3 Fitting In the UK, adults are typically fitted with a monaural CI (unless suffering from concurrent impairment to other senses) whilst children are eligible for binaural implantation (Raine, 2013). However, adults with a CI may also be fitted with an HA for the contralateral ear, where appropriate. The implantation procedure itself takes place under general anaesthetic, and typically lasts about two hours, assuming that there are no complications (Clark, 2003).

As with the HA, once a cochlear implant has been fitted, audiologists follow numerous procedures to ensure that benefit to the patient is maximised. Immediately after implantation, neural response telemetry (NRT) is used to assess the electrically-evoked compound action potential (ECAP) a characteristic neural event arising from the auditory nerve which is expected to occur in response to an auditory stimulus (Hughes, 2010). This is analogous to the Auditory Brainstem Response (ABR), and essentially is measured in order to determine that auditory nerve is receiving adequate stimulation from the CI (Hughes, 2010). Once the audiologist has obtained a clear picture of the functionality of the implanted electrode array, adjustments may be made as required. After initial programming of the CI, patients typically return to the clinic at least six times within the first year, so that the device settings may be optimised for the patient over time (BCIG, 2016). Typically, CI recipients also undergo a process of rehabilitation during this time, which is aimed at assisting users in acclimatising to their new hearing sensation, as well as training specific skills, such as speech intelligibility (Müller & Raine, 2013).

1.4 Other options

Although this thesis is concerned primarily with cochlear implants and hearing aids (the most common treatments for moderate and profound hearing loss, respectively), for the sake of completeness, it should be noted that there are various alternative options for the treatment of hearing loss. The bone-anchored hearing aid (BAHA) is a relatively common alternative to the conventional hearing aid discussed above, suitable for individuals with chronic ear disease or atresia (absence of the auditory canal) (Bess & Humes, 2008). The BAHA provides a sensation of sound via mechanical vibrations, which are delivered to the mastoid bone, thereby circumventing the

auditory canal – therefore, this option bypasses more of peripheral auditory system than the HA, but less than the CI.

More rarely, vibrotactile devices are a treatment option that aim to entirely replace the sense of hearing, by instead encoding the acoustic signal as patterns of vibration at the surface of the skin. This method is usually reserved for extreme cases, where even the cochlear implant does not provide adequate benefit, for example due to particularly extensive cochlear damage. Performance on tasks like speech recognition tends to be far worse than with a hearing aid or cochlear implant, and extensive rehabilitation is required for participants to become familiar with this new mode of 'hearing' (Geers, 1986; Levine, Miyamoto, Myres, Wagner, and Punch, 1987).

Lastly, as mentioned previously, there are various assistive listening devices (ALDs) available for the treatment of hearing loss that is too mild to necessitate a hearing aid. These include portable wireless receivers for hearing loop systems; FM systems, which work in a very similar way; and personal amplifiers, which enable users to amplify desired sounds and have these signals sent to an earpiece-based receiver (Bess & Humes, 2008).

Hearing aids and cochlear implants are far from the end of the story when it comes to treating hearing loss. However, they are by far the most commonly used and widely manufactured devices, and therefore have been the target of much greater research effort in recent years. For this reason, the remainder of this thesis focusses chiefly on these two devices. The next section will consider in more detail the effects of HL itself, and also the effects that the HA and CI have upon the auditory signal. Although these devices are able to circumvent the effects of hearing loss to a large extent, they also create additional, inherent challenges, which are an important consideration when trying to improve users' experience. It should be noted that, in the next section,

and thereafter, the primary emphasis when discussing sound is upon the perception thereof via the sense of hearing. Though sound can in principle be perceived via various other means such as vibrotactile stimulation or bone conduction, a thorough discussion of these modalities lies outside the scope of the thesis.

1.5 How do solutions to hearing loss affect the auditory signal?

1.5.1 Effects of hearing loss

Thus far, hearing loss has been mostly discussed only in terms of its overall extent (i.e. 'moderate', 'severe', etc.). However, hearing loss is a far more complex phenomenon than a mere reduction in the overall audibility of sound. In addition to this, individuals with sensorineural hearing loss must overcome several other obstacles in order for auditory perception to proceed optimally. In fact, various concomitant deficits originate from the hearing loss itself, including sensitivity to a reduced dynamic range, decreased temporal resolution, and decreased frequency resolution (Moore, 1996).

Cochlear damage associated with sensorineural hearing loss is often accompanied by a reduction in frequency selectivity (Moore, 2003). Zwicker and Schorn (1978) demonstrated this by comparing psychophysical tuning curves (PTCs) for normal-hearing listeners and several hearing-impaired groups. In this classic psychoacoustical paradigm, individuals are presented with a probe tone at a given frequency P, and simultaneously a masker tone at frequency $P \pm Q$. Afterwards, subjects are asked to report whether or not they were able to hear the probe tone. Typically, as Q gets larger (i.e. the difference between the probe and masker frequencies increases), the intensity of the masker must be greater, in order to obscure the probe. Therefore, in

normal-hearing listeners, these experiments tend to produce characteristic, narrow 'V' shaped curves – the efficacy of masker tones clearly depends on both presentation level, and their proximity to the probe (see Figure 3). In Zwicker and Schorn's (1978) experiment – and various similar studies (e.g. Carney and Nelson, 1983, Florentine, Buus, Scharf, and Zwicker, 1980), much broader tuning curves were elicited from hearing-impaired participants, indicating successful masking by comparably broad ranges of masker tones. In particular, tuning curves tend to be differentially broader towards the lower-frequency side, i.e. when masking tones are lower-frequency than probes (Glasberg & Moore, 1986). These findings suggests that the cochlea, previously specialised for responding to relatively narrow frequency ranges, affectively becomes 'tuned' to much broader ranges, following hearing loss – most likely as a consequence of deteriorated functioning of outer hair cells (Moore, 1996).

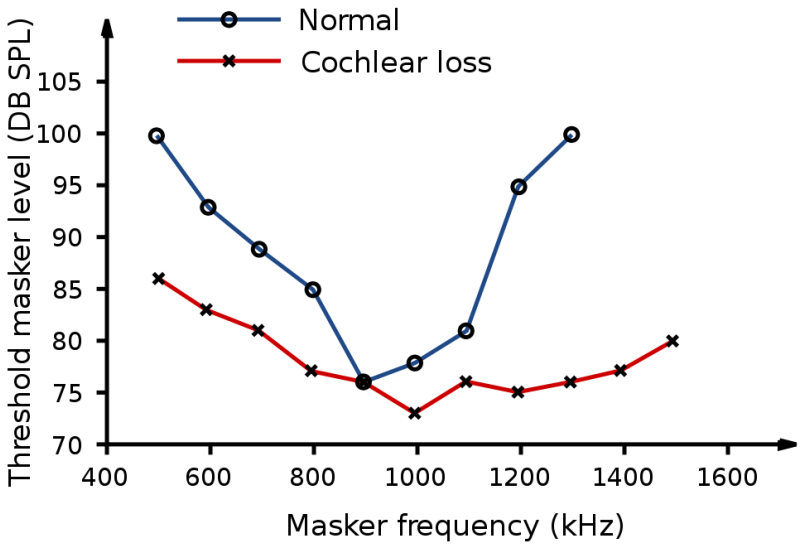


Figure 3: Examples of characteristic psychoacoustic tuning curves for both a normal-hearing subject, and a subject with cochlear damage. Reproduced from https://commons.wikimedia.org/wiki/File:Psychoacoustical_tuning_curves.svg. Public domain.

In addition to this, individuals with hearing loss often display 'loudness recruit-

ment', which refers to an elevated absolute threshold for the perception of sound, accompanied by an abnormally steep loudness growth function (Moore, 2003). That is, supra-threshold increases in stimulus intensity lead to proportionally greater increases in one's perception of loudness, compared to a normally-hearing individual. Perceptually, loudness recruitment is associated with great difficulty hearing quiet sounds, but normal, or near-normal performance with louder sounds (see Figure 4). This is in contrast with a flat pattern of loss, in which sensitivity over the entire dynamic range is attenuated more uniformly, as is usually the case with conductive hearing loss. Loudness recruitment may effectively shrink the dynamic range that one is sensitive to, since the absolute threshold is increased, while the threshold of pain (i.e. the loudest sound that an individual can tolerate) remains largely unchanged. This phenomenon is typically thought to be caused by a decreased cochlear hair cell population – the gradation of different intensities of sound is impaired, since there are fewer intact cells available to represent them (Joris, 2009). The exact prevalence of loudness recruitment in patients with sensorineural HL is not known, although it has been claimed that the majority will experience it to some extent (Plomp & Mimpen, 1979).

Of course, to varying degrees, the treatments available for hearing loss may overcome these perceptual deficits. However, these prostheses also have their own consequences for the sense of hearing, which present additional challenges for hearing-impaired listeners. The effects of hearing aids and cochlear implants on the auditory signal are outlined next.

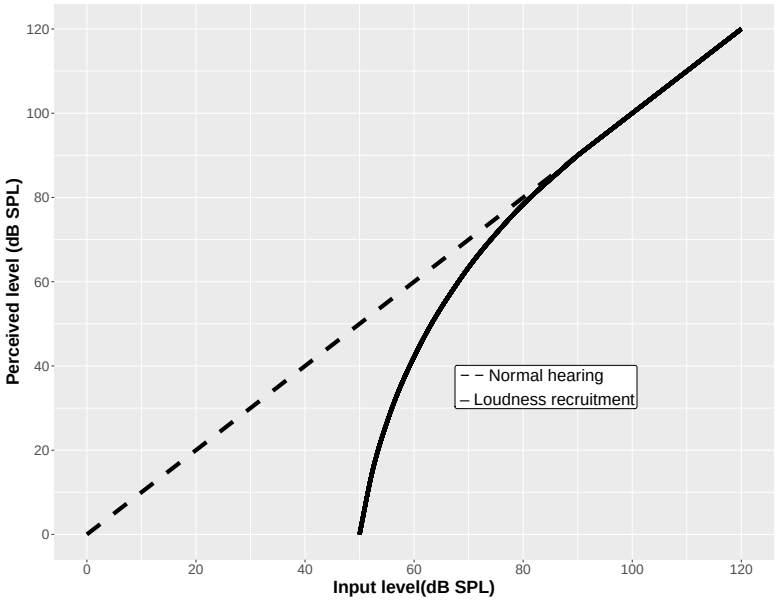


Figure 4: Illustration of the effect of loudness recruitment, relative to the expected loudness growth function for a normal-hearing listener.

1.5.2 Hearing aids

Typically, hearing aids provide amplification across a range from 125 Hz to 8 kHz, although this may vary to a small extent depending on the prescriptive procedure used (Keidser, Dillon, Flax, Ching, & Brewer, 2011). Therefore, frequencies falling outside of this range remain very difficult to hear for those with hearing loss – particularly presbycusis (age-related hearing loss), which is characterised by increased loss for higher frequency sounds. This is much more problematic for the perception of music, relative to speech, since the frequency spectrum associated with music is typically both wider and more variable (Chasin, 2003a). For example, very high-frequency information (>8 kHz) may make an important contribution to certain aspects of music listening (e.g. the appraisal of timbre, or overall sound quality), whilst being relatively uninformative for speech perception. That is, although timbre plays a vital role in the recognition of talker identity/ gender, this is usually of

secondary importance, compared to the actual linguistic content of the speech.

Analogously to the use of WDRC to accommodate a reduced dynamic range, frequency compression (FC) is often used in order to 'squeeze' the broad spectrum of sounds picked up by the HA into the frequency range for which the HA provides amplification (Madsen & Moore, 2014). A side-effect of FC is that it introduces inharmonicity (i.e. upper harmonics may be perceived to be 'out of tune') (Madsen & Moore, 2014). However, the net effect of FC upon music perception may still be positive, most likely because the benefit of additional high-frequency information outweighs any detriment caused by inharmonicity (Uys, Pottas, Vinck, & Van Dijk, 2012).

Additionally, for HA users, squeezing the normal dynamic range of speech into the reduced dynamic range that a person with hearing loss is sensitive to (i.e. WDRC), may result in complaints of excessive loudness, distortion or deterioration of overall sound quality (Hornsby, 2012). In fact, Studebaker, Sherbecoe, McDaniel, and Gwaltney (1999) demonstrated that speech intelligibility is reduced when stimuli are presented at the intensity levels commonly experienced in aided listening, even when the SNR remains constant. There may also be adverse effects of WDRC for the perception of music, particularly because music typically has a much greater dynamic range to begin with (Chasin, 2003a). Indeed, in a study by Madsen, Stone, McKinney, Fitz, and Moore (2015), individuals reported reduced clarity when identifying different musical instruments, when stimuli sounds were processed with WDRC, as opposed to linearly amplified. Particularly in the case of fast-acting WDRC, there is also the concern that overall perceived sound quality may suffer since, when an HA user listens to recorded music, it is essentially being compressed twofold: once as per industry-standard compression limiting for popular music, and then separately

by the HA (Croghan, Arehart, & Kates, 2014).

Lastly, with any digital signal processing, it is important to note that delay introduced by the processing itself can create an audible lag between the occurrence of a sound and the perception thereof, by an HA user. For example, even a delay of 20 ms may be sufficient to adversely affect the intelligibility of speech (Whitmer, Brennan-Jones, & Akeroyd, 2011).

To summarise, the net effects introduced by the HA are relatively subtle, and unlikely to have a major impact upon one's communicative ability (e.g. by impairing the perception of speech content). However, these effects can still have a deleterious effect upon the overall experience of HA users. In particular, they are likely to be more noticeable in the case of music perception, where the perception of attributes like timbre or overall sound quality are of relatively greater importance. Because the aesthetic experience is usually of central importance when listening to music, any perceptual disruptions may be much less tolerable for the listener, if they are considered to impair this experience. In other words, if music does not sound *good* to the listener, then there is rarely any secondary motivation to continue listening.

1.5.3 Cochlear implants

Transfer of the auditory signal through a CI typically has a much greater negative impact on that signal, by comparison to the effects of the HA, discussed above. Like the HA, the CI may also introduce undesirable artefacts via on-board digital signal processing (e.g. WDRC). However, various core limitations of the CI itself have a much more detrimental influence upon the auditory signal.

Most notably, spectral resolution is degraded, albeit to quite varying degrees. Though

some sensory hair cells usually remain after implantation, the procedure essentially replaces many thousands of inner cochlea hair cells with a maximum of twenty-two electrodes (Landsberger, Padilla, & Srinivasan, 2012). Since, in CI listening, pitch height is perceived primarily according to the location along the cochlea that spiral ganglion (auditory nerve) cells fire (determined by which electrodes are stimulated), the limited number of electrodes rather drastically limits spectral resolution. Put simply, the audible frequency spectrum must now be represented internally not by hair cells, but by electrodes, of which there are approximately one thousand times fewer – it is not surprising that this results in a much coarser-grained percept of pitch.

Spectral degradation is exacerbated further by the spread of current away from target electrodes (Kral, Hartmann, Mortazavi, & Klinke, 1998). In other words, the different physical channels, though discriminable, are not fully independent – each channel is subject to interaction with nearby channels stimulated simultaneously (McDermott & McKay, 1994). Therefore, spectral resolution is limited in CI users, not only by the limited spectral content transmitted by the device, but also by the spectral content received by the listener, which is affected by current spread and also by the health and quantity of residual spiral ganglion cells (Fu & Nogaki, 2005), which varies with age and both duration and cause of deafness (Nadol, Young, & Glynn, 1989). In summary, the use of a 22-electrode-array for encoding frequency resolution leads to drastically impaired frequency resolution even in a best-case scenario – however, inter-individual variance determines the extent to which frequency resolution is degraded even further.

As well as disrupting place coding of sound frequency, the CI further distorts the perceived signal by disrupting temporal coding. Since the stimulation rate (and thereby

the firing rate of the auditory nerve) is essentially unrelated to the frequency of the perceived signal, CI users miss out on another potentially rich source of information about the incoming sound. This temporal encoding of frequency is believed to provide more fine grained information about very precise changes in frequency, and is therefore important for the perception of both music, and speech in noise (Moon & Hong, 2014).

In addition to the mechanism for electrical stimulation, the nature of this stimulation itself also has adverse consequences for auditory perception. In CI users, implanted electrodes interface with and directly stimulate spiral ganglion cells, as opposed to these being stimulated via hair cells (Clark, 2003). Logistically, this has consequences for the dynamic resolution of electrical hearing. While the varying thresholds of individual hair cells lead to smooth, gradated responses in corresponding spiral ganglion cells in response to acoustic stimulation, electrical stimulation affects a large population of spiral ganglion cells with much less precision, leading to much more deterministic, or 'all or nothing' firing patterns (Javel & Shepherd, 2000). Accordingly, with electrical relative to acoustic stimulation, the spike rate of spiral ganglion cells is influenced much more drastically by small changes in input amplitude, leading to a steeper loudness growth function and reduced dynamic range (Javel & Shepherd, 2000).

Lastly, temporal resolution is limited for CI users as a consequence of the rate of stimulation. Commonly-used stimulation rates are more than sufficient for the perception of timing differences such as those present in speech prosody or musical rhythm, but perform much worse for more subtle distinctions, i.e. differences on the order of milliseconds. (Duarte, Gresele, & Pinheiro, 2016). Unfortunately, this cannot be circumvented by linearly increasing the stimulation rate, since this can have

unwanted effects on an individual's perception of loudness – more overall stimulation tends to be interpreted as corresponding to a louder signal (ASHA, 2003). Of course, the stimulation rate can influence pitch perception too, because of temporal coding (Reiss, Turner, Erenberg, & Gantz, 2007).

In summary, the impact of the CI on the auditory signal is much more noticeable, compared to the effect of the HA. Chiefly, frequency selectivity and sensitivity to fine structure is reduced enormously, and users also hear over a reduced dynamic range. By contrast, temporal resolution is relatively well-preserved (though still impaired relative to NH listeners). Since the sound sensation attainable via CIs, compared to HAs, is further away from that of normal-hearing individuals, there has been a greater need to prioritise amplification of the most pertinent auditory inputs, and therefore to optimise the device for speech signals. Accordingly, though the net effect of the CI is typically much more severe than the HA, users are nonetheless able to perceive speech content relatively accurately. However, the disadvantages for music perception, and for non-linguistic aspects of speech perception (e.g. emotional prosody), are much more pronounced, and less straightforward to overcome. This is because both non-linguistic speech components and music depend on the listener's awareness of relatively subtle fluctuations in the acoustic signal, and are therefore less robust when distorted by the CI.

In the next section, the complications arising from the use of HAs and CIs are contextualised via discussion of the ultimate 'goals' of these prostheses, and how effectively they are fulfilled.

1.6 What are the goals of hearing loss treatments?

The ideal outcome of any hearing loss treatment is the complete restoration of the entire spectrum of sonic experience, such that the patient is in every sense – functionally speaking – a normal-hearing listener. Unfortunately, as long as this ambitious goal remains unrealisable, treatments for hearing loss must make decisions about which types of sounds are most important to be restored. Therefore, in relation to the goals of hearing loss treatments, it is useful to first consider which 'classes' of sounds exist, and what the functional relevance of each is to the listener.

The spectrum of sonic experience, at least in everyday life, may be broadly confined to three functional categories: speech (and all associated vocalisations), music, and environmental sounds (Zhang and Kuo, 1999; Zhang and Kuo, 2001). This distinction is based upon the typical utility or purpose of these types of sounds, rather than how they are processed by the human auditory system. In fact, evidence suggests that at least speech and music, and possibly environmental sounds too, share substantial cognitive mechanisms and make use of common macro-anatomical neural structures (Handel, 1993; Price, Thierry, and Griffiths, 2005). Peretz, Vuvan, Lagrois, and Armony (2015) considers that neural circuitry developed for language perception may be 'recycled' for music, or vice versa, but also cautions that overlapping neural activation does not offer conclusive proof of this, and does not preclude neural separability. Indeed, this sharing of resources for different cognitive functions, or 'many-to-one' mapping may be a brain-wide organisational characteristic, as opposed to a property specific to speech and music (Anderson, 2010). Physiological implementations notwithstanding, it has been argued that music, speech and environmental sounds rely upon a shared set of cognitive mechanisms for communication (including evaluative conditioning, episodic memory, visual imagery and expectancy)

(Truax, 2016), and thereby constitute a continuum of sonic experience (Truax, 2001). Nonetheless, in terms of the functional goals of hearing loss treatments, it is helpful to treat speech, music and environmental sounds as separable categories.

Importantly for hearing loss treatments, the three broadly-classified categories of sounds identified above are not equivalent in terms of their net communicative utility in everyday life. Ordinarily, speech is most relevant for interpersonal communication, followed by music, and then environmental sounds (though important for personal safety, and for helping individuals orientate themselves within their immediate environment, these do not usually facilitate interpersonal communication per se). Concomitantly, speech is almost always considered the most important 'target' for hearing loss treatments, and is rated as being the most important attribute of sound, in terms of impact on quality of life, by both hearing aid and cochlear implant users (Meister, Lausberg, Kiessling, Walger, and von Wedel, 2002; Stainsby, McDermott, McKay, and Clark, 1997). Logically therefore, it follows that the fitting of both HAs, CIs, and indeed any other hearing prosthesis, is almost always aimed at maximising speech understanding (McFarland, 2000). To a lesser extent, HAs and CIs have began to target music as an important auditory input, which also has a substantial effect upon hearing impaired individuals' quality of life (Petersen, Hansen, Sørensen, Ovesen, & Vuust, 2014). Lastly, environmental sounds, while important to restore, have typically received less consideration. It is usually considered sufficient that important environmental sounds are audible but not aversive, and do not disrupt the perception of speech or music (Cox & Alexander, 1995). The primary purposes of environmental sounds – to orient oneself, to signal danger etc. – do not normally necessitate the same degree of expressive nuance as is present in speech and music.

Accordingly, the following chapter – and indeed much of this thesis – will focus on

speech and music as specific targets for the treatment of hearing loss. The next section contains a brief overview of the functional commonalities and distinctions between speech and music in communication, and their (putatively) shared evolutionary trajectory, arguing that speech and music are essentially similar, and are both specialised for communication.

2 Hearing loss and the perception of speech and music

2.1 Focus on speech and music

With respect to these sonic categories, Adorno (1992) has described both spoken language and music as constituting a 'temporal sequence of articulated sounds which are more than just sounds' (p. 1); in other words, they are auditory means of communication, consisting of the structured arrangement of sounds which hold some significance for the listener. Although this points to an essential similarity betweens speech and music, which is not shared with environmental sounds, the designation of the latter as 'just sounds' is a rather vague and unhelpful notion – environmental sounds frequently disclose information about some object or event, often in interaction with listeners' stored knowledge. Perhaps the distinction between 'speech and music' as one category, and 'environmental sounds' as another is better framed in Gibsonian terms. Since most environmental sounds stem from non-human agents, they typically have just one primary 'affordance' – for example, the sound of an alarm might afford evacuation from a building (Greeno, 1994). By comparison, speech and musical sounds usually originate from a human agent, and therefore may have any

number of different affordances, depending on the intent of the talker or musician (and of course the perception thereof by the listener) (Greeno, 1994). Therefore, it is fair to consider that speech and music are well-suited for more complex, nuanced communication, whereas environmental sounds are not.

Indeed, commonalities observed between language and music led Brown (2000) to argue that the two are derived from a common communicative predecessor, whilst Mithen (2009) suggested that 'musical' variations in pitch, timbre and rhythm may have been used as a rudimentary, pre-linguistic method of communication. Similarly, Darwin (1871) postulated that speech prosody (i.e. expressive, non-linguistic features of speech such as rhythm and intonation) predates the establishment of formal language, based on the observation of frequency-based affective cues in monkey vocalisations. Indeed, whilst language might be better-suited for conveying referential information, a widely accepted commonality of speech prosody and music is their aptitude for the transmission of affective content (Coutinho & Dibben, 2013). Recently, research has suggested that music and language evoke an emotional response via shared cognitive mechanisms, based upon the observed comorbidity of emotional prosody detection deficits in congenital amusia (Thompson, Marin, & Stewart, 2012), musical syntactic processing deficits in Broca's aphasia (Patel, Iversen, Wassenaar, & Hagoort, 2008), and enhanced emotion perception in speech documented in expert musicians (Lima & Castro, 2011). Additionally, functional neuroimaging (fMRI) has implicated overlapping neural structures in the perception of emotion in each domain (Nair, Large, Steinberg, & Kelso, 2002). Therefore, it is likely that expression of emotion is a communicative 'goal' that is achieved in comparable ways via both speech and music.

In terms of evolutionary purpose, Brown (2000) suggests that speech and music later

became functionally separated, such that speech developed with literal, referential communication as its primary goal, while music may have developed specifically to facilitate interpersonal cooperation and intra-group emotional unity. Conducive to this, Brown (2000) highlighted several salient features of music, for example: compared to all other natural forms of vocal communication, music is distinguished by its prominent use of pitch-blending and temporal synchronisation, both of which encourage cooperation, and are conducive to group-level 'performance'. Therefore, although music and speech may share a common preverbal predecessor, they might also have become specialised according to different communicative needs. From this perspective, speech and music might be considered complementary mechanisms, fulfilling a partially-overlapping set of communicative and evolutionary goals, whereby music is better-suited to more abstract types of inter- and intra-group communication, and speech better-suited for more direct inter-individual communication. In the next section, the perception of speech and music by hearing-impaired listeners – both of which will be central concerns of this thesis – will be systematically evaluated.

2.2 How well do hearing-impaired listeners perceive speech?

2.2.1 Hearing aids

To answer the question of how well hearing-impaired listeners perceive speech, the primary tool available is known as speech audiometry. Broadly speaking, this consists of measurement of an individual's ability to recognise and understand excerpts of speech, as a means of evaluating their hearing. For at least seventy years, variants of speech-based audiometry have been incorporated within screening and fitting processes for HAs, beginning with simple aided vs. unaided word-recognition tests (Carhart, 1946). In 2005, Kirkwood found that 92% of audiologists reported using

some form of speech audiometry. In short, accurate perception of speech has long been both a primary goal, and an important method of assessment, for HAs.

The benefit for speech recognition accuracy with HAs, compared to without, is highly variable, based on the duration and extent of hearing loss, individuals' cognitive abilities, and the specifics of the HA used. When listening in a quiet setting, it is possible for HA users, like NH listeners, to achieve speech recognition performance of 90-100% (Metselaar et al., 2008). For each individual however, the overall improvement in speech recognition, aided vs. unaided, may vary from approximately 3% to 20%, with poorer unaided performance typically predicting greater improvement (Metselaar et al., 2008).

Despite this, many HA recipients report imperfect speech understanding (especially in noisy environments), and in fact cite this as a reason for rarely using their devices (Kochkin, 2000; Lupsakko, Kautiainen, and Sulkava, 2005). A more recent survey has suggested that end-user satisfaction is increasing slowly over time, although listening in noisy environments continues to pose a significant challenge (Kochkin, 2010).

However, as modern HA technology continually improves, it is conceivable that speech recognition with HAs could achieve parity with normal hearing, even in more challenging listening scenarios. Indeed, in two recent independent trials, Powers and Fröhlich (2014) reported that HA users with mild-moderate loss, equipped with a novel binaural beamforming (spatial filtering) technique, were able to outperform NH control subjects in a speech recognition in noise task, by 2.1-2.9 dB SNR. Similarly, the emerging use of deep neural networks (DNNs) may offer a solution to the so-called 'Cocktail party problem', by allowing HAs to effectively isolate relevant speech sounds from various kinds of competing signals and/or background noise (Wang, 2017). By having the HA 'learn' how to accomplish this feat, rather than

implementing a specific, predefined signal processing algorithm, it is able to perform well even when faced with background noises that have never been encountered before, thereby achieving a much more 'human-like' level of performance (Wang, 2017).

Lastly, consideration should be given to the non-linguistic elements present in speech, often collectively referred to as 'prosody'. Prosodic elements of speech may carry important information about an individual's tone of voice, or emotional state, but are characterised by relatively subtle variations in the acoustic signal, and therefore can present difficulty for HA users (Most & Aviner, 2009). Of consequence here, for example, is that HA amplification occurs over a limited range of frequencies (up to a maximum of approximately 100 Hz to 10 kHz), relative to the range that is audible by the human ear (Anders H. Jessen, 2014). Moreover, hearing loss is rarely uniform across the frequency spectrum – often individuals have relatively good residual hearing at lower frequencies and very little at higher frequencies, i.e. 'useable' hearing is restricted to a reduced bandwidth (Davis, 2004). For this reason, frequency compression is often used, so that a greater proportion of sound is amplified. However, since frequency information is not preserved verbatim, the process risks introducing distortion, particularly in the higher-frequency harmonics of a signal, which may be perceived as dissonance (Uys et al., 2012).

Therefore, even if an HA user is able to recognise speech with one hundred percent accuracy, they might still miss out on or misinterpret essential non-linguistic information, which could alter the meaning of a given utterance (e.g. in the case of irony). For example, research has documented impairment for HA users, relative to NH participants, in tasks involving the detection of a talker's emotional state (Most & Aviner, 2009), or whether their intent is sarcastic or sincere (Stiles, 2013).

In summary, HA users are typically impaired to some extent in speech recognition, especially when listening in a noisy environment. With this said, HA technology is constantly improving, meaning that the best performance attainable is becoming closer to that of a normal-hearing listener, even when listening to speech in noise. However, while the content of speech is usually its most important attribute, there are also relevant, non-linguistic aspects of speech that have received relatively little attention, and might not be restored as adequately by the HA (Schmidt, Herzog, Scharenborg, & Janse, 2016). Therefore, even those HA users scoring within the top percentiles for speech recognition with/ without noise, it should not be assumed that perception of speech has been restored to a 'normal-hearing' level.

2.2.2 Cochlear implants

As with HAs, in recent years there has been tremendous progress made to improve the perception of speech by CI users. In 1977, Bilger, Black, and Hopkinson, investigating the efficacy of the first generation of single-channel CIs to be fitted in the United States, reported that these prostheses were insufficient to facilitate speech understanding, but that they did appear to convey some supplementary auditory information that led to significantly improved lip-reading scores. By 1995, a National Institute of Health (NIH) consensus statement on CIs was released, which proclaimed that most postlingually-deafened recipients of modern CIs would be expected to score 80% or above when tested for understanding of high-context sentences, presented in quiet and with no accompanying visual information (Wilson, 2004).In fact, speech perception with the CI is now sufficient for many users to understand and conduct telephone conversation (Cray et al., 2004; Helms et al., 2001) – that is, to recognise speech using only auditory cues, without lip-reading.

Thirteen years after the 1995 NIH consensus, Gifford et al. (2008) presented data showing that over 25% of users achieved perfect scores in standard sentence batteries to measure speech recognition in quiet. The authors further stated that more difficult speech testing materials must be developed, due to the prevalence of patients scoring 90-100% rendering it difficult to meaningfully track individuals' progress with the CI. This report was considered a major milestone in the development of CI technology to facilitate speech perception (Wilson & Dorman, 2008).

With this being said, just as with HAs, performance in speech recognition with the CI can be highly heterogeneous, varying according to myriad factors, including: age at implantation (Blamey et al., 1996; Geier, Barker, Fisher, and Opie, 1999; Shipp and Nedzelski, 1995), duration of deafness prior to implantation (Blamey et al., 1996; Geier et al., 1999; Gantz, Woodworth, Knutson, Abbas, and Tyler, 1993; Rubinstein, Parkinson, Tyler, and Gantz, 1999), degree of residual hearing (Gantz et al., 1993; Rubinstein et al., 1999), duration of implant use (Blamey et al., 1996), and cognitive abilities (Gantz et al., 1993).

However, speech perception is much more difficult for CI users when it takes place in a noisy environment, for example in a restaurant, or a busy street. Several studies have documented drastically reduced speech reception thresholds for CI users listening in noise, as opposed to quiet. For example, Tobey, Shin, Prashant, and Geers (2011) found that presenting sentences with accompanying multi-talker babble (i.e. competing, usually unintelligible, speech from multiple voices) significantly reduced the speech intelligibility scores of adolescent CI users.

CI users also tend to be somewhat impaired in the perception of prosodic elements of speech: an impairment that appears to be relatively independent of performance in speech recognition (Chin, Bergeson, & Phan, 2012). Nakata, Trehub, and Kanda

(2012) found that children who used CIs performed significantly poorer than NH controls in both perception and production tasks related to speech prosody. Similarly, in a study that included CI and HA users, Kalathottukaren, Purdy, and Ballard (2017) found that both groups achieved significantly lower scores, compared to their NH peers, on two standardised batteries designed to assess perception of speech prosody (Nowicki and Duke, 1994; Peppé and McCann, 2003). In another study involving both HA and CI users, Most and Peled (2007) observed that the latter group performed significantly worse on tasks involving the identification of stress and intonation patterns in spoken sentences. Lastly, at least two studies have shown that CI users have some level for difficulty in classifying utterances as statements or questions. Peng, Tomblin, and Turner (2008) found that children with CIs scored an average of 70% correct in a question/ statement discrimination paradigm, compared to 97% for NH controls. Likewise, in a similar experiment with adults, Green, Faulkner, Rosen, and Macherey (2005) reported an average of 69% correct for CI users.

To summarise, in recent years, enormous progress has been made towards improving the quality of speech perception of CI users. Because of this, it is now possible for some users to perform relatively well in speech recognition tests, even without lip-reading, and even in the presence of background noise. Unfortunately, in the majority of cases, speech perception performance appears to be significantly worse with the CI than with the HA. This appears to be compounded further by a greater difficulty in perceiving paralinguistic auditory attributes present in speech, facilitating recognition of the talker's tone of voice, or emotional state, for example. Just as with HAs, emerging technological advances paint an optimistic picture of the CI landscape for the near future, with automatic auditory scene classification being implemented in

order to improve the perception of speech in noise (Mauger, Warren, Knight, Goorevich, & Nel, 2014), and researchers exploring the feasibility of light-based stimulation techniques, intended to circumvent the problems of limited frequency resolution and electrical current spread, which are inherent in current CI technology (Johannsmeier et al., 2017; Parveen, 2017).

2.3 How well do hearing-impaired listeners perceive music?

2.3.1 Hearing aids

Until very recently, there has been much less interest in restoring music perception, compared to speech, for people with hearing impairment – music tends to be more heterogeneous as a type of signal, and is less immediately relevant to the end-goal of restoring communication. Consequently, it is harder to reliably ascertain exactly how satisfactorily individuals with hearing impairment perceive music, since there has been less motivation to develop objective measurement paradigms (this specific issue is explored in much greater detail in Chapter 8: 'A novel, objective assessment for music perception in aided listening'). Perhaps unsurprisingly therefore, compared to speech, many HA users are much less satisfied with their perception of music, often choosing to either remove their HAs for music listening, or simply avoid music altogether (Fitz and McKinney, 2010; Madsen and Moore, 2014).

In addition to preference- or self-report-based measures of HA users' overall experience, it is also useful to consider some of the ways in which the HA may affect the perception of specific musical attributes. Because of technological constraints imposed by the hearing aid, the acoustic signal arriving at the user's ear may differ somewhat, compared to a normal-hearing individual. Perhaps the most well-documented source

of this variation is the difference between the dynamic range usable by the hearing aid, compared to the range of intensities that occur in the world. In particular, this can be detrimental for the perception of music, since it typically comprises a much greater dynamic range (Greasley, 2016) – approximately 80 dB on average, compared to 40 dB for speech sounds (Eargle, 2012). For this reason, for many hearing aid fittings, the peak input limiting level – typically optimised for speech – may be set too low for music, and therefore distortion may be introduced, especially during the loudest portions of a musical piece (Chasin, 2003b). Accordingly, (Madsen & Moore, 2014) have suggested that increased input and output dynamic range might improve listeners' enjoyment of music. It should be noted, however, that the dynamic range for recorded music (which will typically have been processed with WDRC) is not necessarily the same as live music, and may even be smaller than that of speech (Kirchberger & Russo, 2016). The dynamic range of music is highly variable, depending on the environment and/or mode of listening, thereby creating an inherent difficulty in achieving an effective 'universal' HA configuration for music. Indeed, most users report that their HAs are more effective when listening to recorded, rather than live, music (Madsen & Moore, 2014).

In the past, music perception with HAs may also have been negatively impacted by digital signal processing intended to improve the audibility of speech (Fitz & McKinney, 2010). Concretely, when optimised for speech perception, processing stages such as noise reduction, peak input limiting, and wide dynamic range compression can have unwanted, disruptive effects when listening to music (Chasin & Russo, 2004). Because of this, modern HAs typically include a 'music program' – a set of digital signal processing strategies specifically intended to improve perception of music. Unfortunately, there appears to be little consensus as to how exactly this should be

implemented, leading to widely varying results. In a simulation-based study, comparing HA processing by three different manufacturers, Fitz and McKinney (2010) showed that each of the HAs led to a noticeably different estimated perceived loudness contour (up to a maximum of 20 sones difference between HAs) for a given musical stimulus, and none well approximated the unaided loudness level for an NH listener (Figure 5). Therefore, even when special consideration is made for music as an input to the HA, 'normal' perception of loudness in music is not restored.

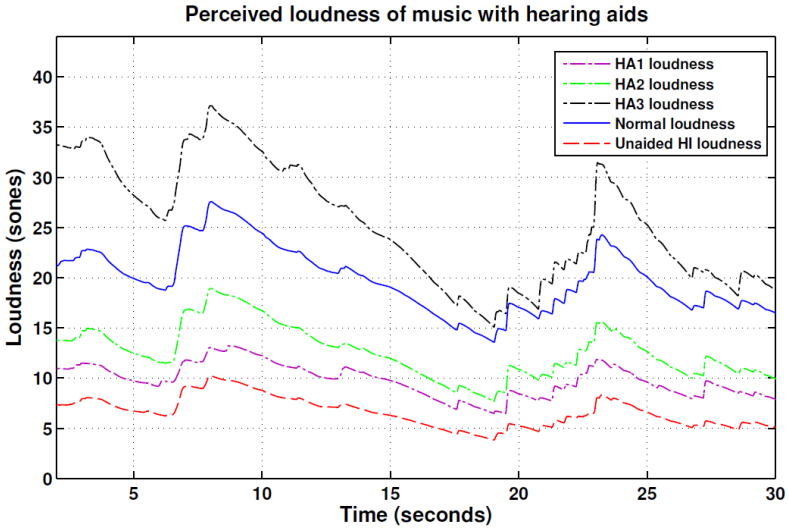


Figure 5: Estimated perceived loudness of a musical stimulus, as processed by three different HAs, by comparison to unaided listening by an NH and an HI person. Reproduced from Fitz and McKinney (2010).

As with the perception of speech prosody, the limited range of frequencies over which HA amplification occurs also presents a problem for music perception. In this case, the use of frequency compression (FC) means that specific frequency information is not always optimally preserved – a detriment that may be perceived as a musical sound's constituent tones being essentially 'out of tune' with each other, and therefore registering as dissonant or otherwise unpleasant (Uys et al., 2012). As alluded to previously however, FC appears to have a net positive effect on music enjoyment,

most likely because the minor frequency distortion is preferable to having altogether missing information at higher frequencies. In addition to FC, frequency transposition may also be used – in which all frequencies above a certain threshold are transposed downwards by a uniform amount – in order to effectively reduce the frequency range over which amplification is required (Kuk, Keenan, Korhonen, & Lau, 2009). However, Chasin (2014, para. 43) cautions that this approach is 'not something that is acceptable to musicians'.

Due to mis-tuned frequency information, and also because of deficits in processing temporal fine structure (Sęk & Moore, 2012), HA users are typically somewhat impaired in the perception of timbre, relative to NH listeners, as indexed by lower mean accuracy in musical instrument identification tasks (Fitz, Burk, & McKinney, 2009). However, the two groups appear to utilise the same acoustic cues in order to make these judgements, implying that the impairment is not severe enough to necessitate a fundamentally different approach to the task. Likewise, (Fitz & McKinney, 2010) showed that the perception of brightness (i.e. the relative prevalence of high-frequency spectral energy) is disrupted non-linearly when listening with an HA – such that perceived brightness throughout a musical piece fluctuates between close to 'normal' at some points, and highly discrepant in others – likely causing the piece to sound somewhat unnatural.

In addition to all of the above, there may also be some more 'mechanical' limitations of the HA when listening to music. For example, in a survey by Madsen and Moore (2014) of over 500 hearing aid users, over one third reported that acoustic feedback from the hearing aid (i.e. a characteristic, high-frequency 'whistling' sound) disrupted their perception of music. Whilst this particular effect is not specific to music listening, it is more difficult to overcome in this case because the adaptive feedback

cancellation implemented by modern HAs runs the risk of mistakenly attenuating high-pitched musical tones (Moore, 2016).

To summarise, music perception is usually found to be less satisfactory for HA users, compared to speech perception. This is due to myriad factors: the historical priority given to restoring speech perception; speech-centric signal processing having inadvertent, adverse affects for music; music typically being more heterogeneous as category of acoustic signals, and music being more 'volatile' to disturbance by relatively small deviations in acoustic parameters. Functionally, current limitations with the HA manifest as various subtle deficits in music perception. In practical terms, a survey by Leek, Molis, Kubli, and Tufts (2008) reported that approximately 17% of elderly HA users experienced difficulty in recognising musical melodies, while Looi, McDermott, McKay, and Hickson (2008) found that HA users were able to achieve only 69% accuracy in the identification of isolated musical instruments, concluding that they were 'largely unable to perceive music accurately' (Looi et al., 2008, p. 433). However, this situation is beginning to improve, with HA manufacturers launching dedicated signal processing prescriptions for music incorporating, for example, higher sample rates and 'twin' compressors to allowing different parts of an auditory signal to be processed simultaneously, but differently (Bricker, 2016).

2.3.2 Cochlear implants

By contrast to the great successes seen in speech perception with the CI, most users do not hear music well (Roy, Jiradejvong, Carver, & Limb, 2012b). In fact, Limb (2016) declared music the 'greatest challenge' for the CI, thereby holding the key to achieving what could be considered 'perfect hearing' – that is, deficits in music perception may reveal limitations of the CI that conventional speech testing does not.

Further, being able to hear music well with a CI should in principle mean that any possible sound is perceived at least adequately (Limb, 2016). It is worth noting that the goals associated with speech perception via the CI are somewhat different to those of music perception. In the former case, the ultimate aim is usually intelligibility – to facilitate understanding of some specific message being communicated. By contrast, when restoring music perception, the objective is usually to provide an aesthetic experience – something that 'sounds good' to the listener (Limb, 2011). In common with speech, research has shown that successful restoration of music perception by the CI is associated with increased quality of life (Lassaletta et al., 2007; Petersen et al., 2014). However, in common with the HA, is the unfortunate (but of course understandable) historical lack of research concerned with assessing and improving music perception for CI users. As with the HA, it is useful therefore to begin by discussing the limitations of the CI, and the consequences that these have for the perception of various musical parameters.

As a result of both cochlear hearing loss and the limitations of the implant, CI users typically display significantly reduced accuracy in pitch discrimination, relative to the normally-hearing (NH) population (Moore and Carlyon, 2005; Pick, Evans, and Wilson, 1977; Townshend, Cotter, Van Compernolle, and White, 1987). In music perception specifically, poorer intervallic discrimination is likely to cause problems with identifying higher-level emotional cues, such as mode – although an event-related potential (ERP) study has demonstrated some remnant sensitivity to the functionally regularity of chords (Koelsch, Wittfoth, Wolf, Müller, & Hahne, 2004). Particularly problematic are complex tones, due to the difficulties of resolving higher-order harmonics with such limited spectral resolution (Galvin, Fu, & Nogaki, 2007), leading to impairment in perception of timbre and fine structure (Gfeller, Knutson, Woodworth,

Witt, and DeBus, 1998; Heng, Cantarero, Elhilali, and Limb, 2011). Converging evidence from electroencephalography (EEG) has documented the absence of a typical mismatch negativity (MMN) response in adolescent CI users, for pitch deviations of four semitones (Petersen, Weed, et al., 2015), providing a neurophysiological index of the aforementioned perceptual deficits. This study utilised Vuust's musical multi-feature paradigm, in which listeners are presented with deviations (i.e. unexpected events) in different musical features, within a complex musical context (Vuust et al., 2011) – interestingly, CI users did demonstrate significant MMN responses to deviations in intensity, rhythm and timbre, denoting some preservation of the neural mechanisms underpinning musical feature processing. Similarly, cortical activity, indexed using positron emission tomography (PET), appears to be much greater when CI users are presented with rhythmic stimuli compared to melodic stimuli, and much less in the latter case, compared to normal-hearing controls (Limb, Molloy, Jiradejvong, & Braun, 2010).

Beyond simple pitch discriminability, Cousineau, Demany, Meyer, and Pressnitzer (2010) showed that, relative to NH listeners, CI users have increased difficulty in processing pitch sequences, even when their constituent pitch differences are above individuals' discrimination thresholds. The researchers speculated that the resolution of individual harmonics, which are not well-conveyed by the CI, may be important for accurate performance on this task. Consistent with a previous study, NH listeners showed greater accuracy in same-different discrimination for pitch sequences, compared with similar sequences that instead varied in intensity (Cousineau, Demany, & Pressnitzer, 2009). Interestingly, overall accuracy for intensity sequence processing did not differ significantly across groups of NH listeners, CI users, and NH users listening with a noise-band vocoder CI simulation. This further corrob-

orates the observation that intensity information is relatively well-preserved by the CI – sufficiently so to permit normal sequence processing in this feature dimension.

Drennan et al. (2015) found that, although CI users showed a mean pitch discrimination threshold of 2.5 semitones, their performance on a melody recognition task was much worse, with participants scoring on average less than 20% correct. Therefore, simple pitch discrimination tasks may overestimate an individual's ability to perceive musical melody accurately – the results suggest that the CI is effectively 'mistuned', i.e. that relative frequency information, sufficient for pitch discrimination, is preserved, whilst absolute frequency information is not. With this said, depending on the nature of the melodic recognition task, access to rhythmic information may bolster performance. Unsurprisingly however, relative to NH listeners, CI users show poorer open-set recognition of previously familiar musical excerpts (Gfeller et al., 2005), although this may be ameliorated to some extent by focussed musical training (Looi, Gfeller, & Driscoll, 2012).

Using the Multiple Stimulus with Hidden Reference and Anchor (CI-MUSHRA) paradigm – an objective methodology designed to assess the perceived sound quality of musical stimuli – Roy et al. (2012b) showed that, relative to NH listeners, CI users were impaired in their ability to distinguish between unaltered and altered (low-pass filtered) versions of a stimulus. The researchers attributed this deficit to the characteristic loss of high frequency information caused by the CI. Using the same paradigm, Roy, Jiradejvong, Carver, and Limb (2012a) found that CI users similarly struggled to perceive musical sound quality deterioration resultant the loss of bass frequencies (i.e. high-pass filtering) – stimuli with frequencies less than 400 Hz filtered were not rated as significantly lower quality. Therefore, musical sound quality impairment caused by the CI is a function of the loss of both high- and

low-frequency spectral content.

Perhaps not surprisingly, CI users typically report reduced enjoyment associated with music listening after implantation than prior to their hearing loss (Petersen et al., 2014; Gfeller et al., 2000), and a large majority of users report post-implantation declines in their music-listening habits (Gfeller et al., 2000; Leal et al., 2003).

However, there are some CI users who are able to enjoy listening to music, particularly after repeated exposure (Gfeller and Lansing, 1991; Gfeller et al., 2005), and several studies have documented a positive correlation between individuals' musical backgrounds and performance in perception of musical features and/ or music recognition tasks (Leal et al., 2003). Similarly, Witt, Murray, and Tyler (2000) reported a positive correlation between CI users' time spent listening to music, and their subjective enjoyment of it – of course, this should be interpreted with caution, since it is unclear whether repeated listening increased enjoyment, or conversely if those that already enjoyed music listening were simply motivated to listen more often.

Further, focussed musical training has been demonstrated to improve specific, relevant perceptual abilities, for example timbre discrimination (Gfeller et al., 2002), melodic discrimination (Gfeller et al., 2000), melodic contour identification (Galvin et al., 2007) and even rhythm discrimination (Petersen, Mortensen, Hansen, & Vuust, 2012). Importantly, these training effects imply that, even though the auditory information available to CI users is degraded, listeners do not typically extract all of this information, or at least do not do so in a way that is optimal (Moore & Shannon, 2009). Therefore, along with the continued development of CI technology itself, there is considerable scope for rehabilitation to maximise performance by improving the ability of the brain to work with the impoverished input provided (Fu & Galvin, 2008).

In summary, much like HA users, CI users tend to be less satisfied with their perception of music, by comparison to speech. Current limitations of the CI mean that the accurate perception of music is inherently very difficult, and users typically struggle to a greater extent than those with HAs. In fact, the experience of music might be considered fundamentally different when listening via a CI, due to the severity of impairment in the perception of music's fundamental acoustic attributes (in particular, pitch). Because of this, users must typically 'relearn' how to listen to music, which unfortunately is off-putting for many, leading to a reduction in music listening for this population. Accordingly, in addition to musical training-based rehabilitation, recent research in this area has focussed on simplification of the musical input to the CI. As an example of this, Nemer et al. (2017) has demonstrated that musical enjoyment reported by CI users can be increased via harmonic reduction of musical tones. Similarly, when given the opportunity to manipulate existing pieces of music – simulating the 'mixing' stage of music production – CI users tended to prefer fewer instruments, clear vocals and prominent bass and drum parts (Buyens, van Dijk, Moonen, & Wouters, 2014). These preferences appear to be especially strong when the original music is more compositionally complex (Buyens, van Dijk, Wouters, & Moonen, 2015). By retaining the vocals, bass and drums, CI users essentially simplify the music, whilst maintaining the most strongly rhythmic aspects, along with the lyrical content.

The next chapter contains a continuation of the discussion of speech and music perception in hearing-impaired listeners. Specifically auditory emotional expression is considered as an important commonality shared by speech and music, and the extent to which the perception thereof is impaired for both HA and CI users is examined. In the next section, emotion perception is defined more precisely, and

behavioural evidence relating to the perception of emotions by hearing-impaired subjects is evaluated.

3 Auditory expression and perception of emotion

3.1 How can we best describe emotions?

Emotional expression is a major component of spoken language – a 'necessary ingredient for natural two way human-to-human communication' (Khanna and Sasikumar, 2011, p. 219), the perception of which is demonstrably impaired for hearing-impaired listeners. Likewise, emotional expression is ubiquitous in music (Lindström, Juslin, Bresin, & Williamon, 2003), and our perception thereof is often considered an important motivation for listening to music (Ahtisaari & Karanam, 2015). However, as with speech, the perception of emotion in music is typically impaired to some extent in individuals with hearing loss. In summary, emotion perception is a major communicative component present in both speech and music, that is almost always affected to some degree by hearing impairment. Accordingly, a central concern of this thesis is the investigation of emotion perception ability in hearing-impaired listeners, and how this might be improved.

Prior to any discussion about the perception of emotions, it is important to establish what exactly is referred to by the term 'emotion'. Unfortunately, within the relevant cognitive, neurobiological and philosophical disciplines, a single, satisfactory, all-encompassing definition of emotion has proved both necessary and elusive (Cabanac, 2002; Griffiths, 1997). In fact, it has been argued that emotions may inherently be indescribable by one unified theory, owing to both the heterogeneity of different emotions, and the tendency for emotional states to vary enormously from one instance

to another (Griffiths, 1997; Griffiths, 2004). Even if one assumes that the entirety of emotional experience is explainable by a single theory, there is further confusion as to the level at which emotion as a concept should be apprehended, since it is inherently multifaceted, containing ,at least: cognitive, behavioural, functional, physiological, phenomenological and neurochemical components. Despite, or perhaps because of, these ontological concerns, many psychologists have attempted to provide definitions for emotion, with over ninety allegedly proposed during the 20th century (Plutchik, 2001). A sensible approach, therefore, is to examine some of the various definitions given and attempt to establish congruities. Leaving aside philosophical reservations about the validity of a single definition of emotion, it would be useful to have such a pragmatic definition for the sake of parsimony and ease of reference.

In the simplest terms, emotion can be understood as a phasic, physiological phenomenon, usually with some kind of evolutionary relevance, that occurs as a reaction to the appraisal of a stimulus (Damasio, 1999; Plutchik, 1980). Damasio (1999) argues that while the emotion itself is essentially constituted by coordinated changes in various physiological systems, the sense of 'feeling' an emotion depends on a – not necessarily conscious – awareness and recognition of these changes. Emotions also include some type of behavioural component, whether explicitly manifest or not (Plutchik, 2001), and, though perhaps not considered 'rational' in themselves, are demonstrably useful in informing practical, rational decision-making (Damasio, 1994). More specifically, Damasio (1994) rejected the idea of a dualistic dichotomy of emotion and reason, and instead argued that one's emotional response can provide invaluable additional information when deciding how best to respond to a given stimulus. An important point highlighted here is that emotions are *for* something, serving a clear, functional purpose as a component of a wider system designed to

inform human-environment interaction. Similarly, Eickers, Loaiza, and Prinz (2017) has emphasised the practical nature of emotions, highlighting that they are both context-specific and goal-directed, and therefore quite highly variable experientially.

In terms of a phenomenological description, Cabanac (2002) conceptualised emotion as a kind of mental experience that is relatively intense and has high hedonic content (i.e. definably signifies pleasure or displeasure). However, several models of emotion argue that, rather than being a simple prerequisite for emotion, 'intensity' is instead a dimension along which different emotional states may vary (Graham, Priddy, and Graham, 2014; Russell, 1980). For example, this dimension allows one to differentiate between melancholy and despair, as varying extents of sadness. Summarising much of the above, Hockenbury and Hockenbury (2007, p. 326) defined emotion as 'a complex psychological state that involves subjective experience, a physiological response, and a behavioral or expressive response'. Similarly, albeit with the addition of two additional components, Scherer (2005) has characterised emotion as a synchronised reaction involving: cognitive appraisal of a stimulus, physiological response, action tendencies, expressive communication of the emotional state, and the subjective experience of the state itself. The most important contribution of this definition is that it widens the scope to include the appraisal of a stimulus, thereby framing emotion as an active, constructive process, as opposed to a mere 'response'. In order to provide further clarity, Fox (2008) sought to distinguish emotion in terms of what it is *not*. Specifically, Fox defined emotion in relation to several related and often-confused concepts: feelings, moods and affect. According to Fox (1996), 'feelings' describe an individual's subjective experience of an emotional state – i.e. their recognition and interpretation of physiological changes elicited by some stimulus (Damasio, 1999). Moods are distinguished from emotions by not necessarily being

tied to a discernible stimulus, and therefore being more diffuse and longer-lasting (Fox, 2008; Hume, 2012). Finally, 'affect' is advocated as an inclusive, umbrella term for emotions, feelings and moods (George, 1996).

Synthesising the commonalities present in the various aforementioned definitions, one may conclude at least the following. 'Emotion' refers to a multifaceted phenomenon that is context-specific, goal-directed and synchronised across distributed systems, involving minimally: appraisal of a stimulus (consciously or otherwise), physiological response to a stimulus, and some kind of behavioural or expressive-communicative response (whether actual, intended or implied). While this definition is doubtlessly imperfect, it shall suffice as an approximate clarification of the concept of emotion, for the purposes of this thesis.

3.1.1 Perceived vs. felt emotions

An important component of emotion understanding, relevant for this thesis, is the distinction drawn between 'perceived' and 'felt' emotions (Gabrielsson, 2009; Truong et al., 2009). Concretely, the former denotes an emotional state that is seen, heard, or otherwise experienced by the senses as arising from a stimulus (e.g. an excerpt of music or speech), while the latter denotes an emotional state that is actually *induced* in the listener, i.e. what one 'feels', following exposure to some stimulus. Of course, despite the inherently multifaceted, multimodal nature of emotion, only a subset of this information is available in the case of emotion perception. For example, when listening to emotional speech, the only component of emotion conveyed directly is the expressive-communicative response. Therefore, emotion perception essentially involves inferring a wider emotional state, as implied by one or more aspects of its manifestation.

Although perceived and felt emotion often co-occur – and indeed may constitute a continuum as opposed to a strict dichotomy – they are also numerous occasions where they might proceed independently (Gabrielsson, 2002). For example, Zentner, Meylan, and Scherer (2000) reported significant differences in the emotions that participants considered to be either expressed or felt whilst listening to their favourite music – specifically, negative emotions (sadness, anger, fear, etc.) were far more frequently rated as being expressed by the music than felt by the listener. Similarly, studies concerned with rating emotional speech corpora have argued for a distinction between perceived and felt emotion, finding that ratings of the latter tend to be more specific and also more difficult to predict via computational modelling (Busso and Narayanan, 2008; Truong et al., 2009). This is likely because the perceived emotion is, to a greater extent, a product of the acoustic properties of a given stimulus, whereas felt emotion involves a greater degree of interactivity with the listener. For example, it would be impossible to predict that an individual might feel sadness in response to a conventionally 'happy' piece of music, because of a prior association between this music and some former, traumatic life event. The contrast between emotions perceived and felt is further supported by the observed double dissociation of dysfunctions in cognitive empathy (i.e. perceiving another's emotional state) and emotional empathy (i.e. feeling another's emotional state) observed in autism and psychopathy, respectively (Blair, 2005). Findings such as this suggest that the experience of perceived and felt emotion depend, at least partially, upon separate neurocognitive mechanisms. Hereafter, this thesis shall focus exclusively on perceived emotions – that is, the emotions that listeners perceive to be communicated by auditory stimuli, irrespective of their own affective response. The reason for this is that perceived emotions are more immediately relevant for interpersonal communication, more likely to be impaired in individuals with hearing loss, and comparably easy

to investigate, since 'normal' emotion perception performance is relatively simply to establish.

Next, some consideration should be given to the different 'types' of emotional states, and how – or indeed, if – these should be categorised.

3.1.2 Taxonomy of emotions: discrete vs. dimensional approaches

With respect to the conceptualisation of different kinds of emotions, two dominant approaches have emerged: discrete and dimensional accounts. The former posits the existence of distinct, separable emotional states (i.e. states that are nameable and therefore may be readily labelled when experienced (Adolphs, 2002)). These states are considered to be distinct in terms of both their experiential aspect, and their origin. For example, a person receives some piece of good news, is asked about their emotional state at that moment, and may declare 'I am happy'. In particular, some proponents of the discrete approach have advocated a small number of 'basic' emotional states (Ekman & Friesen, 1971), biologically constituted and therefore pancultural, that have distinctive and innate adaptive functions that are conducive to survival (Tomkins, 1962). In other words, in response to a vast range of different environmental and/ or interpersonal scenarios, a set of qualitatively separate emotional states have arisen as adaptive 'mechanisms' that offer (or at least once offered) some type of immediate benefit relevant for survival. More recently, proponents of the discrete approach to emotion have conceded that emotions are context-dependent and highly variable across individuals, but emphasised that they are nonetheless, by nature, strategic, embodied responses to external stimuli, best describable by discrete emotion labels because of their functional coherence (Eickers et al., 2017).

This argument has received support from functional localisation studies in cognitive

neuroscience, which have demonstrated some level of specialisation in different brain areas for discrete emotions (Barrett, Gendron, & Huang, 2009). For example, the experience of fear tends to be particularly associated with increased blood-oxygen-level dependent (BOLD) activity centred around the amygdala (LeDoux, 2003), while similar research has demonstrated selective processing of disgust at the insula (Wright, He, Shapira, Goodman, & Liu, 2004). However, contradictory findings suggest that the specificity of particular neural structures for individual emotional states may have been overemphasised, and that discrete emotional states may arise from a shared, distributed network (Schienle et al., 2002).

Furthermore, discrete accounts of emotion, have been criticised for failing to adequately describe the variability and gradation characteristic of emotional expression (Barrett, 2009) and, particularly with respect to the musical domain, for being overly simplistic and neglecting more nuanced 'aesthetic' emotions, for example 'nostalgia' and 'transcendence' (Zentner, Grandjean, & Scherer, 2008). The logical endpoint of this argument is thus: as the number of discrete emotional states accepted to exist is inflated further and further, so too is the amount of labels necessary to describe these states, and therefore the practical utility of these labels is diminished (since the differences between emotional experiences (from a similar type) may be as relevant and meaningful as the distinctions between emotions). Additionally, the assumption of universality of categorical emotion descriptors may be problematic, since it is not always clear what constitutes a 'category'. For example, such descriptors are not necessarily uniform cross-culturally (Russell, 1991). For example, unlike English, Luganda typically makes no formal distinction between the concepts of sadness and anger (Leff, 1973). Research suggests that this discrepancy exists not only linguistically but also phenomenologically – when prompted to describe a personal

event involving anger, Bugandan adolescents were far more likely than Americans to report having cried in response (Davitz, 1969). Therefore, even so-called 'basic' emotion categories might not always provide a valid description of individuals' emotional states.

By contrast, dimensional accounts conceptualise emotional states as non-discrete phenomena, with individual emotions instead describable as points along two or more continua. For example, Russell's (1980) widely influential 'circumplex' model characterised emotion in terms of two dimensions: valence and arousal. According to this model, any possible emotional state may be mapped on a two-dimensional plane with 'arousal' and 'valence' axes, as illustrated in 6. Alternatives to this model have typically added one or more additional dimensions, in order to more effectively disambiguate between different emotional states. For example, Mehrabian (1996) advanced the three-dimensional Pleasure-Arousal-Dominance (PAD) model, in which the latter dimension captures the degree of dominance or submissiveness that one feels, as part of the emotional state. This is useful for distinguishing between emotions that would occur in very similar Arousal/ Valence space; for example, anger and fear are typically associated with negative valence and high arousal, but anger is characterised as dominant while fear is submissive (Broekens, 2012). Even if additional dimensions such as dominance do not correspond to the acitvity of any particular phsyiological system, Broekens (2012) argues that they can be useful in theoretical terms, or for computational modelling purposes. In contrast to the claims made by discrete accounts, such models suggest that all possible affective states are likely underpinned by a common neurophysiological network (Posner, Russell, & Peterson, 2005). On a neurochemical level, Lövheim (2012) posited that a variety of 'distinct' emotional states might be explainable in terms of a three-dimensional model, considering the

relative activity of three monoamine neurotransmitters: dopamine, noradrenaline and serotonin. Cognitively, Lövheim (2012) posited that the activity or inactivity of each of these neurotransmitters approximately corresponds to the dimensions 'activation, vigilance and attention', 'self-confidence, inner strength and satisfaction' and 'reward, motivation and reinforcement', respectively. The first of these dimensions is roughly analogous to arousal, whereas the latter two can be considered as subdivisions of valence.

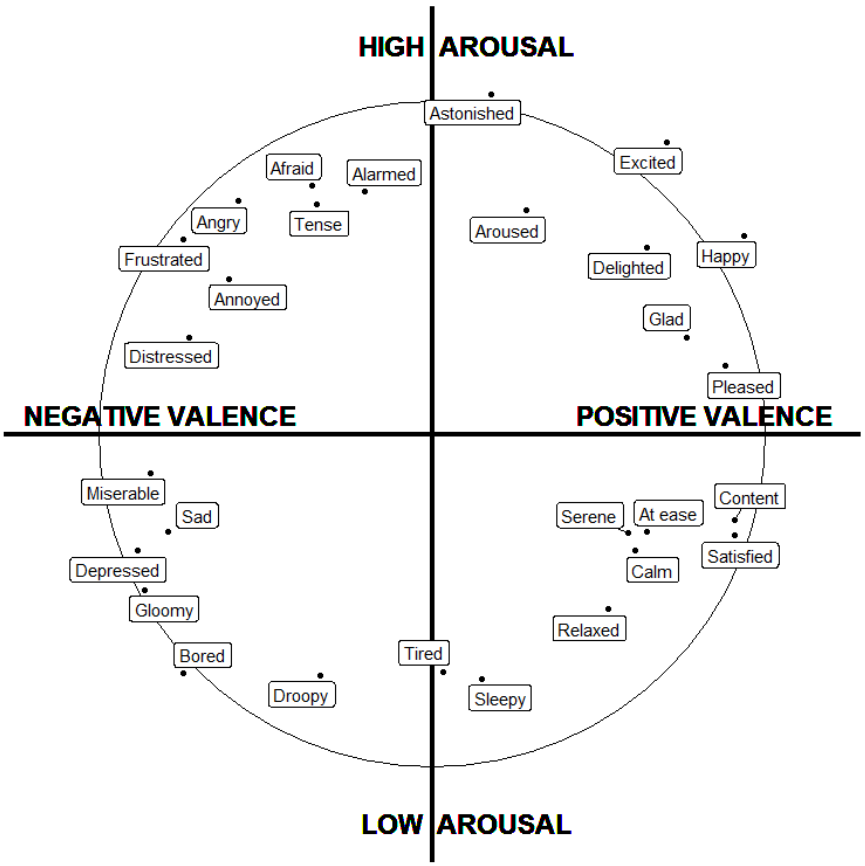


Figure 6: Illustration of Russell's (1980) 'circumplex' model of emotion. Various emotions are plotted according to their constituent arousal and valence dimensions, with the resulting mapping forming an approximate circle. Reconstructed from Russell (1980).

In recent years there has been something of a shift towards dimensional representation of emotions, particularly in the field of affective computing (Gunes, 2010), although

the approach is not without disadvantages. The principle drawback to dimensional conceptualisation of emotion is that emotions that are highly proximal in terms of the resultant dimensional space (e.g. anger and fear) may be in practice quite distinct (Larsen and Diener, 1992; Barrett, 1998). For example, physiological modifications pertaining to anger differ only quantitatively from those associated with happiness, and yet anger does not in practice constitute an elevated or more intense form of happiness (Sartre, 2002). That is, reduction to a limited set of dimensions appears to disregard qualitative variance between different emotions, which may be rather important. In addition, it could be argued that an aforementioned disadvantage of discrete emotions – that they may not be well agreed upon cross-culturally – is also a problem for dimensional accounts. Specifically, the relevance of different dimensions may not be universal, and it is difficult to know whether a given dimensional model is valid cross-culturally. For example, in the case of Luganda mentioned previously, dimensional models might struggle to explain the grouping of anger and sadness, since the 'arousal' dimension should distinguish these emotions quite clearly.

It is worth noting that dimensional and discrete accounts of emotion need not be mutually exclusive – for example, categorical emotion labels may be conceptualised in terms of Russell's (1980) four quadrants of arousal/ valence space (Thayer, 1989). According to this type of framework, emotions are not necessarily 'natural kinds', but instead perceptual 'gestalts' (Bach, 2012), in which a discrete emotion labels refers to a particular set of parameters in a multidimensional space (Barrett, 2006). That is, emotional 'meaning' emerges and is characterised and referred to using discrete emotional labels, even though on a purely physiological level its instantiation is a configuration of continuous parameters. Thus, Barrett (2006) considers that the 'common-sense' conceptualisation of emotions as distinct entities may not

correspond well with the underlying reality. This perspective is reconcilable with Darwinian ideas about the evolutionary utility of so-called 'basic' emotions. Instead of a set of discrete emotions being directly relevant to survival, specific configurations of underlying, dimensional parameters – according to one's environment – happen to be especially relevant to survival. Because of this, there is a need to refer to these states, and therefore linguistic labels are created to describe them. One might therefore expect to see cultural and/ or geographical differences in which emotions are most relevant and, concomitantly, how they are referred to. Summarising this position, discrete emotions provide a convenient way to talk about emotions, even if they are not an accurate reflection of the underlying mechanisms responsible for their production and/or perception. However, because there is nothing sacrosanct about this way of referring to emotions, the exact terms, or groupings of terms used should be expected to vary across individuals or cultural groups. Therefore, emotion perception might be interpreted as a more general phenomenon of categorical perception, in which continuous signals are interpreted with respect to cognitively superimposed stepwise boundaries. Considering the example of colour perception, the underlying external reality consists of light, either omitted or reflected, with continuously-varying wavelengths. No physical property of these stimuli determines the grouping of distinct categories such as 'red' or 'green', and in fact these groupings vary culturally – for example, the Russian language makes a mandatory distinction between light blue and dark blue, creating a direct perceptual advantage for Russian speakers in colour discrimination, relative to native English speakers (Winawer et al., 2007).

Indeed, Barrett (1998) has argued that whether emotions are thought of as discrete or dimensional varies between individuals, and therefore that any nomothetic account

of emotion might unfairly discard inter-individual variability in affective experience. As evidence of this, Barrett reported that, even individuals from the same cultural background differed when instructed to group various emotions together on the basis of similarity. Specifically, some individuals grouped similarly-valenced emotions together (corresponding more closely with predicted groupings from the dimensional account), while others grouped emotions that were similar in terms of arousal (corresponding with predicted groupings from discrete accounts).

To conclude, Barrett's position can be consolidated with the notion of 'basic emotions', by accepting that evolutionary response trends might form another basis for particular emotions to be generally distinguished. Patterns in both conceptualisation and physiological and behavioural responses play a role in defining emotions, and it is common in many cultures to give special status to a smaller set of emotions that are recognised from an early age onwards, despite the variations and differences that may exist within each of the categories.

3.1.3 Categorisation of emotional states in this thesis

Hereafter, throughout this thesis, emotions shall largely be referred to by discrete categorical labels. This should strictly be considered a functional, pragmatic position, and is adopted for several key reasons. Firstly, discrete labels are a convenient way to ensure consistency when referring to emotions, across different studies, that are generally communicated by broadly similar configurations of acoustic features.

Secondly, the treatment of emotions as discrete entities permits the use of a methodology wherein a talker or musician attempts to communicate some distinct emotional state, and a listener attempts to decode this state from a finite set of response options. Thus, a complex and multidimensional phenomenon is reducible to a simple n-

alternative-forced-choice (nAFC) paradigm, in which the emotion detection paradigm has a definably 'correct' answer. By contrast, with a dimensional paradigm, it would be difficult to decide upon an allowable level of discrepancy between the emotion conveyed by a stimulus and the response given by the participant, in order for the response to be considered 'correct'. This is due to potentially differing interpretations of any dimensional scales used – i.e. what the creator of the stimulus considers to be a high level of arousal, might only be considered moderate by the listener. Instead, using a small number of discrete emotion labels (excepting the case of cross-cultural research) should guarantee a good level of agreement between talkers/ musicians and participants.

A third motivation for using discrete emotion labels is related to their *functional relevance*. Although this approach invariably eliminates much of the nuance associated with emotional expression (i.e. it is rare that an entire piece of music can be satisfactorily described by a simple label such as 'sad'), it also provides a powerful way to assess listeners' abilities to recognise general emotional intent. In other words, the fundamental functional relevance (evolutionarily speaking) of being able to distinguish between broad emotion categories is usually greater than that of being able to make very fine-grained emotional judgements. For example, mistakenly perceiving a cry of anger as happiness could have immediate consequences for survival, whereas confusion about the specific type or degree of anger expressed is likely to be less consequential. The advantages of this approach are especially pertinent in the cases of real or simulated hearing impairment, in which the additional degrees of response granted by a continuous, dimensional approach are outweighed by the accompanying increase in task difficulty. Hearing-impaired listeners (especially CI users) have difficulties in reliably distinguishing basic emotional categories, even with

very few response options (Volkova, Trehub, Schellenberg, Papsin, & Gordon, 2013), and therefore there is likely to be little benefit, at present, to increasing this number.

In addition to these considerations, emotion labels are well-understood in everyday language, and therefore may be a more ecologically valid way for listeners to respond. That is, people appear to think about emotion in terms of the explicit meaning of the state itself (e.g. anger), and intuitively discern between different emotions on this basis (Hamann, 2012), rather than in terms of more abstract parameters, such as arousal and valence (although it could of course be argued that the former is arrived at by lower-level understanding of the latter). In fact, whether or not discrete emotion labels correspond to a mere subset of infinite possibilities along the various emotional dimensions, is not especially important for the purposes of this thesis. Functionally, it is sufficient that the talker or musician deliberately attempt to communicate some emotional state, and that this state can be decoded by the listener, and the accuracy of their response assessed. In summary, as Eickers et al., (2017, p. 37) puts it, 'discrete emotions are scientifically useful even if variable and constructed'.

Having delineated what exactly is meant by emotion (at least within the context of this thesis), and how different emotional states might be categorised, the next section considers how individuals produce, recognise and perceive emotion. More precisely, the section considers how these tasks are accomplished within the specific domains of speech and music perception.

3.1.4 How is emotion communicated and perceived in speech and music?

To understand how listeners go about perceiving and identifying an emotion conveyed by an audio stimulus, it is useful to begin by considering how individuals approach perceptual identification tasks more generally. Within cognitive psychology, cogni-

tive neuroscience, and their related fields, the most prevalent account explaining the processes of discrimination and identification (e.g. of different emotions) is the information processing approach (Lindsay & Norman, 1977). Broadly speaking, this perspective assumes a several-stage process of analysis, matching and association – a sound is transduced, informative properties of the auditory signal are extracted, and these are then compared with stored representations in long-term memory, in order to assign semantic significance (McAdams, 1993).

Although often referred to as one complete process, it is worth noting that listening to a piece of music and ascribing an emotion to it may be described in terms of two different processes: perception and recognition. Where such a distinction is made, perception refers to a series of lower-level processes, concerned with the sensory transduction of a stimulus, whereas recognition denotes a process by which meaningful attributes are assigned to this stimulus, typically by comparing what is perceived to some stored mental representation (Adolphs, 2002). Therefore, recognition may be said to depend on information external to the stimulus, at least to a greater extent (discounting, for the sake of simplicity, constructivist accounts of perception, which would claim that the incorporation of top-down information is an essential part of perception itself (Gregory, 1970)). Strictly speaking, some experimental psychologists have drawn additional distinction between recognition and identification, where the latter task constitutes a more narrowly-focussed variant of the former (McAdams, 1993). Specifically, identification requires that stimuli be labelled according to some lexicon of different names – in this case, emotions – whereas recognition may denote mere familiarity with a stimulus (i.e. seen vs. unseen). Therefore, the most commonly used paradigms in the study of emotion perception can usually be referred to as *identification* tasks.

Considering the case of emotion identification, and assuming that we want to categorise a perceived emotional stimulus as one or more of several emotion categories, there are two predominant strategies available: rule-based categorisation and similarity-based categorisation (Smith, Patalano, & Jonides, 1998). In the former, rules of the form 'if X then Y' are applied to a set of stimuli, where X denotes some particular feature configuration, and Y denotes the resultant classification. As an example, one might learn that, if a piece of music is composed in a major key, then it conveys happiness. In fact, in the field of artificial intelligence, this rule-based approach was at the heart of the 'expert systems' popular in the 1970s and 1980s, in which domain-specific expert knowledge is implemented computationally by a series of if-then rules (Russell & Norvig, 1995). By contrast, similarity-based categorisation entails a computation of the similarity of each stimulus-to-be-classified with stored exemplars of the various emotional categories. In particular, relevant stimulus parameters are extracted and compared with acquired knowledge about the distributions of these parameters in each of the categories-to-be-classified. This method is analogous to more modern machine learning approaches to classification, for example involving the use of neural networks (Galushkin, 2007).

For the task of emotion recognition, similarity-based classification is likely to predominate for two key reasons. Firstly, owing to the complex and multidimensional nature of emotional stimuli, it is unlikely that simple rules would have much more predictive power (Smith et al., 1998). That is, a set of rules such as 'if the stimulus is very loud, then the emotion communicated must be anger' would be too inflexible and would likely encounter too many counterexamples to be effective. Secondly, with increased experience, individuals tend to gravitate towards similarity-based categorisation. This approach inherently becomes more accurate when more exemplars are

available for comparison (Allen & Brooks, 1991), and is also typically faster and less cognitively demanding (Smith and Kemler, 1984; Smith and Shapiro, 1989). Since most individuals will have had extensive experience in perceiving and recognising myriad emotional states, it seems fair to assume that they will have long abandoned rule-based categorisation.

It should be noted that the Information Processing account of emotion perception and recognition is not necessarily as fixed or linear as the above (simplified) account implies. In reality, situational factors, individual differences, task effects and personal goals may all interact and exert influence upon perception and identification. Additionally, higher-order 'meaning' may not necessarily arise simply from comparison of a stimulus to some stored representation(s), but instead may be constructed or inferred based on various different items of existing experience and/ or knowledge.

To summarise, in order to perform well in auditory emotion identification tasks featuring speech or music, most individuals will: gather perceptual information by listening, organise this information in terms of the relevant acoustic features, and then compare observed patterns of features with stored representations acquired via previous experience. Of course, this process requires some awareness (though perhaps not necessarily a conscious awareness) of the ways in which emotions are usually conveyed in speech and music. In other words, an individual will need to have already experienced a wealth of emotional speech and music, in order to determine which acoustic parameters are most relevant, and which configurations of these parameters tend to represent particular emotions. Accordingly, the next section is concerned with exactly this: establishing which auditory features are used for the communication of emotion in speech and music, and how these features interact in order to convey distinct emotional states.

3.1.5 Commonalities in the expression of emotion in speech and music

Fortunately, research concerning the transmission of emotion via the auditory domain has typically placed primary emphasis on speech prosody and music, due to their prevalence in social situations and ability to convey emotion effectively and quickly (Coutinho and Dibben, 2013; Bhatara, Laukka, and Levitin, 2014). Communication via this modality is an important aspect of social interaction (Ekman, 1992) and, in most individuals, develops at a young age. Indeed, five-month-old infants demonstrate rudimentary sensitivity to vocal expression of emotion (Fernald, 1993). In normally-hearing children, vocal emotion recognition is typically well-developed by five years old (Sauter, Panattoni, & Happ, 2013), while musical emotion recognition tends to emerge slightly later, by around six to seven years, as children develop awareness of culture-specific musical structures, e.g. Western major and minor tonality (Trainor & Corrigall, 2010).

In both speech and music, emotion is communicated through subtle fluctuations in the auditory signal, characterised as variations in different psychoacoustic attributes, that are associated with the affective state of the speaker or performer. In speech and vocal music, for example, emotion-specific acoustic variance is produced by physiological changes in respiration (the flow of pulmonic air), phonation (modification of the air stream by the larynx) and articulation (modification of the air stream by the lips, teeth, tongue etc.) that accompany one's emotional state (Scherer, 1986; Scherer, 2003). Indeed, using electroglottography (EGG) Johnstone and Scherer (1999) showed that distinct emotional states are associated with different patterns of vibratory vocal fold activity. Therefore, at least to some extent, one's emotional state may be disclosed by so-called 'gestural primitives' associated with the production of sound, i.e. the articulatory mechanisms underlying variations in the acoustic signal

(Rosenblum, 2004). In a Gibsonian sense, this means that characteristic acoustic feature configurations associated with different emotional states may arise directly from the physiological bases of these states (Gibson, 1966). As evidence that listeners are sensitive to such information, Ladefoged and McKinney (1963) showed that loudness judgements were predicted more accurately by using an index of vocal effort (subglottal pressure multiplied by air velocity across the glottis) than by using the actual intensity of the acoustic signal. In the case of instrumental music, a performer's affective state may instead be disclosed via their interaction with their instrument. For example, Gabrielsson and Juslin (1996) found that listeners were sensitive to small modulations in several musical parameters (e.g. tempo, dynamics), which were used to express emotion for a range of different instruments.

It should of course be noted that emotions can also be expressed in the absence of genuine underlying affective states (and their associated physiological indices). This is evident in the case of acted speech (which, in fact, makes up many databases of emotional speech stimuli, e.g. Burkhardt, Paeschke, Rolfes, Sendlmeier, and Weiss (2005)), and is especially important in the performance of music (Lindström et al., 2003). In both scenarios, one can express an emotional state simply by knowing how this state would usually affect an individual's speech or musical performance, and mimicking these effects. However, there may be individual differences regarding the extent to which emotion in music is 'embodied' by the musician, as opposed to merely 'performed' (Van Zijl & Sloboda, 2011). Of course, unlike speech, music can also express emotion via higher-level compositional cues, (e.g. tonality), which are essentially unrelated to the physical manifestations of emotional states. To maintain comparison with speech, however, this thesis is primarily concerned with performance-based cues.

It is argued that associations between patterns of modulations in these acoustic parameters and distinct emotional states may constitute a universal in human communication (Scherer, Banse, and Wallbott, 2001; Thompson and Balkwill, 2006). Indeed, evidence has shown that different acoustic configurations are employed to communicate specific emotions during speech, irrespective of syntactic or semantic context, and that these configurations are very similar across different individuals (Cosmides, 1983). However, more recent research has acknowledged that, at least in the case of music, there may also be cross-cultural differences in terms of the 'weighting' given to specific acoustic parameters when identifying an emotion (Lennie, 2017).

Importantly, in any case, research shows that the acoustic profiles characterising distinct emotions in speech and music reliably lead to perceptual bias towards the relevant emotion in recognition tasks (e.g. Mozziconacci and Hermes, 1999). In other words, distinctive sets of acoustic attributes used by talkers/ musicians to convey specific emotions are 'recognised' by listeners, and are predictably associated with perception of the corresponding emotions. As such, it is possible for computer models to mimic human performance relatively closely in such emotion discrimination tasks, by utilising statistical pattern recognition to extract relevant prosodic features from the acoustic signal, and learning to associate these with emotional states (Dellaert, Polzin, and Waibel, 1996; Petrushin, 1999).

In speech, with respect to specific cues, Petrushin (1999, p. 4) noted that 'All studies in the field point to the pitch (fundamental frequency) as the main vocal cue for emotion recognition'. This being said, in addition to frequency, emotional expression in speech is typically thought to be comprised of paralinguistic variance in intensity, temporal pattern and spectral content (Juslin and Laukka, 2001; Scherer, Banse, Wallbott, and Goldbeck, 1991). The latter of which may be further subdi-

vided taxonomically into timbre- and dissonance-based cues (Coutinho and Dibben, 2013; Gabrielsson and Lindström, 2010). Indeed, such variables as sharpness (the proportion of high frequency, relative to low frequency, content in a signal) and spectral centroid (the weighted mean of a signal's constituent frequencies) demonstrably influence ratings of stimuli arousal-level and valence (Coutinho & Dibben, 2013). Importantly, each of the aforementioned cues may be differentially important for the decoding of emotional expression, depending upon the emotion to be identified (Banse & Scherer, 1996). For example, the expression of fear is most often associated with a higher than average mean pitch, whereas the expression of sadness may be characterised by a lower mean pitch and also a reduced speech-rate.

As in speech, musical emotion is conveyed by variations in structural acoustic features (Juslin & Sloboda, 2010), which may be recognisable universally (Balkwill, Thomspon, and Matsunaga, 2004; Fritz et al., 2009). Research has suggested that common acoustic features are implicated in the processing of prosodic structure in both speech and music (Fedorenko, Patel, Casasanto, Winawer, and Gibson, 2009; Falk, Rathcke, and Dalla Bella, 2014). Furthermore, specific emotional states, for example sadness, may sometimes be communicated by highly similar patterns of acoustic variation in both speech and music – i.e. not only are similar features utilised for the communication of emotion in the two domains, these features also appear to be utilised in similar ways (Curtis & Bharucha, 2010). As an example of this, anger is typically characterised by increased tempo or speech rate, greater intensity and greater high-frequency content in both speech and music (Juslin & Laukka, 2003). Similarly, increased intensity in both speech and music generally results in greater perceived valence and tension (Ilie & Thompson, 2006). However, Ilie and Thompson also documented some contradictory examples – increased pitch height

appeared to be associated differentially with greater valence ratings in speech and lower valence ratings in music, while slow rate was associated with positive valence for speech but had no clear effect for music. To some extent, discrepancies such as these may reflect the fact that music (typically more so than speech) is considered a creative art form. Therefore it may be that variation in emotional expression is a part of this art. That is, emotional expression may interact with and be influenced by more general artistic expression on the part of the performer, in a way that does not usually occur in speech. In fact, since this aspect of music tends to be ignored in experimental contexts, it is possible that the overlap between speech and music – in terms of the ways that acoustic cues are utilised for emotional expression – has been somewhat exaggerated.

In any case, as in speech, emotion in music appears to be principally communicated by variance in psychoacoustic cues that fall into four broad categories: frequency, intensity, tempo/ rate and spectral composition (with higher-order musical features such as articulation being explainable in terms of these more basic features). Reflecting this, a recent computational modelling study investigating the acoustic features associated with emotion communication used overlapping input vectors for predicting emotion in both speech and music (Coutinho & Dibben, 2013). For both, information about loudness, tempo/ rate, pitch contour, spectral centroid and sharpness were deemed to be important – in fact, the speech and music models only differed with respect to the contribution of cues related to spectral content, with roughness (associated with narrow harmonic intervals), for example, being relatively more important in speech than in music (Coutinho & Dibben, 2013).

In addition to the above, there is also evidence that training in music may benefit individuals' ability to decode emotion in speech, (Strait, Kraus, Skoe, and Ashley,

2009; Thompson, Schellenberg, and Husain, 2004) – an effect that is demonstrable even in hearing-impaired populations (Good et al., 2017). That is, the acoustic commonalities shared by speech and music in the expression of emotion (i.e. Bregman's (1994) 'primitive' processes) are substantial enough such that training in emotion-decoding with one medium confers a benefit when tested with the other.

3.1.6 Problems with the comparison of emotion in speech and music

It should be kept in mind however, that, although the expression of emotion via each medium depends upon variation in common acoustic parameters, the communicative purposes of emotional speech and music are usually quite different and subject to varying cultural influence. That is, the observation that speech and music share lower-level cognitive resources should be qualified by the consideration that the higher-order functional attributes of each are likely processed separately (Dalla Bella, Berkowska, & Sowiński, 2011). For example, recent research suggests that the superior temporal sulcus (STS) is sensitive specifically to the detection of speech rather than musical sounds (Overath, McDermott, Zarate, & Poeppel, 2015). This evidence supports a theoretical account in which sounds share overlapping resources for the processing of their constituent features, but are then categorised as particular 'classes' of sounds (e.g. speech, music), for which the processing of higher-level attributes (e.g. their meaning) proceeds relatively independently.

Additionally, although meta-analysis has demonstrated comparable decoding accuracy for emotion in speech and in music (Juslin & Laukka, 2003), the prevalent use of fixed melodies to convey different emotions may result in a reduced range of modifiable acoustic features in music stimuli, leading to relatively poorer recognition accuracy, and further complicating the comparison with speech stimuli (Livingstone,

Thompson, Wanderley, & Palmer, 2015). Of course, speech is not usually thought of as being 'composed' in advance in the same way as music – in fact, a fairer comparison in this regard might be made between speech and improvised music. Given a particular sentence, a talker can expressively alter their pitch contour without disrupting the sentence's semantic content (though the interpretation thereof might differ). In music, however, given a specific melody – which is the closest analogy one can make to a sentence – pitch contour cannot be changed substantially while preserving musical 'meaning'. Fundamentally, if one accepts that, in speech and music respectively, sentences and melodies are the meaningful 'units' that must be preserved when expressing emotion (a heavily context-dependent assumption), then it follows that the content of speech is more robust to 'disruption' introduced by expressive fluctuation of acoustic parameters. Within this admittedly specific context, speech can be considered as facilitating a greater degree of expressive variance in acoustic parameters (since pitch contour is free to vary). Since this extra 'feature' is present, one might expect decoding of emotion in speech to be more accurate than in music. Conversely, in music, timbre might be considered relatively 'free' to vary note-by-note, without disrupting the overall melody. By contrast, in speech, variation in timbre (excluding differences between different talkers) must be smaller in order to preserve the meaning of the words spoken.

Relatedly, although music and speech often make use of similar auditory features in order to signal meaning, the ways in which these features are utilised can vary considerably. For example, in non-tonal languages, the magnitude of relevant pitch variations is much greater in speech than in music (Peretz & Hyde, 2003). When forming a question, speakers tend to accentuate the final word with a pitch increase of up to or exceeding twelve semitones (Patel, Wong, Foxton, Lochy, & Peretz, 2008),

whereas in music, differences of even one semitone often carry meaningful expressive information (Vos & Troost, 1989). Similarly, the use of timbre as an expressive device in speech is of course limited by the capabilities of the human voice, whereas in music the voice is one of very many instruments. In fact, given the capabilities of modern sound synthesis, the scope for expressive timbre variation within music is essentially infinite. One may conclude in any case that, depending upon the setting, speech and music may not necessarily be equally expressive media, or at least may not always achieve emotional expression in exactly the same way.

3.1.7 Is there a 'gold standard' for emotion perception?

Considering all of the above, and anticipating the discussion to follow on auditory emotion perception by hearing-impaired listeners, it should be noted that error rates tend to be relatively high for these types of auditory emotional judgement tasks (Bachorowski, 1999), even for normal-hearing listeners. Although most individuals can consistently decode emotion at a level significantly exceeding chance, a relatively large degree of inter-individual variance has been observed. There are at least two factors that could explain why emotion identification paradigms should be characterised by large error rates and highly variable performance across individuals.

Firstly, substantial individual differences exist in both expression and perception of emotion. For example, differences in various contextual factors, including: listener emotional state, language abilities, and degree of musical training have all been shown to influence emotion perception (Bouhuys, Bloem, and Groothuis, 1995; Saheer and Potard, 2013; Strait et al., 2009). In fact, emotion perception abilities may also vary between individuals as a function of the particular emotion(s) to be identified (Bachorowski, 1999).

Secondly, it could be argued that emotion perception tasks lack ecological validity. In everyday life, it is rare to experience strongly-expressed, isolated emotions, divorced from context, and unrelated to the sentence or melody being communicated. In fact, in speech perception, this can lead to confusion because of the 'irony effect', in which the combination of intense emotional expression and banal semantic content result in a perception that the talker is being insincere or ironic (Saheer & Potard, 2013).

For these reasons, one might ask the question: 'What is the "gold standard" for auditory emotion perception?', or more precisely: 'If emotion perception is at best difficult and rather varied for NH listeners, then what is a reasonable target for those with hearing impairment'. In the case of speech recognition, for example, it is relatively clear that the standard of performance to aim for with hearing restoration is 100% accurate decoding of presented words or sentences. Of course, it is less straightforward than this, and there are various considerations to be made which might affect this standard – e.g. acoustic environment, extent of background noise – but nonetheless it is relatively easy to imagine what ideal performance should entail. In emotion perception, since 100% accuracy appears to be a less realisable target, it is much more difficult to know what the 'gold standard' should be, in terms of restoring hearing function (Luo, 2016). This is compounded by the argument that music might be associated with a different 'mode' of listening altogether, which may be less well-suited to objective assessment (Limb, 2016).

The issue of 'gold standards' in auditory emotion perception – and more generally in perceptual tasks with hearing-impaired participants – is returned to later. For now, it is sufficient to note that, even for NH listeners, emotion perception constitutes a somewhat difficult task, which is something that should be kept in mind when considering the emotion perception abilities of hearing-impaired listeners. The next

section begins this discussion, with an overview of auditory emotion perception in HA and CI users.

3.2 How do hearing-impaired listeners perform in emotion perception tasks?

3.2.1 Hearing aids

Although hearing aid digital signal processing (DSP) is generally very successful in facilitating accurate speech recognition, it may be less helpful for making judgements of emotion from speech. Several studies have documented poorer performance in emotion recognition by individuals with hearing loss listening with hearing aids, compared to age-matched normal-hearing controls (Most, Weisel, and Zaychik, 1993; Most and Aviner, 2009; Rigo and Lieberman, 1989). In a recent study addressing this phenomenon, Goy, Pichora-Fuller, Singh, and Russo (2016) found that listeners were, unsurprisingly, significantly better at word recognition when listening with their hearing aids vs. unaided; however, there was no such benefit for the recognition of emotions. That is, hearing aids were neither detrimental nor conducive to accurate emotion perception. Additionally, by contrast to normal-hearing subjects, past research has shown that hearing aid users do not show improvement in emotion recognition when presented with audio and visual information, as opposed to purely visual information (Most et al., 1993; Rigo and Lieberman, 1989). This implies that either: A) the HA does not provide any helpful auditory cues to emotion, once visual cues have been considered (or perhaps provides unhelpful or conflicting cues), or B) HA users assumed that the auditory cues to emotion would not be useful, and therefore focussed solely on the visual information.

Recently however, (Schmidt et al., 2016) has pointed out that several of the above studies specifically investigated emotion perception by children with HAs, and that the results might be less applicable to adults. The principle reason for this is that adults may have acquired hearing impairment later in life, and therefore had the benefit of normal hearing when learning about the relationships between different emotional states and configurations of acoustic features Schmidt et al. (2016). Of the studies carried out with adults, results have been less consistent than for children. For example, Rigo and Lieberman (1989) found that low frequency hearing losses in particular were associated with greater difficulty in emotion perception tasks. Conversely, Orbelo, Grim, Talbott, and Ross (2005) found that, for elderly subjects, extent of age-related hearing loss was not a significant predictor of performance in the perception of emotional speech. Additionally, with respect to perception of arousal and valence in emotional speech, Schmidt et al. (2016) found that HA users did not differ significantly from NH listeners, although HA users' arousal ratings reflected a slightly greater sensitivity to small differences in intensity.

Specifically considering musical emotion, the difference between HA users and NH listeners appears to be somewhat stronger, although the number of studies is too small to be certain. In a pre-validated five-alternative forced-choice paradigm (anger, fear, happiness, sadness, tenderness), Russo and Fanelli (2016) documented significantly worse emotion recognition accuracy for HA users, relative to an NH control group. Interestingly however, performance did not differ significantly between HA users and non-aided HI listeners (Russo & Fanelli, 2016). Therefore, with respect to this paradigm, there is clearly improvement to be made for HI listeners, and this is not currently being achieved by the use of HAs. The authors note that this need not necessarily denote a failure of the HA per se, but might instead be a consequence of

DSP geared primarily towards speech perception.

In summary, emotion perception by HA users appears to be impaired relative to NH listeners, but the exact extent of this deficit is unclear, and is complicated by the fact that studies have included both children and adults, leading to inconsistent results. Since at least a handful of studies have shown no (or very little) significance difference between HA and NH listeners, it is probable that any differences are relatively small. There has been less research into emotion perception in music, but the available evidence suggests a more apparent deficit here. This is most likely because HAs are not always well-optimised to deal with music as an input.

3.2.2 Cochlear implants

CI users typically perform relatively poorly on tasks requiring sensitivity to pitch-dominant features of speech (Chatterjee et al., 2015) and music (Kong, Mullangi, Marozeau, and Epstein, 2011; Tao et al., 2015), including vocal and musical expression of emotion (Luo, Fu, and Galvin, 2007; Nakata et al., 2012; Volkova et al., 2013). As with HA users, emotion perception via other modalities is preserved (Hopyan-Misakyan, Gordon, Dennis, & Papsin, 2009), but appears not to be enhanced by simultaneous auditory input (Most & Aviner, 2009). In most cases however, CI users are able to identify basic emotions in speech at a level significantly greater than chance, particularly where exaggerated acoustic cues are used (Chatterjee et al., 2015) or response options are limited, e.g. to a binary happy or sad judgement (Volkova et al., 2013). In addition, the ability of both CI users and NH participants listening with CI-simulation to discriminate between different talkers suggests some residual sensitivity to speaker-specific phonetic detail (van Heugten, Volkova, Trehub, & Schellenberg, 2014). In music also, CI users are able to perceive emotion at a level

greater than chance. In fact, research has reported recognition accuracy as high as 87.5% correct, when using binary happy/ sad judgment paradigms (Hopyan, Gordon, and Papsin, 2011; Hopyan, Manno III, Papsin, and Gordon, 2015) – for comparison, emotion recognition accuracy has been estimated at 84% (House, 1994) to 90% correct (Luo et al., 2007) in NH listeners. In another recent experiment, Ambert-Dahan, Giraud, Sterkers, and Samson (2015) demonstrated CI users' above-chance performance in emotion discrimination of short musical excerpts, using a forced choice of: fear, happiness, peacefulness or sadness, in addition to generic arousal and valence ratings.

It is possible that this performance might arise as a result of compensatory attention to relatively preserved psychoacoustic features. For example, Shannon, Zeng, Kamath, Wygonski, and Ekelid (1995) demonstrated that, when spectral information is artificially attenuated, normally-hearing listeners can decode speech accurately by attending primarily to its temporal features. Recently, Tao et al. (2015) suggested that CI users may utilise similar compensatory strategies to achieve adequate performance in lexical tone perception and Meng, Zheng, and Li (2016) showed that tone recognition in Mandarin might be improved by the use of an algorithm to artificially represent f_0 contour information as loudness variation. Additionally, research suggests that both adult CI users and NH controls listening to CI simulated speech and music similarly shift their attention away from pitch-based features and towards relatively preserved acoustic features, such as intensity and timing/ rate-based cues (Peng, Lu, and Chatterjee, 2009; Peng, Chatterjee, and Lu, 2012). Giannantonio, Polonenko, Papsin, Paludetti, and Gordon (2015) showed that children using CIs tend to rely more heavily on temporal rather than mode-based cues when decoding emotion in music, though the tendency to focus on mode was more prevalent in

participants to whom residual acoustic information was available (i.e. via a contralateral hearing aid). NH participants listening via NBV-based CI simulation displayed a similar reliance on temporal cues, regardless of inter-individual differences in musical training. Likewise, Caldwell, Rankin, Jiradejvong, Carver, and Limb (2015) found that adult CI users tended to base judgements of musical emotion on tempo rather than mode, significantly more so than a group of NH controls.

The results described above accord well with the observations that, in both speech and music, temporal and intensity-related acoustic cues are relatively well preserved in CI users (Volkova, Trehub, Schellenberg, Papsin, and Gordon, 2014; Hopyan, Peretz, Chan, Papsin, and Gordon, 2012; Shannon, 1989; Shannon, 1992), since these features are relatively well delivered by electrical stimulation (Cooper, Tobey, and Loizou, 2008; Drennan and Rubinstein, 2008). Encouragingly, these results suggest that accurate auditory perception of emotion (or rather, accuracy comparable to NH listeners) may be an achievable target for CI users. To this end, recent research has suggested that children implanted at a very early age and receiving intensive rehabilitation may be able to reach levels of emotion detection performance equivalent to those observed in NH controls (Mildner & Koska, 2014). Once again, the researchers speculated that different recognition strategies were employed for the two groups, due to discrepancies in confusion matrices indicating diverging patterns of errors made. Thus, by adapting divergent listening strategies, CI users may have the potential to perceive emotion auditorily at a level comparable to the NH population.

In summary, emotion perception is, on the whole, noticeably impaired in CI users. Relative to both NH listeners and to HA users, CI users tend to decode emotional expression in both speech and music with reduced accuracy. The level of performance that is obtained by CI users – which is typically above chance-level – is likely to be

achieved by a different underlying listening 'strategy', in which relatively preserved components of the auditory signal are preferentially attended to. This is encouraging in terms of clinical rehabilitation, and suggests that a level of performance comparable to NH listeners might be achievable. However it remains to be seen exactly how this performance might be realised, and whether this could apply in more challenging emotion recognition paradigms.

Addressing some of the gaps in the literature reviewed thus far, Chapter 4 introduces the first two empirical experiments – behavioural listening studies examining the perception of emotion in both speech and music by CI-simulated listeners. Specifically, these experiments aimed to elucidate the listening strategies underpinning above-chance performance by CI users in emotion perception tasks. In the next chapter, the rationale for these studies is outlined in more depth, the experimental paradigm (including the use of CI simulation) is described in detail, and the results from both studies are discussed.

4 Studies 1 and 2: Examining emotion perception in speech and music by cochlear implant-simulated listeners

4.1 Overview

This chapter reports two listening experiments, in which NH participants were presented with either emotional speech or music stimuli, both with and without a cochlear implant (CI) simulation. Participants also heard digitally-manipulated stimuli, in which different auditory feature cues (e.g. variation in stimulus intensity) were

systematically attenuated. For both speech and music, participants' performance was evaluated with and without the CI simulation, as a function of their performance using the various auditory feature cues, and as a function of the particular emotion being judged.

4.2 Introduction

The purpose of conducting Studies 1 and 2 was to examine auditory perception of emotion in CI users, in both speech and music, respectively. These studies were intended as the starting point of an investigation aimed at addressing several gaps in prior knowledge in this area, as uncovered by the review of previous literature. Most significantly, and as mentioned during the conclusion of the previous chapter, there has been relatively little attention paid by previous studies to the underlying listening strategies utilised by CI users when decoding emotion auditorily – therefore it is unclear how exactly above-chance performance occurs. For clarity, 'listening strategies' hereafter refer to distinct modes of, or approaches to, listening – encompassing listeners' understanding of task demands, selective direction of attention towards relevant acoustic attributes, and metacognitive awareness of perceptual goals (Vandergrift, Goh, Mareschal, & Tafaghodtari, 2006).

Secondly, several studies have made especial effort to ensure that expressed emotions are easily discriminable – for example, via either limiting the response options available, using deliberately-exaggerated expressive intonation, or both of the aforementioned. This usually means that NH participants are able to perform at ceiling level, thereby complicating the comparison of CI and NH groups. Although, this approach clearly has merit, and has been valuable in showing that CI users can indeed decode some aspects of emotion, it remains to be seen how well CI users might

perform when allowed more response options, and presented with more 'natural', ecologically-valid stimuli.

Turning specifically to emotion expressed in music, much of the literature previously discussed has focussed on compositional components (e.g. major/ minor tonality, rising/ falling melodic contour etc.), while there has been relatively little attention given to performance-based cues, making it difficult to discern how sensitive CI users might be to these. Lastly, there appear to have been only a small handful of studies that have examined the perception of emotion in music by CI-simulated listeners (Ahmed, 2017; Giannantonio et al., 2015). This line of enquiry is worth pursuing further, since simulation studies may be informative as to the extent that CI users' difficulties in emotion perception are a product of distorted input signals (since CI-simulated participants experience similarly distorted signals, but otherwise are NH listeners).

Therefore, the present studies are distinguishable from much of the previous literature by the following points of divergence: the emotion discrimination paradigm employed was more difficult, stimuli were systematically manipulated to probe the listening 'strategies' underlying performance, music performance cues were investigated specifically (as opposed to composition cues), and CI simulation was used with both speech and music stimuli.

It has been well established already that cochlear implant users perform worse than normally-hearing listeners, but above chance-level on auditory emotion recognition tasks. This finding has been documented with both speech and music stimuli, and has also been observed with NH participants, listening with a CI simulation. However, what is not currently known is the extent to which this performance originates from residual frequency information conveyed by the implant, or by diversion of attention

towards better-preserved features (i.e. differences in timing/ rate or intensity).

Studies thus far have tended to use naturalistic stimuli (typically recorded, unprocessed excerpts of speech and music), improving ecological validity, but rendering it difficult to assess the reliance on different cues and the extent to which features are sufficient for emotion recognition. Therefore, a primary aim of the current study is to more closely examine these factors by evaluating emotion recognition accuracy with various auditory feature-attenuated stimuli. For the studies reported in this paper, frequency, intensity and temporal variation were chosen as cues because of their prevalence in the auditory emotion recognition literatures relating to both speech and music, as well as their relative ease of manipulability. Cues relating specifically to spectral/ timbral content (roughness, spectral centroid, spectral flux etc.) were not investigated in the current study because of the difficulties inherent in attenuating these features satisfactorily.

4.3 Aims

To summarise, the studies aimed to build upon and extend the work of Chatterjee et al. (2015) and other similar studies, by including both emotional music and speech stimuli in a five-alternative forced-choice (5-AFC) emotion discrimination paradigm, investigating in greater depth the listening strategies used and the contributions of various acoustic parameters to emotion recognition. Specifically, by selectively attenuating different acoustic features, the study aimed to uncover which types of information are attended to during CI simulation. Accordingly, these studies aimed to shed some light upon the kinds of listening strategies that might be used under conditions of degraded spectro-temporal resolution. Considering music specifically, Study Two aimed to ascertain the extent to which CI-simulated listeners were able

to decode emotion using only performance cues, rather than compositional ones.

Further, the studies were intended to serve as a 'platform' from which to launch a follow-up investigation, incorporating real CI users in addition to NH participants. Taken together, it is hoped that these studies may help to determine the extent to which different listening strategies are utilised by CI users during emotion perception, and also to what extent CI users' deficits in emotion perception are a product of degradation to the input signal. In essence, the overarching goal of this research is to unearth as much information as possible about the nature of CI users' auditory emotional judgements, which may find wide-ranging applications in approaches to rehabilitation and/or CI digital signal processing.

4.4 Methods

4.4.1 Participants

In total, thirty-two participants were recruited: seventeen for Study 1 (11 female; mean age = 31.81 years, SD = 8.40) and fifteen for Study 2 (9 female; mean age 32.69 years, SD = 13.01), via a combination of University of Sheffield student and staff volunteer mailing lists and opportunity sampling. All participants reported having normal hearing and normal or corrected-to-normal vision, and all provided fully-informed consent prior to participation. Of those participants recruited for Study 1, all reported non-fluency in German, which was important since this language was used for the speech stimuli.

4.4.2 Materials

During both experiments, participants were presented with numerous auditory stimuli and were asked to report which emotions they perceived on each occasion. In total, two hundred speech excerpts were presented in Study 1 and two hundred musical melodies were presented in Study 2.

The speech stimuli comprised: four different sentences × five emotions × five auditory feature conditions × two CI conditions. Music stimuli comprised four different melodies × five emotions × five auditory feature conditions × two CI conditions. All stimuli were encoded in single-channel (mono), 16-bit Audio Interchange File Format (AIFF) with 16,000 Hz sampling frequency.

In Study 1, speech stimuli were derived from a subset of the Berlin Emotional Speech Database (Burkhardt et al., 2005), which consists of microphone recordings of emotional utterances by non-professional actors, selected by a panel of expert listeners. From the database, four different sentences were chosen, spoken by ten different speakers (five female) each intentionally intonated to express five different emotions: anger, happiness, sadness, fear and neutrality. In terms of the literal, semantic content conveyed however (i.e. the actual words spoken), all utterances were emotionally neutral sentences that could realistically occur in everyday life (e.g. "Der Lappen liegt auf dem Eisschrank" – "The cloth is on the refrigerator"). German sentences of a mean 2.59 (SD = 0.57) seconds in length. Previous research has confirmed that speech excerpts of this duration or even shorter are sufficient to convey differences in emotional expression – (Laukkanen, Vilkman, Alku, & Oksanen, 1996) found that talkers were able to annunciate isolated vowel sounds such that five different emotions were distinguishable. For the stimuli used in the present study, Burkhardt et al. (2005) verified that the intended emotions expressed by each of the utterances

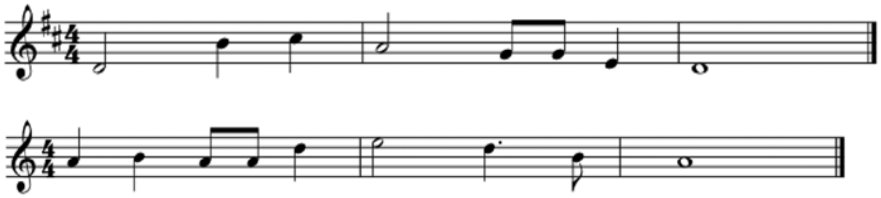


Figure 7: Musical notation, illustrating two of the emotionally ambiguous stimulus melodies used. Adapted from Quinto et al. (2014)

were recognisable with at least 80% accuracy.

In Study Two, musical stimuli were derived from a recent study in which highly-trained musicians (mean 15 years formal training) performed four different, pre-composed melodies (Quinto, Thompson, & Taylor, 2014). These melodies were composed by the study's first author, and were intended to be unfamiliar and emotionally ambiguous. As such the melodies were rhythmically simplistic and were not readily identifiable as being in either the major or minor mode, since the mediant (third scale degree) was omitted (Figure 7). The melodies consisted of between seven and nine notes (mean = 7.75) and were on average 6.31 (SD = 1.97) seconds in length. The four melodies were performed by expert musicians, who tailored their playing in order to express one of five emotions: anger, happiness, sadness, fear or neutrality (a sixth emotion, 'tenderness' was also included in the original study by Quinto et al., but is not used here since it was not satisfactorily well-decoded by listeners). It should be emphasised that the melodies themselves were not designed to convey any particular emotion, and as such the emotional conditions described hereafter for musical stimuli denote differences in performance expression, i.e. subtle variations in timing, dynamics, timbre or pitch (Palmer, 1997). The stimuli used comprised two timbres: half of the stimuli were performed by violinists, whilst the rest were performed by vocalists.

As with the speech stimuli, prior research has confirmed that listeners can adequately identify emotional expression from very short excerpts (i.e. Vieillard et al. (2008) found that five musical events were sufficient to discriminate between four different emotions). Moreover, emotion judgements (discrimination between five emotions) based upon 300-400 ms show significant correlation with judgements based upon much lengthier stimuli (Krumhansl, 2010). In the original study from which the stimuli were derived, average recognition accuracy of 41.64% was observed for the five emotions included here – well above chance level of 16.67% (since the study included six response options) (Quinto et al., 2014).

4.4.3 Auditory feature conditions

In addition to the original stimuli, four additional versions were created per stimulus, each preserving a particular auditory feature (frequency, intensity, duration or articulation), whilst attenuating the others (see Table 2). Note that, in speech, articulation usually refers to the resonant qualities of the vocal tract, as determined by the position of vocal articulators (lips, tongue, jaw etc.) (Sundberg, 1999), whereas in music articulation most commonly refers to the amplitude envelope that is characteristic of a specific instrument or style of playing. For the purposes of these studies (for both speech and music), unless otherwise specified, 'articulation' shall henceforth refer to the residual acoustic characteristics present in a sound following attenuation of frequency, intensity and temporal variance.

Variation in stimulus fundamental frequency (f0) was attenuated using Praat (Boersma & Weenink, 2009) in conjunction with the Vocal Toolkit plugin (Corretge, 2012), by using the *Monotonize* function to flatten the extracted frequency contour to the estimated stimulus median (derived by dividing the stimuli into 0.01 second-long frames

Table 2: Overview of different auditory feature conditions for speech and music stimuli. * Note that the isolated Frequency, Intensity and Duration stimuli still contained residual spectral and amplitude envelope information.

Condition	**Feature(s) preserved**	**Features attenuated**
Original	All	None
Frequency	Frequency variation*	Duration, intensity
Intensity	Intensity variation*	Duration, frequency
Duration	Temporal variation*	Frequency, intensity
Articulation	Spectral envelope	Duration, frequency, intensity

and calculating the fundamental frequency for each). To check that this manipulation to attenuate stimulus frequency variation was successful, measures of variation in fundamental frequency (F_0) were calculated for both the original and processed versions of the stimuli. Frequency variation was quantified as the standard deviation of estimated F_0 across each stimulus. F_0 was estimated within Sonic Visualiser (Cannam, Landone, & Sandler, 2010) using the YIN algorithm – an autocorrelation-based fundamental frequency estimator – which essentially works by examining small 'windows' of a given signal, and assessing the correlation between these and the same windows offset by some amount (τ) (de Cheveigne & Kawahara, 2002). The algorithm then searches for the non-zero value of τ that leads to the greatest correlation value, which is used to derive the F_0 estimate. The effects of this manipulation upon frequency variation, for both the speech and music stimuli, are illustrated in Figure 8. For all stimuli, frequency variation was substantially reduced. Music had greater frequency variation on average to begin with, but both stimuli sets were reduced by a similar magnitude.

As would be expected, however, variation in overtones across stimuli was largely preserved by this manipulation, as shown in Figure 9.

Intensity variation was attenuated using a dynamic range limiter within FL Studio, with a 42 ms attack, 21 ms release, and 0.0 dB ceiling. These parameters were cho-

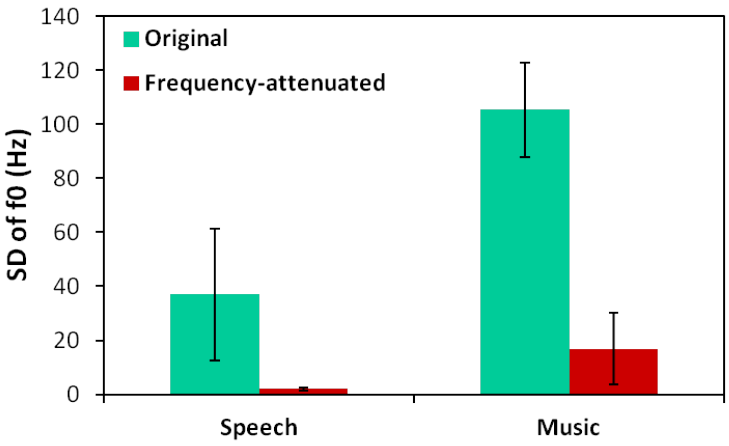


Figure 8: Mean frequency variation (mean SD of f0) for the speech and music stimuli, before and after the frequency attenuation process. Error bars indicate ±1 standard deviation.

sen in order to minimise dynamic range, without introducing unwanted distortion. Essentially, dynamic range compression reduces the volume of the loudest parts of an acoustic signal, so that the overall dynamic range is also reduced. To check that this manipulation was successful, intensity variation was estimated for both original and processed stimuli. This measure was calculated in MATLAB by first deriving an analytic representation of each signal via the Hilbert transformation, which returns both the original signal and an imaginary component, comprising a 90° phase-shifted version of the original. This is achieved by computing a Fourier transform of the original signal, rejecting the negative frequencies, and then computing an inverse Fourier transform (Liu, 2012). Essentially, this provides a way to represent a given signal in terms of both amplitude and frequency modulation components. In this case, the resulting instantaneous amplitude (envelope), was smoothed using a fourth-order Butterworth low-pass filter (100 Hz cut-off) and the resulting values were converted to dB re 1 (dB with reference value of 1, i.e. $20\log_{10}(env)$). For each stimulus, the standard deviation of these values was calculated and used as an estimate of intensity variation. The effects of this manipulation upon intensity variation, in both

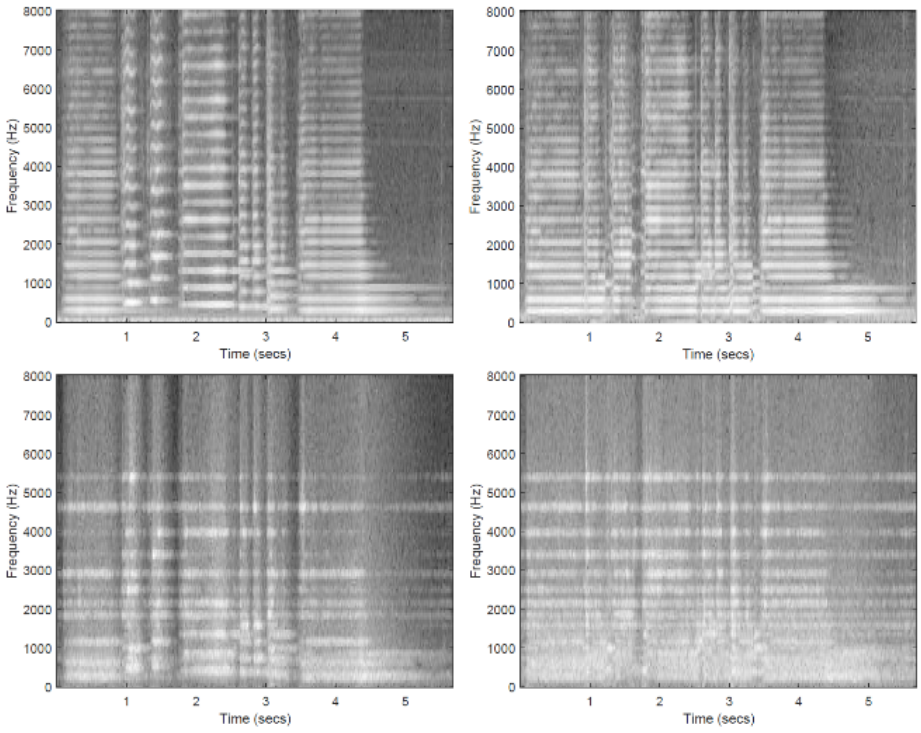


Figure 9: Spectrograms depicting residual harmonic content following frequency normalisation. Left panels = original stimuli, right panels = frequency-normalised stimuli. Upper panels = without CI simulation, lower panels = with CI simulation.

the speech and music stimuli, are illustrated in Figure 10. As expected, all stimuli showed a reduction in intensity variation following the manipulation – speech stimuli initially contained greater variability, but both speech and music were affected to a similar extent by the manipulation (approximately 4 dB).

Temporal variation was attenuated using the Vocal Toolkit for Praat, using the dynamic time-warping (DTW) algorithm, which aims to non-linearly map two signals to each other, (Ratanamahatana & Keogh, 2004). In this case finding an optimal alignment between the time series of two stimuli. Given time series vectors, A and B, of length n and m respectively, an $n \times m$ matrix is constructed, consisting of the distances between each point in vector A and every point in vector B. A 'warping path', W, is then computed, which denotes a sequence of grid points that minimise the

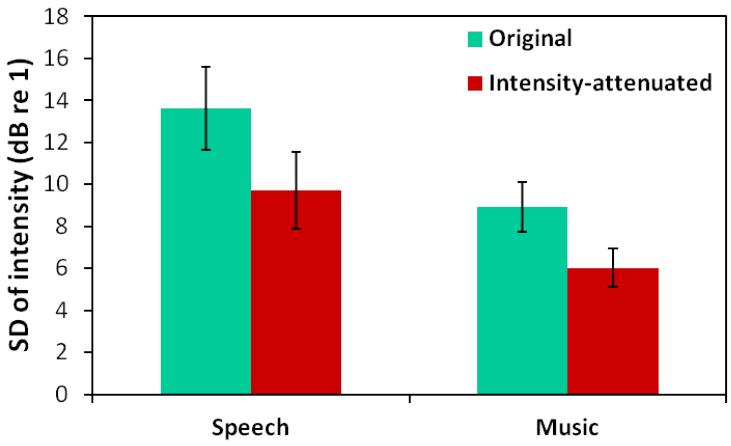


Figure 10: Mean intensity variation (mean SD of dB re 1) for the speech and music stimuli, before and after the frequency attenuation process. Error bars indicate ±1 standard deviation.

Euclidean distance between the corresponding points in sequences A and B (Berndt & Clifford, 1994). The algorithm can be formulated as such, where δ denotes some measure of distance:

$$DTW(A, B) = min_w[\sum_{i=1}^{n} \delta(w_i)]$$

In this case, DTW was used to time-align each emotionally expressive sentence or melody with its neutral-intonated counterpart (a corresponding version of the same sentence or melody, spoken/ performed with emotionally neutral expression). The emotional sentences and melodies were resynthesised according to the DTW vector, W, resulting in emotionally expressive stimuli whose temporal modulation profiles were shifted towards those of the corresponding neutral stimuli. Results of the DTW procedure to attenuate temporal variation in the speech and music stimuli are illustrated in Figures 11 and 12, respectively.

For all stimuli versions, where applicable, auditory features were normalised in the following order: Duration, Frequency, then Intensity. After processing to attenu-

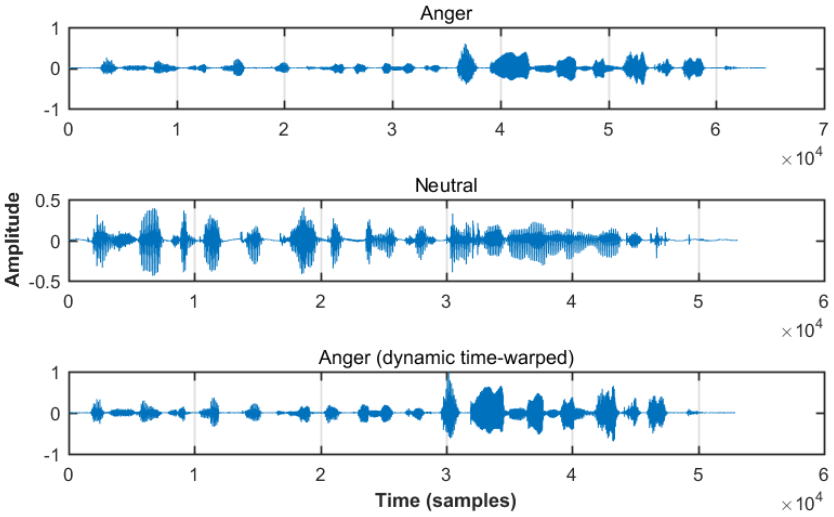


Figure 11: Illustration of the effects of DTW on the speech stimuli. Waveforms are shown for an example original emotional (in this case, Anger) stimulus, corresponding neutral stimulus, and temporal-variance-attenuated stimulus, resultant from the DTW process.

ate variation in the different auditory features, the mean intensities of all stimuli were scaled to 70 dB SPL, removing any unwanted inter-stimulus intensity variation resulting from the processing.

4.4.4 Notes about the stimuli used

Although the aforementioned processing inevitably leads to reduced ecological validity of the stimuli, research has suggested that auditory cue utilisation during emotion detection is very similar, even in synthesised speech-like tone sequences, when compared to natural speech (Scherer & Oshinsky, 1977). Despite the reduction in ecological validity inherent in using such stimuli, the results obtained were informative with respect to listening to normal speech, since listeners approached the task in a very similar way. Additionally, research has shown that speech stimuli remain intelligible, even when drastic alterations are made to the acoustic signal, e.g. resyn-

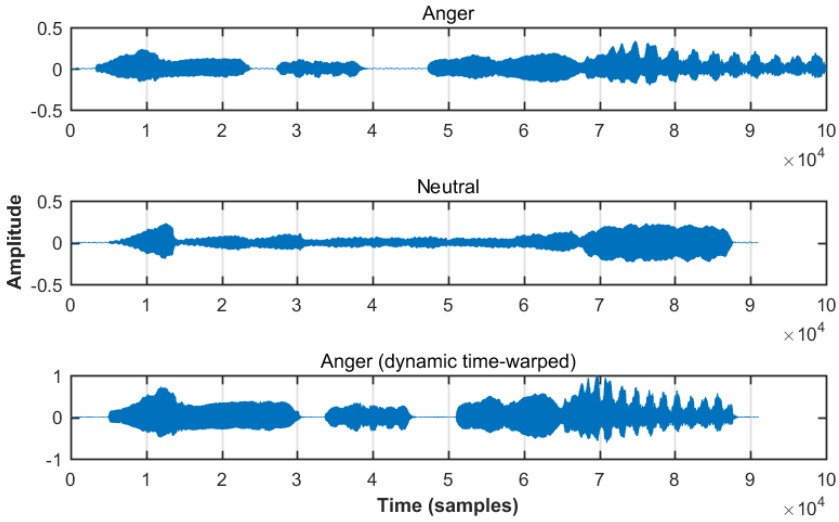


Figure 12: Illustration of the effects of DTW on the music stimuli. Waveforms are shown for an example original emotional (in this case, Anger) stimulus, corresponding neutral stimulus, and temporal-variance-attenuated stimulus, resultant from the DTW process.

thesis of formants as sinusoids (Remez, Rubin, Berns, Pardo, & Lang, 1994) or time reversal of stimulus sub-segments (Saberi & Perrott, 1999). In music perception, too, melodic recognition in aided listening may be effectively evaluated using sinusoids in place of traditional musical instruments (e.g. Digeser, Hast, Wesarg, Hessel, and Hoppe, 2012). Therefore, despite the processing to the speech and music rendering them less natural-sounding, the underlying mechanisms used for the task in the current study are expected to be similar to their unprocessed equivalents (excepting of course, that the different feature-attenuation conditions should affect participants' listening strategies).

Importantly, attenuation of physical acoustic properties is not necessarily synonymous with attenuation of their perceptual equivalents. This is due to the inherent discrepancy between raw physical signals emitted by external stimuli and our perception thereof. For example, selective attenuation of intensity variation does not preclude

variation in stimuli loudness (the perceptual 'homologue' of intensity), which may be affected by both physical acoustical properties e.g. frequency (Fletcher, 1934), and higher-level cognitive factors such as stimuli meaningfulness (Mershon, Desaulniers, Kiefer, Amerson, & Mills, 1981). Similarly, the perceived 'virtual pitch' of a complex tone, as present in speech and musical sounds, does not correspond exactly to F_0, but is also influenced by harmonic content, among other factors (Terhardt, 1979). Therefore, although physical auditory features were demonstrably attenuated in the stimuli described above, it is possible that a portion of the variance in corresponding perceptual auditory features may have remained, which would potentially be capable of influencing participants' emotion judgements.

Unfortunately this represents a necessary compromise, since attenuation of both physical and perceptual attributes simultaneously and in exactly-equal proportion is not realisable (e.g. removing variation in stimuli loudness, as much as possible, would entail making many intricate modifications to stimulus intensity, to account for frequency variations within the signal). Nonetheless, for the simple acoustic feature processing performed here, one can be relatively confident that attenuation of physical stimulus properties led to approximately similar attenuation of their perceptual correlates. To illustrate this, modulations in instantaneous loudness (measured in phons) were reduced in the intensity variation-attenuated stimuli, in a very similar way to intensity itself (Figure 13). This calculation was performed in MATLAB, following the time-varying loudness model of Glasberg and Moore (2002), which works by dividing the signal into groups of similar frequencies, called Equivalent Rectangular Bandwidths (ERBs), and summing the loudness values produced by each group (Moore, 2003).

Similarly, while, in their respective conditions, auditory information about variation

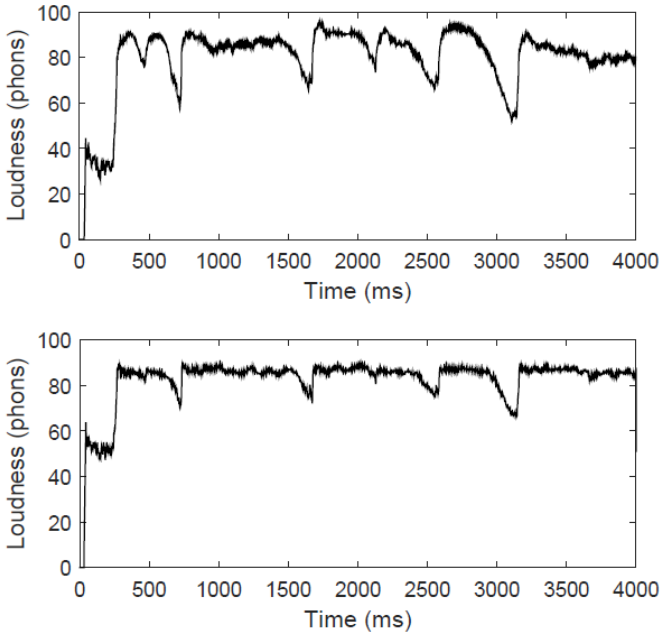


Figure 13: Illustrative example of loudness variation in an intensity-attenuated music stimulus. Upper panel = pre-attenuation, lower panel = post-attenuation.

in frequency, intensity and timing was preserved, this does not guarantee that these attributes were perceived in the same way across the various conditions. Auditory perception does not correspond with absolute congruity to the available acoustic cues (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). For example, although the perception of loudness is reliably associated with signal intensity, it may also be informed by subtle spectral changes, resultant from differential vocal characteristics associated with articulating a louder/ quieter sound (Rosenblum, 2004). Concretely, when a listener attempts to decode an emotion based on perceived loudness, they are likely to consider concomitant factors such as the spectral consequences of vocal effort (Ladefoged & McKinney, 1963). Therefore, where intensity information was preserved, and other features attenuated, one cannot necessarily infer that the perception of loudness was preserved verbatim. In summary, the acoustic features manipulated in these studies do not exist in a vacuum – rather, the perception

of various attributes of sound depends upon interactions between different acoustic features, and therefore the selective attenuation of these features may have had unintended consequences for the perception of the stimulus in question. This problem is related fundamentally to the nature of auditory perception, and would be difficult to circumvent without the use of artificially synthesised stimuli. Nonetheless, it should be kept in mind when interpreting the results of these studies.

Finally, a caveat of the stimuli created is that global differences in stimulus duration and frequency remained. For example, whilst intra-stimulus variation was attenuated, the 'angry' speech stimuli may have had a higher mean frequency than the sad speech stimuli. The primary motivation for retaining these differences was to ensure that the stimuli sounded somewhat 'natural'. While it is trivial to present all stimuli at the same mean intensity, changing the overall duration and median frequency of each sound too drastically would have produced a much more artificial-sounding stimulus set, thereby reducing ecological validity. Additionally, models of emotion perception in speech typically demonstrate better performance when using local rather than global prosodic features (Rao, Koolagudi, & Vempada, 2013) – in fact, the latter may instead encode task-irrelevant characteristics of the speaker, e.g. gender, identity etc. (Nogueiras, Moreno, Bonafonte, & Mariño, 2001). In music too, local structures may be more perceptually relevant for the expression of emotion than global structures (Tillmann & Bigand, 2004). This applies strictly to the performance, rather than composition of music – global compositional factors like mode also play an enormous role in conveying emotion (Juslin & Laukka, 2003). In any case, though global differences between stimuli might be sufficient for emotion recognition, the current study was primarily concerned with investigating the effects of attenuating the microstructure.

4.4.5 Simulation of cochlear implants

Because a sample of CI users was not readily available for participation in this study, a simulation approach was used in order to approximate the effects of CI processing using NH listeners. Therefore, for each of the stimuli resulting from the aforementioned stimuli manipulations, two versions were created: the original stimulus and a CI-simulated stimulus.

In previous research, the effects of the CI have most often been modelled using noise-band vocoding (NBV) (Friesen, Shannon, Baskent, & Wang, 2001), a method in which a given input signal is represented by broad bands of noise centred at a specified number of frequencies, which are modulated temporally by envelopes extracted from corresponding frequency bands in the original signal (Shannon et al., 1995). Although more detailed, physiologically-accurate models (i.e. containing a representation of the human auditory system and accounting for inter-individual differences) have been proposed with some success (Fredelake and Hohmann, 2012; Stadler and Leijon, 2009), research has shown that NBV-based modelling provides a good general approximation of CI users' performance in emotion recognition, and is additionally able to predict the specific patterns of errors typically made by CI users (Chatterjee et al., 2015). The NBV approach is therefore suitable for the current studies for two reasons: A) the studies aim to provide NH listeners with a general experience of the effects of the CI, as opposed to modelling the experience of any specific CI user; and B) it is important that the patterns of errors in emotion perception made by CI users will be comparable to those made by simulated CI listeners, which the available evidence suggests is the case.

However, the use of NBV-based CI simulation is not without limitations. Winn, Chatterjee, and Idsardi (2012) found that, in a speech recognition task, CI listeners

used spectral cues such as formant structure, formant change and consonant voicing significantly less than NH controls listening via an NBV CI simulation, instead relying more heavily upon durational cues. Although this task was not related to emotion perception per se, it is worth considering the researchers' caution that, although NBV models with NH individuals might adequately approximate CI users' performance, the underlying listening strategies adopted by each group may differ somewhat. With this said, the similarities between CI users and CI-simulated listeners, in terms of confusions made during emotion discrimination, does suggest comparable underlying strategies (Chatterjee et al., 2015).

A further caution that should be borne in mind when considering CI simulation is that no model is likely to fully account for the highly heterogeneous capabilities of CI users. Due to variance in implant model, processing strategy, age at implantation, general cognitive ability and a wealth of other factors, CI users as a group tend to display vastly varying capabilities, even in relatively simple auditory tasks. For example, in a sample of eleven CI users, accuracy in melodic contour identification varied between 14% and 91% (Galvin et al., 2007). Thus, whilst simulation of the CI provides a satisfactory estimate of the broad deficits experienced by users, the approach is ill-equipped to capture inter-individual differences, and is not a substitute for the study of real users. However, simulation studies can provide a useful starting point, offering valuable information about how, on average, listeners cope with the kinds of signal degradation associated with the CI.

For these studies, CI simulation was achieved via noise-band vocoding, as developed for this purpose by Shannon et al. (1995). This approach was implemented using MATLAB. First, based upon physiological parameters obtained from CI users, input signals were divided into twenty-one logarithmically-spaced frequency bands, begin-

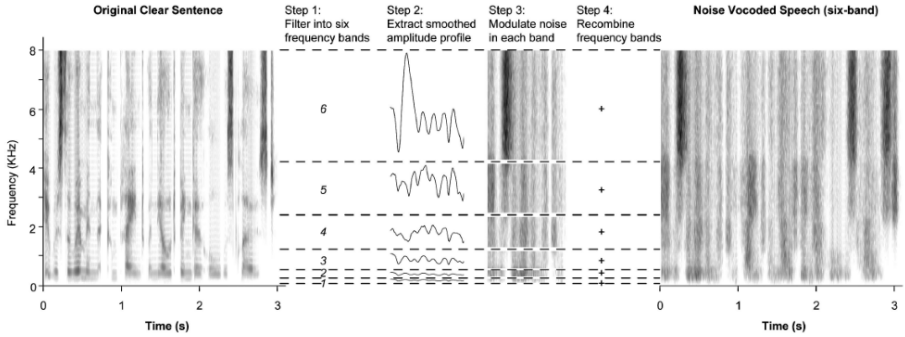


Figure 14: Illustration depicting original and noise-band vocoded speech as spectrograms, and the processing steps involved in this transformation. Note that, in this example, only six frequency bands were used. Reproduced from Davis et al. (2005).

ning at centre frequency 250 Hz and ending at centre frequency 5,400 Hz. Each of these bands was filtered using a fourth order Chebyshev Type II low-pass, with 200 Hz cut-off, in order to remove high-frequency energy, leaving only the amplitude envelope. For each frequency band, a band of white noise centred at the corresponding frequency (i.e. the 'carrier') was modulated according to the extracted envelope (Figure 14). The different bands of noise were then recombined, and the output signal was scaled to the same RMS as the input.

To briefly recap, in total there were two hundred total speech stimuli, consisting of: four different sentences × five emotions (Anger, Fear, Happiness, Neutral, Sadness) × five auditory feature conditions (Articulation, Duration, Frequency, Intensity, Original) × two CI conditions (CI-simulated, non-CI-simulated). Likewise, there were two hundred music stimuli, comprised of: four different melodies × five emotions × five auditory feature conditions × two CI conditions.

4.4.6 Experimental procedure

With the exception of the stimuli used, Studies 1 and 2 were methodologically identical, and therefore are described here simultaneously.

Prior to both listening experiments, two brief questionnaires were administered: the Forty-item Empathy Quotient (Baron-Cohen & Wheelwright, 2004), and a brief demographic questionnaire including questions pertaining to musical experience (i.e. instrument played, years' experience, years' music lessons etc.). These measures were included in order to examine potential interactions with listening strategies in emotion perception, since previous studies have documented correlation between auditory emotion perception accuracy and both empathic responding (Philip et al., 2010; Stevens, Charman, and Blair, 2001) and musical ability (Lima and Castro, 2011; Thompson et al., 2004).

Stimulus presentation was controlled using Psyscope, installed on a 13.3-inch Apple Macbook computer. Participants were presented with auditory stimuli via AKG K550 headphones (speech excerpts in Study 1 and short melodies in Study Two), and were asked to indicate the emotion expressed by each stimulus via key-press (making a 5-AFC judgement of either: anger, sadness, happiness, fear or neutral), along with a confidence rating related to each judgement, using a five-point Likert scale. For the sake of simplicity, neighbouring keys on the computer keyboard: Q, W, E, R and T were used for emotion judgements (with labels: A, F, H, N and S, attached to the respective keys to denote the different emotions). This ensured that participants' reaction times would not be less affected by their degree of familiarity with the QWERTY keyboard layout. To record confidence ratings for each judgement, numerical keys 1 to 5 were used.

Prior to any of the experimental trials, participants were given explicit instruction to focus on the emotions that they perceived, irrespective of their own potential affective responses. Participants first completed a block of ten practise trials, to familiarise them with the nature of the stimuli. The main experiments each consisted of two hundred trials, divided into two blocks of one hundred, between which participants were encouraged to take a five-minute break in order to prevent fatigue. The emotion judgement task was self-paced, advancing automatically after each judgement was registered, though participants were encouraged to respond as quickly and accurately as possible. Judgement accuracy, confidence ratings, and reaction times for each response were recorded as outcome measures.

4.5 Hypotheses

In line with previous research, it was hypothesised that emotion perception accuracy in Studies 1 and Two would be markedly reduced by the CI simulation, but would remain above chance level. In both studies, it was also hypothesised that detection accuracy would vary significantly across the different auditory feature conditions, and that these conditions would be affected differentially by the CI simulation. Firstly, it was expected that the non-feature-attenuated 'Original' stimuli would be associated with the greatest emotion recognition accuracy, both with and without the simulation. More specifically, it was predicted that emotion recognition for 'Frequency' stimuli (signals with preserved frequency variation but attenuated intensity and temporal variation) would be affected to a larger extent by the CI simulation, whilst 'Duration' and 'Intensity' stimuli would be less strongly affected. This was because the CI simulation, like the CI itself, has a much stronger deleterious effect upon frequency information, compared to the other acoustic cues. Accordingly, as-

suming that participants would adopt listening strategies focussed upon relatively preserved components of the CI-simulated stimuli, it was predicted that performance in the CI condition would be better for Duration and Intensity stimuli, relative to Frequency.

It was expected that decoding accuracy would vary across the different emotions being judged, and further that these may be differentially affected by the CI simulation. This was because of variance in the types of cues that are most strongly associated with different emotions, and their differential preservation in the CI condition. For example, if fear happened to be associated most strongly with an erratic melodic contour, then this emotion might not be clearly conveyed by the CI simulated stimuli. For this reason, it was considered that a three-way interaction between auditory feature processing, emotion and CI conditions might occur, although the lack of previous literature rendered it difficult to make more specific predictions regarding this. In general, it was expected that task performance would be best in cases where emotions were easily identifiable via non-frequency-based cues, and where these cues were preserved by the auditory feature processing. Therefore, for example, sadness was expected to be relatively well-decoded.

Lastly, in line with previous research, and because emotion recognition in speech is likely a more pervasive task in everyday life – as compared with emotion recognition in music – it was predicted that task performance would be worse on the whole in Study Two. However, it was also expected that scores may be correlated with measures of musical experience (particularly in Study Two), and also that those individuals self-reporting greater empathic capabilities might be more accurate in decoding emotion.

4.6 Results

4.6.1 Data processing

Prior to statistical analysis, emotional judgement responses for each stimulus were converted to values of -1 or 1, based on whether they were incorrect or correct, respectively (i.e. whether the response given matched the emotion intended by the talker/ musician). Each value was then multiplied by the confidence ratings associated with that judgement (which varied from 1-5), such that responses ranged between -5 to 5.

4.6.2 Study 1: Speech

Preliminary analyses revealed no significant effects of stimulus sentence or speaker gender on emotion detection accuracy, hence these variables were collapsed for subsequent analyses. Participants' scores were summed across these variables to produce only one value for each unique combination of: Acoustic feature processing condition $\times$ Emotion $\times$ CI condition (so that the two hundred responses originally recorded were re-encoded as only fifty). As a result, these scores (correctness multiplied by confidence rating) varied between -20 to 20 for the following analyses, and are hereafter referred to as 'judgement scores'. Before running further analyses, it was confirmed that these judgement scores were normally distributed.

For the interpretation of the results to follow, chance level performance was estimated as corresponding to -7.2. This was based on the assumption that chance level for confidence ratings would be 3 (i.e. the midpoint of the 1-5 scale used). For each of the four sentences in each condition (after collapsing across sentence and speaker gender), the odds of choosing the correct emotion by chance were 20% (since there were five response options), and the odds of choosing an incorrect emotion 80%.

Therefore, multiplying these odds by the response 'correctness' and the assumed confidence rating of 3, chance performance for the judgement score was considered equal to $3(.2 \times 1 \times 4) - 3(.8 \times -1 \times 4) = -7.2$.

Three-way repeated measures ANOVA revealed a significant effect of auditory feature processing condition upon participants' judgement scores, $F(4, 64) = 61.77$, $p < .001$, $\eta_p^2 = .79$. There were also significant main effects of emotion, $F(4, 64) = 19.28$, $p < .001$, $\eta_p^2 = .55$ and CI condition, $F(1, 16) = 80.04$, $p < .001$, $\eta_p^2 = .83$. Additionally, ANOVA revealed a significant interaction between auditory feature processing condition and CI condition, $F(4, 64) = 27.47$, $p < .001$, $\eta_p^2 = .63$ (Figure 15). Planned contrasts determined that, relative to the Original condition, scores in the Articulation, Duration and Intensity conditions were affected significantly less by the CI manipulation, whereas scores in the frequency condition were not (Table 3).

The ANOVA also revealed a significant interaction between emotion and CI condition, $F(4, 64) = 15.15$, $p < .001$, $\eta_p^2 = .49$ (Figure 16). Planned contrasts showed that, compared to Neutral stimuli, judgement scores for Anger, Fear and Happiness were affected significantly less by the CI manipulation, whereas scores for sadness were not (Table 4). Of the different emotions, Neutral and Sadness stimuli were best recognised without the CI simulation, and therefore the reduction in judgement scores with simulation appeared more drastic.

Table 3: Planned contrasts illustrating how participants' emotion judgement scores in the different auditory feature processing conditions were affected by the CI manipulation, relative to the Original condition. * = significant at .050 alpha-level.

Condition	F	**df**	p	η_p^2
Articulation	44.62	1, 16	<.001*	.74
Duration	27.30	1, 16	<.001*	.63
Frequency	4.30	1, 16	.055	.21
Intensity	49.04	1, 16	<.001*	.75

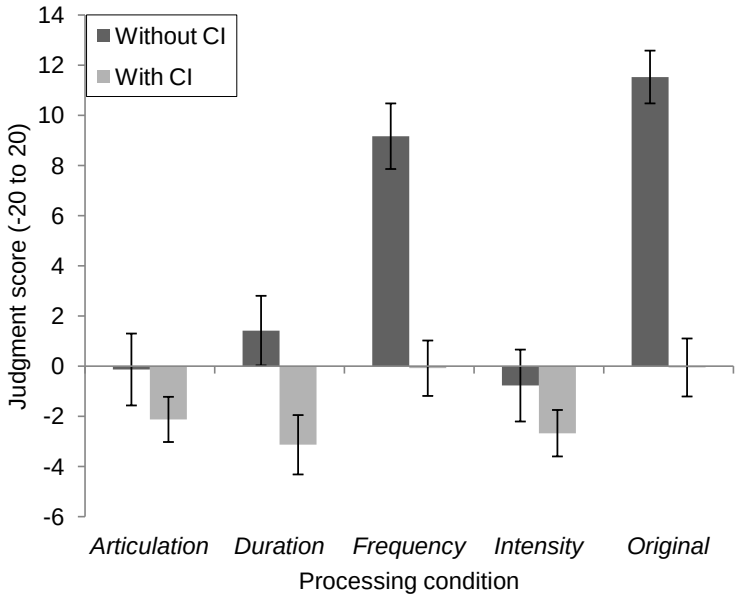


Figure 15: Interaction between auditory feature processing condition and CI condition, for participants' emotion judgement scores. Error bars denote standard error of the mean.

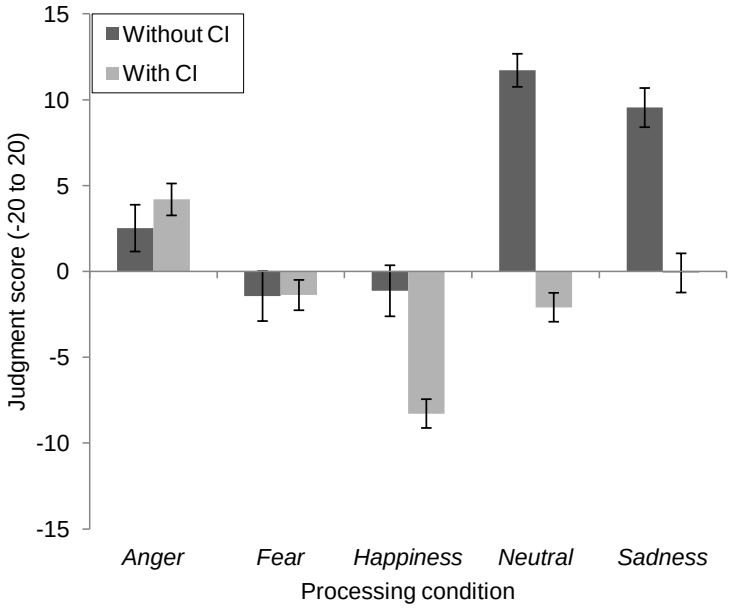


Figure 16: Interaction between emotion and CI condition, for participants' judgement scores. Error bars denote standard error of the mean.

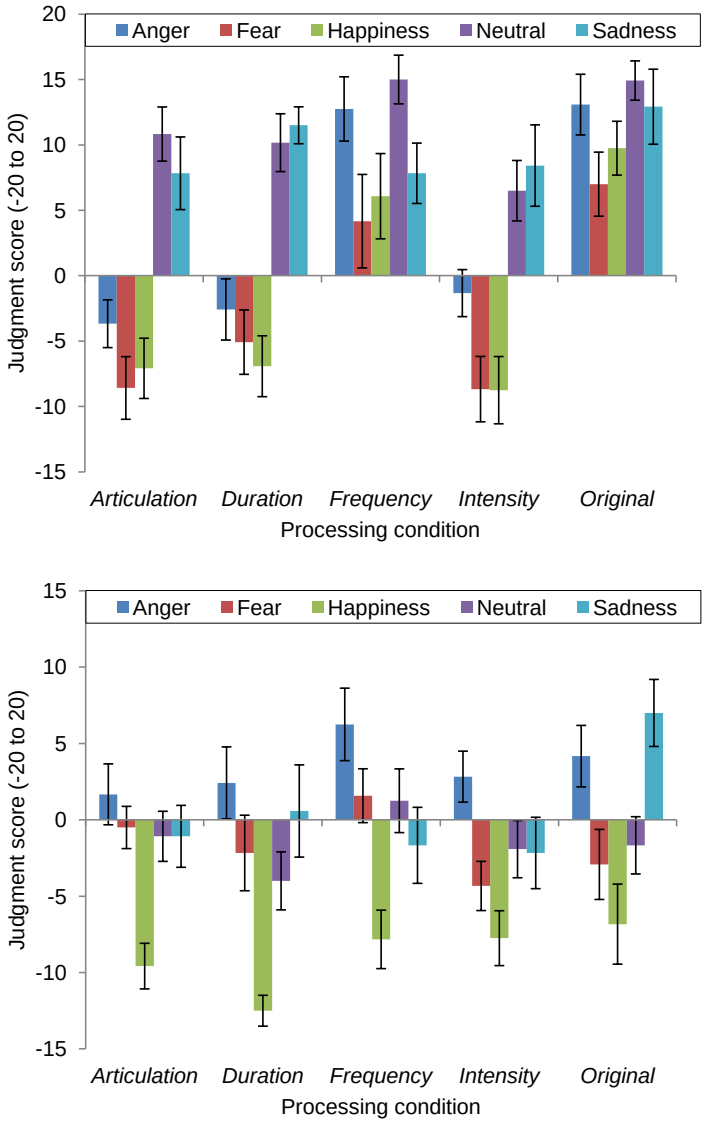


Figure 17: Interaction between auditory feature processing condition, emotion and CI condition, for participants' emotion judgement scores. Upper panel = non-CI stimuli, lower panel = CI stimuli. Error bars denote standard error of the mean.

Table 4: Planned contrasts illustrating how participants' judgement scores for the different emotions were affected by the CI manipulation, relative to the Neutral stimuli. * = significant at .050 alpha-level.

Emotion	F	**df**	p	η_p^2
Anger	33.26	1, 16	<.001*	.68
Fear	30.51	1, 16	<.001*	.66
Happiness	10.29	1, 16	.005*	.39
Sadness	3.47	1, 16	.081	.18

Table 5: Planned contrasts illustrating how participants' emotion judgement scores were affected by the CI manipulation in the different auditory feature processing conditions, relative to the Original condition, for each of the different emotions. * = significant at .050 alpha-level.

Condition	**Emotion**	F	**df**	p	η_p^2
Articulation	Anger	17.47	1, 16	.001*	.52
	Fear	126.09	1, 16	<.001*	.62
	Happiness	12.24	1, 16	.003*	.43
	Sadness	0.15	1, 16	.703	.01
Duration	Anger	12.84	1, 16	.002*	.45
	Fear	12.62	1, 11	.003*	.44
	Happiness	9.49	1, 16	.007*	.37
	Sadness	0.69	1, 16	.417	.04
Frequency	Anger	0.01	1, 16	.940	<.01
	Fear	4.01	1, 16	.062	.20
	Happiness	0.17	1, 16	.685	.01
	Sadness	0.83	1, 16	.777	.01
Intensity	Anger	7.69	1, 16	.014*	.33
	Fear	8.91	1, 16	.009*	.36
	Happiness	10.32	1, 16	.005*	.39
	Sadness	1.90	1, 16	.187	.11

Lastly, the ANOVA revealed a significant three-way interaction between Emotion, Auditory feature processing condition and CI condition, $F(16, 256) = 4.77$, $p < .001$, $\eta_p^2 = .23$ (Figure 17). Planned contrasts were computed to inspect this relationship in more detail (Table 5). In the Articulation, Duration and Intensity conditions, relative to the Original condition, emotion judgement scores were affected significantly differently by the CI manipulation for Anger, Fear and Happiness stimuli compared to Neutral. In the Frequency condition, judgement scores were not affected signifi-

cantly differently by the CI manipulation in any of the emotional conditions.

To examine in more detail the patterns of errors made by participants during the emotion judgement task, confusion matrices were constructed using the R package *caret* (Kuhn, 2008). These depict participants' judgement accuracy for the different emotions in each of the auditory feature precessing conditions, and for both non-CI-simulated and CI-simulated conditions (Figure 18). These matrices were constructed considering only the correctness of participants' responses – confidence ratings were not taken into account here.

Figure 18: Heat-mapped confusion matrices for the speech stimuli, depicting the percentage of responses in each emotion category, for each stimulus emotion. Columns denote presented emotions, rows denote emotion judgement responses. Red= higher values, green lower values.

In the Original, Non-CI condition, emotion judgement accuracy was comparable to

that reported by Burkhardt et al. (2005) for the same stimuli, with the exception that Fear was slightly less well-recognised here. Without CI simulation, the Frequency condition produced the pattern of errors most similar to the Original stimuli. However, specific emotions were affected differentially by the feature attenuation conditions. For example, Anger and Happiness were well-decoded in the Frequency condition, but often confused otherwise. Conversely, Sadness stimuli were relatively robust to distortion, but appeared to be better-preserved in the Duration condition.

The matrices associated with the CI-simulated stimuli showed a less clear pattern. Again, emotions were decoded most accurately in the Original and Frequency conditions. Sadness and Anger were relatively well-decoded across the different processing conditions, although the latter was very commonly confused with Happiness. As with the non-CI-simulated stimuli, Sadness was recognised more often in the Duration condition, compared to the other feature-attenuation conditions, although the differences between conditions were relatively small.

Interestingly, the confusion matrices also revealed some more general response biases caused by the CI simulation. While Anger appeared to be recognised primarily by frequency information in the non-CI-simulated stimuli, the emotion was consistently well-recognised across all feature-attenuation conditions for the CI-simulated stimuli. Therefore, the simulation itself may have been interpreted as a cue denoting Anger. Across all stimuli and all feature-attenuation conditions, participants were much more likely to respond with Anger, Fear or Neutral for the CI-simulated stimuli. On the contrary, respondents were less likely to judge these stimuli as conveying Happiness.

Lastly, no significant correlations were found between the overall emotion judgement scores achieved by each participant and responses to any of the ad hoc measures of

musical training or musical engagement. Likewise, no significant correlation was observed between overall judgement scores and participants' results for the self-report-based empathy measure.

4.6.3 Study 2: Music

Similarly to Study 1, there were no significant effects of stimulus melody or instrumentation, therefore these variables were collapsed for subsequent analyses. Thus, participants' scores were summed across these variables to produce only one value for each unique combination of: Acoustic feature processing condition × Emotion × CI condition. As in Study 1, judgement scores (correctness multiplied by confidence rating) varied between -20 to 20 for the following analyses. Prior to running further analyses, it was confirmed that these judgement scores were normally distributed.

As in Study 1, chance level performance was estimated as corresponding to -7.2. For each of the four melodies in each condition (after collapsing across melody and musical instrument), the odds of choosing the correct emotion by chance were 20% (since there were five response options), while the odds of choosing an incorrect emotion were 80%. Multiplying these odds by response 'correctness' and the assumed confidence rating of 3, chance performance for the judgement score was $3(.2 \times 1 \times 4) - 3(.8 \times -1 \times 4) = -7.2$.

Overall, participants' average reaction times in Study 2 (global mean RT from stimulus onset = 8.47, or mean stimulus length + 1.84 seconds) were very similar to Study 1 (mean RT = 5.04, or mean stimulus length + 1.92 seconds), once differences in stimuli lengths were taken into account. However, confidence ratings were on average half a point lower in Study 2 (mean = 2.89, compared to 3.44), suggesting that participants had greater difficulty in decoding emotion from musical stimuli.

A three-way repeated measures ANOVA revealed significant effects of both auditory feature processing condition, $F(4, 56) = 4.19$, $p = .005$, $\eta_p^2 = .23$, and emotion, $F(4, 56) = 6.35$, $p < .001$, $\eta_p^2 = .31$ upon participants' judgement scores. However, the effect of CI condition upon judgement scores was not significant, $F(1, 14) = 3.66$, $p = .076$, $\eta_p^2 = .21$, nor was the interaction between CI condition and Auditory feature processing condition, $F(4, 56) = 2.44$, $p = .058$, $\eta_p^2 = .15$.

Nonetheless, relative to the Original condition, judgement scores in the Duration conditions were affected to a noticeably lesser extent by the CI manipulation, whilst scores in the Articulation, Frequency and Intensity conditions all showed at least some deterioration as a consequence of CI simulation (Figure 19). Notably, compared to the results of Study 1, scores were generally negative, reinforcing the observation that participants found this task much more difficult with music stimuli. It should be noted, however, that scores were mostly above the estimated level of performance expected by chance (-7.2).

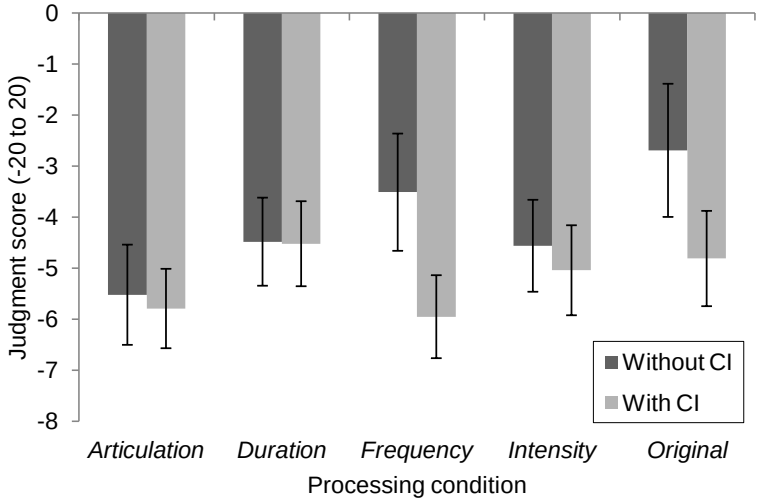


Figure 19: Interaction between auditory feature processing condition and CI condition, for participants' judgement scores. Error bars denote standard error of the mean.

The ANOVA also revealed a significant interaction between Emotion and CI condi-

tion, $F(4, 56) = 7.34$, $p < .001$, $\eta_p^2 = .34$ (Figure 20). Planned contrasts determined that, compared to Neutral stimuli, judgement scores for Anger and Fear were affected significantly differently by the CI manipulation (i.e. performance was better, rather than worse, with CI simulation), whereas scores in the remaining emotional conditions were not (Table 6).

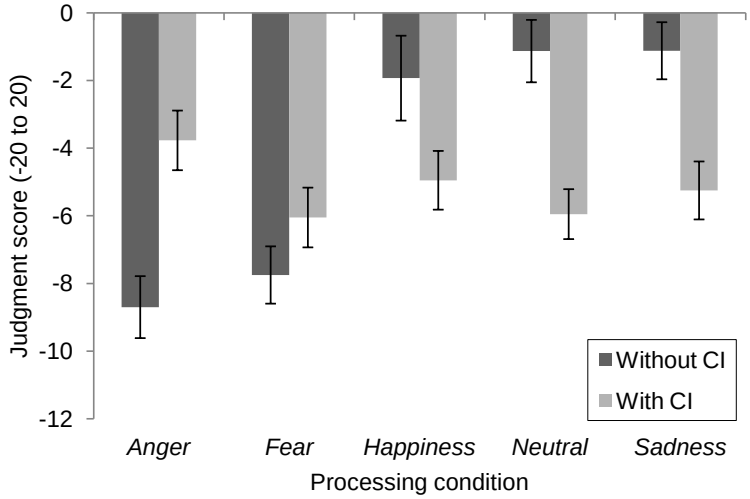


Figure 20: Interaction between Emotion and CI condition, for participants' emotion judgement scores. Error bars denote standard error of the mean.

Table 6: Planned contrasts illustrating how participants' judgement scores for the different emotions, relative to the Neutral stimuli, were affected by the CI manipulation.

Emotion	$\boldsymbol{F}$	**df**	$\boldsymbol{p}$	$\boldsymbol{\eta_p^2}$
Anger	12.92	1, 14	.003*	.48
Fear	5.24	1, 14	.038*	.27
Happiness	0.92	1, 14	.354	.06
Sadness	0.03	1, 14	.877	<.01

* = significant at .050 alpha-level.

Lastly, the ANOVA revealed a significant three-way interaction between emotion, auditory feature processing condition and CI condition, $F(16, 224) = 2.76$, $p < .001$, $\eta_p^2 = .17$ (Figure 21). Planned contrasts were computed to inspect this relationship in more detail (Table 7). In the Articulation condition, relative to the Original condi-

tion, judgement scores were affected significantly differently by the CI manipulation only for Happiness and Sadness stimuli, compared to Neutral. In the Duration and Frequency conditions, judgement scores were affected significantly differently by the CI manipulation only for Happiness stimuli. By contrast, judgement scores were affected significantly differently by the CI manipulation for Anger, Fear and Happiness stimuli in the Intensity condition. In summary, without CI simulation, in the Original and Frequency conditions Happiness and Sadness were by far the emotions most often identified, and therefore identification accuracy in these particular conditions appeared to be especially affected by the CI manipulation.

To examine more closely the patterns of errors made by participants during the emotion judgement task, confusion matrices are shown below for each of the auditory feature precessing conditions, and for both non-CI-simulated and CI-simulated conditions (Figure 22).

Foremost, not all of the emotions were consistently decoded, with Anger and Fear stimuli never reaching above 40% recognition accuracy (irrespective of the feature-attenuation and CI simulation conditions), in contrast to the results reported by Quinto et al. (2014) with the same stimuli. In fact, Anger and Fear were seldom chosen as responses, while Neutral was overestimated – most likely an index of general uncertainty.

As with speech stimuli in Study 1, without CI simulation, the Frequency condition produced the pattern of errors most similar to the Original stimuli. In these conditions, relative to Articulation, Duration and Intensity, participants tended not to overestimate the Neutral stimuli, suggesting that participants perceived these stimuli as being more expressive. However, specific emotions were again affected differentially by the feature attenuation conditions. For example, Sadness was equally

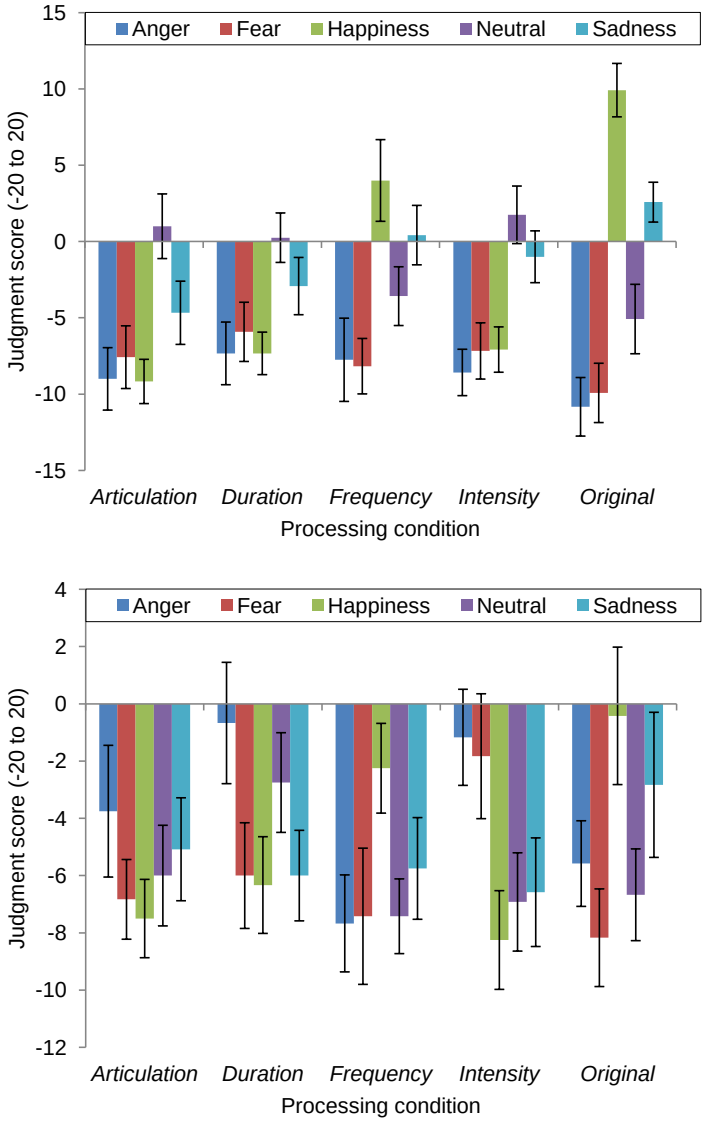


Figure 21: Interaction between auditory feature processing condition, emotion and CI condition, for participants' emotion judgement scores. Upper panel = non-CI stimuli, lower panel = CI stimuli. Error bars denote standard error of the mean.

well-decoded in both the Intensity and Frequency conditions, whilst Happiness was decoded much better in the latter condition. As in Study 1, Sadness stimuli were on the whole the most robust to distortion, but appeared to be better-preserved in the Duration condition.

The matrices associated with the CI-simulated stimuli showed a somewhat differ-

Figure 22: Heat-mapped confusion matrices for the music stimuli, depicting the percentage of responses in each emotion category, for each stimulus emotion. Columns denote presented emotions, rows denote emotion judgement responses. Red cells = higher values, green cells = lower values.

ent, more complicated pattern. Overall, emotions were decoded most accurately in the Original condition. Elsewhere, judgement accuracy varied heavily depending upon both the emotion and the feature-attenuation condition. Happiness was well-decoded in the Frequency condition, but otherwise was often confused, particularly with Anger. Conversely, Anger was decoded with greater accuracy in the Intensity and Duration conditions, even compared to the Original condition. Fear was also relatively well-recognised in the Intensity condition, although only 30% accuracy was achieved.

There were also some more general response biases caused by the CI simulation.

By comparison to the non-CI-simulated stimuli, Anger and Fear were much more readily offered as responses in the CI simulation condition. By contrast, Sadness was a far less common response in this condition. As with the non-CI simulated stimuli, Neutral stimuli tended to be overestimated, especially in the Articulation, Duration and Intensity conditions.

As in Study 1, there were no significant relationships between participants' emotion judgement scores and any measures of musical training or musical engagement. There was also no significant relationship between judgement scores and participants' self-reported empathy scores.

Table 7: Planned contrasts illustrating how participants' emotion judgement scores were affected by the CI manipulation in the different auditory feature processing conditions, relative to the Original condition, for each of the different emotions.

Condition	**Emotion**	F	**df**	p	η_p^2
Articulation	Anger	2.67	1, 14	.125	.16
	Fear	3.36	1, 14	.088	.19
	Happiness	30.87	1, 14	<.001*	.69
	Sadness	5.58	1, 14	.033*	.29
Duration	Anger	1.66	1, 14	.22	.11
	Fear	1.18	1, 14	.296	.08
	Happiness	11.51	1, 14	.004*	.45
	Sadness	2.63	1, 14	.127	.16
Frequency	Anger	0.70	1, 14	.417	.05
	Fear	0.03	1, 14	.856	<.01
	Happiness	5.79	1, 14	.031*	.29
	Sadness	<0.01	1, 14	.984	<.01
Intensity	Anger	11.02	1, 14	.005*	.44
	Fear	7.56	1, 14	.016	.35
	Happiness	25.25	1, 14	<.001*	.64
	Sadness	3.73	1, 14	.074	.21

* = significant at .050 alpha-level.

4.7 Discussion

4.7.1 Summary of results: Study 1

As hypothesised, the auditory feature processing had a significant impact on the accuracy with which participants were able to judge emotions conveyed in speech. Specifically, performance was best overall in the Original and Frequency conditions, supporting the primacy of frequency information when perceiving emotional speech (Petrushin, 1999).

Emotion judgement accuracy was significantly reduced by the CI simulation, as predicted by studies comparing the perception of emotional speech by NH and CI listeners (Chatterjee et al., 2015; Luo et al., 2007). However, performance in the CI condition remained generally above chance level, in line with studies evaluating the ability of CI listeners to recognise emotion in speech (Chatterjee et al., 2015; Volkova et al., 2013). The results obtained here extend these findings further, demonstrating that above-chance performance is not only possible with CI simulation, but can occur even with non-binary response options and even with additional processing to attenuate different auditory features.

As expected, auditory feature processing condition and CI condition interacted significantly with respect to judgement scores. This interaction appeared to be largely driven by better recognition performance in the Frequency and Original conditions to begin with, meaning that the CI manipulation was comparatively more detrimental. In other words, because performance was better in these two conditions without the CI simulation, the addition of the simulation made a more drastic difference than for the other conditions, which were already recognised comparably poorly. Somewhat surprisingly however, judgement scores remained highest for Frequency and Original

stimuli in the CI condition. This suggests that participants did not shift their attention to those acoustic features relatively better-preserved by the simulation, as has been reported previously (Peng et al., 2009; Peng et al., 2012), but instead relied upon the residual frequency information available (which may not, in principle, be the most effective listening strategy).

Emotional judgement accuracy also varied depending on the emotion conveyed. That is, some emotional states were more readily perceived than others. Specifically, judgement scores were significantly worse for Fear and Happiness stimuli, whilst Anger and Sadness were comparably well-recognised. Previous research has documented similar patterns of performance in the perception of emotional speech (Scherer et al., 1991), by both Western and non-Western listeners (Scherer et al., 2001), suggesting that the acoustic features contributing to the perception of Fear and Happiness may be inherently less salient. Although it is not immediately obvious why this should be the case, this observation was corroborated by the current study.

The auditory feature processing conditions had different effects upon judgement scores for each of the different emotions conveyed. In general, participants were similarly accurate and made similar errors in the Frequency and Original conditions, although participants were significantly less accurate in the former condition when decoding Sadness stimuli. This accords with the notion that frequency information is especially important for the identification of emotion in speech (Petrushin, 1999), though the reduced speech rate that is characteristic of sadness (Breitenstein, Lancker, & Daum, 2001) may have been a more important cue for this particular emotion. Accordingly, the confusion matrices showed that Sadness stimuli were more accurately decoded in the Duration condition. However, relative to the Original condition, the Articulation, Duration and Intensity conditions were associated with

significantly reduced judgement accuracy overall, and in particular for Fear and Happiness. This most likely occurred because participants found these emotions most difficult to disambiguate when frequency variation was not preserved.

An interaction between Emotion and CI condition showed that judgement scores for some emotional stimuli were affected differently by the CI simulation than others. While Happiness, Neutral and Sadness stimuli showed predictably reduced performance in the CI condition, Anger and Fear stimuli were relatively unaffected. In fact, judgement performance for the Anger stimuli was better on average in the CI condition compared to the non-CI condition. This might be explainable by a tendency of participants to appraise the CI-simulated stimuli in general as conveying anger, due to the harsh, distorted timbre characteristic of the manipulation. Therefore, what appears to be an increase in performance in the CI condition may be a result of many type 1 (i.e. 'false positive') errors in responses. Supporting this suggestion, the confusion matrices showed that, in the CI condition, participants overestimated Anger to a large extent, while also overestimating Fear and Neutral, and underestimating Happiness. Therefore, the CI simulation had two generic effects on responses: a bias in favour of Anger/ Fear because of the more dissonant timbre created, and a bias in favour of Neutral – most likely because of increased task difficulty, making it harder to perceive any particular emotion.

Lastly, there was a three-way interaction between Auditory feature processing condition, Emotion and CI condition, demonstrating that, in the different auditory processing conditions, judgement scores were not only affected differently by the CI simulation, but these differences varied depending on the emotion being conveyed. In the Articulation, Duration and Intensity conditions (by contrast to the Original condition), the CI simulation was associated with improved judgement scores for

Anger stimuli. Again, this is likely due to the characteristic timbre caused by the simulation being mistakenly interpreted as anger. Since Anger was not very well decoded in these conditions, the extent to which Anger was overestimated for CI-simulated stimuli meant that the emotion was actually correctly identified more often with the simulation. In fact, in the Articulation, Duration and Intensity conditions (relative to Original), judgement scores for Anger, Fear and Happiness stimuli were affected significantly less than Neutral stimuli by the CI simulation. Stimuli in these conditions were perceived relatively poorly on the whole, irrespective of the CI condition, implying that frequency information was highly important for the recognition of these emotions. By contrast, Sadness stimuli were affected by the CI condition fairly uniformly across the different feature attenuation conditions. This is likely because these stimuli were relatively robust to the feature attenuation processing, meaning that they were well-recognisable in all conditions prior to the CI simulation, and thereby were affected by it to similar degrees. In the Frequency condition, scores did not differ from Original for any emotion (in terms of the effect of CI simulation), implying that frequency was an important carrier of affective content across all of the stimuli.

There were no significant effects of musical training or experience upon emotion recognition accuracy, either generally or for any of the individual conditions. This appears to contradict the suggestion that musicianship confers an advantage in emotion detection in speech (e.g. Thompson et al., 2004). However, it may simply be that musical experience is less advantageous in CI-simulated listening conditions and/ or with auditory feature-attenuated stimuli. The lack of an effect may also reflect a limited diversity among participants in terms of extent of musical training. Most participants reported at least some musical experience, and therefore an effect might

be more readily observable with a binary comparison of musicians vs. non-musicians.

Similarly, there were no effects of individual differences in empathy, somewhat contradicting previous research which has implied a link between empathic abilities and proficiency in auditory emotion recognition (Philip et al., 2010; Stevens et al., 2001). However these studies considered cases of disordered empathic responding (autistic spectrum disorders and psychopathy, respectively), and so it is possible that a clear relationship was not observed here because all participants Empathy Quotient scores fell within a 'normal' range, hence the data were limited in variability.

4.7.2 Summary of results: Study 2

As hypothesised, the auditory feature processing had a significant effect upon the accuracy with which emotions were decoded in music stimuli – judgement performance was best overall in the Original and Frequency processing conditions. This was surprising, since the performance expression literature does not imply the same predominance of pitch-based information that is posited in the literature pertaining to emotional speech. Indeed, whilst pitch variations are acknowledged as an important component of expressive performance (Juslin, 2001), several studies have placed greater emphasis on other factors such as temporal or dynamic variance (Bhatara, Tirovolas, Duan, Levy, and Levitin, 2011; Juslin, 1997). However, it is possible that the preservation of frequency information, apart from its role as an expressive cue, was also particularly important in making the music sound 'natural', thereby (indirectly) facilitating judgements of emotional expression.

As in Study 1, emotion judgement accuracy was reduced by the CI manipulation, consistent with prior research pertaining to the perception of emotion in music by CI users (Hopyan et al., 2011; Hopyan et al., 2015; Luo et al., 2007). However, this

effect was not statistically significant, which may have been due to the relatively poor performance with non-CI stimuli leading to a floor effect. More specifically, confusion matrices showed that participants tended to respond with Happiness, Sadness or Neutral (which was drastically overestimated) across all feature attenuation conditions. Therefore, there was no 'main effect' of CI simulation per se, since some of the emotions were already very poorly decoded, thus the simulation did not appreciably disrupt recognition in these cases.

However, while performance was not above chance level in all conditions, recognition accuracy was relatively good for Happiness and Sadness stimuli in the original condition, corroborating the previous observation that CI users are able to make this binary distinction (Hopyan et al., 2015). Moreover, the current study extended this finding, showing that listeners were able to make this distinction with only performance cues available (as opposed to compositional cues), and even in some cases of auditory feature attenuation.

There was also an interaction between auditory feature processing and CI condition, such that performance in the Frequency and Original conditions was better overall and therefore was affected more strongly by the CI simulation. As in Study 1, the most parsimonious explanation for this interaction is that, since task performance was better in these conditions to begin with, there was therefore greater potential for the simulation to have a disruptive effect. By contrast, Articulation and Duration stimuli were only influenced minimally by the CI condition. As was the case in Study 1, participants may have been attending primarily to frequency-based cues across all of the stimuli, as opposed to selectively focussing on relatively preserved aspects of the acoustic signal.

There was a main effect of emotion, with Happiness, Neutral and Sadness stimuli

being comparably better-recognised than Anger or Fear. This is consistent with the prior observation that happiness and sadness are comparatively well-decoded in this type of paradigm (Quinto et al., 2014), and also with the observation that Neutral was overestimated in participants' responses. Concomitantly, the influence of Emotion interacted with the CI manipulation, such that judgement scores were reduced to a relatively greater extent for Happiness, Neutral and Sadness stimuli. In contrast, Anger stimuli were significantly better-recognised in the CI condition. As in Study 1, the most plausible explanation for this is in terms of type 1 (false positive) errors caused by the nature of the CI-simulated stimuli, with the distorted timbre caused by the simulation being misinterpreted as denoting anger.

Again, there was a three-way interaction between Auditory feature processing condition, Emotion and CI condition – i.e. the aforementioned interaction between Auditory feature processing and CI conditions varied across the different emotions. This effect appeared to be primarily driven by the much better performance recorded for Happiness and Sadness stimuli in the Original, non-CI condition. With the CI simulation, there was a tendency towards Anger and Fear being most accurately decoded in the Intensity and Duration conditions, which might reflect the relative importance of temporal or dynamic cues for this emotion. In fact, these emotions tended to be decoded best when frequency information was *not* preserved, implying that this information had some distracting, counteractive effect. However, the confusion matrices were also indicative of a general bias towards Anger (and to a lesser extent Fear) for all CI-simulated stimuli.

As in Study 1, there were no significant effects of musical training or experience upon emotion decoding accuracy. While musicians were expected to have an advantage in this paradigm, this result might reflect the observations that emotional responses to

music are only weakly influenced by expertise (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005), and that music performance rules (and interpretations thereof) vary vastly between different performers and listeners (Gabrielsson & Juslin, 1996). As alluded to in Study 1, the observed results may also stem from the fact that most participants were musically trained to at least a small extent – an effect might be more noticeable in a direct comparison between musicians and non-musicians.

Lastly, there was no significant effect of empathy upon judgement scores, which appears to contradict previous research linking the perception of emotion in music to wider empathic and socio-emotional capabilities (Saarikallio, Vuoskoski, & Luck, 2014). However, only relatively small effect sizes have been documented, relating to specific emotions (e.g. tenderness, which was not included in the current study), which may explain why these results did not translate to the current study.

4.8 Conclusions

Taken together, Studies 1 and 2 provide broad support for the idea that, under CI-simulated listening conditions, participants can identify expressions of emotion in speech and music at a level that is greater than chance. Moreover, these studies extend existing findings by demonstrating above-chance performance in a paradigm incorporating multiple response options, and including multiple auditory feature attenuation conditions.

Decoding accuracy appeared to vary across the different emotions being appraised, with Study 1 finding that emotion judgement performance with speech was best for Anger and Sadness stimuli. To some extent, this may be due to the fact that these emotions are characterised by particularly strong non-frequency-based cues, i.e. anger by increased intensity and sadness by reduced rate (Juslin and Laukka,

2003; Banse and Scherer, 1996). Since these cues are preserved relatively well by the CI simulation, the emotions associated with them may have been relatively well-decoded.

However, other than in the Original condition, participants in Study 1 demonstrated the greatest emotion judgement accuracy in the Frequency condition – irrespective of CI simulation. This suggests that participants did not adapt their listening strategies between the different conditions, i.e. to focus on acoustic attributes better-preserved by the CI, but instead continued to base their judgements largely upon frequency-based cues. A possible explanation for this is that, due to the purported primacy of pitch in emotional speech perception (Petrushin, 1999), participants were accustomed to a mode of listening that prioritised pitch cues, and did not have sufficient time or incentive (for example, in the form of feedback on their performance) during the study to adapt their listening strategies. Furthermore, because the different auditory feature conditions were intermixed in stimulus presentation, it may have been too difficult for participants to constantly shift their attention towards better-preserved features.

The three-way interaction in Study 1 showed that Frequency and Original stimuli were affected very similarly by the CI manipulation for all of the emotions studied. However, relative to the Original condition, Sadness was decoded less accurately in the Frequency condition, implying that for this emotion participants relied instead on intensity or temporal cues. This finding is encouraging in terms of the ability of listeners to switch listening strategies to suit the demands of the task, and implies that this tendency might vary depending on the emotion being appraised. To be clear, it is not claimed that listeners become aware of the emotion being expressed and *then* change their listening strategies. Rather, listeners might adapt their listening

'on the fly', attending to more salient cues, which may themselves be determined by the emotion being expressed.

Participants in Study 2 demonstrated significantly less accurate emotion perception in general, compared to Study 1, and also were less confident about their responses, implying that the task was more difficult with music than with speech. Aside from the fact that most of the population are likely to have more extensive experience from everyday life of decoding emotion in speech compared to music, a possible explanation for the discrepancy is that acoustic features varied relatively less in Study 2. That is, since the melodies were fixed across the different emotions, there were fewer modifiable acoustic parameters in the music stimuli compared to the speech stimuli (Livingstone et al., 2015).

Possibly due to generally increased task difficulty, participants' emotion judgements were relatively good for only the Happiness and Sadness stimuli, suggesting that the information preserved was sufficient only for a binary discrimination. Accordingly, emotion judgements in these conditions were more drastically affected by the CI simulation. Examination of confusion matrices implied that this was indeed a specific 'Happy vs. Sad' discrimination – as opposed to a more general appraisal of valence – since various cross-valence confusions were made (most prominently between Anger and Happiness).

In both studies, it appeared that participants were predisposed to attend primarily to frequency-based sound features. However, both of the studies suggest that emotion recognition accuracy might conceivably be improved by switching listening strategies for specific combinations of auditory feature conditions and emotions (e.g. Intensity and Anger, respectively). With this said, examination of confusion matrices in both studies showed that, although some response patterns may denote direction

of attention to relatively preserved acoustic features, they may also stem from more general response biases associated with CI simulation. Therefore, it would be useful for future research to eliminate these biases (as much as is possible) by allowing participants more time for acclimatisation with the CI-simulated stimuli.

Lastly, leaving aside the CI manipulation, participants' generally above-chance performance across the different feature attenuation conditions suggests that the features preserved might be more relevant for emotion identification than has previously been recognised. For music in particular, some emotions were able to be decoded with accuracy well above chance, using only information about articulation and/ or intensity modulation. Therefore, these studies draw attention to the importance of these types of microstructural variation for emotional expression, in both speech and music.

4.9 Limitations

Perhaps the most important limitation of Studies 1 and 2 is that listeners were not sufficiently incentivised to adapt their listening strategies to the different stimuli, since no feedback was given about participants' performance at any stage of the experiment. Therefore, it would have been very difficult for participants to gauge the effectiveness of their current listening strategy at any moment. For example, a participant may have begun the experiment by primarily listening out for frequency variation, in order to judge the different emotions expressed. Although this approach would not likely be optimal when listening to frequency-attenuated stimuli, the participant would have been given no feedback to indicate this, and therefore might have persevered with the strategy. Even if the participant felt dissatisfied with their performance and decided to attend to a different cue (or combination of cues), there would

have been no way for them to ascertain whether or not this led to any improvement. Since the experiment was relatively difficult in general, the lack of feedback would also have made it difficult for participants to distinguish between true 'guesses' (i.e. those leading to approximately chance-level performance) and responses perceived as such, which nonetheless lead to above-chance performance (Nisbett & Wilson, 1977). Therefore, participants may have become demotivated if they perceived their judgements as mere guesses, and received no feedback to contradict this.

Relatedly, participants may have found it difficult to adapt their listening strategies because stimuli were presented in random order across each of the different auditory feature conditions. Therefore, the task demands may have been too high, since different listening strategies were potentially required from every stimulus to the next. For example, trial 1 might be an Intensity stimulus (meaning that attending to the intensity contour might be a good strategy), whereas in trial 2, intensity information might be attenuated, and so on. In other words, it is quite likely that participants were never exposed to a particular feature-attenuation condition for long enough to deduce the most relevant auditory features for the task, and to update their listening strategy accordingly. Instead, it is likely that participants either stuck with the listening strategy that they would normally use (e.g. tending to focus on frequency content), or switched strategies erratically due to the unpredictable nature of stimulus presentation (particularly considering the intermixed presentation of original and CI-simulated stimuli).

More generally, since the emotion judgement was a 5-AFC task, with no option to 'pass', it is possible that participants sometimes chose 'Neutral' as a 'pass' option. Across all participants, 'Neutral' was the most common response in the emotion judgement task, and was also associated with the lowest average confidence ratings.

Unfortunately, it is difficult to effectively address this issue, since the inclusion of a pass option might have encouraged participants to respond more conservatively, thereby avoiding potentially correct 'guesses' that are based upon higher order cognitive processes eluding verbal report (Nisbett & Wilson, 1977). Given the inherently difficult nature of the task, this type of response may be rather informative. In any case, the prevalence of neutral responses is somewhat inevitable considering the removal of acoustic feature cues in Studies 1 and 2. As for the previous limitation, this issue might be circumvented to some extent by the use of a training-based paradigm. In principle, if participants were given time to 'acclimatise' to the stimuli, and become more adept at attending to better-preserved acoustic features, over-estimation of 'Neutral' might be reduced.

Finally, it is problematic to draw detailed comparisons between speech and music from the studies reported (which, in any case, was not the primary aim of these studies), since the two tasks were not equivalent in difficulty. Therefore, it is unclear to what extent differences between Studies 1 and 2 arose due to the different stimuli types, or merely due to a difference in overall task difficulty. As discussed previously, the two most probable explanations for this discrepancy are: participants' relatively greater experience of decoding emotion from speech, and the comparative lack of acoustic feature variation in the music stimuli. Unfortunately this was an inevitability of the experimental design, since the music stimuli used varied in performance expression rather than compositional cues. This was considered the most suitable approach however, since the use of different melodies for each emotion may have led participants to simply memorise the different melodies and respond accordingly, rather than considering the other acoustic features.

4.10 Following up on Studies 1 and 2

To address some of the aforementioned limitations of Studies 1 and 2, the next studies carried out introduced both a blocked design for stimulus presentation, and a feedback component. Summarily, Studies 1 and 2 were concerned with how participants might update their listening strategies when presented with different feature-attenuated stimuli, in order to prioritise those auditory features encoding the most relevant information for emotion classification. To an extent these studies were successful, providing some initial insights into which auditory features may be most important for the recognition of different emotions in speech and music when listening with CI simulation. However, as already discussed, there were two primary features of the previous paradigm that might have prevented or otherwise discouraged participants from changing listening strategies: randomly intermixed presentation of different feature-attenuated stimuli, and a lack of feedback about performance. With this in mind, Chapter 5 describes two follow-up studies, which aimed to build upon the findings outlined in this chapter by adapting the experimental paradigm to include a training element. Chapter 5 describes in detail the changes made to the experimental procedure, and discusses the results that were obtained from these studies.

5 Studies 3 and 4: Emotion perception training with speech and music for cochlear implant-simulated listeners

5.1 Introduction

The rationale for conducting studies 3 and 4 was chiefly to address the previously-noted shortcomings of Studies 1 and 2. Studies 3 and 4 represent a continuation of the line of enquiry pursued by the previous studies, albeit with an updated experimental paradigm. As described in the previous chapter, there were two primary drawbacks to Studies 1 and 2: A) participants had little awareness of their performance, and therefore limited incentive to switch listening strategies during the task, and B) the randomly intermixed presentation of stimuli might have unnecessarily increased overall task difficulty.

To resolve the first issue, separate phases of 'training' were incorporated within the experimental procedure for studies 3 and 4, during which participants received feedback on the accuracy of the emotion judgements, and therefore could discern the efficacy of their listening strategies. To resolve the second issue, a blocked design for stimulus presentation was adopted – i.e. groups of stimuli from only one feature-attenuation condition were presented at a time – in principle allowing much more time for participants to 'acclimatise' to the feature-attenuated stimuli, thereby facilitating the adaptation of listening strategies, as well as making the overall task slightly easier. To further facilitate acclimatisation to the cochlear implant (CI) simulated stimuli more generally, only these stimuli were presented in these studies – no non-CI-simulated stimuli were included, unlike in Studies 1 and 2. This was intended to

bring the experimental paradigm closer in principle to a process of 'rehabilitation', in which participants practised perceiving auditory emotion with CI simulation, without being distracted by randomly interspersed presentation of non-CI-simulated stimuli.

Thus, the central question that Studies 3 and 4 attempted to answer, for speech and music respectively, was: 'can participants be trained to change their approach to the CI-simulated emotion identification task and, if so, how will this affect performance?'. In fact, the training paradigm introduced here had two primary purposes.

Firstly, it was intended to allow participants to learn about the way the CI-simulated stimuli sound, thereby improving performance generally. Providing more time for NH listeners to become familiar with the stimuli is also advantageous in that it facilitates more accurate comparison with CI users', who typically demonstrate improved performance in music and speech perception tasks, following aural rehabilitation (Gfeller, Mehr, and Witt, 2001; Wei et al., 2000). More specifically, increasing familiarity with the CI simulation was expected to reduce the kinds of response biases seen in Studies 1 and 2, such as the tendency to over-estimate Anger. Further, training or 'rehabilitation' may have an effect on the way in which listeners approach the emotion judgement task, in a wider sense. Recall that there are essentially two strategies available for the completion of any given recognition task: similarity-based categorisation and rule-based categorisation (Smith et al., 1998). Because most participants listening with CI simulation will have been very unfamiliar with the way that the stimuli sounded, it is likely that they would not have been able to rely upon a similarity-based strategy for recognition. That is, since recognition relies on a correspondence between what is currently being heard and what has already been heard in the past (McAdams, 1993), there must be some period of acclimatisation, during which participants 're-learn' how the different emotional sentences or melodies

sound, so as to have some reference point with which to compare novel stimuli.

Secondly, the training paradigm aimed to investigate the extent to which participants were able to adapt their listening strategies to 'overcome' the different feature-attenuation conditions. Although participants were able to decode emotion at a level generally greater than chance with CI simulation in Studies 1 and 2, the evidence implied that this performance was largely achieved via attending to residual frequency information, as opposed to better-preserved features such as intensity and duration. However, it was not clear whether participants were unable (or unwilling) to adopt a different listening strategy, or simply did not have sufficient incentive to do so. In presenting only one feature-attenuation condition per participant, it was hoped that there would be more incentive for participants to shift towards a theoretically 'optimal' listening strategy (e.g. for Intensity stimuli, attending primarily to intensity-based cues, and so forth).

The final significant amendment made to the experimental procedure for studies 3 and 4 was the omission of the Articulation condition (which comprised stimuli with attenuated frequency, intensity and temporal variation). The original rationale for including this condition was to address the somewhat tangential question of how emotion recognition might proceed in even more severely deteriorated stimuli (i.e. whether emotion identification might be possible even with only cues relating to amplitude envelope and temporal fine structure preserved). However, since performance in this condition was consistently poor across both speech and music stimuli, with or without CI simulation, it was considered to offer very little additional insight, above and beyond the other conditions.

5.1.1 Summary

In summary, the studies carried out in this chapter build upon Studies 1 and 2, addressing key limitations by adding a training component, as well as making other minor refinements to improve the experimental procedure. Hence, the key research questions addressed are very similar to those tackled in the previous chapter, with the addition of a more specific line of enquiry into the role of practice (i.e. 'rehabilitation') in CI-simulated emotion perception.

In addition to addressing limitations of Studies 1 and 2, these studies were intended to investigate the extent to which CI users' performance in emotion recognition is relatively contributed to by both cognitive abilities and impoverished auditory input. Concretely, the degradation to the input stimuli remained constant throughout, while listening strategies could develop and adapt during learning to bolster performance. In particular, this study was important in facilitating future comparison with CI users, since NH participants' listening strategies with CI simulation may be more likely to resemble CI users' strategies after a period of acclimatisation.

In the next section, the updated experimental paradigm is outlined in detail. In most respects (e.g. stimuli used), it is very similar to that outlined in the previous chapter, and therefore particular attention is given to describing the novel training aspect.

5.2 Methods

5.2.1 Participants

Forty-four normally-hearing participants were recruited in total, via student and staff volunteer lists at The University of Sheffield: twenty-four in Study 3 (17 female; mean age = 25.96 years, SD = 10.21, mean 6.88 years experience with a musical instrument) and twenty in Study 4 (15 female; mean age = 22.90 years, SD = 5.04, mean 7.85 years experience with a musical instrument). All participants reported having normal hearing and normal or corrected-to-normal vision, and provided fully-informed consent prior to participation. Participants recruited for Study 3 additionally reported non-fluency in German, which was important since German speech stimuli were used.

5.2.2 Materials

Twenty CI-processed speech and twenty music stimuli from Studies 1 and 2 were used (i.e. all stimuli excluding those from the Articulation condition), along with twenty novel CI-processed speech and music stimuli, extracted from the same databases and processed in the same ways. That is, in Study 3, twenty additional speech stimuli were taken from the Berlin Emotion Speech Database (Burkhardt et al., 2005) and, in Study 4, twenty extra music stimuli were taken from Quinto et al.'s (2014) investigation of emotion communication via musical performance. The stimuli consisted, respectively, of short excerpts of speech and brief musical melodies, each expressing one of five emotions (Anger, Fear, Happiness, Sadness or Neutral). As described in Studies 1 and 2, these were processed in order to produce different versions in which Frequency, Intensity or Duration cues were systematically attenuated.

In total, there were 160 speech stimuli (four different sentences × two speaker genders (male/ female) × five emotions × four auditory feature conditions) and 160 musical stimuli (four different melodies × two musical instruments (violin/ voice) × five emotions × four auditory feature conditions). All stimuli were encoded in single-channel (mono), 16-bit Audio Interchange File Format (AIFF) with 16,000 Hz sampling frequency.

5.2.3 Procedure

Prior to the main experiment, all participants completed the demographic and empathy questionnaires as described in Studies 1 and 2 (Baron-Cohen & Wheelwright, 2004). However, the ad hoc evaluation of musical experience was replaced with the more sophisticated Music Use questionnaire (MUSE) (Chin & Rickard, 2012), facilitating a more thorough investigation of the putative effects of musical engagement upon emotion perception. The rationale for including the MUSE was that more nuanced effects of musical aptitude may not have been well-captured by the very brief, more general measure used previously. If there was indeed a relevant influence of musical expertise upon performance in emotion discrimination, then it was hoped that an established inventory would be more likely able to detect it. An additional measure of music perceptual ability (e.g. Law and Zentner, 2012; Müllensiefen, Gingras, Musil, and Stewart, 2014) may also have been useful in this regard, but was not included in order to restrict the experiment to a reasonable length – while the relationship between emotion perception and music perception skills may be a fruitful avenue for research, it is not directly relevant to the aims of these studies.

Studies 3 and 4 were designed identically, except that participants listened to speech stimuli in the former and musical stimuli in the latter. Therefore, both experimental

procedures are described simultaneously here. As in Studies 1 and 2, the experimental paradigm involved making perceptual judgements about the emotion expressed by each stimulus. This time however, each participant experienced only one of the auditory feature attenuation conditions, along with the original stimuli for comparison. In Study 3, there were six participants in each of the four feature attenuation conditions, and in Study 4 there were five per condition. All stimuli were processed with the same CI simulation using MATLAB, as described in the previous chapter.

As a pre-test, participants first listened to a subset of forty stimuli, containing Original and feature-attenuated versions of one sentence or melody, spoken by two genders (male, female) or performed using two instruments (voice, violin), each intonated in order to express one of the five emotions and each repeated twice. This subset was divided into two sections, containing Original stimuli and feature-attenuated stimuli, with the order of presentation counterbalanced between participants. During this task, participants made a 5-AFC judgement about the emotion conveyed by each stimulus, and provided an associated confidence rating, as in Studies 1 and 2. During the pre-test stage of the experiment, participants received no feedback about their performance.

Next there was a training phase, which was developed following relevant studies investigating auditory learning with spectral degradation (e.g. Burkholder, Pisoni, and Svirsky, 2004; Driscoll, Oleson, Jiang, and Gfeller, 2009). During this phase, participants performed the same task as described above, but this time received on-screen text-based feedback after each trial, indicating which emotion the talker or musician had intended to express (i.e. the 'correct answer'). Participants were also given auditory feedback: each stimulus was played a second time, after the correct answer had been revealed. This was intended to allow participants to re-

appraise the acoustic features of each stimulus in a top-down fashion, having been made aware of the emotion communicated. The rationale was that participants might reconsider which auditory feature(s) were used to express the emotion, thereby potentially altering their listening strategy for subsequent stimuli. The training phase consisted of two blocks of thirty stimuli. Each block contained one sentence or melody $\times$ two genders or instruments $\times$ five emotions $\times$ three repetitions. These blocks were comprised solely of the feature-attenuated stimuli, with no Original stimuli presented (with the exception of participants in the 'Original' control condition, who did not hear any feature-attenuated stimuli).

After the first block of training stimuli, the experiment included a post-test block, using previously-heard stimuli. This phase was identical to the pre-test, with the same stimuli used – participants again heard both Original and feature-attenuated stimuli and did not receive any form of feedback. Following this, participants were required to take a short break for five to ten minutes, to prevent fatigue.

The training phase was then repeated, using the same stimuli as presented during the first block of training. Finally, participants completed a post-test with stimuli that had not been heard before. This phase was identical to the intermediate-test, but with novel stimuli based on a different sentence or melody. The aim of this phase was to assess how well participants' learning would generalise beyond the stimuli encountered previously, to novel stimuli. In other words, this phase measured how well participants were able to learn underlying correspondences between patterns of acoustic features and specific emotions, and then apply these to stimuli as they were heard for the first time.

For clarity, Table 8 illustrates the specific stimuli that were presented in each stage of the experiment. Note that the positions of the different sentences/ melodies were

counterbalanced to eliminate order effects, such that each stimulus set occurred in each position equally often. That is, one quarter of participants heard sentence/ melody A in the pre-test, while another quarter heard sentence/ melody B, and so on.

Table 8: Information about stimuli included at each stage of the experiment.

Stage	**Stimuli presented**	**Description**
Pre-test	Sentence/ melody A	No feedback
Training 1	Sentence/ melody B	Audio and visual feedback
Training 2	Sentence/ melody C	Audio and visual feedback
Post-test	Sentence/ melody A	Identical to Pre-test
Training 3	Sentence/ melody B	Identical to Training 1
Training 4	Sentence/ melody C	Identical to Training 2
Post-test (transfer)	Sentence/ melody D	No feedback

During all phases of the experiment, data were gathered about whether or not participants' judgements were correct, along with the confidence ratings and reaction times associated with each judgement made.

5.2.4 Hypotheses

Firstly, in line with previous research, and with the results of Studies 1 and 2, it was hypothesised that emotion identification performance with CI-simulated speech and music stimuli would be above chance level. Additionally, it was predicted that emotion identification accuracy would increase significantly as a function of experimental phase, representing improved task performance with more practise. It was predicted that performance would be better in the two post-test phases, by comparison to the pre-test phase. Although it was assumed that performance might be better in the first post-test (with already-encountered stimuli), judgement accuracy was still expected to be significantly better in the second post-test phase compared to the pre-test, denoting generalisation of learning to stimuli previously unencountered

stimuli.

It was also hypothesised that detection accuracy would vary significantly across the different auditory feature conditions, as was the case in Studies 1 and 2. Specifically, it was predicted that Frequency stimuli would be affected to a greater extent by the CI simulation, whilst Duration and Intensity stimuli should be less affected. It was expected that participants would adopt listening strategies focussing upon relatively preserved components of the CI-simulated stimuli, and therefore that overall emotion identification accuracy would be better for Duration and Intensity stimuli, relative to Frequency. Although Frequency stimuli were associated with more accurate emotion recognition in Studies 1 and 2, it was hypothesised that the process of training, along with blocked presentation of stimuli would overturn this effect. Since the updated experimental paradigm provided a larger incentive for participants to switch listening strategies, it was hypothesised that they would be more likely to move away from making judgements based primarily on frequency-based cues, thereby achieving better judgement scores in the Duration and Intensity conditions.

It was also expected that decoding accuracy would vary across the different emotions. This was because A) there appears to be general variability in the decoding of different emotions, as described by previous literature (e.g. Scherer et al., 2001), and B) as documented in the previous chapter, different emotions vary in the types of cues that are most strongly associated with their expression, and these cues are preserved to varying extents by the CI simulation.

Considering the above hypotheses relating to effects of Auditory attenuation condition and Emotion, it was expected that a Feature-attenuation condition $\times$ Emotion interaction might occur, since participants would learn which features were best preserved by the CI simulation in each of the auditory feature conditions, and these

features would be more or less informative, depending on the specific emotion being judged.

Lastly, it was predicted that scores may be correlated with one or more components of musical experience, since a validated questionnaire was used for this study – an improvement upon the ad hoc measurement taken in Studies 1 and 2. Specifically, the Music Listening, Musical Instrument Playing, and Musical Training sub-components of the MUSE were considered most likely to be associated with improved performance, since previous research has linked auditory emotion perception with musical training and expertise (Lima and Castro, 2011; Thompson et al., 2004). It was hypothesised that self-reported empathic capabilities would not significantly influence participants' performance, since no effect was found in the previous studies. Nonetheless, given the updated paradigm used, it was considered possible that empathy could influence performance to some degree. For example, highly empathic individuals might be better a noticing subtle acoustic changes used to express emotion, but be hampered by unfamiliarity with the sound of the CI simulation. Thus, a period of acclimatisation and training to recognise emotion with CI-simulated stimuli may render any influence of empathy more easily detectable.

5.3 Results

5.3.1 Statistical methods

For the main statistical analyses, a different approach was taken than in Studies 1 and 2. Instead of computing a metric based on the multiplication of 'correctness' (1 or -1) by confidence rating, the frequency of correct responses (regardless of confidence) was treated as the dependent variable. The primary motivation for

this development was that it provides more easily interpretable results. Specifically, chance performance, quantified as a of percentage correct responses, is intuitively understandable: since there are five potential responses, chance performance = 20% correct. However, when confidence ratings are also considered, it becomes more difficult to determine convincingly what constitutes chance performance. In Studies 1 and 2, the solution proposed was to hold chance level constant at 3 (the midpoint of the 1-5 scale used), and sum participants scores across each of the different sentences/ musical melodies within a particular condition. Therefore, across the four different melodies/ sentences, participants were assigned a score between -20 and 20. With an average confidence level of 3 assumed, chance performance was therefore equal to $3(.2 \times 4) - 3(.8 \times 4) = -7.2$.

This is problematic for two reasons. Firstly, it contains an in-built, a priori assumption about how participants' confidence ratings will be distributed, that is not necessarily valid. Secondly, the resulting measure is more complicated to interpret, since it conflates both judgement accuracy and judgement confidence. Therefore, considering the case that an individual scores just above the 'chance level' of -7.2, it is impossible to tell whether they were actually more accurate in decoding emotion, or merely more conservative overall in their confidence ratings.

For the analyses to follow, this problem was circumvented by the use of loglinear analysis – a statistical procedure for assessing the relationship between two or more categorical variables. This method works analogously to the Chi-squared test, in that log values of observed frequencies are used, effectively representing the categorical variables such that it is possible to fit a linear model to the data (similarly to the use of the logarithmic transformation for categorical data used in logistic regression) (Field, Miles, & Field, 2012). Loglinear analysis is characterised by a hierarchical approach,

in which a so-called 'saturated' model is first built, in which all of the available variables and all interactions thereof are included. Then, the model components are removed one-by-one, beginning with the highest-order interactions, and are compared to the saturated model using ANOVA. This process stops at the point where the removal of a component significantly reduces the explanatory power of the model.

The confidence rating data were analysed separately, using rank-based factorial ANOVA, since these data were ordinal and therefore non-normally distributed. Lastly, parametric ANOVA was used to explore the overall emotion identification data by participant (i.e. percentage correct responses). This approach was used to more generally analyse the effects of training over the different experimental phases, whereas the loglinear analysis was implemented to examine data on a stimulus-by-stimulus basis, to explore the effects of feature attenuation and emotion.

In this section, rank-based and parametric ANOVA were implemented using R packages *ARTool* (Wobbrock, Findlater, Gergle, & Higgins, 2011) and *ez* (Lawrence, 2016), respectively. Loglinear analysis and relevant follow-up Chi-square tests were also implemented in R, using *MASS* (Venables & Ripley, 2002) and *gmodels* (Warnes, Bolker, Lumley, & Johnson, 2015).

5.3.2 Study 3: Speech

On average, across each of the conditions, participants made their emotion judgements 5.10 seconds after stimulus onset (corresponding to mean stimulus length + 1.93 seconds). Reaction times were not compared across the different feature-attenuation conditions or emotions, since this measure was confounded by differences in stimulus length.

The average confidence rating associated with participants' judgements, across all of the experimental trials, was 3.68. Rank-based factorial ANOVA revealed a very small but significant effect of Emotion upon confidence ratings, $F(4, 820) = 2.61$, $p = .034$, $\eta^2 = .01$. Post-hoc, Bonferroni-adjusted Wilcoxon-Mann-Whitney tests showed that the only significant pairwise difference in confidence ratings was between Sadness and Neutral, with the former being higher (see Table 9). There was no significant effect of Feature attenuation, $F(3, 820) = 1.56$, $p = .194$, $\eta^2 < .01$, nor was there a significant interaction with Emotion, $F(12, 820) = 0.90$, $p = .551$, $\eta^2 = .01$.

Table 9: Pairwise, Bonferroni-adjusted Wilcoxon-Mann-Whitney tests, comparing differences in participants' confidence ratings, according to the emotion judged.

	Anger	**Fear**	**Happiness**	**Neutral**
Fear	1.000			
Happiness	1.000	1.000		
Neutral	0.528	0.058	0.307	
Sadness	0.100	0.351	0.071	$<.001^*$

* = significant at .050 alpha-level.

Overall performance (mean percentage correct) in the pre-test, post-test and post-test-transfer blocks is summarised for each condition in Figure 23. In general, participants in each of the different feature processing conditions showed an effect of training with the CI-simulated stimuli, performing better, in all conditions, in the post-test and post-test transfer stimuli blocks, compared to pre-test. In all conditions except for Duration, performance also improved from the post-test to the post-test-transfer phase, implying that participants' learning was not restricted to the specific stimuli experienced during training. ANOVA confirmed that the effect of test phase upon emotion identification accuracy as significant, $F(2, 46) = 20.52$, $p < .001$, $\eta^2 = .47$. Planned contrasts revealed that the differences between pre-test and the two post-test phases were significantly different ($p < .001$ for both), whilst the difference between the post-test and post-transfer phases was not significant.

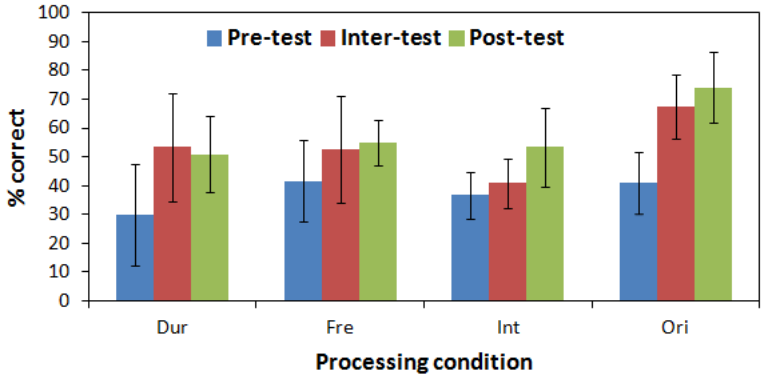


Figure 23: Mean percentage correct emotion recognition, for each condition, across the different testing phases. Error bars show ± 1 standard deviation.

Looking at post-test performance, participants did best in the Original condition, as expected, but elsewhere the differences between feature attenuation conditions were minimal. Even though, for example, accuracy in the Duration condition at pre-test was eleven percentage points worse than in the Frequency condition, these differences were largely eliminated by the training phases. By contrast, the training created a noticeable discrepancy between the feature attenuation conditions and the Original condition, with the latter associated with much greater emotion judgement accuracy.

Loglinear analysis, examining the effects of feature processing condition and emotion upon judgement accuracy, produced a final model that retained all effects – likelihood ratio $\chi^2\ (0) = 0$, $p = 1$ – indicating that the Processing × Emotion interaction was significant, $\chi^2\ (12) = 26.36$, $p < .010$.

Follow-up chi-square analyses revealed significant effects of Emotion upon judgement accuracy in all processing conditions except for Frequency, as summarised in Table 10. This appeared to reflect the fact that the different emotions were recognised with relatively uniform accuracy in the Frequency condition, whereas in the Duration condition, for example, Sadness was distinguished with almost four times the accuracy of Happiness.

Table 10: Chi-square tests, examining the influence of stimulus emotion class on emotion recognition accuracy in each of the feature processing conditions.

Condition	χ^2	**df**	p
Duration	21.14	4	$< .001^*$
Frequency	3.16	4	.531
Intensity	13.19	4	$.010^*$
Original	40.34	4	$< .001^*$

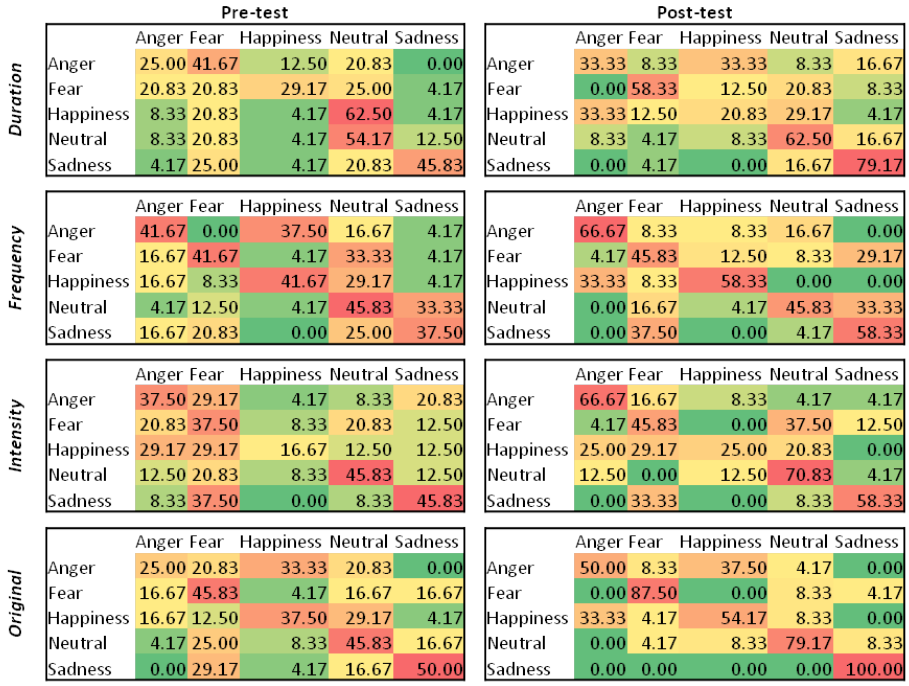
* = significant at .050 alpha-level.

To explore the patterns of errors in emotion judgements made by participants, confusion matrices were created for each of the feature processing conditions (Figure 24). These show, for each stimulus emotion, the frequency with which participants selected each of the five response options.

Firstly, it is clear that there was a general effect of learning - accuracy for each of the emotions largely increased from pre-test to post-test-transfer. Additionally, the overestimation of Neutral, as observed in Study 1, declined after the training, as indicated by lower 'false positive' response rates for Neutral in the post-test-transfer phase compared to the pre-test. There was also some evidence of reduced false positive responding for Anger, although confusions with Happiness increased in all conditions, implying that this particular confusion was highly pervasive and difficult to overcome.

Sadness and Neutral appeared to be the most robust emotions to disruption overall, with these emotions being relatively well-decoded across all feature attenuation conditions, and also showing quite clear improvement as a result of training. Although there was some universal improvement as a result of acclimatisation to the CI-simulated stimuli, there were also specific combinations of feature attenuation conditions and emotions that showed particularly good improvement with training. For example, recognition accuracy for Sadness stimuli improved by 20.83 and 12.50

percentage points in the Frequency and Intensity conditions, respectively, but by 41.85 in the Duration condition. Similarly, Fear stimuli were better recognised in the Duration condition, and showed more improvement from training compared to the Frequency and Intensity conditions. Conversely, Anger improved to a greater extent and was much more accurately identified in the Frequency and Intensity conditions.



Pre-test

	Anger	Fear	Happiness	Neutral	Sadness
Duration					
Anger	25.00	41.67	12.50	20.83	0.00
Fear	20.83	20.83	29.17	25.00	4.17
Happiness	8.33	20.83	4.17	62.50	4.17
Neutral	8.33	20.83	4.17	54.17	12.50
Sadness	4.17	25.00	4.17	20.83	45.83
Frequency					
Anger	41.67	0.00	37.50	16.67	4.17
Fear	16.67	41.67	4.17	33.33	4.17
Happiness	16.67	8.33	41.67	29.17	4.17
Neutral	4.17	12.50	4.17	45.83	33.33
Sadness	16.67	20.83	0.00	25.00	37.50
Intensity					
Anger	37.50	29.17	4.17	8.33	20.83
Fear	20.83	37.50	8.33	20.83	12.50
Happiness	29.17	29.17	16.67	12.50	12.50
Neutral	12.50	20.83	8.33	45.83	12.50
Sadness	8.33	37.50	0.00	8.33	45.83
Original					
Anger	25.00	20.83	33.33	20.83	0.00
Fear	16.67	45.83	4.17	16.67	16.67
Happiness	16.67	12.50	37.50	29.17	4.17
Neutral	4.17	25.00	8.33	45.83	16.67
Sadness	0.00	29.17	4.17	16.67	50.00

Post-test

	Anger	Fear	Happiness	Neutral	Sadness
Anger	33.33	8.33	33.33	8.33	16.67
Fear	0.00	58.33	12.50	20.83	8.33
Happiness	33.33	12.50	20.83	29.17	4.17
Neutral	8.33	4.17	8.33	62.50	16.67
Sadness	0.00	4.17	0.00	16.67	79.17
Anger	66.67	8.33	8.33	16.67	0.00
Fear	4.17	45.83	12.50	8.33	29.17
Happiness	33.33	8.33	58.33	0.00	0.00
Neutral	0.00	16.67	4.17	45.83	33.33
Sadness	0.00	37.50	0.00	4.17	58.33
Anger	66.67	16.67	8.33	4.17	4.17
Fear	4.17	45.83	0.00	37.50	12.50
Happiness	25.00	29.17	25.00	20.83	0.00
Neutral	12.50	0.00	12.50	70.83	4.17
Sadness	0.00	33.33	0.00	8.33	58.33
Anger	50.00	8.33	37.50	4.17	0.00
Fear	0.00	87.50	0.00	8.33	4.17
Happiness	33.33	4.17	54.17	8.33	0.00
Neutral	0.00	4.17	8.33	79.17	8.33
Sadness	0.00	0.00	0.00	0.00	100.00

Figure 24: Heat-mapped confusion matrices for speech stimuli during the pre-test and post-test-transfer phases, depicting the percentage of responses in each emotion category, for each stimulus emotion. Columns denote presented emotions, rows denote emotion judgement responses. Red = higher values, green = lower values.

Lastly, the potential relationships between musical ability (MUSE) and empathy (EQ) measures, and emotion identification accuracy were explored. Controlling for the effect of Feature attenuation condition, multiple linear regression analysis found that none of the MUSE sub-scores, nor the EQ scores, were able to explain a significant portion of the variance in the percentage of correct emotion judgements at any of the test phases. However, the Pearson correlation between the 'Index of

Musical Training' subcomponent of the MUSE and emotion judgement accuracy became increasingly strong with training, and reached statistical significance for the post-test-transfer results (Table 11). Specifically, more highly musically-trained individuals appeared to benefit to a greater extent from the training paradigm, tending to register better emotion identification accuracy in the final, post-test-transfer phase.

Table 11: Pearson correlations between MUSE subcomponent 'Index of Musical Training and overall emotion identification accuracy.

Phase	***r***	**df**	***p***
Pre-test	.29	22	.163
Post-test	.38	22	.063
Post-test-transfer	.42	22	.040*

* = significant at .050 alpha-level.

5.3.3 Study 4: Music

On average, across each of the conditions, participants made their emotion judgements 6.73 seconds after stimulus onset (corresponding to mean stimulus length + 1.63 seconds). As in Study 3, reaction times were not compared across the different feature-attenuation conditions or emotions, since this measure was confounded by differences in stimulus length. Compared to Study 3, participants responded very slightly faster on the whole.

The average confidence rating associated with participants' judgements, across all of the experimental trials, was 3.04, which was slightly lower than in Study 3. Rank-based factorial ANOVA revealed a very small but significant effect of Feature attenuation condition upon confidence ratings, $F(3,\ 680) = 11.52,\ p < .001,\ \eta^2 = .05$. Post-hoc, Bonferroni-adjusted Wilcoxon-Mann-Whitney tests showed that only two pairwise comparisons between Feature attenuation conditions were significant: Frequency and Intensity, and Frequency and Original – Frequency being the less

confidently-recognised condition in both cases (see Table 12). There was no significant effect of Feature attenuation, $F(3, 820) = 1.56$, $p = .194$, $\eta^2 < .01$, nor was there a significant interaction with Emotion, $F(12, 820) = 0.90$, $p = .551$, $\eta^2 = .01$.

Table 12: Pairwise, Bonferroni-adjusted Wilcoxon-Mann-Whitney tests, comparing differences in participants' confidence ratings, according to the emotion judged.

	Duration	**Frequency**	**Intensity**
Frequency	.068		
Intensity	1.000	<.001*	
Original	1.000	<.001*	1.000

* = significant at .050 alpha-level.

Participants performed slightly better than in the previous chapter, in terms of the mean percentage of correct responses and, as hypothesised, performance was above chance level (20% correct) for all of the feature-attenuation conditions, even at pre-test (Figure 25). However, performance was quite variable on the whole, as indicated by the relatively large standard deviations.

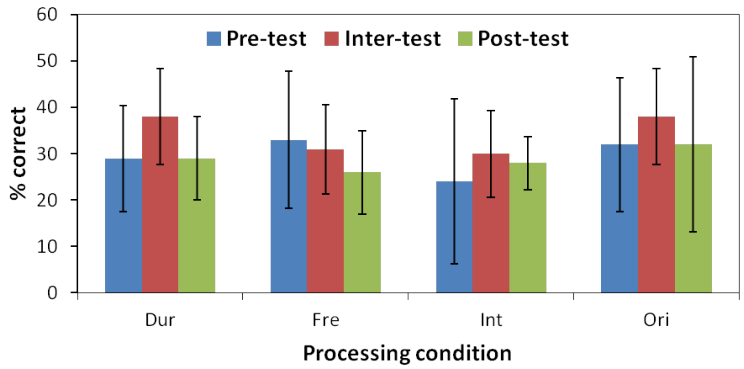


Figure 25: Mean percentage correct emotion recognition, for each condition, across the different testing phases. Error bars show ± 1 standard deviation.

Contrary to expectation, the training paradigm appeared to have relatively little effect – ANOVA revealed that there was no significant effect of test phase upon participants' emotion identification accuracy, $F(2, 38) = 1.18$, $p = .317$, $\eta^2 = .06$. Despite this, the graph shows evidence of some modest improvement between the

pre-test and post-test blocks. This effect was noticeably stronger in the Duration condition, compared to the other feature attenuation conditions, indicating that training had the greatest effect when stimuli duration cues were preserved. However, across all of the feature attenuation conditions, learning did not appear to generalise to previously-unheard stimuli: emotion identification accuracy was unanimously lower during the post-test-transfer block, compared to the post-test block. In fact, only in the Intensity condition did performance improve at all from pre-test to post-test-transfer.

Three-way loglinear analysis produced a final model that retained only the main effect of Emotion, $\chi^2(16) = 45.51, p = .001$, indicating that the emotion being judged significantly influenced identification accuracy. Conversely, neither Feature attenuation condition, $\chi^2(15) = 21.55, p = .120$, nor the interaction between the latter and Emotion had a significant impact upon identification accuracy, $\chi^2(12) = 16.85, p = .160$.

To examine the effect of Emotion upon identification accuracy in more detail, the equivalent chi-square test was computed. This test produces standardised residuals associated with each level of an experimental condition, thereby providing an approximate index of the extent to which each emotion contributed to the overall effect of Emotion reported above (Sharpe, 2015). According to established convention, standardised residuals of more than $\pm$ 2 indicate that values for that level of the variable are less compatible with the null hypothesis (Agresti, 2007; Sharpe, 2015). Based on this criterion, Fear (Std. residual = -2.036) and Sadness (Std. residual = 3.504) appeared to primarily drive the main effect of Emotion upon identification accuracy. Specifically, Fear tended to be recognised less accurately than expected by chance, whilst Sadness was recognised more accurately.

Confusion matrices showed that, contrary to the experimental hypothesis, there was

no appreciable effect of the training upon emotion identification accuracy (Figure 26). Certainly, the effect of training was much less strong than with the speech stimuli in Study 3. As in the previous study, however, acclimatisation to the CI-simulated stimuli appeared to reduce the extent to which Neutral was overestimated, as evidenced by fewer false positive responses for this emotion in the post-test-transfer block. Conversely, the overestimation of Anger did not clearly show a decline. As was the case in Study 3, confusions between Anger and Happiness were the most pervasive, taking into account all of the different conditions.

Between the pre-test and post-test-transfer phases there also appeared to be greater consistency in terms of the specific errors made. For example, Sadness was clearly mistaken for Fear more often than any other emotion during the post-test-transfer, whereas at pre-test there was little evidence of any pattern.

Unfortunately, none of the individual emotions were consistently associated with improved identification accuracy as a result of the training paradigm. In fact, substantial practice effects were limited to only two cases: Neutral stimuli in the Duration condition, and Sadness stimuli in the Original condition.

Sadness appeared to be the most robust emotion to disruption overall, being relatively well-decoded across all feature attenuation conditions. Anger stimuli were also relatively well-identified, although with far more false positive errors.

Lastly, inter-individual differences in empathy (EQ) and musical training and expertise (MUSE) appeared to exert little influence on overall task performance. A multiple linear regression analysis found that neither Empathy Quotient scores nor any of the MUSE sub-components explained a significant proportion of the variance in emotion judgement accuracy, once Feature attenuation condition was controlled

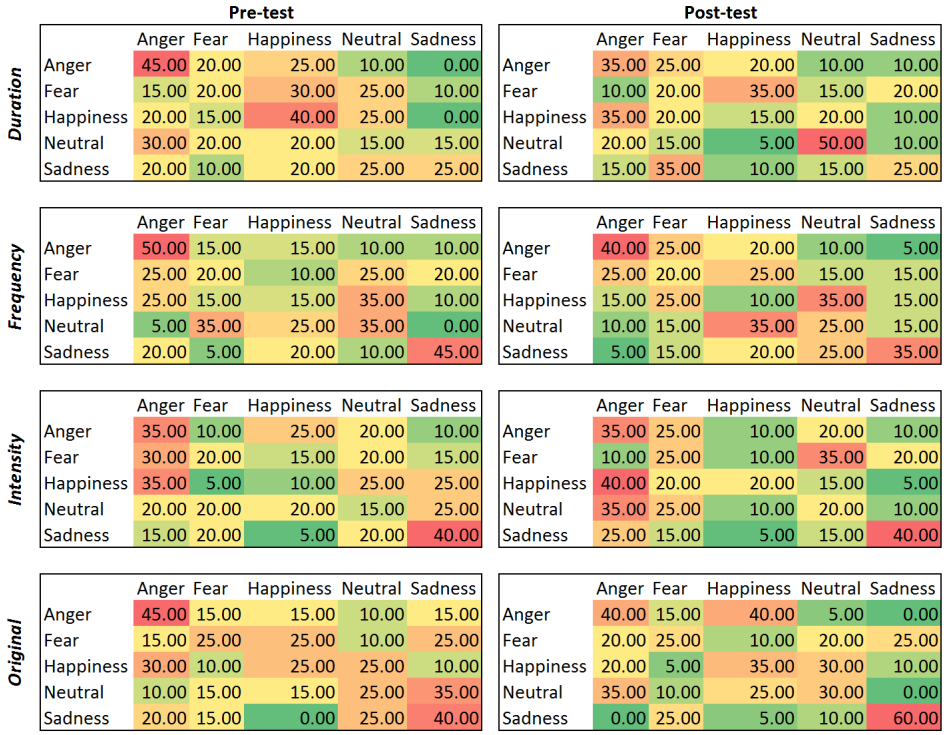


Figure 26: Heat-mapped confusion matrices for music stimuli during the pre-test and post-test-transfer phases, depicting the percentage of responses in each emotion category, for each stimulus emotion. Columns denote presented emotions, rows denote emotion judgement responses. Red = higher values, green = lower values.

for. This was true for emotion identification accuracy at the pre-test, post-test and post-test transfer phases. Additionally, no significant correlations were found with any of the MUSE components or EQ scores.

5.4 Discussion

5.4.1 Summary of results: Study 3

Firstly, as hypothesised, emotion identification performance was generally above chance level. In fact, as in Study 1, mean accuracy was above 20% correct in all of the different feature attenuation conditions, for all testing phases. This is con-

cordant with the results obtained in the previous chapter, as well as the literature reviewed therein.

Secondly, and also as hypothesised, there was a significant effect of testing phase upon emotion identification accuracy, suggesting that the training paradigm was successful in improving participants' ability to decode emotion from CI-simulated stimuli. Recognition accuracy was significantly enhanced in both the post-test and post-test-transfer phases. This implies that participants not only learned how to decode emotion in familiar stimuli, but were able to apply their learning to previously-unheard stimuli. Therefore, one can conclude that the training helped participants to learn the underlying associations between different auditory features and different emotions, rather than simply memorising these associations for specific stimuli.

The results also showed clear variability in terms of identification accuracy across the different emotions. This was expected on the basis of both prior research (Scherer et al., 2001) and the results documented in the previous chapter. However, this effect was also quite variable, depending on the auditory feature attenuation condition considered. While Sadness was generally the best-decoded emotion (and indeed was recognised with the greatest level of confidence), other emotions were particularly well-decoded only in certain conditions. On the whole, the tendency for participants to overestimate Neutral and Anger in their responses (as noted in the previous chapter) was reduced by the training. However, the specific confusion between Anger and Happiness remained somewhat prevalent. This most likely occurred because Anger and Happiness are conveyed by overlapping features – both tend to be fast rate and high intensity (Juslin and Laukka, 2003; Yang and Lugger, 2010). Therefore, what ordinarily separates them is frequency-based information – for example, Anger is usually associated with greater sensory dissonance (Juslin & Laukka, 2003). In the

case of CI simulation, the disambiguating frequency information is distorted, hence the two emotions are frequently confused.

There appeared to be little direct influence of Feature attenuation condition itself on emotion identification accuracy. Frequency stimuli were decoded with greater proficiency at pre-test, but this difference was eliminated by the training paradigm, such that participants in each of the feature attenuation conditions performed roughly equivalently during the post-test-transfer phase. Somewhat surprisingly, performance in the Original condition was very similar to the various feature attenuation conditions at pre-test, although the effect of training was far greater for these stimuli, such that they were the best-decoded stimuli during both post-test blocks. Summarily, training seemed to alter the observed pattern of results in two noticeable ways. Firstly, Original stimuli clearly benefited most from the training. However, differences between the remaining feature attenuation conditions were minimised by the training, with participants in each condition achieving an average emotion identification accuracy of approximately 50% correct. This implies that relatively good performance was attainable in each of the feature-attenuation conditions, and that the 'primacy' of frequency-based cues for decoding emotional speech may disappear after a period of training and/ or acclimatisation. However, the results also do not indicate that selectively attending to duration- or intensity-based cues (i.e. cues preserved better by the CI) should necessarily improve performance.

As was anticipated, there was a significant interaction effect of Emotion $\times$ Feature attenuation condition, upon emotion identification accuracy. This interaction was manifest as differential improvements for specific combinations of Feature attenuation condition $\times$ Emotion. Stimuli in the Frequency condition were recognised relatively equivalently, regardless of the emotion being judged, whereas all of the

other feature attenuation conditions (including Original) were associated with significantly varying recognition accuracy, dependent on the emotion. For example, the Duration condition appeared to be particularly effective at preserving the expression of Sadness, and to a lesser extent Fear.

The most likely explanation for these results is that some emotions are better conveyed by certain features than others, and hence benefit to a greater extent when these features are preserved and trained with (assuming that these features are also well-enough preserved by the CI simulation). According to this explanation, an emotional stimulus should be more accurately decoded (within the current paradigm) when the following are true: A) the stimulus is conveyed clearly by salient auditory features; B) these features are adequately preserved by the feature attenuation manipulation; C) these features are adequately preserved by the CI simulation. Considering the aforementioned example, Sadness was well decoded in the Duration condition because it is portrayed clearly by temporal cues (i.e. slow speech rate), which were not attenuated, and were relatively unaffected by the CI simulation.

Lastly, as predicted, empathy (EQ) was not significantly related to inter-individual differences in emotion identification accuracy, dispelling the suggestion that the lack of an effect in the previous chapter might have been due to participants unfamiliarity with the CI-simulated stimuli. It remains a possibility, however, that an effect was not observed here because the data lacked sufficient variability, with participants' EQ scores all falling well within a 'normal' range (from a possible range of 0-80, participants' scores ranged from 23-61). In any case, this possibility lies outside the scope of the current investigation.

By contrast, the hypothesis that musical training might influence emotion identification was partially supported. Significant correlation was found between the 'Musical

Training' sub-component of the MUSE and individuals' overall accuracy in emotion identification. However, this was only true when considering data from the post-test-transfer phase of the experiment. There are two potential explanations for this. Firstly, participants perhaps needed to acclimatise to the sound of the CI-simulated stimuli before musical expertise could confer an advantage. Therefore, with additional test phases, participants had more time to adjust to the stimuli, until the effect of musical expertise became noticeable. Alternatively, those with greater musical expertise may have specifically benefited more from the training so that, even though they were equivalent at the start, the training led to a greater level of improvement. For example, it could be that these individuals were better able to learn about the underlying correspondences between specific auditory features and different emotions, thereby having greater success when appraising novel stimuli.

5.4.2 Summary of results: Study 4

As with the speech stimuli in Study 3, emotion identification performance was generally above chance level for music stimuli. Mean accuracy was above 20% correct in each of the different feature attenuation conditions, across all of the testing phases. This represents a small improvement compared to the results obtained in the previous chapter, and is consistent with the broader notion that above-chance performance is attainable in musical emotion identification tasks, with CI-simulated stimuli. However, compared to the Study 3, participants tended to be slightly less confident overall in their emotion judgements. This is consistent with the findings of the previous chapter, and suggests that participants had greater difficulty in decoding emotion with music stimuli than with speech.

In contrast to what was hypothesised, the training paradigm (indexed by perfor-

mance in the different test phases) made no significant difference to participants' emotion identification accuracy in Study 4. The most likely explanation for this is that emotion identification accuracy was almost unchanged between the pre-test and post-test-transfer phases. In other words, the training paradigm appeared to be far less effective with music stimuli, compared to speech. Although participants showed modest improvement between the pre-test block and the first post-test, this improvement was lost at the post-test-transfer stage. Therefore, rather than being entirely ineffective, it appeared that the training paradigm was only effective in improving identification of previously-experienced stimuli. This is non-trivial, since participants still showed improved identification of stimuli that they had not previously been told the 'correct answers' for. Therefore, the results do not denote a failure to generalise beyond the training set per se, but rather reflect an inability to apply the training to previously unencountered stimuli.

Nonetheless, this pattern of results is disappointing, suggesting that the training paradigm may only be effective in facilitating emotion identification once participants are already familiar with the musical stimuli in question. Given the very small magnitude of improvement with training, however, it is possible that a comparable effect could have been achieved simply by assessing participants twice, without training in between. A potential explanation for this lack of an effect might have been a lack of diversity in terms of the stimuli presented. In other words, one might ask: 'Were enough musical examples presented, such that an effective strategy could be developed that would help listeners infer emotion in novel stimuli?' To this end, presenting a greater variety of musical melodies may have made more apparent any aspects of performance expression that were invariant across compositions. The limited number of different melodies was motivated by the consideration that listeners

might try to learn correspondences between individual melodies and expressed emotions, although there may be a better balance to be found.

As hypothesised, a significant effect of Emotion upon overall identification accuracy was found. This finding is consistent with all of the results reported thus far, and also with previous literature. Specifically, Sadness was recognised substantially more accurately than would have been expected by chance, whilst Fear was recognised substantially less accurately. Since this effect occurred independently of Feature attenuation, the most likely explanation for this pattern of results is that Sadness stimuli were especially robust to disruption by the CI simulation and feature attenuation manipulations (as found in Study 3, for speech stimuli). Conversely, Study 2 showed that Fear was the most difficult emotion to identify in music, without the CI simulation. Therefore, it is not surprising that participants struggled to identify Fear accurately in this case, regardless of the training paradigm.

The training had several other, smaller effects on the way that the different emotions were recognised. Firstly, Neutral stimuli were overestimated at pre-test, but the effect of training appeared to counteract this somewhat. By contrast, Anger was overestimated even after training, and continued to be confused with Happiness, as was also the case for speech stimuli. In fact, some confusions appeared to 'crystallise' as a result of the training. For example, Sadness and Fear emerged as commonly-confused emotions during the post-test-transfer phase. This was most likely a reflection of 'faulty' learning during the training phase – i.e. learning to associate some pattern of features uniquely with expression of a particular emotion, when in actuality these are shared with a different emotion.

Contrary to the hypothesis, no significant influence of Feature attenuation condition upon emotion identification accuracy was found. Confidence ratings indicated that

participants were less certain about their emotion judgements when only frequency information was preserved, although this effect did not translate to a difference in decoding performance. This finding appears to contradict the result obtained in Study 2, which showed a significant main effect of the feature attenuation manipulation. However, Study 2 included both CI-simulated and non-CI-simulated stimuli, whereas the current study only included the latter, which may explain this discrepancy. In fact, Study 2 documented a (non-significant) trend towards the effect of Feature attenuation condition varying according to the CI simulation condition. Therefore, the current results may be best explained in terms of the extent to which musical stimuli are affected by the CI simulation in general. In other words, it is possible that the simulation had such a large negative impact upon emotion expression, that any further deterioration caused by feature attenuation had only relatively minor consequences for emotion identification. With this said, performance did appear to be slightly better in the Original condition, compared to the different feature attenuation conditions. That is, there was a noticeable benefit to emotion identification by having multiple acoustic cues, which was not surprising, since any identification task should be easier when additional relevant information is provided.

Additionally, no significant interaction effect of Emotion $\times$ Feature attenuation condition upon emotion identification accuracy was found. The primary reason for this was that, as noted above, the different feature attenuation conditions appeared to make very little difference to recognition accuracy. While there was no systematic interaction, however, the results showed that decoding accuracy was substantially better for some isolated combinations of Emotion and Feature attenuation. For example, Neutral stimuli were recognised relatively well in the Duration condition, at post-test. Though, in the absence of a significant interaction effect, such potentially-

spurious patterns should be interpreted with caution. The overarching result was that Feature attenuation condition made very little difference, and this was largely uniform for each of the different emotions. It is difficult to ascertain precisely why this was the case, though this may indicate that were global characteristics which were A) informative to some extent about expressed emotion, and B) relatively unaffected by the feature attenuation manipulation.

Lastly, the hypotheses relating to musical training and empathy were not supported. More precisely, no component of the MUSE, nor the EQ, was significantly associated with individual differences in emotion identification accuracy, during any of the test phases. This was contrary to previous findings which have suggested relationships with emotion decoding abilities for both musical expertise (Lima and Castro, 2011; Thompson et al., 2004) and empathy (Philip et al., 2010; Stevens et al., 2001). In the previous chapter, it was suggested that the lack of effects for musical expertise and empathy might stem from participants' relative unfamiliarity with the CI-simulated stimuli. This interpretation is less likely here however, given that participants had much more opportunity to acclimatise to the stimuli. Thus, the alternative possibility offered in Study 2 – that the lack of results denote an underlying lack of diversity within the sample, in terms of musical and empathic abilities – should be considered the more plausible interpretation.

5.4.3 Conclusions

These studies added further support to the finding that emotions can be identified, above chance level, in CI-simulated speech and music stimuli. As in the previous chapter, this level of performance was attained in a 5-AFC task, and even with acoustic feature attenuation, although performance varied according to the combi-

nation of emotion and feature attenuation condition.

Studies 3 and 4 were successful in addressing the limitations of Studies 1 and 2, via the implementation of a training paradigm with blocked presentation of stimuli. Considered together, these studies showed that auditory emotion identification with CI-simulated stimuli may be improved by a perceptual learning paradigm. However, the effect of training was much stronger for speech stimuli, and therefore may require refinement in order to be useful for music. With speech stimuli, the training appeared to assist in acclimatising participants to the CI simulation, since the extent to which Anger and Neutral were overestimated was reduced over time. However, this effect was much less clear for music stimuli.

However, even with speech stimuli, there appeared to be a limit approached in terms of the identification accuracy that was attainable – in Study 3, mean percentage correct scores for each of the feature attenuation conditions (excluding Original) were approximately 50%, and improved little from post-test to post-test-transfer. Though it is difficult to speculate about what might have happened with further training, it is potentially very difficult to achieve better accuracy than this, with only one 'category' of auditory features fully preserved. In other words, the results suggest that the combination and integration of different auditory features is especially important for the perception of emotion (at least in the case of CI simulation), rather than any one class of features being differentially important. Though the CI simulation effects different auditory features to varying degrees, it may be that the different auditory features investigated here are each either more or less useful, depending on the emotion being judged. For example, Duration and Intensity appeared to be especially salient cues, respectively, for the perception of Sadness and Anger. Therefore there may be no universal 'optimum' listening strategy for CI-simulated

emotion decoding, but rather a set of strategies, the utility of which depends on the stimulus being considered.

This account is tenable for Study 3, but less so for Study 4. As in the previous chapter, participants appeared to find the task more difficult overall with music stimuli. In Study 4, both the training paradigm itself and the feature attenuation manipulation failed to substantially impact participants' identification accuracy. However, there was a small extent to which emotion decoding accuracy was improved for specific emotions in certain feature attenuation conditions, as in Study 3.

In any case, the primacy of frequency information observed in the previous chapter was not present to the same degree in either of these studies, once participants had undergone training. Therefore, these studies support the notion that is no special reason to focus on frequency as an auditory cue. However, no clear advantage was observed for the Duration or Intensity conditions either (excluding some specific emotions), which were better preserved by the CI simulation. Therefore, Studies 3 and 4 imply that the best listening strategy would most likely take into account all of the available cues, prioritising these dynamically according to whichever cues are most salient (i.e. depending on the emotion being expressed).

In summary, the studies carried out in this chapter build upon the account advanced in the previous chapter in several important ways. Emotions were identifiable above chance still for CI-simulated speech and music, but training appeared to substantially benefit only the former. With respect to the underlying listening strategies responsible, the picture seems to be more complicated than initially assumed. On one hand, the results corroborate Juslin (2000)'s 'Lens' model: the Brunswikian notion that musical emotions are perceived via utilisation of a diverse set of cues, which may be flawed to varying extents, but nonetheless are informative when considered together.

That is, optimal emotion decoding with CI-simulated stimuli is not achieved simply by attending preferentially to better-preserved features. On the other hand, the results also showed that emotion identification *is* possible with a focus on one particular feature, but that generally listening strategies are A) adaptive to the emotion being expressed, and B) include multiple features in a non-linear manner.

5.4.4 Limitations

Although Studies 3 and 4 were largely successful in addressing the limitations of the studies in the previous chapter, these experiments were not without shortcomings.

Firstly, it is possible that better emotion identification performance might have been attained by using a longer training paradigm. The whole experiment took up to ninety minutes for participants to complete, leaving little room to extend the training session further, though the study could have incorporated more training by distributing it over multiple experimental sessions. While several studies investigating auditory perceptual learning have favoured a single-session design (e.g. Burkholder et al., 2004; Loebach, Pisoni, and Svirsky, 2009), longitudinal approaches have also been adopted. For example, in a study investigating the recognition of musical instruments via CI-simulated stimuli Driscoll et al. (2009) administered fifteen training sessions over a period of five weeks. Although significant improvements were not observed from the third week onwards, the results suggest that multiple training sessions, distributed over time, may be beneficial. Since this study specifically related to the perception of music, and considering the results of Study 4, it is possible that the use of a more extended paradigm might be especially helpful for emotion identification with musical stimuli.

Secondly, the diversity of the stimuli presented in these experiments was relatively

limited, with only four different sentences in Study 3 and four melodies in Study 4. Providing more variation in the training sets might have made it easier for participants to perceive commonalities in emotional expression across different stimuli. In other words, facilitating awareness of underlying correspondences between different auditory features and emotions, that were not tied to the specific stimuli presented. In principle, one would expect this to improve the ability of participants to identify emotion in previously-unheard stimuli. Again, this limitation might be more relevant for musical stimuli than speech, since the training paradigm was notably less effective for the former.

Lastly, it is possible that individual differences in empathic responding and musical abilities were not well-captured by the particular measures of empathy and musicality that were employed. Different results may be obtained if other measures of musicality were taken into account – particularly if more objective assessments were used. For example, to address the question of whether musical perceptual abilities confer an advantage for emotion recognition with CI-simulation, it may therefore be preferable to include a sophisticated assessment of various musical perceptual abilities (e.g. Law and Zentner, 2012; Müllensiefen et al., 2014).

Similarly, the EQ (Baron-Cohen & Wheelwright, 2004), as well as being inherently subjective, might be limited by reducing empathy to a single component. Some researchers have argued, for example, for a distinction between affective and cognitive aspects of empathy (Shamay-Tsoory, Aharon-Peretz, & Perry, 2009), and perhaps only the latter is relevant to the current studies (participants were explicitly instructed to *identify* emotions, regardless of how they themselves may have felt). Not only does the EQ fail to disentangle these two components, it also contains several questions that relate to social desirability (Lawrence, Shaw, Baker, Baron-Cohen, &

David, 2004) and several others that appear to be irrelevant, according to an internal consistency assessment (Wakabayashi et al., 2006). Therefore, it remains possible that at least one aspect of empathy is related to performance in CI-simulated emotion decoding, and that this relationship might have been revealed by the use of a more objective measure. However, such measures were not incorporated within these studies primarily due to time constraints. Potential influences of both empathy and musical expertise upon CI-simulated emotion decoding were a more peripheral strand of the current investigation, and therefore it was not feasible to incorporate detailed and objective measures of both within the experimental paradigm used. Unfortunately, this means that all conclusions made about these potential covariates must be tempered by the fact that they were based solely on self-reported data.

5.5 Further exploration of these data

In order to explore the data gathered in this chapter in greater detail, the next chapter documents several computational modelling approaches that were undertaken, intended to supplement and add additional insights to the results from Studies 3 and 4. Summarily, the aim was to uncover which auditory features were most likely used by participants when identifying emotion in CI-simulated speech and music, and whether or not these combinations of features were optimal. In others words, given the results that have been obtained thus far, the aim of the computational modelling was to uncover the processes potentially responsible. Chapter 6 describes in depth the modelling work carried out, and discusses the outcomes in relation to the studies carried out previously.

6 Computational modelling of cochlear implant-simulated emotion identification

6.1 Introduction

6.1.1 Aims and rationale

Although the previous chapter provided valuable insight into the effects of training upon cochlear implant (CI) simulated emotion recognition, it also left unanswered the question of *how* this performance was achieved. However, a central objective of this thesis has been to investigate the listening strategies used by participants when decoding emotional, CI-simulated stimuli. Therefore, the primary objective of this chapter was to uncover the auditory features that contributed most substantially to participants' emotion judgements in Studies 3 and 4. To achieve this, no new data were gathered in this chapter, but instead the data collected in the previous chapter were analysed using a variety of computational modelling techniques.

While the preceding chapter attempted to make inferences about participants' likely listening strategies based upon identification accuracy in the different feature attenuation conditions, the current chapter re-framed the question somewhat. Specifically, the primary question asked in this chapter was: 'Given non-attenuated stimuli (CI simulation notwithstanding), which features are most important for emotion identification?'. Therefore, rather than considering scenarios where specific listening strategies were artificially encouraged by feature attenuation, the current chapter was concerned with which listening strategies participants adopted 'spontaneously'.

Accordingly, the current chapter focussed only on emotion identification performance in the Original (i.e. non-feature attenuated) condition. In Studies 3 and 4, emotion

decoding in the Original condition was consistently more accurate than in any of the feature attenuation conditions. In this condition, all of the auditory features preserved by the CI-simulation were available to listeners, but it is not known which features were actually used. The observation of increased accuracy with Original stimuli suggests that performance is better when all auditory features are preserved, as opposed to only one – but this does not necessarily mean that all of the preserved features were actually utilised (or at least utilised to the same degree).

To provide additional insight into participants' listening strategies, and thereby build a better understanding of the results from Studies 3 and 4, a series of computational modelling techniques were implemented to explore these data further.

6.1.2 Outline of the approach taken

The purpose of this chapter was to investigate the relationships between lower-level acoustic properties of speech and music stimuli, and the ways that emotional states were ascribed to these stimuli. Therefore, to begin with, various auditory 'features' were extracted from the stimuli presented in Studies 3 and 4. These were used to provide descriptions of the different stimuli, in terms of their underlying auditory properties, for all of the subsequent analyses.

Broadly speaking, there were two separate strands of enquiry pursued in this chapter, concerned with A) modelling the listening strategies most likely to have been used by participants, and B) discovering the 'optimal' listening strategies that could potentially be used instead.

Firstly, by using the emotion identification data from Studies 3 and 4 (in the form of the confusion matrices generated), it was possible to estimate a 'likely' pattern of

features used by human listeners, in order to make the emotion judgements. That is, a model was trained, not to achieve optimal emotion classification accuracy, but to produce a pattern of classification results that emulated as closely as possible the typical pattern of judgements errors made by human listeners.

By knowing which acoustic features best predicted the average error distribution characterising human performance, one can infer that the human listeners achieved those results via a listening strategy prioritising similar features (Peng et al., 2012). Of course, this approach discounts individual variance in listening strategies, and additionally provides no guarantee that human performance did indeed arise from the listening strategy identified – only that it was statistically the most likely model. Unfortunately, this method is susceptible to noise in cases in which listeners used inconsistent or random 'strategies' (e.g. guessing). Nonetheless, this technique was expected to provide at least a reasonable approximation of the acoustic features likely to have been preferentially attended to by human listeners. Therefore, this approach provides an approximate answer to the question of which auditory features participants attended to spontaneously in Studies 3 and 4.

Secondly, by using overall emotion classification accuracy as the criterion by which to optimise a model's performance, it was possible to discern a theoretically 'ideal' set of features relevant for the judgement task, and therefore an ideal listening strategy for emotion perception (at least within this paradigm). That is, the subset of acoustic features that leads to the least overall error in emotion classification represents, in principle, the best possible features which could be prioritised by human listeners, in order to maximise performance. It follows logically that selectively attending to the features mostly strongly associated with variance in stimulus emotion should improve one's ability to make accurate emotion judgements.

Additionally, comparing this feature set to the one previously derived by considering actual human performance should provide an indication of whether the latter might be improved by direction of attention towards different stimulus features. In other words, these two approaches should provide an indication of an 'average' human listening strategy, and a theoretically optimal one. Any disparity observed between the two therefore denotes potential for improvement in human performance.

Addressing a slightly different question, cluster analysis was utilised, in order to determine the extent to which the speech and music stimuli effectively expressed the five distinct emotions intended (Anger, Fear, Happiness, Neutral and Sadness). In Studies 3 and 4, participants appeared to find the identification task rather difficult, especially with music stimuli, and made several pervasive confusions (e.g. Anger and Happiness). Therefore, this unsupervised learning approach investigated the degree to which the auditory features of the various stimuli could be satisfactorily divided into five different 'clusters' (in principle corresponding to the five emotions). This was performed for both CI-simulated and non-CI simulated stimuli, in order to investigate how the simulation may have disrupted the structure present in the auditory feature data (thereby making the judgement task more difficult).

6.1.3 Summary of research questions

In summary, the computational modelling carried out in this section attempted to provide insight with respect to several important research questions. Firstly, an important aim was to explain which auditory features, when used as input data for an emotion classifier, would lead to performance most similar to that observed for human listeners. This was intended to provide an approximation of the listening strategies most likely to have been used by these participants. A second aim was to

find the set of auditory features that would lead to the most accurate classification of emotions, irrespective of what human listeners did. This was intended to uncover what the optimal listening strategy might have been, and how closely this resembled the approach taken by human participants. Lastly, this section examined how effectively the five distinct emotional states were conveyed by acoustic features of the stimuli, and to what extent this was disrupted by the CI simulation. The aim of this was to uncover how much overall task difficulty was increased due to the simulation, compared with how effectively the stimuli were in expressing separable emotional states.

6.1.4 Hypotheses

The analyses carried out in this chapter were largely exploratory, and therefore there were few formal hypotheses. In fact, null hypothesis significance testing was not utilised in relation to any of the analyses performed. Nonetheless, some general predictions were made.

Firstly, it was predicted that there would be at least some level of discrepancy between the auditory features associated with human performance, and those selected for optimum performance. In particular, since frequency-based cues are preserved less well in the CI-simulated stimuli, it was hypothesised that the optimal model might assign less importance to these features.

In both cases, it was expected that the feature sets chosen would contain more than one type of cue (i.e. at least two of duration, frequency, intensity). Via the feature attenuation manipulation, Studies 3 and 4 implied that a listening strategy prioritising just one type of auditory feature would not explain human performance, nor would it lead to the best possible identification accuracy.

It was assumed that there would be some underlying subset of features reliably associated with optimal emotion identification performance. Therefore, it was expected that the different computational modelling techniques used would converge upon approximately similar (or at least overlapping) sets of features, enabling conclusions to be drawn about which auditory features were most important.

It was predicted that the cluster analysis would find less substantial structure in stimuli disrupted by the CI simulation. Since this processing disrupted various cues pertinent for emotion perception, it was expected that the different emotional states would be less readily separable, when the simulation was applied. Additionally, it was considered that the music stimuli might be less easily separable than the speech stimuli, which would explain the disparity in emotion identification difficulty, as experienced by participants. Finally, it was expected that the subset of features producing the best clustering should be similar to those selected by the previous models for optimal emotion identification accuracy.

The next section provides an overview of the different computational modelling methods used, beginning with a detailed description of the extraction of auditory features from the stimuli.

6.2 Methods

6.2.1 Dataset

All of the subsequent analyses reported in this chapter are based on the data obtained in Studies 3 and 4. Specifically, the stimuli used were those in the Original condition (no feature attenuation), presented during the training phases of these experiments. As in Studies 3 and 4, all of these stimuli were processed with the NBV-based CI

simulation.

For speech, forty stimuli were used in total: four different sentences × two speaker genders (female, male) × five emotions.

Likewise, for music, forty stimuli were used in total: four different melodies × two instruments (violin, voice) × five emotions.

For modelling related to the estimation of human listening strategies, unless otherwise specified, confusion matrices used were derived from the post-test-transfer phases of Studies 3 and 4, again from only the Original condition. Therefore, the confusion matrices used denote the average errors made by a group of five participants.

6.2.2 Feature extraction

Two different input feature vectors were created – one each for speech and music – derived directly by analysing these stimuli. Analysis to derive the input acoustic features was based upon known parameters implicated in the communication of emotion in speech and music, corresponding to a large degree with those used in a previous study by Coutinho and Dibben (2013). Some of the features extracted were the same for both speech and music, while in other cases separate measures were used, as appropriate.

For all of the stimuli (including both speech and music), mean intensity was calculated in MATLAB by averaging the absolute values of the intensity spectrum.

Intensity variation was calculated by computing the amplitude envelope for each stimulus using the Hilbert Transform, smoothing this using a second-order Butterworth low-pass filter (100 Hz cut-off), converting to dB re 1 and taking the standard deviation.

Median frequency was calculated using the *medfreq* function provided in MATLAB.

Various measures relating to the spectral content of the different signals were derived using MIRtoolbox for MATLAB (Lartillot & Toiviainen, 2007). Firstly, Brightness was calculated using the *mirbrightness* function, and indexes the proportion of energy in frequencies above 1.5 kHz, compared to below.

Similarly, Spectrum centroid, calculated by *mircentroid* provides a weighted mean of signal energy at different frequency bands.

Roughness (*mirroughness*) denotes the amount of sensory dissonance present in the signal, caused by close frequency peaks occurring simultaneously.

Spectral flux was estimated using the *mirflux* function, and describes how quickly the power spectrum of a given signal changes over time.

Sharpness was measured by first generating an instantaneous loudness contour (phons) for each stimulus, following Glasberg and Moore (2002)'s time-varying loudness model. This contour was then used as an input to Zwicker and Fastl's 1990 sharpness calculation. Roughly, this measure corresponds to the proportion of a signal's loudness is determined by its high-frequency energy.

Modulation spectrum information was calculated, in order to describe the extent to which changes in amplitude envelope occurred over time (Greenberg and Kingsbury, 1997; Hermansky, 1997). This measure was derived by first extracting the amplitude envelope from each signal using the Hilbert transform, and smoothing it using a fourth-order Butterworth low-pass filter (100 Hz cut-off). The smoothed envelope was then converted to the frequency domain using the fast Fourier transform (FFT) so that the magnitude of amplitude modulations could be compared across different frequencies. Lastly, the resulting modulation spectrum amplitude was converted to

a logarithmic scale, in order to compress the values into a narrower range. From this information, ten separate measures were derived, denoting the amount to which amplitude modulation occurred in each signal at linearly spaced frequencies from 1.5 to 15 Hz.

For speech, prosody contours for each stimulus were generated using Prosogram, a prosodic intonation transcription program for Praat (Mertens, 2004). From this data, frequency range was extracted, along with the total number of rises and falls present in the prosodic contour. No equivalent melodic contour information was extracted for the music stimuli, since this feature would likely have captured irrelevant variation between the different melodies included (musicians did not utilise melodic variation to produce emotional expression, since they were restricted to the use of only performance cues).

For speech, rate was estimated using De Jong and Wempe (2009)'s algorithm within Praat, which detects syllabic nuclei within stimuli and from these calculates the number of syllables per minute (SPM). For music stimuli, tempo in beats per minute (BPM) was estimated using the *mirtempo* function within MIRtoolbox for MATLAB (Lartillot & Toiviainen, 2007).

Lastly, additional variables, 'Pulse clarity' and 'Event density', both derived via MIRtoolbox, were also included for the music data. Respectively, these features provide estimates of the salience of a the musical pulse or 'beat' and the number of musical 'events' occurring per second.

6.2.3 Screening of features

Prior to carrying out any computational modelling, the input features described above were screened, in order to check for multicollinearity. The presence of highly correlated predictor variables makes it very difficult to assess the independent contribution of those variables to a given model – which would have been especially problematic in this case, given the aim to illuminate the relative importance of difference acoustic features.

Multicollinearity was assessed via calculation of the Variance Inflation Factor (VIF) associated with each input variable, using the *usdm* package for R (Naimi, a.s. Hamm, Groen, Skidmore, & Toxopeus, 2014). In accordance with the threshold suggested by Myers (1990), any variables for which the VIF was greater than 10 were considered problematic with respect to multicollinearity. For this reason, three variables were removed from both the speech and music input vectors: Brightness, Sharpness and Spectral Flux.

For speech, the resulting input vector, used for all subsequent analyses, is illustrated in Table 13. The final input vector for music is summarised in Figure 14.

For each input vector, there was also a corresponding vector containing the expected values, i.e. a list of 'ground truth' values, denoting the emotions that each stimulus was intended to express (Anger, Fear, Happiness, Neutral, Sadness).

Table 13: Description of the input feature vector used in order to assign speech stimuli to different emotion categories. From this list of features, various selection methods were utilised in order to choose subsets of features most relevant for this classification.

Feature	**Description**
Mean intensity	Mean of absolute values of signal intensity
Intensity variation	Standard deviation of signal intensity contour
Median frequency	Calculated using *medfreq* function in Matlab
Frequency range	Calculated using *Prosogram* for Praat
Spectrum centroid	Calculated with MIRtoolbox (*mircentroid*)
Roughness	Calculated with MIRtoolbox (*mirroughness*)
Rises	Total 'rises' in the prosodic contour (*Prosogram*)
Falls	Total 'falls' in the prosodic contour (*Prosogram*)
Length	Length of signal, divided by sampling frequency
Speech rate	Calculated with *Prosogram* for Praat
Event density	Calculated with MIRtoolbox (*mireventdensity*)
Pulse clarity	Calculated with MIRtoolbox (*mirpulseclarity*)
Mod. spectrum 1	Proportion of temporal dynamic variation at 1.5 Hz
Mod. spectrum 2	Proportion of temporal dynamic variation at 3 Hz
Mod. spectrum 3	Proportion of temporal dynamic variation at 4.5 Hz
Mod. spectrum 4	Proportion of temporal dynamic variation at 6 Hz
Mod. spectrum 5	Proportion of temporal dynamic variation at 7.5 Hz
Mod. spectrum 6	Proportion of temporal dynamic variation at 9 Hz
Mod. spectrum 7	Proportion of temporal dynamic variation at 10.5 Hz
Mod. spectrum 8	Proportion of temporal dynamic variation at 12 Hz
Mod. spectrum 9	Proportion of temporal dynamic variation at 13.5 Hz
Mod. spectrum 10	Proportion of temporal dynamic variation at 15 Hz

Table 14: Description of the input feature vector used in order to assign music stimuli to different emotion categories. From this list of features, various selection methods were utilised in order to choose subsets of features most relevant for this classification.

Feature	**Description**
Mean intensity	Mean of absolute values of signal intensity
Intensity variation	Standard deviation of signal intensity contour
Median frequency	Calculated using *medfreq* function in Matlab
Spectrum centroid	Calculated with MIRtoolbox (*mircentroid*)
Roughness	Calculated with MIRtoolbox (*mirroughness*)
Length	Length of signal, divided by sampling frequency
Tempo	Calculated with MIRtoolbox (*mirtempo*)
Pulse clarity	Calculated with MIRtoolbox (*mirpulseclarity*)
Event density	Calculated with MIRtoolbox (*mireventdensity*)
Mod. spectrum 1	Proportion of temporal dynamic variation at 1.5 Hz
Mod. spectrum 2	Proportion of temporal dynamic variation at 3 Hz
Mod. spectrum 3	Proportion of temporal dynamic variation at 4.5 Hz
Mod. spectrum 4	Proportion of temporal dynamic variation at 6 Hz
Mod. spectrum 5	Proportion of temporal dynamic variation at 7.5 Hz
Mod. spectrum 6	Proportion of temporal dynamic variation at 9 Hz
Mod. spectrum 7	Proportion of temporal dynamic variation at 10.5 Hz
Mod. spectrum 8	Proportion of temporal dynamic variation at 12 Hz
Mod. spectrum 9	Proportion of temporal dynamic variation at 13.5 Hz
Mod. spectrum 10	Proportion of temporal dynamic variation at 15 Hz

6.3 Logistic regression classification

6.3.1 Procedure

In order to avoid redundancy in methodological description, the results from the various computational modelling approaches explored are henceforth organised firstly by the technique used, and secondly by the stimulus type (speech, music).

The first modelling approach taken was the construction of a multinomial logistic regression classifier. The overarching aim was to 'learn' the five emotion labels corresponding to the stimuli described above. For both speech and music, this analysis was run two times: once to emulate human performance as closely as possible (i.e. by producing similar patterns of errors and correct responses), and once to achieve optimal emotion identification.

Multinomial logistic regression works in a very similar way to binomial logistic regression, in that various different input features are used to predict some valued categorical outcome (in this case, emotion). However, instead of using a sigmoid function to choose the best prediction from one of two options, this process occurs N times, where N is the number of levels of the variable to be predicted. For each option, the model's confidence in predicting that option is compared to that of the remaining options (i.e. option x versus 'other', option y versus 'other', and so on). The prediction strengths for all of the available options are then evaluated, and the option with the highest value is outputted by the model as the prediction.

For any given vector of input features, it is unlikely that each feature will be equivalently useful in predicting the outcome variable. In fact, some features may encode irrelevant information, in which case classification accuracy may be negatively affected by their inclusion in the model. Further, once a certain number of input

features have been considered, there may be a point reached at which the addition of subsequent features does not substantially improve performance. Therefore, in order to determine which of the input features to use, feature selection was performed.

All methods for feature selection method include both a criterion (to assess how well the model performs with different feature combinations), and a method for choosing which features to test (Mao, 2004). In this case, two different criteria were chosen. Firstly, features were chosen according to their contribution to the model's overall accuracy in emotion identification (i.e. percentage correct responses). Secondly, features were chosen according to the correlation between the confusion matrix associated with the model's classifications, and the average confusion matrix for human responses.

To decide which combinations of features to test, two selection methods were considered. Firstly, the sequential forward selection (SFS) algorithm was used (Devijver & Kittler, 1982). This approach assesses the predictive value of each of the input features, considered individually, and chooses the feature that is the best predictor as its starting point. With this predictor entered into the model, the improvement in prediction associated with the addition of each remaining feature is assessed, in order to find the next-best predictor. This process continues until the addition of subsequent features no longer produces an improvement in prediction accuracy, at which point the set of features selected can be used to train the model.

To complement this approach, sequential backward selection (SBS, also known as backward elimination) was also utilised (Devijver & Kittler, 1982). This algorithm operates in essentially the same way as SFS, albeit in reverse – that is, all possible features are included in the initial model and, with each iteration, the feature making the weakest contribution to the performance of the model is removed. This process

continues until either only one feature remains, or the procedure is otherwise terminated. In principle, SBS is more likely to choose a solution with more features, since more solutions are considered with a greater number of features, while the opposite is true for SFS. Therefore, SBS less optimal if we are expecting a solution with only a small number of features, and vice versa (Guyon & Elisseeff, 2003). For the current application there was no clear theoretical motivation to favour one approach over the other, and so the results of both were inspected, as is relatively common practice (Guyon and Elisseeff, 2003; Mao, 2004).

To validate the logistic regression classification model, leave-one-out cross-validation (LOOCV) was used, meaning that the model 'learned' the relevant acoustic feature configurations associated with the different emotions by analysing $N - 1$ of the 40 stimuli (Refaeilzadeh, Tang, & Liu, 2009). The one remaining stimulus was then used to validate the model, i.e. a prediction was made about the emotion expressed by this stimulus, based on its auditory features. This procedure was carried out $N = 40$ times, such that each stimulus was used for cross-validation data exactly once. For small samples, LOOCV is an efficient way to utilise the available data, since only one data point is omitted at each step. Using the largest possible training set leads to greater stability of the validation estimate, while reducing bias that might arise from the partitioning of a dataset into 'training' and 'validation' data (Cawley, 2006).

6.3.2 Results: Speech

For the speech stimuli, optimising the classifier according to percentage accuracy in emotion identification, using both SFS and SBS, produced the results shown in Figure 27. In this graph, classification accuracy is plotted as a function of the number of features chosen by both SFS and SBS. The best emotion classification

achieved was 55% correct, which was obtained using four features chosen by SFS: Median frequency, Roughness, Length and Modulation spectrum component 2. SBS also suggested a four-variable model, albeit with slightly different features: Median frequency, Roughness, and Modulation spectrum components 8 and 9. This model achieved 52.50% correct emotion identification.

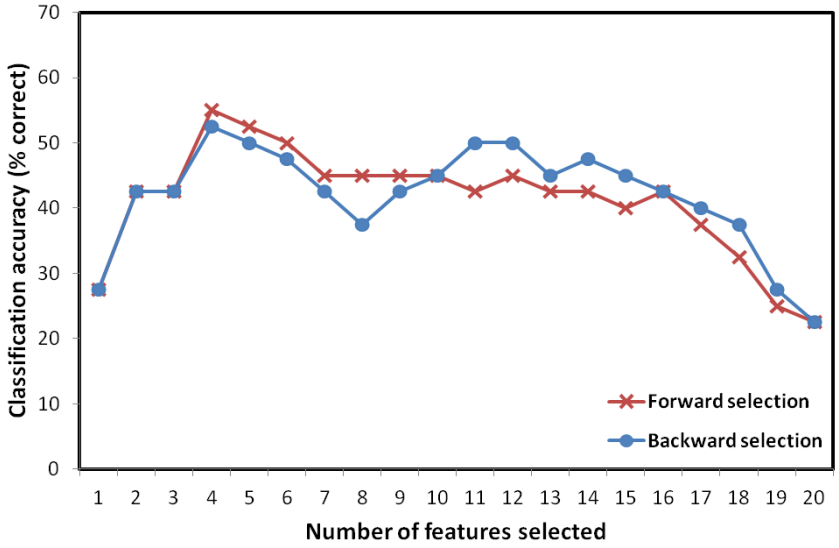


Figure 27: Prediction accuracy of the multinomial logistic regression classifier for speech stimuli, as a function of the number of input features considered.

Optimising the model according to the confusion matrix generated from human performance produced the results shown in Figure 28. In this case, SFS and SBS yielded very similar results, which was reassuring – while these algorithms rarely produce identical results, greater similarity indicates a higher likelihood that a globally 'optimum' model was found. Both methods suggested an optimal model consisting of the same four features previously selected by SFS in order to optimise performance: Median frequency, Roughness, Length and Modulation spectrum component 2. Using this model for emotion classification produced a confusion matrix that was strongly, positively correlated with the averaged human confusion matrix, $r = .89$.

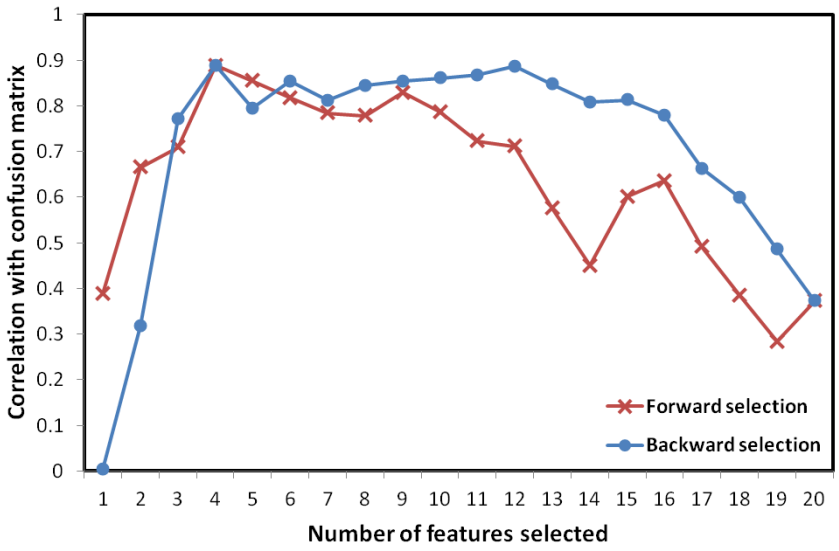


Figure 28: Prediction accuracy of the multinomial logistic regression classifier for speech stimuli (with respect to human performance), as a function of the number of input features considered.

Compared to the aforementioned data for the post-test-transfer phase of Study 3, using the confusion matrix for the pre-test phase resulted in a relatively similar model. This time, Median frequency, Roughness, Intensity variation and Modulation spectrum component 10 were chosen. Again, the outcome confusion matrix was strongly correlated with that of human listeners, $r = .80$.

6.3.3 Results: Music

For the music stimuli, optimising the model according to percentage correct emotion classification produced the results shown in Figure 29. Using SFS, the best emotion identification attained was 52.50% correct, which was possible with five features: Intensity variation, Roughness, Pulse clarity, and Modulation spectrum components 4 and 6. SBS was able to achieve slightly performance (55% correct), although this required ten input features: Roughness, Length, and Modulation spectrum compo-

nents 2 through to 9.

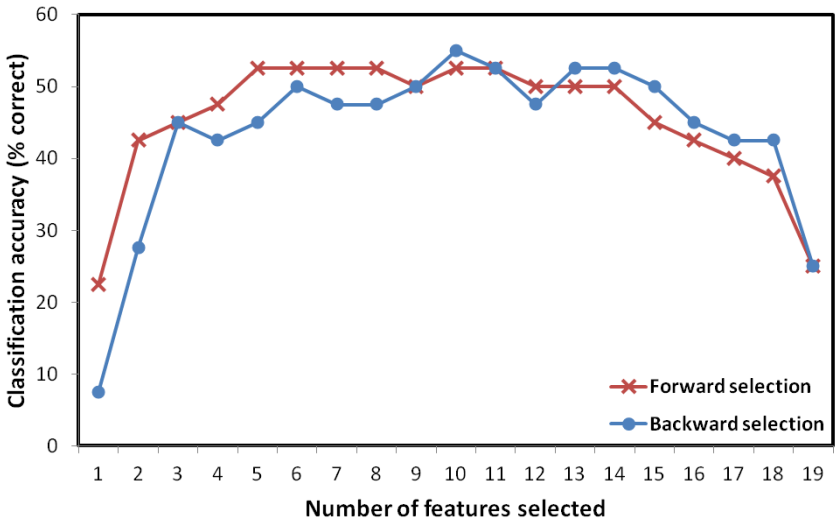


Figure 29: Prediction accuracy of the multinomial logistic regression classifier for music stimuli, as a function of the number of input features considered.

Optimising the model according to the confusion matrix generated from human performance produced the results shown in Figure 30. Using SFS, the strongest correlation with human performances was achieved using eighteen features. However, very little substantial improvement occurred once more than five features were considered. In fact, a strong, positive correlation of $r = .84$ was achieved using five features: Roughness, Pulse clarity, Event density, and Modulation spectrum components 2 and 9. Using SBS produced a visible peak in correlation strength with the use of fifteen features, although the performance was very close to the aforementioned five-feature model, $r = .86$. The fifteen-feature model included largely the same features, with the addition of Intensity variation, Spectral centroid, Length, and all but one of the Modulation spectrum components.

Compared to the aforementioned data for the post-test-transfer phase of Study 4, using the confusion matrix for the pre-test phase resulted in a somewhat similar model.

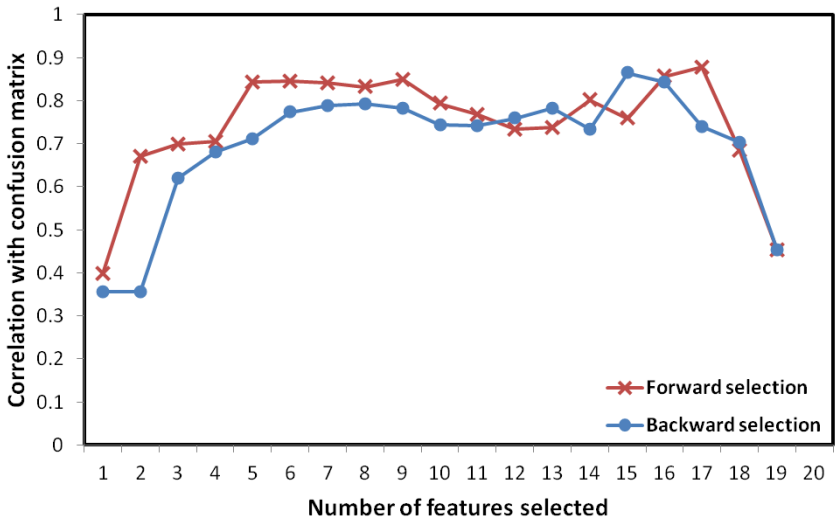


Figure 30: Prediction accuracy of the multinomial logistic regression classifier for music stimuli (with respect to human performance), as a function of the number of input features considered.

This time, seven features were chosen: Intensity variation, Roughness, Spectral centroid, Tempo, Pulse clarity, Event density and Modulation spectrum component 10 were chosen. Again, the outcome confusion matrix was strongly correlated with that of human listeners, $r = .80$.

It is worth noting that, in describing the models chosen by SFS/ SBS (for both speech and music) in this section, no p-values were included. This is because traditional hypothesis testing assumes only two 'models': the Alternative hypothesis (H_1) and the Null hypothesis (H_0). However, in the case of model selection, far more candidate models are evaluated (essentially many different instantiations of H_1), and therefore a conventional null hypothesis test, using the optimum model selected a posteriori, would have been inappropriate for this application (Raftery, 1995).

In the next section, a slightly different approach was taken, investigating whether similar models would be chosen if an alternative, more comprehensive method for model selection was used.

6.4 Criterion-based model selection

6.4.1 Procedure

To complement the approach described above, this section explores the use of information criterion-based, rather than stepwise, model selection. Again, multinomial logistic regression was used, though this time there was no predictive element involved. The aim was simply to find the logistic regression model(s) best able to explain the observed data. Therefore, unlike the previous section, this approach was intended only to provide insight into the best possible listening strategies that participants might utilise, irrespective of the previously observed patterns of errors characterising human performance.

Essentially, rather than considering different models by adding or removing one feature at a time, this criterion-based approach exhaustively evaluated every possible unique combination of features. The most immediate advantage of this method is that the search space is wider – therefore, with more models considered, the likelihood of finding the best available model(s) is increased (Whittingham, Stephens, Bradbury, & Freckleton, 2006). In fact, this approach also places less emphasis on uncovering a single 'best' model (Whittingham et al., 2006). While SFS and SBS naturally result in just one model being chosen, criterion-based selection is easily adaptable to output the best N models. This is advantageous considering the exploratory nature of this research. That is, the goal was not necessarily to build a de facto model to best predict emotion, but rather to investigate which auditory features consistently emerged as being important. In other words, this approach is "more about hypothesis generating than testing, providing a basis for subsequent work and data collection" (Symonds and Moussalli, 2011, p. 15).

By comparison to the previous section, this approach adopts both a different method for choosing which features to test, and different criteria by which to evaluate the resulting models. For the former, an exhaustive screening approach was utilised, such that every possible combination of potential features was tested. Compared to sequential feature selection, this method was much more comprehensive – for the speech stimuli, a total of 209 models were evaluated using the sequential selection approach, whereas 1,048,575 models were evaluated with exhaustive screening. To evaluate each of these models, the Akaike Information Criterion (AIC) was computed – a measure of the relative quality of each regression model. AIC indicates the likelihood of information loss resulting from choosing a particular model, and therefore a lower AIC usually indicates a better model. The AIC is reduced by goodness of fit, but is increased by the addition of more input features, in order to prevent overfitting (Symonds & Moussalli, 2011).

Additionally, the Bayesian Information Criterion (BIC) was examined as an alternative model selection criterion. The BIC is very similar in principle, but tends to penalise model complexity more severely, and consequently is likely to choose a smaller model (Dziak, Coffman, Lanza, & Li, 2012). Although there are theoretical considerations relating to the prevalence of type I and type II errors, which might lead one to favour the AIC over the BIC (or vice versa), it is often suggested that both criteria be used in a complementary fashion in order to establish a range of plausible models (Dziak et al., 2012). Accepting this suggestion, and without strong reasons to favour either a larger or smaller model, both the AIC and BIC are used here. However, in the case of large discrepancies between the models chosen, slight favour should perhaps be given to those chosen by the BIC. The primary goal of this research is not to find the model that minimises mean squared prediction error for

the observed data, but to uncover which acoustic features are most important for accurate emotion perception emotion in general, which is more closely aligned to the ethos of the BIC (Burnham & Anderson, 2004).

To implement this automated model selection procedure, the R package *glmulti* was used (Calcagno & de Mazancourt, 2010). For both speech and music, the same input vectors were used as described in the previous section.

6.4.2 Results: Speech

For the speech stimuli, the three most optimal models, chosen according to the AIC, are summarised in Table 15. The first model included a total of fifteen features, whilst the second- and third-best models were much more parsimonious. However, according to the guidelines outlined by Burnham and Anderson (2004), models with a difference in AIC of $4 \leq \Delta_i \leq 7$ (where Δ_i denotes the difference between the i^{th} model and the optimum model), have 'considerably less support'. Since $\Delta_i = 4.67$ and 5.57 for the second- and third-best models, respectively, these must be considered inferior, in terms of the AIC.

Table 15: Summary of the three best multinomial logistic regression models for predicting emotion in speech, selected according to the AIC.

Model	**AIC**	**Features**
1	32.00	Intensity variation, Median frequency, Frequency range, Spectral centroid, Roughness, Rises, Falls, Length, Speech rate, Modulation spectrum components 3, 4, 5, 8, 9 and 10
2	36.67	Mean intensity, Intensity variation, Modulation spectrum component 9
3	37.57	Mean intensity and Intensity variation

As expected, using the BIC Information Criterion as a measure of model quality yielded a slightly different result, tending to select less complex models. The top three most optimal models, chosen according to the BIC, are summarised in Table 16. In

this case, the BIC values for the three models are all very similar, and therefore there is no clear reason to prefer one over the other. In terms of Burnham and Anderson (2004)'s rule of thumb, the second- and third-best models have 'considerable' support, relative to the optimum model, and should receive due consideration.

Table 16: Summary of the three best multinomial logistic regression models for predicting emotion in speech, selected according to the BIC.

Model	**BIC**	**Features**
1	41.39	Intensity variation
2	42.64	Mean intensity, Intensity variation
3	42.92	Roughness

6.4.3 Results: Music

For the music stimuli, the three best models, selected according to the AIC, performed equally well (each AIC = 22.00, and each contained ten features). The three most optimal models, selected on the basis of the AIC, are summarised in Table 17. Since the models were equally preferred, one can be confident that features common to all three are likely to be important. Features occurring in at least two of the models were as follows: Intensity variation (2), Length (3), Tempo (3), Event density (2), Modulation spectrum components 1 (2), 2 (2), 4 (3), 5 (3), 7 (2), and 8 (2).

Table 17: Summary of the three best multinomial logistic regression models for predicting emotion in music, selected according to the AIC.

Model	**BIC**	**Features**
1	22.00	Mean intensity, Intensity variation, Length, Tempo, Modulation spectrum components 4, 5, 6, 7, 8 and 10
2	22.00	Intensity variation, Spectral centroid, Length, Tempo, Event density, Modulation spectrum components 1, 2, 4, 5 and 8
3	22.00	Median frequency, Roughness, Length, Tempo, Event density, Modulation spectrum components 1, 2, 4, 5 and 7

As with the speech data, using the BIC as the models selection criterion, resulted

in much simpler models being chosen. The three most optimal models chosen using the BIC are summarised in Table 18. According to Burnham and Anderson's (2004) criteria, the second model has considerable support, relative to the optimum, while the third model has less support. Therefore, features common to the first two models may be interpreted as being most important. Accordingly, Intensity variation and Event density (which were in fact present in all three models) were important features, while Modulation Spectrum component 5 may also have been informative.

Table 18: Summary of the three best multinomial logistic regression models for predicting emotion in music, selected according to the BIC.

Model	**BIC**	**Features**
1	31.77	Intensity variation, Event density
2	33.02	Intensity variation, Event density, Modulation spectrum component 5
3	34.44	Intensity variation, Event density, Tempo

Next, cluster analysis was used to investigate how effectively the auditory feature could be partitioned into five 'clusters' corresponding to the different emotions.

6.5 Cluster analysis

The motivations for performing cluster analysis this were twofold. Firstly, this approach aimed to investigate how difficult the emotion identification was, on principle, based on how effectively the auditory stimuli could be placed into different groups based on their auditory features. In principle, the more easily separable the stimuli were, the easier the emotion identification task would be. Secondly, the approach investigated which specific subsets of auditory features produced the best clustering solutions, for comparison with the results of the previous modelling work, as well as human data from the previous chapter.

To answer these questions, k-means clustering (i.e. Lloyd's algorithm) was used.

This algorithm aims to partition the data space into a specified number of clusters, based on the input features provided, such that each within-cluster sum of squares (WCSS – i.e. the sum of the distances between each point in the cluster from its respective centre) is minimised. Mathematically, this problem may be expressed as minimising the following function, J, where k = number of clusters, n = number of data points, x denotes an individual data point, and c denotes the respective centroid:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

k-means clustering works by first initialising a given number, k, of 'means', which act as the centres of these k clusters. In this case, these initial values were chosen according to the k-means++ method (Arthur & Vassilvitskii, 2007) – the first mean is chosen at random from the data space, and then means up to the k^{th} are chosen from the data points that remain, with probability proportional to their squared distance from the nearest existing mean. The aim of this strategy is to ensure that the initial means are sufficiently spread out, such that the likelihood of the clustering algorithm finding a sub-optimal solution is minimised (Arthur & Vassilvitskii, 2007).

Once k 'means' have been initialised, the algorithm assigns each data point (i.e. vector of input features) to whichever mean leads to the lowest WCSS. Once all of the data points are assigned to these means – thereby forming clusters around each mean – the 'means' are updated to be the centroids of the new clusters. This procedure is performed repeatedly, until the subsequent iterations no longer cause the positions of the centroids to change (Figure 31).

For the current application, k-means clustering was run with five clusters, corresponding in principle to the five emotions that were expressed by the talkers/ mu-

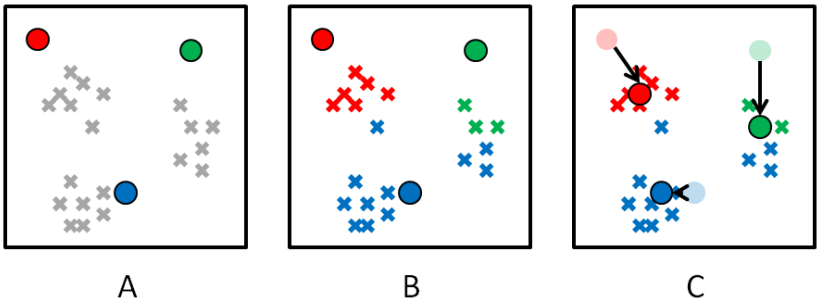


Figure 31: Illustration of the processes of initialisation (A), assignment (B) and updating (C) in k-means clustering, for an example two-dimensional dataset.

sicians. In order to determine the best subset of acoustic features to use for the clustering, the sequential forward selection algorithm was used. For each candidate feature subset, the corresponding clustering solution was evaluated by considering the extent to which stimuli from the same emotion class were assigned to common clusters. Thus, feature subsets that resulted in a greater proportion of like-stimuli being placed into the same clusters were preferred.

The k-means algorithm is liable to generate a different clustering solution each time it is run, owing to the random initialisation of the cluster centroids. Therefore, the optimum set of features to use, and the overall pattern of cluster assignment is likely to vary somewhat each time the k-means procedure is run. This limitation was circumvented by running the k-means algorithm (with SFS) a total of 1,000 times. The optimum feature subset for clustering was then taken as the most commonly chosen feature subset across all of the runs, and the pattern of cluster assignment for each of the different emotions was recorded cumulatively over all runs.

Prior to running any of the cluster analyses, all data pertaining to the speech and music stimuli were standardised and mean centred, to improve clustering accuracy, and ensure that differences in variance or magnitude of measurement for particular

features did not have any undue effects (Mohamad & Usman, 2013). This process, along with the 1,000 k-means runs with SFS, was implemented within R.

6.5.1 Speech

For the speech data, the most commonly selected feature subset included seven features: Mean intensity, Intensity variation, Median frequency, Frequency range, Roughness, Length and Modulation spectrum component 3.

For this particular clustering, the mean Dunn index (DI) over 1,000 runs was 0.27 (calculated using R package *clValid*, (Brock, Pihur, Datta, & Datta, 2008)), indicating that the ratio between the largest intra-cluster distance between two data points to the smallest distance between two data points in different clusters was rather small (Dunn, 1974). This measure is fairly conservative, and is inherently biased by the presence of outliers, but nevertheless reflects a less-than-optimal clustering (Dunn, 1974).

Similarly, the mean silhouette value, indicating the extent to which each data point fits well within its cluster, as compared to neighbouring clusters, was 0.26 (Rousseeuw, 1987). This measure varies between -1 and 1, with 0.25 being the rule-of-thumb threshold indicating that some substantial (albeit weak) structure has been found in the data (Kaufman & Rousseeuw, 2009). Lastly, the between-cluster sum of squares (SS) / total SS was 60.10 %, indicating that this percentage of the global variance in the data was explained by the clustering. Figure 32 illustrates this five-dimensional clustering, using multiple two-dimensional plots. The clusters were not very well separated overall. However, of the features included, Roughness (and to a lesser extent Length) appeared to be most helpful in assigning stimuli to distinct classes.

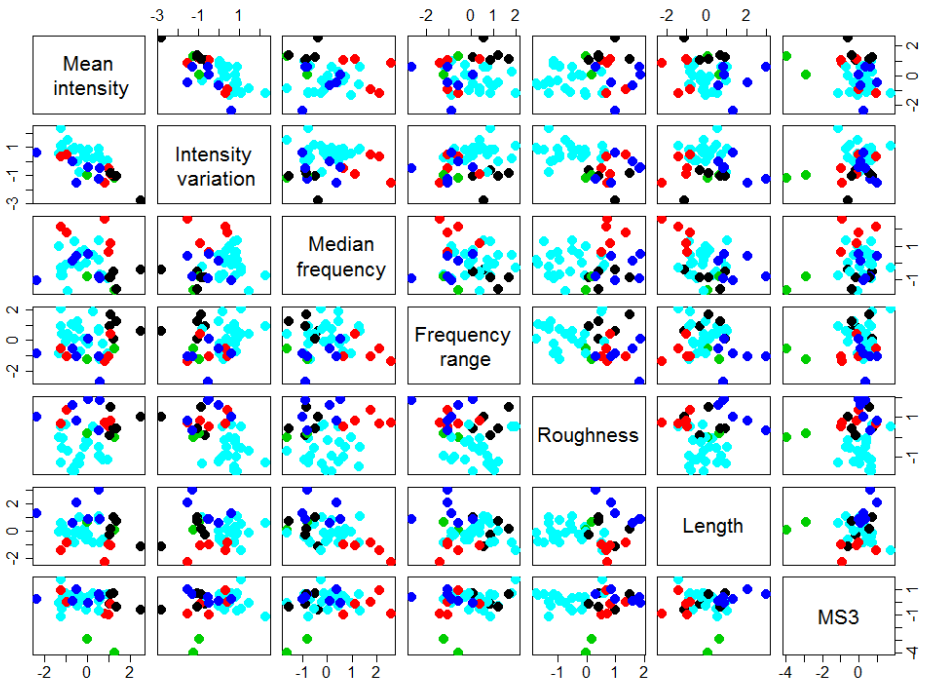


Figure 32: Plots illustrating the seven-feature, five-centre clustering achieved by the k-means algorithm. The seven-dimensional clustering is visualised here by plotting each of the seven features against each other, and colouring the data points according to the cluster to which it was assigned.

Table 19 shows the extent to which stimuli belonging to each of the different emotion classes were assigned to the same cluster. Note that all values along the diagonal are equal to 8,000, since there were a total of eight stimuli per emotion × 1,000 runs, and each stimulus must be in the same cluster as itself. For the remaining elements of the table, values may vary between 0 (indicating that stimuli from the two emotions considered were *never* assigned to the same cluster) and 8,000 (denoting that stimuli from the two emotions considered were *always* put in the same cluster). Hence, excepting values along the diagonal, better, more distinct clustering performance is indicated by smaller numbers.

The most frequent clustering 'confusion' occurred between Anger and Happiness, with the eight stimuli from these emotion classes being assigned to the same cluster a total of 6,770 times over the 1,000 runs. Sadness was the most distinctive emotion class since, of the different emotion classes, sad stimuli were assigned to the same cluster as the other emotional stimuli least often.

Table 19: Cluster 'confusion matrix', depicting the extent to which speech stimuli from the different emotion classes were assigned to the same cluster (cumulative over 1,000 k-means runs).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**
Anger	8000	4551	6770	2917	1230
Fear		8000	4235	4954	1827
Happiness			8000	2743	1147
Neutral				8000	2413
Sadness					8000

6.5.2 Music

For the music data, the most commonly-chosen feature vector for optimal clustering consisted of only two features: Pulse clarity and Modulation spectrum component 8.

For this clustering, the mean DI over 1,000 runs was 0.24, again indicating that the

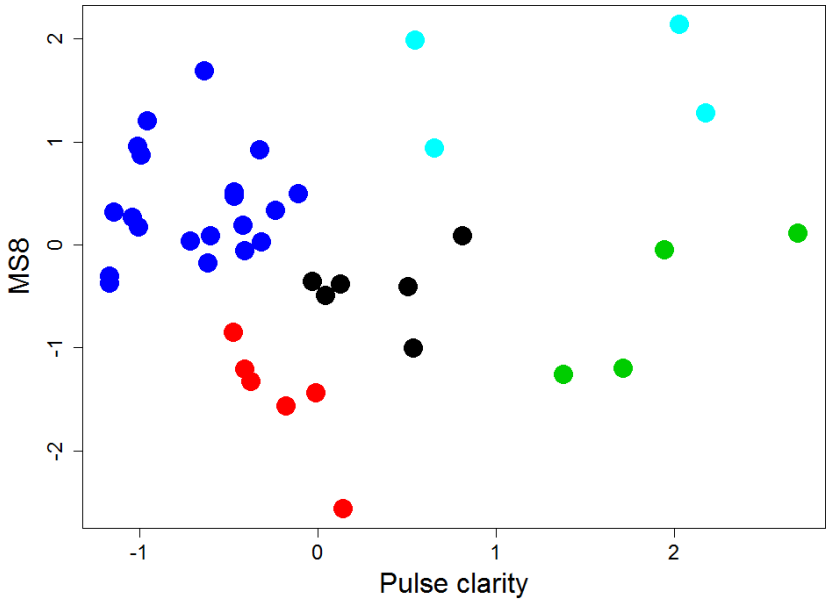


Figure 33: Plot illustrating the two-feature, five-centre clustering achieved by the k-means algorithm, for the music stimuli. Data points are coloured according to the cluster to which it was assigned.

clustering was less than optimal. Similarly, the mean silhouette value for all data points was 0.40, indicating that some substantial but weak structure was found in the data (Kaufman & Rousseeuw, 2009). Lastly, the between SS / total SS was 78.90 %, indicating that this proportion of variance was explainable by the clusters. Figure 33 illustrates the two-dimensional clustering with respect to Pulse clarity and Modulation spectrum component 8. As with the speech stimuli, the clusters constructed for music stimuli did not appear to be linearly separable. Additionally, assignment of stimuli to the different clusters was unevenly distributed, with many more cases being assigned to some clusters compared to others.

Table 20 shows the extent to which stimuli belonging to each of the different emotion classes were assigned to the same cluster. As was the case for speech stimuli, the most frequent clustering confusion occurred between Anger and Happiness, with stimuli

from these emotion classes being assigned to the same cluster a total of 5,357 times. Sadness was again the most distinctive emotion class since, of the different emotion classes, sad stimuli were assigned to the same cluster as the other emotional stimuli least often. However, the differences between emotions were small, indicating that all were confused to a similar degree.

Table 20: Cluster 'confusion matrix', depicting the extent to which music stimuli from the different emotion classes were assigned to the same cluster (cumulative over 1,000 k-means runs).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**
Anger	8000	3663	5357	3820	2772
Fear		8000	3583	3914	4555
Happiness			8000	3192	3005
Neutral				8000	4102
Sadness					8000

6.6 Cluster analysis with non-CI simulated stimuli

The clustering solutions achieved were less than optimal, but it was unclear whether this was due to the stimuli themselves not being distinctive enough (in terms of the different emotions expressed), or whether this was a consequence of the CI simulation applied. In order to explore this, cluster analysis was performed with the same emotional speech and music stimuli as before, this time without the CI simulation. The exact same procedure was used as for the previous cluster analyses: 1,000 k-means runs with SFS were performed, and the results aggregated over all of the runs.

6.6.1 Speech

Interestingly, for the unprocessed speech stimuli, the most commonly-chosen feature vector for optimal clustering consisted of only a single feature: Median frequency.

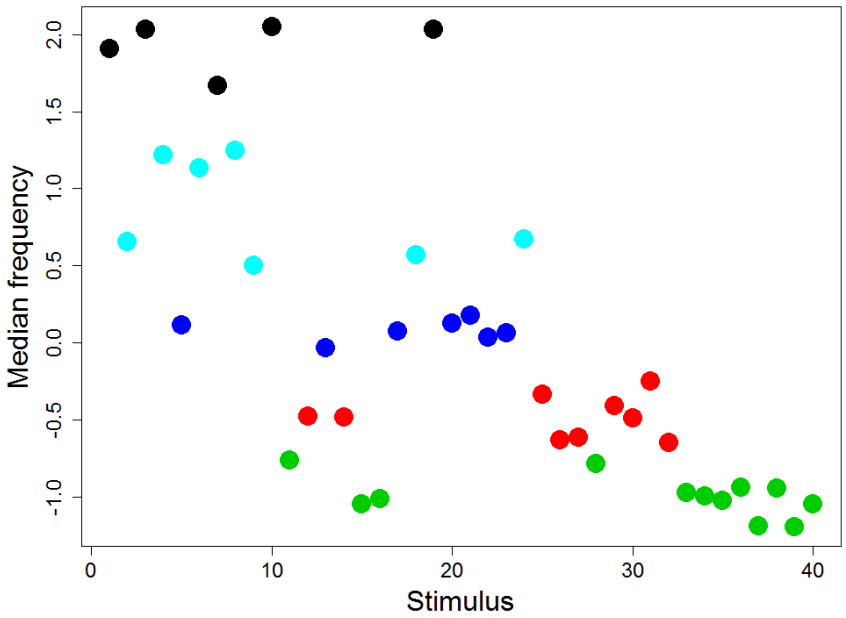


Figure 34: Scatter-plot illustrating the one-feature, five-centre clustering achieved by the k-means algorithm. Data points are coloured according to the cluster to which it was assigned.

For this clustering, the mean DI over 1,000 runs was 0.24, again indicating that the distance between clusters were not very well separated or compact. However, this metric is potentially misleading, since the data used were one-dimensional. The between SS / total SS was 97.20 %, indicating that the vast majority of the total variance was explainable by the clusters. Figure 34 illustrates the one-dimensional clustering with respect to Median frequency. As opposed to the non-simulated speech stimuli, the data were more clearly separated, and the clusters constructed appeared to be linearly separable.

Table 21 shows the extent to which speech stimuli belonging to each of the different emotion classes were assigned to the same cluster. Compared to the CI-processed speech stimuli, there were fewer overall 'confusions'. Unlike the CI-processed stimuli, the most frequent clustering confusion for original stimuli occurred between Fear

and Happiness, with stimuli from these emotion classes assigned to the same cluster a total of 4,376 times. As with the CI-simulated stimuli, sadness was the most distinctive emotion class, as sad stimuli were assigned to the same cluster as the other emotional stimuli least often.

Table 21: Cluster 'confusion matrix', depicting the extent to which speech stimuli from the different emotion classes were assigned to the same cluster (cumulative over 1,000 k-means runs).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**
Anger	8000	2812	4282	1140	352
Fear		8000	4376	4330	2492
Happiness			8000	3272	1097
Neutral				8000	2917
Sadness					8000

6.6.2 Music

For the unprocessed music stimuli, the most commonly-chosen feature vector for optimal clustering consisted of four features: Intensity variation, Roughness, Event density, and Modulation spectrum component 1.

For this clustering, the mean DI over 1,000 runs was 0.22, again indicating that the clustering was less than optimal. Similarly, the mean silhouette value for all data points was 0.31, indicating that some substantial but relatively weak structure was found in the data. Lastly, the between SS / total SS was 66.50 %, indicating that this proportion of variance was explainable by the clusters. Figure 35 illustrates this four-dimensional clustering, using multiple two-dimensional plots. Unlike the speech stimuli, the overall quality of the clustering appeared to change very little for non-simulated stimuli. None of the features included appeared to be especially effective in distinguishing between different clusters.

Table 22 shows the extent to which music stimuli belonging to each of the different

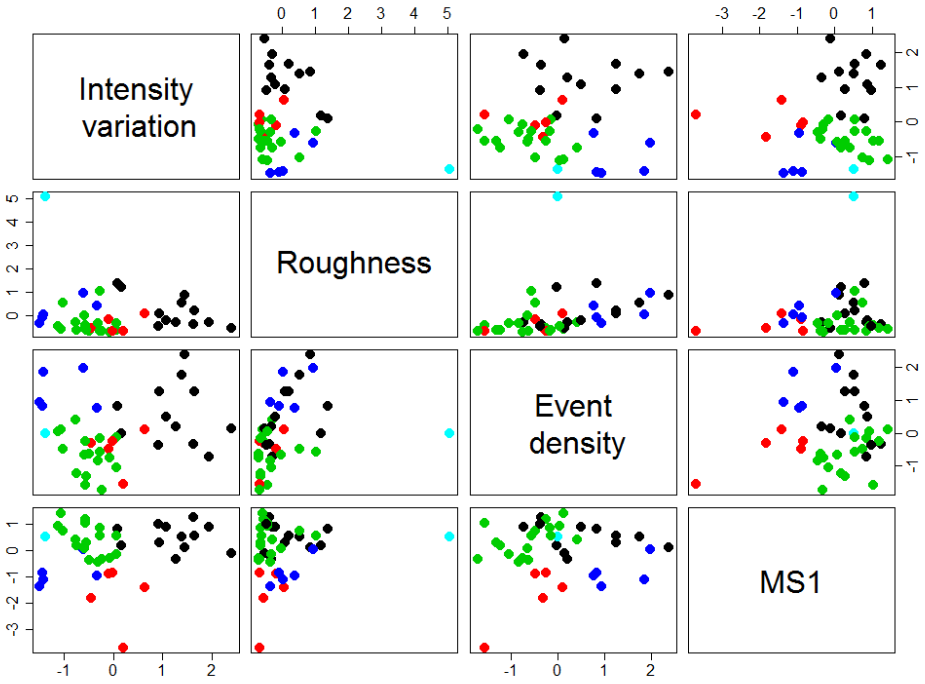


Figure 35: Plots illustrating the four-feature, five-centre clustering achieved by the k-means algorithm. The four-dimensional clustering is visualised here by plotting each of the four features against each other, and colouring the data points according to the cluster to which it was assigned.

emotion classes were assigned to the same cluster. The overall number of 'confusions' made was similar to the result obtained using CI-processed music stimuli. As with the CI-processed stimuli, the most frequent clustering confusion for original stimuli occurred between Anger and Happiness, with stimuli from these emotion classes clustered together a total of 5,727 times. However, unlike the CI-simulated stimuli, anger rather than sadness was the most distinctive emotion class: angry stimuli were assigned to the same cluster as the other emotional stimuli least often.

Table 22: Cluster 'confusion matrix', depicting the extent to which music stimuli from the different emotion classes were assigned to the same cluster (cumulative over 1,000 k-means runs).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**
Anger	8000	3454	5727	3169	2896
Fear		8000	3696	5054	5044
Happiness			8000	3369	3221
Neutral				8000	4932
Sadness					8000

In all, the clustering analysis had limited success, even for non-CI-simulated speech and music stimuli: the clusters constructed were relatively poorly separated and most analyses implied that the clustering solutions found were less than optimal (Dunn, 1974). Typically, failure to find a well-separated clustering solution indicates that the data in question are not linearly separable (Elizondo, 2006). In other words, stimuli belonging to the different emotion classes may not be separated by drawing straight lines (or more accurately, hyperplanes, due to the high dimensionality of the data).

However, the classification techniques used thus far have aimed to build linear models to predict and explain variance in expressed emotions as a function of underlying auditory features. Accordingly, the next section considered whether (and if so how) the relative importance of the different input features would change if a linear rela-

tionship between acoustic parameters and emotions was not presupposed.

6.7 Non-linear emotion classification

6.7.1 Procedure

In this section, based on the results obtained from the cluster analyses, a non-linear approach to classification was employed. In principle, it was considered that this method might result in a more accurate classification solution for the apparently non-linearly separable data. Specifically, random forest classification was performed in order to explore both: A) how accurately the different emotional stimuli could be identified using a non-linear methodology, and B) how the 'most important' auditory features identified by this method would compare to those chosen using linear modelling techniques.

Random forest is a non-linear, 'ensemble' approach that can be applied to classification problems, and is so-named because it essentially consists of a large collection of decision trees, each acting as a classifier (Breiman, 2001) (see Figure 36).

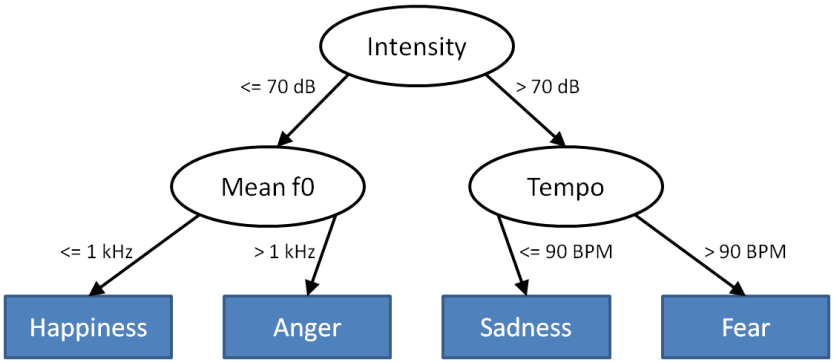


Figure 36: Illustrative example of a hypothetical decision tree that might be used for the classification of musical stimuli according to expressed emotion. Note: decision criteria included are arbitrary

In random forest classification, a number of subsets are first created from the dataset,

containing random combinations of both observations and variables. For example, if a dataset contained 1,000 observations of 50 variables, a subset might contain only 100 observations of 10 variables (with both observations and variables chosen at random). For each of these subsets, a unique decision tree is created, in order to best classify the data in terms of the outcome variable.

A decision is created by first calculating the overall entropy (a measure of uncertainty or unpredictability) of the dataset. This is given by the following equation, where n is the number of data points, and p denotes a particular point:

$$Entropy = \sum_{i=1}^{n} p_i \log(p_i)$$

Next, for each variable, the reduction in overall entropy associated with splitting the dataset according to that variable is calculated. This is usually referred to as 'information gain'. For continuous variables, the threshold used to divide the data in two is selected according to which value would cause the greatest reduction in overall variance.

With each iteration of the decision tree algorithm, a new 'branch' is added, and the variable associated with the greatest information gain is used to partition the data. This process continues until there are no data left to partition. At this point, the groups of data at the ends of these branches all belong to the same class of the output variable, and are fittingly labelled 'leaves'. The decision tree is now completed and can be used to make (non-linear) predictions, classifying new stimuli according to the hierarchical rule structure established.

Once all of the decision trees have been created in this way, predictions may be made using a process called bootstrap aggregating (or 'bagging'), in which each decision tree independently makes a prediction, and the accepted outcome is the classification

with the most 'votes' (Breiman, 2001). As with the multinomial logistic regression classifier described previously, LOOCV was incorporated within the random forest classification procedure. Thus, for both speech and music, of the forty stimuli, thirty-nine were used to train the classifier, and predictions were made about the remaining one. This was repeated forty times, such that all of the stimuli were predicted exactly once.

In this case, a drawback to random forest classification, is that the element of randomness involved means that the relative importance of features for classification will not necessarily be the same each time the procedure is run. Thus, it would be difficult to make definitive assertions about which features contribute the most to emotion classification accuracy, since the overall contribution of any given feature is subject to change each time the algorithm is run (depending upon which subsets of observations/ variables happen to be chosen). In order to circumvent this problem, the entire random forest classification procedure was run one hundred times.

Random forest classification was implemented in R, using the *randomForest* package (Liaw & Wiener, 2002). Default parameters were used, meaning that five hundred trees were constructed, each based on a subset of four variables ($\sqrt{}$ of the total number of variables, rounded down) and twenty-six stimuli ($N \times 0.632$, rounded up). For each tree, given the use of LOOCV, these stimuli were randomly selected from the subset of thirty-nine training stimuli. The logic required to run random forest classification with LOOCV one hundred times was coded in R, and the results were visualised using the *ggplot2* package (Wickham, 2009).

6.7.2 Results: Speech

For the speech data, random forest classification with LOOCV produced the pattern of predicted emotions shown in Figure 37, recorded cumulatively over one hundred 'runs' of the random forest algorithm.

		Response				
		Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**
Emotion presented	**Anger**	16.12	8.88	51.88	21.75	1.38
	Fear	13.00	37.50	12.50	36.50	0.50
	Happiness	81.75	0.00	17.38	0.88	0.00
	Neutral	18.38	35.50	8.12	12.88	25.12
	Sadness	0.00	14.37	0.00	7.88	77.75

Figure 37: Heat-mapped confusion matrices, depicting the percentage of predictions by the random forest classifier in each emotion category, for each stimulus emotion, for the speech data. Columns denote presented emotions, rows denote 'responses'. Red = higher values, green = lower values.

Figure 38 shows the most important variables for this model, quantified in terms of the mean decrease in the Gini coefficient associated with each variable, averaged over one hundred random forest runs. This measure corresponds approximately with the notion of 'information loss', and therefore a greater decrease in Gini denotes more information lost as a result of removing a particular variable from the model.

6.7.3 Results: Music

For the music data, random forest classification with LOOCV produced the pattern of predicted emotions shown in Figure 39, recorded cumulatively over one hundred runs of the of the random forest algorithm.

Again, Figure 40 shows the ten most important variables for this classification task, in decreasing order of their associated mean decrease in Gini, averaged over one

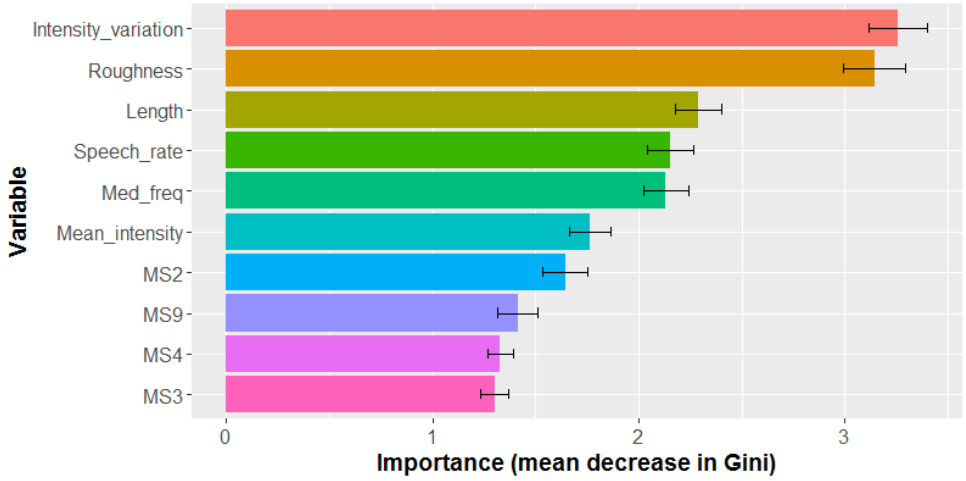


Figure 38: Mean decrease in Gini for each of the ten 'most important' input variables (for the speech data), averaged over one hundred random forest runs. Greater decrease in Gini corresponds to greater 'importance'.

		Response				
Emotion presented		**Anger**	**Fear**	**Happiness**	**Neutral**	**Sadness**
	Anger	18.25	21.50	16.00	32.12	12.12
	Fear	2.00	47.12	12.50	22.50	15.88
	Happiness	29.00	0.00	45.00	1.00	25.00
	Neutral	28.12	0.00	0.00	55.50	16.38
	Sadness	11.75	14.37	7.62	19.75	46.50

Figure 39: Heat-mapped confusion matrices, depicting the percentage of predictions by the random forest classifier in each emotion category, for each stimulus emotion, for the music data. Columns denote presented emotions, rows denote 'responses'. Red = higher values, green = lower values.

hundred random forest runs.

For both the speech and music datasets, an alternative approach to training the random forest classifier was considered, in which the model was trained using data for three of the sentences/ melodies (N = 30), and then tested on the remaining, unseen sentence/ melody (N = 10). The rationale was that the training/ test sets would essentially be identical to those experienced by the human listeners in Studies 3 and 4. However, this approach did not provide substantial additional insight – the

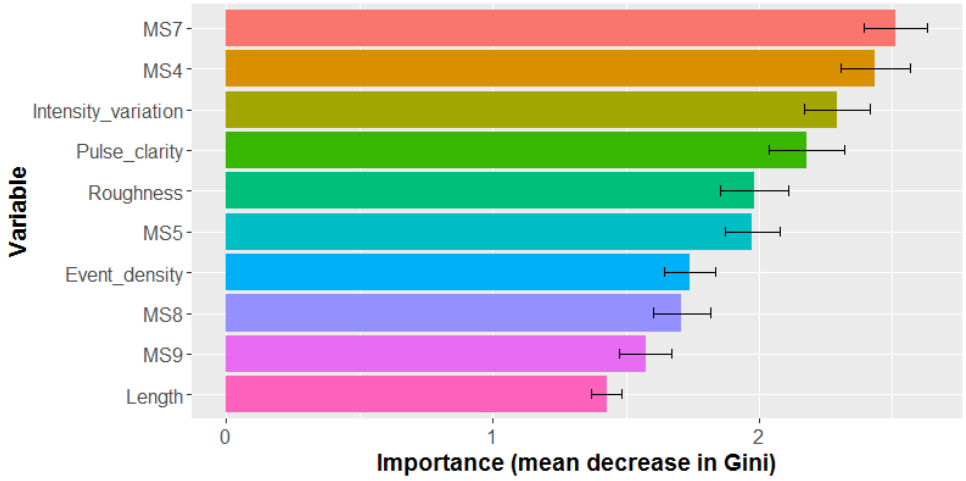


Figure 40: Mean decrease in Gini for each of the ten 'most important' input variables (for the music data), averaged over one hundred random forest runs. Greater decrease in Gini corresponds to greater 'importance'.

results obtained, in terms of emotion classification accuracy and relative importance of features, were almost identical to the findings already described using the leave-one-out method.

6.8 Discussion

6.8.1 Summary of results: Speech

The first aim of this chapter was to establish which auditory features would be most informative for emotion identification, and thereby which features would theoretically be prioritised by an ideal listening strategy. Optimising for maximum emotion identification accuracy, multinomial logistic regression classification with stepwise feature selection achieved a maximum of 55% accuracy for the CI-simulated speech stimuli. This was somewhat lower than the emotion identification accuracy achieved by participants in Study 3, who averaged approximately 74% accuracy during the post-test-transfer testing phase. This model chose four features – Median frequency,

Roughness, Length and Modulation spectrum component 2.

Criterion-based model selection led to somewhat different results. This approach identified Intensity variation as being the most important auditory feature, on the basis that this feature appeared in the greatest number of models. Mean intensity and Roughness were also commonly-chosen features. Therefore, intensity-based cues were valued more highly according to this approach.

Random forest classification produced sub-optimal identification results, with only Fear and Sadness correctly classified at a level greater than expected by chance. Interestingly, however, this classification procedure prevalently produced the Anger-Happiness confusion that was characteristic of human performance. Moreover, Sadness was by far the most accurately-decoded emotion by both the random forest classifier and by human listeners. The most important features identified by this approach corroborated the results of criterion-based model selection: Intensity variation and Roughness were clearly the most important two features. After these, Length, Speech rate and Median frequency were very similarly important.

As hypothesised, across the different modelling methods employed, no single 'type' of auditory feature was preferred. In other words, the best possible identification accuracy was not achieved via a listening strategy prioritising just one feature type. Instead, features relating to duration, frequency and intensity information were chosen.

Taking into account the three modelling approaches considered, Roughness and Intensity variation emerged as the most important auditory features, followed by Length and Median frequency. Thus, consistent with the conclusions of Study 3, optimum performance was not achieved simply by attending to or prioritising better-

preserved auditory features. Instead, residual frequency information appeared to play an important role also. To some extent this contradicts previous findings, which have reported a tendency for both real and simulated CI users to focus less on frequency-based cues and instead attend more to intensity and temporal information (Peng et al., 2009; Tao et al., 2015). This apparent discrepancy is consolidated by the conclusions of a similar study by Peng et al. (2012). The authors showed that, when the 'primary' cue type (in this case, frequency) is well-represented, then other cues (i.e. intensity and duration) are essentially ignored. However, these 'secondary' cues become more important when listening with CI simulation. Therefore, there is a 'trade-off' between attending to the most relevant or 'dominant' features, and the features left most intact by the simulation. This may result in frequency, duration and intensity cues being utilised approximately equally, as the results here suggest.

Variability in terms of the features chosen by the different modelling techniques implies that there may not have been a single best 'route' to successful emotion identification. This suggestion is consistent with the results of Study 3, which found that different auditory features were of varying importance for emotion identification, depending on the specific emotion being judged. Therefore, it would be possible in principle for two models to achieve comparable overall accuracy by prioritising different features and therefore recognising different emotions particularly well. This underscores the suggestion made previously that, although the optimal listening strategy might favour a specific handful of features, this strategy should be dynamic, rather than rigid, in order to perform well across different emotions and different stimuli.

A second aim of the modelling work was to estimate the strategies likely used by human listeners. Optimising the logistic regression classifier to emulate human per-

formance led to four features being chosen: Median frequency, Roughness, Length and Modulation spectrum component 2. Contrary to what was hypothesised, this suggests that human listeners were using an almost-optimal listening strategy. In fact, Intensity variation was the only notable feature not included in the model of human performance. Interestingly, human listening strategies did appear to change somewhat from pre-test to post-test-transfer, although there was no evidence that training caused listeners to deprioritise frequency-based cues. Instead, after training, it appeared that Intensity variation was deprioritised in favour of Length, while lower-frequency Modulation spectrum information was utilised.

However, there was no clear indication that the 'optimum' classifications arrived at by the various modelling approaches were more accurate than those of human listeners. Therefore, it is difficult to say what might have motivated participants' apparent deprioritisation of Intensity variation. Although computational modelling showed that Intensity variation was one of the most informative features with respect to emotion identification, human participants appeared to achieve better performance by focussing less on this feature. It should be noted that the human performance data from Study 3 were based on different stimuli sentences for the pre-test and post-test-transfer phases. Although no significant differences were found between these different sentences, in terms of emotion identification accuracy, it remains a possibility that differences in emotional expression across sentences somehow resulted in participants adopting slightly different listening strategies.

Lastly, as hypothesised, the best solution found using cluster analysis chose very similar features to those identified as being important for optimum classification accuracy. Roughness, Intensity variation, Median frequency and Length were all selected, in addition to three other features. In particular, Roughness and Length

appeared to be effective in separating stimuli into clusters. In all, though substantial structure was found, it was relatively 'weak' (Kaufman & Rousseeuw, 2009). Contrary to the hypothesis, the results did not explicitly indicate more substantial structure in the non-CI-simulated stimuli. Interestingly, however, an approximately similar level of structure was obtained using a one-dimensional clustering, with only Median frequency as a feature. Additionally, fewer cluster 'confusions' were made using the non-simulated stimuli. These findings strongly supports the aforementioned assertion by Peng et al. (2012), that frequency might be a dominant feature which would therefore be selectively attended to.

With both simulated and non-simulated stimuli, Sadness was the least commonly confused emotion, which is concordant with the results of Study 3 which found that Sadness was best recognised. Finally, consistent with the results of Study 3, the Anger-Happiness confusion was prevalent with simulated and non-simulated stimuli, though much more strongly in the former.

6.8.2 Summary of results: Music

Optimising for maximum emotion identification accuracy, multinomial logistic regression classification with stepwise feature selection achieved a maximum of 55% accuracy for the CI-simulated music stimuli. By contrast to the results for speech stimuli, this was substantially higher than the average 38% emotion identification accuracy achieved by participants in Study 4, during the post-test-transfer testing phase. This model chose ten features – Roughness, Length and Modulation spectrum components 2 to 9.

Criterion-based model selection led to quite different results. As with the speech data, this approach identified Intensity variation as being the most important audi-

tory feature, on the basis that this feature appeared in the greatest number of models. Other important features identified were Event Density, Modulation spectrum component 5 and, to a lesser extent, Length and Tempo. Therefore, frequency-based cues were valued less highly by this approach, in favour of intensity and duration cues.

Random forest classification produced much better identification results with music stimuli than with speech stimuli – of the five emotions, only Anger was classified with lower accuracy than expected by chance. Overall accuracy was greater than that attained by human participants, but lower than achieved by the logistic regression classifier. Interestingly, Neutral-Anger was the most common confusion made, which was not the case for the participants in Study 4. However, Neutral was the best-decoded emotion overall by the random forest classifier. By contrast, Sadness was the most accurately-decoded emotion by human listeners. The explanation for this discrepancy was not immediately clear.

The most important features identified by this approach overlapped to some extent with the results of the sequential- and criterion-based model selection procedures. Modulation spectrum (components 4, 5 and 7) were the most important features, followed by Intensity variation, Pulse clarity and Roughness.

As hypothesised, and as observed for the speech data, no preference for a single 'type' of auditory feature was preferred, across the different modelling methods. Therefore the best possible listening strategy was not one that prioritised just one feature type. However, features relating to duration and intensity information were chosen to a much greater extent than those relating to frequency information.

Considering all of the modelling approaches employed, Intensity variation, Length

and Modulation spectrum component 5 appeared to be the most important auditory features, followed by Event density, Roughness and Modulation spectrum components 4 and 7. Therefore, by contrast to the speech data, optimum performance appeared to be primarily achieved via selective attention to better-preserved (i.e. frequency-based) auditory features. The only exception to this was Roughness. This pattern of results is supported by previous findings, which have suggested that CI users shift their attention away from frequency-based cues and towards temporal cues in particular, when decoding emotional expression in music (Caldwell et al., 2015; Giannantonio et al., 2015). Interestingly, both of these studies found that tempo specifically was attended to by participants, while the results of the current analyses suggested that other temporal features were more important. There are two potential reasons for this small discrepancy. Firstly, both of the aforementioned studies presented pieces of music that differed in terms of both composition *and* performance, while the stimuli here differed only in terms of the latter. Therefore, the relative difference in tempo may have been larger in those studies. Secondly, both Caldwell et al. (2015) and Giannantonio et al. (2015) considered only mode and tempo as features. Thus, it is possible that other temporal cues not included might also have been relevant for emotion identification.

By comparison to the results for speech stimuli, frequency cues appeared to be utilised to a lesser extent for music. This was most likely because the expression of emotion in music performance is not underscored by the same 'primacy' of frequency information that is considered to be the case for speech (Petrushin, 1999). Instead, there is no strongly dominant cue for music, and so intensity and duration cues are comparatively more important (Peng et al., 2012). With CI simulation, in the absence of any dominant auditory feature, attention is directed those disrupted least by the

simulation (Peng et al., 2012), which happens to be duration- and intensity-based cues.

As with the speech data, the different modelling techniques produced reasonably variable subsets of 'most important' features, suggesting that there was not one 'optimum' listening strategy for emotion identification. Again, this is consistent with the idea advanced in Study 4, that the best possible emotion identification accuracy would likely be dynamic, attending selectively to different auditory features depending on the specific stimulus and/ or emotion being appraised.

Human listening strategies were also investigated via computational modelling. Optimising the logistic regression classifier to emulate human performance, the most important features appeared to be: Roughness, Pulse clarity, Event density and Modulation spectrum components 2 and 9. To a lesser extent, Intensity variation, Spectral centroid and Length were also important features. Contrary to what was hypothesised, this suggests that human listeners used a listening strategy very similar to what was determined 'optimal'. However, there were two points of divergence. Firstly, human listeners potentially attended to Spectral centroid, which was not included in any of the optimal models. Secondly, participants appeared to attend to high- and low-frequency Modulation spectrum components, whereas the optimal models tended to favour mid-frequency modulation spectrum information.

Aside from Tempo being relevant for predicting pre-test but not post-test-transfer performance, there was little evidence that human listening strategies changed substantially as a function of training. Therefore, the suggestion from Study 4, that listeners would deprioritise frequency-based cues, following training, was not supported. However, this finding is consistent with the actual results of Study 4, which reported no significant overall improvement in emotion identification following train-

ing. Given this lack of improvement, it is perhaps unsurprising that the training did not substantially alter participants' listening strategies. Similarly, it is unsurprising that, unlike the results for speech stimuli, the 'optimum' classifications produced were (to varying extents) more accurate than those of human listeners.

Somewhat surprisingly, the best clustering of the auditory features for music stimuli was based on a subset of only two features: Pulse clarity and Modulation spectrum component 8. These features were chosen by some of the modelling techniques, but did not stand out as being especially important. As with the speech stimuli, only relatively weak structure was found in the data. Contrary to the hypothesis, the results did not indicate the presence of more substantial structure in the non-CI-simulated stimuli. In fact, despite selecting more auditory features, the clusters constructed for non-simulated stimuli were less well-separated and explained a lower proportion of the overall variance in the data. It is unlikely that the CI simulation actually improved the 'structuredness' of the auditory feature data, but, in principle at least, this could have occurred due to frequency-based cues encoding uninformative aspects of the stimuli, and being suppressed by the simulation.

Lastly, corroborating the results of Study 4, Sadness was the least commonly confused emotion CI-simulated music stimuli. However, Anger was the least confused emotion for non-simulated music stimuli. However, consistent with the results of Study 4, the Anger-Happiness confusion was the most prevalent confusion with both simulated and non-simulated stimuli.

6.8.3 Conclusions

This chapter provided many valuable insights into emotion identification with CI-simulated speech and music. Specifically, the computational modelling approach

built upon Studies 3 and 4 by answering questions about the potential listening strategies underlying performance in these experiments.

Firstly, a variety of auditory features were selected for emotion identification with both speech and music stimuli. Therefore, the best strategy in each case consisted of more than just one feature (e.g. spectral centroid) or feature type (e.g. frequency). Additionally, since feature subsets varied across the different modelling approaches, it may be concluded that there was no single 'definitive' feature subset reliably associated with more accurate emotion identification. This is concordant with the conclusions of Studies 3 and 4, which suggested that the combination and integration of different auditory features is essential for the perception of emotion in CI-simulated stimuli. As concluded in Chapter 5, there may be no universally effective listening strategy, but rather a set of overlapping strategies, which may be more or less useful depending on the specific stimulus.

Generally speaking, the best strategy for identifying emotion in speech took into account information about frequency, intensity and duration (i.e. a 'holistic' listening strategy). By contrast, music tended to prioritise temporal information, followed by intensity information, and barely ever frequency. This may be due to differences in emotional expression, since frequency cues are less important for these particular music performance stimuli (since speech stimuli varied in terms of pitch contour, whereas the musical stimuli did not). In both cases, the results corroborated the observation made in the previous chapter, that there was no special reason to focus on frequency as an auditory cue.

For both speech and music stimuli, the models were able to reproduce human performance very well. However, the models had less success in achieving 'optimum' performance. Approximately the same accuracy was achieved for both speech and music

(∼55%). For music stimuli this was substantially better than human performance, but with speech stimuli human listeners were more accurate than the 'optimal' model. This is partially a reflection of human listeners' much poorer performance with music stimuli, but may also reflect something specific about the way that participants made judgements about the speech stimuli, which was not well-modelled by the current approach.

Generally, participants' listening strategies were very similar to the 'optimal' ones, although there were minor differences which may have been important. However, there was little evidence of human listeners systematically attending more to frequency-based cues, or of substantially amending their listening strategies in response to the training.

The patterns of confusions made by humans in Studies 3 and 4 (e.g. Anger and Happiness prevalently confused, Sadness very seldom confused) appeared to have a very tangible basis in the auditory feature data, since these confusions were replicated even by models optimised for best performance, and by unsupervised clustering. Therefore, these patterns of responding appear to be a product of the way emotion is expressed in speech and in music (even without CI simulation).

Overall, the pattern of results obtained suggested that the idea of 'attending to better-preserved features' might be overly simplistic. Instead of focussing on only these features, it appeared that both human and optimal listening strategies incorporated residual frequency information to the extent that it was informative, and otherwise focussed on duration and intensity cues. Therefore, even with CI simulation, there were instances where the frequency information preserved was useful for emotion classification (especially for the speech stimuli, in this case). Taken together, the results suggest that, even with CI simulation, variation in duration-,

frequency- and intensity-based cues is capable of encoding information relevant for emotion identification. Therefore, echoing the conclusions made in Chapter 5, when decoding emotion in CI-simulated stimuli, one should consider each of these features, prioritising them in accordance with their differential contributions to the expression of different emotional states, via different media.

6.8.4 Limitations

Foremost, it should be kept in mind that these analyses used relatively small samples of speech and music stimuli, and assumptions about human performance were based on only a handful of participants. Therefore, the findings from this chapter should be interpreted conservatively, and may not generalise well beyond the speech and music datasets studied.

Additionally, the small number of stimuli investigated here prevented the use of a more powerful machine learning algorithm that might have been able to provide superior insight with respect to 'optimal' accuracy in emotion identification. For example Pentoś (2016) recently documented a method for assessing the importance of different input features to a neural network, which works by comparing several different models to account for the random initialisation of connection weights (similar to the approach used here with random forest classification). Unfortunately, this approach typically necessitates a much larger sample than was available here – recent attempts at emotion classification in speech and music have used datasets with at least three thousand stimuli (Chernykh, Sterling, and Prihodko, 2017; Liu, Chen, Wu, Liu, and Liu, 2017). Nonetheless, this approach does not appear to have been used before with CI-simulated stimuli, and may be an interesting avenue for future research.

The small number of participants upon which the human performance data is based also limits the potential applicability of these findings to a wider population. However, this may be less of a problem, since listening strategies were not expected to vary greatly between participants. This was because, unlike a sample of real CI users, the NH sample recruited all had approximately similar hearing, and all received exactly the same auditory information, courtesy of using the same simulation.

Relatedly, the analyses of human performance assumed a representative 'average' listening strategy although, in reality, this may not exist for CI users. According to Peng et al. (2012), listening strategies represent a 'trade-off' between attending to auditory features that carry the most relevant information, and features that are best-preserved by the CI (or simulated CI). Of course, all features are preserved to (roughly) the same extent for NH, CI-simulated participants, and therefore one would expect these listeners to gravitate towards very similar listening strategies. However, the degree to which different features are preserved may vary extensively among real CI users, due to myriad factors, including residual acoustic hearing, use of a contralateral hearing aid, etc. In addition, inter-individual differences in experience with the CI, age at implant, duration of deafness etc. may influence the listening strategies that are adopted by different individuals. Therefore, the conclusions made pertaining to human performance with CI-simulated stimuli should not necessarily be assumed to hold true for real CI users.

6.9 Revisiting the emotion perception training paradigm

Although this chapter successfully built upon the results of Studies 3 and 4, there emerged a clear need to evaluate the experimental training paradigm with a sample of real CI users. Accordingly, the next chapter returns to the emotion identification

training paradigm presented in Chapter 5, this time recruiting real CI users. The studies presented in the following chapter aimed to build upon both the modelling results and the results from Chapter 5, by investigating how well the insights gained from each would translate from CI-simulated listeners to CI users.

7 Studies 5 and 6: Emotion perception training studies with CI users

7.1 Introduction

This chapter presents two experiments that were very similar to the the training paradigm explored in Studies 3 and 4. Studies 5 and 6 once again investigated perceptual learning with emotional speech and music, respectively. However, unlike the previous studies, these experiments each recruited a sample of real CI users, as opposed to CI-simulated participants. Methodologically, the studies very closely resembled Studies 3 and 4, with the key exception that the stimuli presented to CI users were not processed with the NBV-based CI simulation. Studies 5 and 6 were intended to assess the efficacy of the emotion recognition training paradigm with CI users, as compared to NH participants listening with CI simulation.

As in Studies 3 and 4, the emotion identification paradigm included separate 'training' phases, during which participants received feedback about their emotion judgements, and therefore the effectiveness of their listening strategies. Unlike these studies however, Studies 5 and 6 did not include a feature attenuation manipulation. Instead, all participants heard an identical set of stimuli (albeit with counterbalanced order of presentation), which was the same as presented in the Original conditions

in Studies 3 and 4. There were two primary reasons for this. Firstly, fewer participants were tested in these studies, owing of the greater difficulty of recruiting CI users compared to NH listeners. Therefore, it would not have been possible to make a statistically valid comparison of emotion identification accuracy across the different feature attenuation conditions. Secondly, via consultation with an audiological scientist at the clinic from which participants were recruited, it was determined that feature attenuation may have made the judgement task inappropriately difficult for participants. Unfortunately, the lack of feature attenuation conditions limited the conclusions that could be made with respect to CI users' potential listening strategies, or the relative importance of different auditory features for their emotion judgements. Accordingly, the focus of this chapter was on the potential utility of the training procedure for 'rehabilitation' of emotion identification. CI users' potential underlying listening strategies were probed only indirectly, using the logistic regression modelling approach described in the previous chapter.

Aside from recruiting CI users, presenting non-simulated stimuli and excluding the feature attenuation manipulation, there was only one other difference between the experiments presented in this chapter and the ones presented in Chapter 5. The self-reported measure of empathy (EQ) was not included in Studies 5 and 6 because of a lack of significant relationships found between emotion identification and empathy across all of the previous studies. Because this measure was intended to address a relatively peripheral research question, and did not generate any interesting results, it was omitted for these studies, to help ensure that the experiments were kept to a reasonable length. By contrast, the Music Use questionnaire (MUSE) was retained in Studies 5 and 6. This measure was significantly correlated with emotion identification accuracy in Study 3, and also enabled an interesting comparison of music listening

and engagement between CI users and NH participants.

The central question that Studies 5 and 6 attempted to answer was essentially the same as in Studies 3 and 4, namely: 'can participants be trained to improve their emotion identification accuracy for speech and music?'. Considering both of these pairs of studies together, another research question was: 'how similar are the effects of this training for CI users vs. CI-simulated listeners?'. As already described in Chapter 5, an advantage of the training paradigm was that it provided time for NH participants to 'acclimatise' to the sound of CI simulated stimuli, thereby facilitating comparison with CI users, who may have already improved their perception of speech and music via rehabilitation (Gfeller et al., 2001; Wei et al., 2000).

7.1.1 Summary

In summary, the overarching aim of Studies 5 and 6 was to discover the extent to which the findings of Studies 3 and 4 would be comparable to experiments with real CI users, rather than CI-simulated NH participants. These studies therefore aimed to establish to what degree the conclusions made previously about listening strategies used for emotion identification may be generalisable to CI users.

In the next section, the experimental paradigm, as adapted for testing with CI users, is described in more detail.

7.2 Methods

7.2.1 Participants

Eight CI users were recruited in collaboration with Sheffield Teaching Hospitals NHS Foundation Trust. Participants were notified about the experiment prior to sched-

uled review sessions, and were invited to participate immediately after their next clinical session. For convenience, participants recruited in this manner were tested in a sound-attenuated room within the Hearing Services department. Ethical approval was obtained via the Integrated Research Application System (IRAS), in collaboration with Sheffield Teaching Hospitals NHS Foundation Trust.

Additionally, two participants were recruited via a local, non-NHS-affiliated CI user support group. Volunteers recruited from this group were tested in a quiet laboratory at the University of Sheffield. To facilitate this recruitment additional approval was obtained from the Department of Music ethical review committee.

In total, ten participants were tested: four in Study 5 (3 female; mean age = 61.50 years, SD = 14.90) and six in Study 6 (2 female; mean age = 62.50 years, SD = 16.30). All ten participants were post-lingually deafened, and had used their CI for over a year at the time of testing.

All participants took part voluntarily, although reimbursement was provided for any travel costs incurred. Prior to testing, all participants were given an information sheet, and provided fully informed consent prior to participation. Demographic information for all CI participants and details about their devices are summarised in Table 23. Two participants wore contralateral hearing aids in addition to their CI, though in both cases the hearing aid was turned off prior to the experiment.

7.2.2 Materials

The same set of stimuli from Studies 3 and 4 was used, consisting of short excerpts of speech (Burkhardt et al., 2005) and brief musical melodies (Quinto et al., 2014), each expressing one of five emotions (Anger, Fear, Happiness, Sadness or Neutral).

Table 23: Demographics and device information for each of the CI participants recruited.

ID	Study	Age	Gender	Years use	Device type	Processing strategy
1	6	33	Female	13.9	CI24M	SPEAK
2	6	62	Female	17.3	CI24M	SPEAK
3	6	66	Male	1.4	CI522	ACE
4	6	77	Female	2.5	CI422	ACE
5	6	67	Male	1.3	AB HiFocus MS	HiRes Optima-S
6	5	41	Female	1.7	CI522	SPEAK
7	5	79	Male	1.9	CI522	SPEAK
8	5	60	Female	5.0	AB	SPEAK
9	6	64	Male	4.7	CI422	SPEAK
10	5	70	Male	3.6	CI422	SPEAK

However, feature attenuation was not included as an experimental manipulation, hence only the 'Original' stimuli from the previous experiments were presented.

In total, this meant that there were forty speech stimuli presented in Study 5 (four different sentences × two speaker genders (female, male) × five emotions), and forty music stimuli in Study 6 (four different melodies × two musical instruments (violin, voice) × five emotions).

As in all of the previous studies, stimuli were encoded in single-channel (mono), 16-bit Audio Interchange File Format (AIFF) with 16,000 Hz sampling frequency.

7.2.3 Procedure

Prior to the emotion identification task, the Music Use questionnaire (MUSE) (Chin & Rickard, 2012) was administered, in order to assess potential effects of musical engagement and/ or expertise upon emotion perception.

The emotion identification task was run on a Macbook laptop computer, and all stimuli were presented at approximately 65 dBA SPL via two Wharfedale Diamond 7.1 loudspeakers, connected to an NAD 310 stereo integrated amplifier. All partici-

pants were seated approximately 0.5 metres away from the loudspeakers. Unlike in Studies 1 through to 4, headphones were not used for stimulus presentation for CI users, in the interest of participants' comfort.

The experimental procedure for Studies 5 and 6 was essentially the same as in Studies 3 and 4, with participants listening to the speech or music stimuli and making 5-AFC perceptual judgements about the emotions expressed. Corresponding to each judgement, participants also provided an indication of their confidence on a five-point Likert scale. As in Studies 3 and 4, the experiments were divided into different testing phases: pre-test, training, post-test, additional training, and post-test-transfer. The first post-test phase examined participants' emotion identification proficiency with previously-encountered stimuli, while the post-test-transfer phase examined proficiency with previously unheard stimuli. As in Studies 3 and 4, participants heard different sentences or melodies during this phase (for example, if they trained with melodies B and C, they were tested on melody D). Likewise, as in Studies 3 and 4, participants only received feedback on their performance during the training experiment. During these blocks, 'ground truth' values about the emotions expressed were provided after each judgement, and stimuli were then played a second time. To prevent listener fatigue, all participants took a break for approximately ten minutes, after the first post-test phase had been completed.

For ease of reference, the table from Chapter 5, indicating the specific stimuli presented during each stage of the experiment is reproduced here (Table 24). As in the original emotion identification training studies, the different sentences and melodies were counterbalanced to eliminate order effects.

During all experimental phases, data were gathered about the accuracy of participants' judgements, along with corresponding confidence ratings and reaction times.

Table 24: Information about stimuli included at each stage of the experiment. Reproduced from Chapter 5.

Stage	Stimuli presented	Description
Pre-test	Sentence/ melody A	No feedback
Training 1	Sentence/ melody B	Audio and visual feedback
Training 2	Sentence/ melody C	Audio and visual feedback
Post-test	Sentence/ melody A	Identical to Pre-test
Training 3	Sentence/ melody B	Identical to Training 1
Training 4	Sentence/ melody C	Identical to Training 2
Post-test (transfer)	Sentence/ melody D	No feedback

7.2.4 Hypotheses

Firstly, in line with previous research, and with the results of Studies 1, 2, 3 and 4, it was hypothesised that emotion identification performance would be above chance level with both speech and music stimuli. However, owing to individual differences in level of experience with the implant, device type, and various other relevant factors, it was expected that inter-individual variance in identification accuracy might be greater in Studies 5 and 6.

Based upon the results of Study 3, it was predicted that emotion identification accuracy with speech stimuli would increase substantially as a function of experimental phase, representing improved task performance with more practise. However, emotion identification accuracy with music stimuli was expected to show much more modest improvement, as was found in Study 4. Nonetheless, it was predicted that performance with both stimulus types would be better in the post-test phases, compared to pre-test. Based on the outcomes of Studies 3 and 4, it was assumed that decoding of previously-unencountered stimuli (as presented in the post-test-transfer phase) might improve to a greater degree with speech stimuli than with music.

Assuming that the simulation provided an accurate approximation of signal degradation caused by the CI, it was hypothesised that overall performance in these studies

would be similar to with CI-simulated listeners. However, regarding the potential effects of training upon emotion identification, there were two hypotheses. Firstly, it was assumed that CI users would already be much more accustomed to the sound of the CI, and therefore might have less potential for improvement as a function of training. By contrast, it was assumed that CI users might have more potential for improvement with music stimuli, since they might have been less accustomed to making emotional judgements about music, compared to NH participants.

In the absence of evidence suggesting otherwise, it was predicted that participants' potential listening strategies (as estimated by logistic regression classification) would be approximately the same as those potentially used by NH participants in Studies 3 and 4. Since the previous chapter suggested that CI-simulated participants were already using close-to-optimal listening strategies, it was expected that real CI users would also be.

Lastly, on the basis of previously reported survey data (Roy et al., 2012a), it was expected that, relative to NH participants, CI users would register lower scores on most or all of the sub-scores derived from the MUSE questionnaire. Because most CI users were expected to have much less experience with (and interest in) music, it was considered unlikely that MUSE scores would have any systematic predictive affect with respect to emotion decoding accuracy.

7.3 Results

7.3.1 Study 5: Speech

On average, across each of the conditions, participants made their emotion judgements 6.15 seconds after stimulus onset (corresponding to mean stimulus length +

2.98 seconds) – approximately 1 second slower than the CI-simulated participants in Study 3. Reaction times were not compared for the five different emotions, since they were confounded by differences in stimulus length.

The average confidence rating associated with participants' judgements, across all of the experimental trials, was 2.88 – less confident than simulated CI listeners, who averaged 3.68 in Study 3. One-way rank-based ANOVA (an analysis similar in principle to the Kruskal-Wallis test) revealed that confidence ratings were not significantly influenced by the emotion presented, $F(4, 715) = 1.72$, $p = .144$, $\eta^2 = .01$.

As hypothesised, all participants were able to decode emotion at above-chance level during the final post-test. However, inter-individual variability was high with final-post-test scores varying between 25% and 65% correct judgements.

The effect of training was also quite variable between participants: participant 1 showed a modest improvement (+ 15%), participant 2 a more marked improvement (+ 25%), and participant 3 showed a negative effect of training (- 10%) (Figure 41).

As in Studies 3 and 4, confusion matrices were created, to evaluate in more detail the overall pattern of errors and correct responses made by participants during the pre-test and post-test-transfer phases of the experiment (Figure 42).

The confusion matrices demonstrated that, on average, the positive effect of training upon decoding accuracy was driven by improvements made in the identification of Fear and Sadness. By contrast, Anger and Neutral were decoded well above chance level, but did not show any improvement with training. Happiness showed very slight improvement, but was still well below chance level. In the post-test-transfer phase, Sadness was by far the best-recognised. Conversely, Fear seemed to be particularly

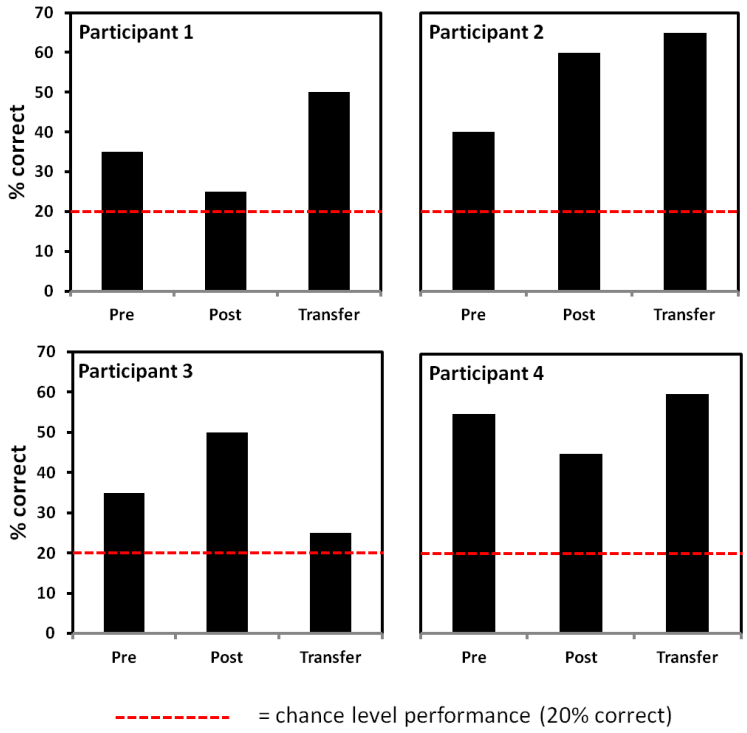


Figure 41: Barplots showing participants' % correct emotion judgements at pre-test, post-test and post-test-transfer, for the speech version of the experiment.

problematic, and Anger and Happiness were often confused.

Pearson correlation showed that, at pre-test, these confusion matrices were moderately positively correlated with those produced by CI-simulated listeners in Study 3 (in the Original condition), $r(23) = .52$, $p = .008$. At post-test-transfer however, the two matrices were very strongly positively correlated, $r(23) = .81$, $p = <.001$. However, in the current study, participants appeared to have greater difficult in correctly identifying Happiness. On the other hand, the Anger-Happiness confusion was much less prevalent in this study than in Study 3. Generally, CI users showed less overall improvement in emotion identification as a function of testing phase (mean 8.75 percentage points increase from pre-test to post-test transfer), compared to CI-simulated listeners (mean 33.33 percentage points increase).

Pre-test

Emotion presented \ Response	Anger	Fear	Happiness	Neutral	Sadness
Anger	56.25	6.25	12.50	18.75	6.25
Fear	18.75	37.50	6.25	31.25	6.25
Happiness	31.25	18.75	6.25	31.25	12.50
Neutral	6.25	0.00	25.00	56.25	12.50
Sadness	6.25	0.00	18.75	25.00	50.00

Post-test-transfer

Emotion presented \ Response	Anger	Fear	Happiness	Neutral	Sadness
Anger	56.25	25.00	12.50	0.00	6.25
Fear	0.00	50.00	6.25	31.25	12.50
Happiness	25.00	31.25	12.50	31.25	0.00
Neutral	0.00	0.00	18.75	56.25	25.00
Sadness	6.25	12.50	0.00	6.25	75.00

Figure 42: Heat-mapped confusion matrices for the speech stimuli, depicting the percentage of responses in each emotion category, for each stimulus emotion. Columns denote presented emotions, rows denote emotion judgement responses. Red = higher values, green lower values.

Following the same procedure outlined in the previous chapter, a multinomial logistic regression classifier was trained to recognise emotions. Auditory features (from the input feature vector outlined in Chapter 6) were chosen using the Sequential Forward Selection (SFS) algorithm. The classifier was optimised according to the human confusion matrix, meaning that features were more likely to be selected if they led to a model that made similar patterns of errors in classification to those made by humans. This was used as a by-proxy method of investigating listeners' potential listening strategies.

The best classification model for reproducing the errors and correct judgements made by CI users' incorporated three auditory features: Spectral centroid, Roughness, and Modulation spectrum component 9. Emotion classification with this model resulted in a confusion matrix that was strongly, positively correlated with that of human

listeners, $r(23) = .73$, $p = .001$.

This result should be interpreted very cautiously, since listening strategies were likely more heterogeneous for CI users compared to simulated listeners, given participants' differing levels of experience with their devices, among other factors. Therefore the notion of an 'average listening strategy' is inherently less meaningful for this population. Additionally, the model used auditory input features that were processed with the CI simulation, which provides only a general approximation of the sounds as experienced by CI users.

Lastly, there appeared to be very little relationship between participants' responses to the MUSE and their overall accuracy in decoding emotions. However, the two participants with the best overall scores were also the only participants to report having had formal musical training.

7.3.2 Study 6: Music

On average, across each of the conditions, participants made their emotion judgements 8.70 seconds after stimulus onset (corresponding to mean stimulus length + 3.60 seconds) – approximately 2 seconds slower than the CI-simulated participants in Study 3. Reaction times were not compared for the five different emotions, since they were confounded by differences in stimulus length.

The average confidence rating associated with participants' judgements, across all of the experimental trials, was 2.76 – slightly less confident than simulated CI listeners, who averaged 3.04 in Study 4. One-way rank-based ANOVA revealed that confidence ratings were not significantly influenced by the emotion presented, $F(4, 1025) = 0.50$, $p = .739$, $\eta^2 < .01$.

Of the six participants tested, five completed all of the trials, whilst the last participant chose to withdraw from the study due to migraine. Of the five participants that completed the final post-test phase of the experiment, three achieved above-chance level task performance (Figure 43).

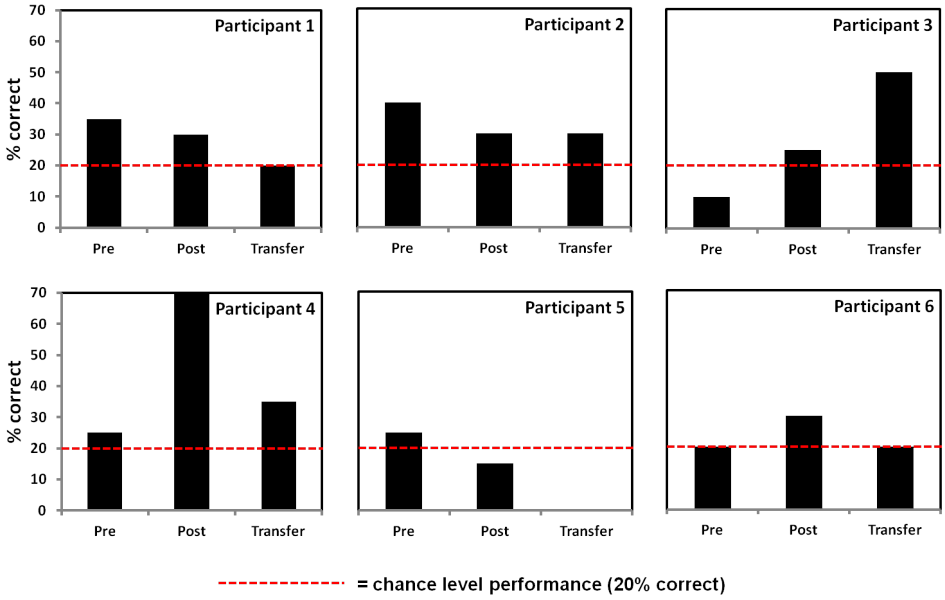


Figure 43: Barplots showing participants' % correct emotion judgements at pre-test, post-test and post-test-transfer, for the music version of the experiment. Note: Participant 5 withdrew from the experiment prior to the post-test-transfer phase.

Performance was highly variable: participants 1, 2 and 6 appeared not to benefit at all from the training sessions, while participants 3 and 4 showed marked effects of practice. In particular, the performance profile of participant 4 was very similar to the data observed for NH listeners – that is, there was a definite practice effect at post-test, but this did not generalise well to novel stimuli. By contrast, participant 3 improved linearly with practice throughout the experiment, and performed very well even in the post-test-transfer phase.

Again, confusion matrices were created in order to examine the overall pattern of errors and correct judgements made by participants during the pre-test and post-

test-transfer phases of the experiment.

Pre-test

Emotion presented \ Response	Anger	Fear	Happiness	Neutral	Sadness
Anger	20.83	4.17	16.67	33.33	25.00
Fear	20.83	20.83	0.00	25.00	33.33
Happiness	12.50	20.83	25.00	20.83	20.83
Neutral	12.50	0.00	12.50	25.00	50.00
Sadness	0.00	20.83	16.67	25.00	37.50

Post-test-transfer

Emotion presented \ Response	Anger	Fear	Happiness	Neutral	Sadness
Anger	35.00	15.00	15.00	15.00	20.00
Fear	25.00	10.00	10.00	25.00	30.00
Happiness	20.00	5.00	30.00	20.00	25.00
Neutral	10.00	20.00	25.00	30.00	15.00
Sadness	5.00	15.00	10.00	20.00	50.00

Figure 44: Heat-mapped confusion matrices for the music stimuli, depicting the percentage of responses in each emotion category, for each stimulus emotion. Y-axes = presented emotions, X-axes = emotion judgement responses.

On the whole, the confusion matrices demonstrated a modest influence of training upon decoding accuracy, although this was somewhat inconsistent across the different emotions being judged. Anger and Sadness showed the most substantial improvement, being recognised much more accurately in the post-test-transfer phase, while identification of Happiness and Neutral improved only to a relatively small degree, and accuracy actually decreased for Fear. In fact, Fear was chosen only 7.5% of the time in the post-test-transfer phase, compared to 13% of the time during pre-test, suggesting that participants largely learned to avoid this response option.

As with the speech stimuli, Sadness was the most accurately recognised emotion in the post-test-transfer phase. At pre-test, Sadness was prevalently confused with Neutral, and in fact false positive errors for Sadness were quite prevalent generally, although the training appeared to alleviate this. Interestingly, compared with the

speech stimuli, all emotions were decoded less accurately with music stimuli, except for Happiness, which was decoded substantially better.

Across all of the different emotions, identification accuracy in this study was very similar to the pattern observed for CI-simulated listeners in Study 4. Confusion matrices for the current study were only weakly positively correlated with those produced in Study 4 at pre-test, $r(23) = .32$, $p = .119$, but were strongly positively correlated at post-test, $r(23) = .66$, $p <= .001$. In both studies, Fear was decoded the least accurately, and Sadness most accurately. However, the Anger-Happiness confusion was not made as often by real CI users.

Multinomial logistic regression with SFS was used to estimate the listening strategy most likely to have been used by participants. The subset of auditory features (from the input vector described in Chapter 6) which led to emotion classification responses most similar to the confusion matrix described above included nine features: Intensity variation, Spectral centroid, Length, Event density, and Modulation spectrum components 2, 4, 5, 8 and 10. This model resulted in a confusion matrix that was strongly, positively correlated with that of human listeners, $r(23) = .84$, $p = .001$. As in Study 5, this should be interpreted very cautiously, owing to potential inter-individual variability in listening strategies, and the fact that auditory input features to the model were derived from analysis of CI-simulated stimuli.

Lastly, most participants reported very low levels of engagement with music, and there was no evidence of a relationship between participants' MUSE responses and their overall proficiency in emotion identification.

7.3.3 Music Use questionnaire data

The MUSE data gathered from CI users in this study was compared to MUSE data gathered from NH participants in Studies 3, 4 and 7. Since this combined sample of NH participants (N = 54) was much larger than the sample of CI participants (N = 10), and the groups were unbalanced in terms of mean age (with CI users being older, on average), propensity score matching was performed in order to create a suitable sub-sample of NH participants for comparison. This process was achieved via the *MatchIt* package for R (Ho, Imai, King, & Stuart, 2011), using the nearest neighbour matching method, with distance estimated by logistic regression (i.e. the extent to which hearing status (CI vs. NH was predictable by participants' age was minimised). From the fifty-four NH participants, ten were selected that most closely matched the CI participants in terms of age. Prior to matching, the CI group had a mean age of 61.90 years (SD = 14.56), whereas the NH group had a mean age of 24.92 years (SD = 8.02). After matching the latter sample to the former, the resultant ten participants from the NH group had a mean age of 37.60 years (SD = 10.72). Figure 45 illustrates the effect of the matching process upon the data. Essentially, the extent the certainty with which group membership (CI or NH) could be predicted by age was reduced by choosing a matched subset of the data, indicating that the confounding influence of age was reduced (though not eliminated entirely).

Following the age-matching process, the NH and CI groups were compared in terms of their responses to the MUSE questionnaire. As hypothesised, CI participants tended to utilise music to a lesser extent than NH participants, across the majority of the sub-components tested by the MUSE (Figure 46). Since the data were not normally-distributed (owing to the small sample sizes in question), permutational multivariate analysis of variance (PMANOVA) was used to make a statistical

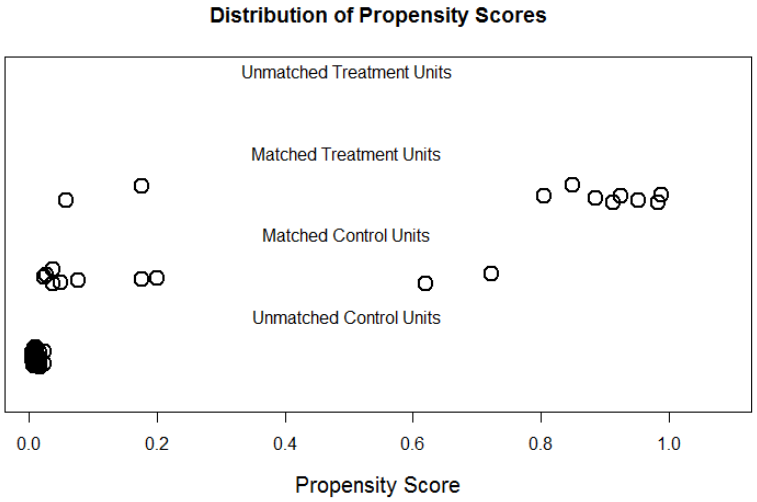


Figure 45: Jitter plot illustrating propensity scores for NH and CI participants both before and after nearest neighbour matching. Scores correspond to the estimated probability of a data point belonging to either the NH group (closer to 0) or the CI group (closer to 1)

comparison about the distribution of scores across all of the MUSE subcomponents, between the two groups. This method was implemented using the *vegan* package for R (Oksanen et al., 2017), and is analogous to traditional multivariate analysis of variance (MANOVA), with the exception that the p-value output is derived from comparison of the observed F-ratio to the F-ratios obtained from random permutations of the data points between each of the experimental groups. Consequently, PMANOVA makes no assumption about the normality of the data. Considering the overall effect of hearing status upon responses to the MUSE, PMANOVA (1,000 permutations) confirmed that the NH and CI groups differed significantly, ($F(1, 18) = 2.67$, $p = .012$, $\eta^2 = .13$).

To investigate differences between NH and CI participants for the individual subcomponents of the MUSE, follow-up Wilcoxon-Mann-Whitney tests were conducted (Bonferroni corrected for multiple comparisons), as summarised in Table 25. As ex-

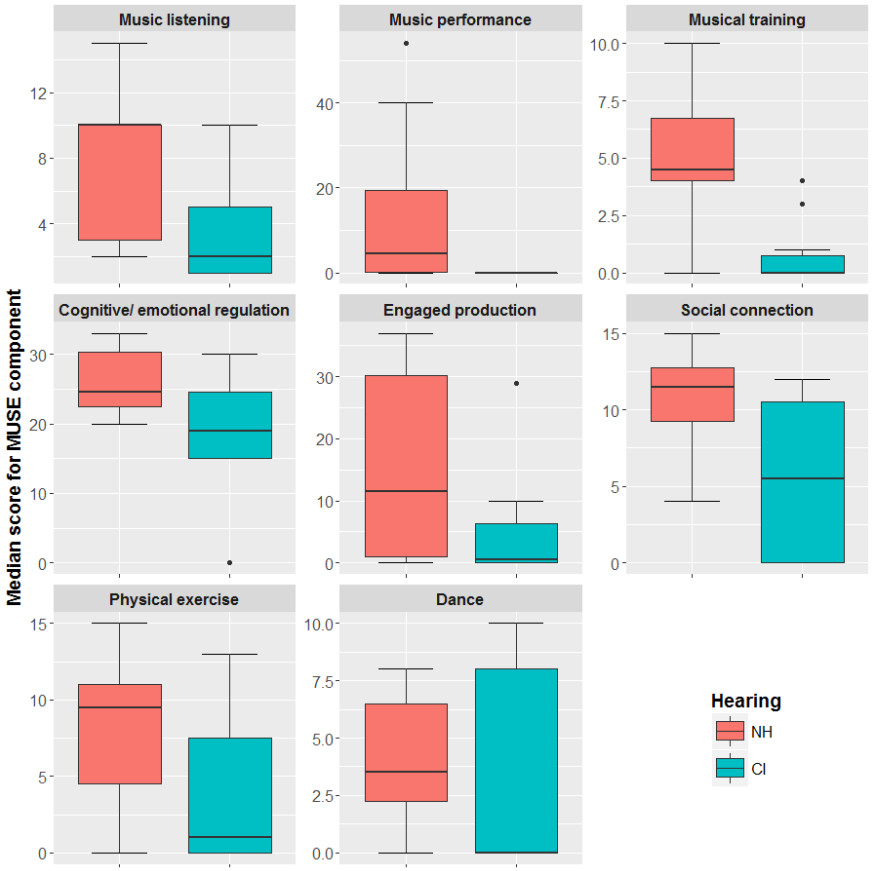


Figure 46: Boxplots illustrating NH and CI participants' median scores for each of the sub-components of the MUSE questionnaire. Centre lines = medians, Boxes = 25th and 75th percentiles, Whiskers = min and max values within 1.5 IQR of the 25th and 75th percentiles, respectively

pected, NH and CI participants responded significantly differently for the majority of MUSE subcomponents, reporting listening to music less often, performing music less often, having less musical training and music to a lesser extent for cognitive/ emotional regulation and social connection. Interestingly, however, the two groups did not differ significantly in their reported use of music for physical exercise or dance. The between-groups difference in engaged production of music was also non-significant, though this appeared to be mostly due to one outlier in the CI group.

Table 25: Bonferroni-corrected Wilcoxon-Mann-Whitney tests, examining the differences between NH and CI participants on the various sub-components of the MUSE questionnaire.

MUSE Component	***U***	***p***	***r***
Music listening	81.50	.015*	.54
Music performance	85.00	.002*	.69
Musical training	90.00	.002*	.70
Cognitive/ emotional regulation	78.00	.034*	.47
Engaged production	68.50	.148	.32
Social connection	81.00	.018*	.53
Physical exercise	69.00	.142	.33
Dance	65.00	.245	.26

* = significant at .050 alpha-level.

7.4 Discussion

7.4.1 Summary of results: Study 5

As hypothesised, overall emotion identification accuracy was above chance for all participants, across all testing phases. With the exception of Happiness, performance was well above chance for all emotions. This was in accordance with various studies that have evaluated the ability of CI listeners to decode emotion in speech (Chatterjee et al., 2015; Volkova et al., 2013). Consistently with Studies 1 and 3, these results show that above-chance performance is possible even using a 5-AFC judgement task, as opposed to a binary one. Therefore, the results corroborate the previous findings which were based on NH participants listening with a CI simulation. Additionally, there was no indication of increased variability in the emotion identification abilities of CI users, relative to CI-simulated listeners. Although it was expected that this population might be more variable, the lack of a difference here most likely reflects the relatively small sample sizes compared.

The hypothesis that decoding accuracy would increase as a function of training was partially supported. While overall accuracy was better for the post-test-transfer

phase than for pre-test, this appeared to be driven only by improvements in identification of Fear and Sadness. Additionally, one participant registered poorer performance at post-test-transfer compared to pre-test. Largely however, improvements made at post-test were maintained at post-test-transfer, supporting the hypothesis that the effects of training with speech stimuli would generalise to previously-unencountered sentences. Broadly, these results were in agreement with those previously reported for CI-simulated listeners, with the exception that simulated listeners showed more uniform improvement across the different emotions after training.

As predicted, overall emotion identification in by CI users was quite similar to the CI-simulated participants tested in Study 3, suggesting that the simulation provided a relatively good approximation of the effects of the CI. However, correlations between confusion matrices in these two studies indicated that the results became more similar with training. This is most likely a reflection of pe-test differences caused by CI users' greater familiarity with the sound of the CI, as hypothesised. The results suggest that the training procedure caused CI and CI-simulated participants to respond more similarly. Therefore, it may be concluded that the training paradigm affected the two groups in a broadly similar way.

The hypothesis that participants would use listening strategies similar to those used by CI-simulated listeners in Study 3 received partial support. Participants' potential listening strategies were probed via two different methods: examination of the distributions of emotion identification errors, and logistic regression modelling aimed at reproducing these response patterns using different auditory feature subsets.

In terms of the patterns of errors and correct responses made, Sadness was reliably identified most accurately, consistent with the results obtained in Studies 1 and 3. However, CI users also displayed a weaker tendency to confuse Anger and Happiness:

emotions that were commonly confused by CI-simulated listeners in Studies 1 and 3. Additionally, as expected, CI users showed a lesser effect of acclimatisation to the stimuli – prior to training, simulated CI listeners showed a strong tendency to overestimate Anger in their responses, which was not so evident with real CI users.

Of the five emotions evaluated, Happiness was decoded with by far the lowest accuracy. This appeared to contradict Volkova et al. (2013)'s (2013) finding that CI users are able to decode Happiness at above-chance level. However, this discrepancy was likely due to the different experimental paradigm employed – since Volkova et al. (2013) asked participants to make a binary distinction of happiness vs. sadness, it is possible that happiness was only recognised inasmuch as it was 'not sadness'. In any case, this result suggested that real CI users had greater difficulty than CI-simulated listeners in distinguishing Happiness.

Using the modelling approach, it was estimated that participants used a listening strategy prioritising information about Spectral centroid, Roughness and Modulation spectrum component 9 (amplitude modulations at approximately 13.5 Hz). This was somewhat different to the strategy estimated in the previous chapter for CI-simulated listeners, which potentially prioritised: Median frequency, Roughness, Length, and Modulation spectrum component 2 (amplitude modulations at approximately 3 Hz).

Therefore, even though the two groups achieved comparable overall emotion identification accuracy, and produced similar patterns of errors, the results suggested that their underlying listening strategies were slightly different. Specifically, it appeared that real CI users prioritised temporal cues to a lesser extent, and tended to focus more on information related to timbre or voice quality. However, both groups appeared to recognise Roughness as an important feature, which could explain why performance was quite similar for both groups. These results imply that either: A)

CI users were more attuned to residual timbre cues, having had longer to acclimatise to the sound of the CI or B) NBV-based CI simulation overestimates the extent to which timbre cues are disrupted.

In any case, these results appear to somewhat contradict previous research, which has suggested that CI users deprioritise frequency-based cues in favour of cues relating to duration and intensity (Peng et al., 2009; Tao et al., 2015). There are two potential explanations for this. Firstly, previous studies have typically focussed on explicitly f0-based cues (e.g. median f0, frequency contour) as opposed to elements of timbre/ voice quality (e.g. Roughness), although timbre cues may be better-preserved by the CI (Vuust et al., 2011). Secondly, as described in the previous chapter, the apparent tendency of CI users to focus to timbre-based cues might result from a 'trade-off' between attending to the most relevant features, and those best conveyed through the CI (Peng et al., 2012). As suggested in the previous chapter, frequency-based cues like Roughness or Spectral centroid appear to be highly relevant for emotion identification in speech, even though they may not be as well-preserved by the implant.

It should be noted that listening strategies were less predictable here than in Study 3 – probably because data from the simulated stimuli were used to approximate strategies applied to non-simulated stimuli. That is, prediction accuracy was constrained by the simulation providing an inherently imperfect representation of the effects of the CI. Furthermore, inter-individual variation in terms of users' experience with the CI, make and model of the device, etc., all likely contributed to higher variation in participants' listening strategies, thereby making the 'average' strategy more difficult to predict.

Taking into account overall emotion identification accuracy and the specific patterns

of errors made by participants, the simulation used in the previous studies seemed to be relatively successful in emulating the CI. This adds support to the suggestion made by other researchers that NBV-based simulation provides a good approximation of listening with CI (e.g. Chatterjee et al., 2015).

Lastly, as hypothesised, there was no reliable indication that emotion identification performance was related to any component of musical expertise or engagement. The 'Musical training' subcomponent appeared to be associated with better decoding accuracy overall, although with such a small sample size it was difficult to be confident about this relationship. As discussed in Chapters 4 and 5, there might have been a more noticeable effect of musical expertise if the experimental design had included a direct comparisons of musicians vs. non-musicians.

7.4.2 Summary of results: Study 6

As was hypothesised, overall emotion identification accuracy was generally above chance level. However, unlike Study 5, this was not the case for all participants and for all experimental phases. Nonetheless, above-chance performance was demonstrably possible, supporting the conclusions of similar studies addressing CI users' ability to decode emotion in music (e.g. Volkova et al., 2013). Consistently with Studies 2 and 4, the results showed that above-chance performance was attainable with a 5-AFC judgement task, as opposed to a binary one, as used by some prior researchers (Volkova et al., 2013). Importantly, this level of performance was possible even using stimuli that varied only in terms of musical performance, rather than composition. Given the relatively low importance ascribed by CI users to compositional cues like mode (Caldwell et al., 2015), this finding is perhaps unsurprising.

The hypothesis that emotion identification accuracy would be improved by training

was only weakly supported. On average, across all participants, accuracy was higher in the post-test-transfer phase than the pre-test. In fact, average accuracy for each of the different emotions saw modest improvement as a function of training, with the exception of Fear, which became worse. However, looking at the individual performance data, two participants registered poorer performance at post-test-transfer than pre-test, one stayed the same, and only two improved. In other words, inter-individual variability in terms of the effects of the training paradigm was much higher for music stimuli than for speech.

Largely, participants made improvements at the first post-test phase, which were not maintained at post-test-transfer. This is concordant with the results of Study 4, and suggests that participants found it difficult to apply their learning to previously-unencountered stimuli. In all, these results were in broad agreement with those reported previously for CI-simulated listeners.

As expected, CI users' overall response patterns for emotion identification were quite similar to those of CI-simulated participants, again suggesting that the simulation was an adequate approximation of the effects of the CI. As with the speech stimuli, correlations between confusion matrices generated in this study and in Study 4 showed that response patterns became increasingly similar with training. As hypothesised, the most likely explanation for this was that groups approached the task very differently to begin with, since CI users were more familiar with the sound of the CI but possibly less accustomed to judging the emotional content of music, and vice versa for CI-simulated listeners. As was the case with speech stimuli, the training seemed to have an 'equalising' effect, causing these two groups to approach the task in a more similar way.

As was the case in Study 5, the hypothesis that participants would use listening

strategies similar to those used by CI-simulated listeners received partial support. In terms of the patterns of errors and correct responses made, Sadness was reliably identified most accurately and Fear least accurately, consistent with the results obtained in Study 4. However, in Study 4, CI-simulated listeners displayed a tendency to systematically overestimate Anger. By contrast, CI users tended to overestimate both Neutral and Sadness, both before and after training. CI users also made fewer confusions between Anger and Happiness, which was a prevalent confusion for CI-simulated listeners. Compared to the speech stimuli in Study 5, Study 6 indicated substantially better decoding of Happiness with music stimuli. It is not clear why exactly this should have occurred – it may have been that CI users were somehow more sensitive to Happiness cues when expressed via music, although no such pattern was observed in Studies 3 and 4.

Interestingly, although the overall distributions of errors and correct responses in emotion identification differed somewhat for the two groups, the underlying listening strategies overlapped to a moderate degree. As with speech stimuli, however, participants' listening strategies were less predictable here than in Study 4. Again, this was most likely due to the of data from CI-simulated stimuli, and perhaps greater inter-individual variation in listening strategies.

Via logistic regression classification, it was estimated that participants selectively attended to nine auditory features: Intensity variation, Spectral centroid, Length, Event density, and five Modulation spectrum components. Four of these features overlapped with those potentially used by CI-simulated listeners in Study 4, and therefore it was assumed that CI users and CI-simulated listeners adopted quite similar listening strategies when identifying emotions expressed by music. The most notable difference was that real CI users seemed to prioritise Modulation spectrum

information to a greater extent, as opposed to other temporal cues, i.e. Pulse clarity and Length. The exact reason for this discrepancy was difficult to ascertain using the current methodology. In any case, participants' listening strategies consisted of attending primarily to temporal cues, which is consistent with the analyses carried out in the previous chapter, and with prior research about musical emotion perception by CI users (Caldwell et al., 2015; Giannantonio et al., 2015).

Lastly, as hypothesised, there was no indication that emotion identification performance was related to any component of musical expertise or engagement.

7.4.3 Summary of results: Comparison of musical engagement

As hypothesised, CI users on average engaged with music to a much lesser extent, had less musical training and made less use of music for emotional regulation and social bonding. This was consistent with recent survey data, in which CI users reported reduced enjoyment of music, and therefore decreased time spent listening to it (Roy et al., 2012a).

Interestingly, however, CI users were much more similar to NH listeners in terms of their use of music for physical exercise and dance. The most plausible explanation for this is that these are ways of engaging with music for which rhythm is the most important content. Since rhythmic components of music are relatively well-preserved by the CI (Limb et al., 2010), participants may have had more motivation to engage with music in these ways. This conclusion is in accordance with Petersen, Sørensen, Pedersen, Parsons, and Vuust's (2015) suggestion that a focus on rhythmic aspects of music and language (in this case, using rap music as a vehicle for musical engagement) might be helpful in encouraging CI users to participate in musical activities.

7.5 Conclusions

These studies successfully built upon the findings of the previous four studies, by incorporating a sample of real CI users within a very similar experimental paradigm. Generally speaking, these studies lent support to the conclusions of Studies 1, 2, 3 and 4, by demonstrating very similar patterns of results to those obtained with simulated CI participants.

Foremost, these studies supported the conclusions of the earlier studies with CI-simulated listeners by showing that real CI users were able to identify emotions expressed via speech and music more accurately than expected by chance. It was demonstrated that this level or proficiency in emotion decoding was attainable, even using a 5-AFC task and, in the case of music stimuli, even when emotions were expressed solely via performance cues, rather than composition. More precisely, with speech stimuli, above-chance performance was reliably achieved by CI users, whereas with music stimuli this level of performance was possible but not guaranteed. Inter-individual differences were far greater, with some individuals improving substantially with training and others remaining the same.

CI users' response patterns showed broad similarities with those made by CI-simulated listeners in previous chapters. Additionally, analysis of both groups' potential underlying listening strategies suggested substantial overlap. Although there were differences at pre-test, training seemed to increase the extent to which CI users and CI-simulated listeners approached the task in a similar way. However, some small differences remained between the two groups, which may be a consequence of CI users' greater experience with the sound of the CI, or of some other discrepancy between CI users' hearing sensation and the approximation provided by the simulation. These small discrepancies might denote effects of the CI (and of the users'

hearing loss) on emotion identification, that are not directly explainable in terms of acoustic signal degradation. For example, having had a much longer time to become accustomed to the sound of the CI may influence the ways in which CI users listen. Equally importantly, modern CIs include increasingly complex DSP designed to improve user experience (e.g. binaural beamforming for noise-reduction (Powers & Fröhlich, 2014)), which is not taken into account by NBV-based simulation – a comparably crude approach. Nonetheless, general similarities observed between these results and those reported in the previous studies suggested that the CI simulation was very successful.

In both studies, identification accuracy varied according to the emotion expressed. With speech stimuli, CI users were able to decode Anger, Fear, Neutral and Sadness relatively accurately, though Happiness proved much more difficult. By contrast, listeners were able to decode Happiness more accurately with music stimuli, but appeared to have particular difficulty in identifying Fear. Considered along with existing literature on this topic, and the previously-reported studies with CI-simulated listeners, these results suggest that the CI does not universally impair perception of any one emotional state (or set of states). There may be theoretical justifications for why certain emotions were more or less discriminable in the speech and music stimuli presented here. At present, however, in the absence of many more studies with different and diverse stimuli sets, it is difficult to know the extent to which the findings obtained for these stimuli were representative. Therefore, it seems more appropriate to conclude that it is *possible* for CI users to decode a range of emotional states, including (but probably not limited to) the five evaluated here. However, the efficacy with which these states can be identified depends on the interaction between the particular listener and the specific stimulus considered.

Lastly, the degree to which individuals were musically trained or engaged with music did not seem to exert a large influence on overall emotion identification, although the small sample sizes studied made it difficult to draw definitive conclusions about this. Compared to NH participants, CI users predictably spent less time engaging in music-related activities. The exception to this was activities involving movement to music, likely because rhythm information is well preserved by the CI. This result suggests that existing descriptions of CI users' diminished overall engagement with music might overlook specific types of engagement that are particularly relevant for this population.

7.5.1 Limitations

The data reported in this chapter were based on relatively few participants, which unfortunately limits the extent to which conclusions can be applied to the wider population of CI users. Primarily this was due to inherent difficulties in identifying and recruiting participants belonging to this population. The lack of participants is a particularly important limitation because of the diversity of CI users as a whole. Due to large variability in myriad factors such as age at implant, device type and degree of residual acoustic hearing, it is difficult to draw strong conclusions relating to this population without studying a much larger sample.

Secondly, as alluded to in Chapter 5, though the training paradigm appeared to be successfully applicable to CI users, it is unclear whether the advantages demonstrated for emotion identification could have translated to substantial longer-term improvement. For this reason, a longitudinal study might be helpful in assessing whether this training procedure is feasible as a means for rehabilitation of CI users' emotion perception. Although the primary purpose of the training paradigm was investiga-

tive rather than rehabilitative, it might be valuable, for both purposes, to determine how many sessions might be required, and how these ought to be structured, in order to facilitate longer-term improvement.

As mentioned earlier in this chapter, the estimation of CI users' listening strategies via computational modelling should be interpreted extremely cautiously. There are two reasons for this. Firstly, by estimating a single listening strategy for a group of CI users, it is logically assumed that a meaningful 'average' listening strategy must exist. This same approach was taken in Chapter 6, although the assumption is more contentious when applied to a sample of CI users, since this group is much more heterogeneous in terms of their sense of hearing. Furthermore, the auditory features used as inputs for the models were derived from analysis of CI-simulated stimuli (as described in the previous chapter). Therefore, successful estimation of listening strategies additionally depends upon the accuracy with which the simulation approximates the effects of the CI.

Lastly, it may be useful for future research to supplement the insights gained from these studies via collection of qualitative data. More precisely, a structured interview-based approach could provide valuable additional information about the ways in which CI users approach the identification of emotion in speech and music. In particular, this methodology may be useful in pinpointing more nuanced aspects of participants' listening strategies, and thereby in delineating differences between CI-simulated listeners and real CI users. In other words, the conclusions made in this chapter about listening strategies could be strengthened and clarified by asking participants about them directly.

7.6 From experimental to clinical assessment of music perception

The next chapter pursues a line of enquiry that is somewhat different to what has been explored thus far, focussing on the measurement of auditory perception in a clinical setting. Specifically, the chapter presents a novel, audiometric approach for the assessment of music perception by hearing-impaired listeners. While the previous chapters have focussed on experimental investigation of speech and music perception, using a training paradigm that might also be useful for rehabilitation, the following chapter aimed to develop a different tool, specifically for use in a clinical context.

8 Study 7: A novel, objective assessment for music perception in aided listening

8.1 Introduction

8.1.1 Rationale

The last two empirical chapters of this thesis (Chapters 7 and 8) had a somewhat different emphasis and set of objectives, compared to the research carried out thus far. Leaving aside emotion identification for now, these studies supported the widely agreed-upon conclusion that hearing-impaired (HI) individuals struggle to a greater degree with perception of music than speech. This discrepancy is described in detail in Chapter 2 of this thesis, but essentially is a product of music being a much more heterogeneous type of stimulus, and research having focussed primarily on restoration of speech. In recent years, the latter factor has been slowly changing and,

increasingly, modern hearing aids (HAs) and cochlear implants (CIs) tend to include a dedicated music 'program', consisting of specialised digital signal processing (DSP) for music perception.

Current assessments of music perception tend to rely on subjective, preference-based measures, however there is currently no standardised, objective measure for the assessment of music perception in HI individuals. Such a measure could be invaluable for both audiological assessment of patients, and for more effective evaluation of different DSP approaches for music. Accordingly, the next two studies were concerned with the development of a new paradigm for objective measurement of music perception in HI listeners. In this chapter, this assessment paradigm was constructed and validated with HA users, and in the following chapter it was adapted and validated for use with CI users.

8.1.2 Overview

In Study 7, a novel test paradigm for music perception in aided listening was created. Summarily, participants listened to brief excerpts of four-part polyphonic music, and on each trial made a two-alternative forced-choice (2-AFC) judgement about whether or not one of the instrumental tracks had a note missing (similar in principle to generic auditory gap detection tasks, e.g. Moore, Peters, and Glasberg (1992)). The difficulty of this task was manipulated by varying the relative intensity level of the relevant instrumental track, compared to the accompanying instruments (referred to as target-to-background ratio TBR).

As participants performed the judgement task, their performance was tracked and difficulty was adjusted using a transformed, 1-up-2-down adaptive procedure. That is, TBR was increased following an incorrect response, and decreased after two consecu-

tive correct responses. On each 'run', three adaptive tracks ran in parallel, permitting the simultaneous estimation of thresholds for multiple experimental conditions. The rationale was that, if an experimental manipulation leads to an objective improvement in music perception, then this should be measurable as a reduced psychometric threshold for the missing note detection task.

In order to trial this paradigm, NH and HI participants were recruited for a small investigation into the effects of different dynamic range compression strategies upon music perception was carried out. Three experimental conditions were investigated: compression of target melody and accompaniment separately ('Independent' compression), compression of mixed ensemble ('Mix' compression), and no compression ('Linear' amplification). It was hypothesised that participants would perform best in the first condition (Independent compression of target and background), because this approach was expected to provide relatively greater amplification to the target melody. Additionally, a subset of normal-hearing (NH) listeners were re-tested, in order to estimate repeatability of the paradigm.

In the next section, the current state of music perception assessment in HI individuals (i.e. HA and CI users) is discussed and some of the shortcomings of current approaches are outlined.

8.1.3 How is music perception currently assessed in aided listening?

In aided listening, different devices or digital signal processing (DSP) strategies are usually evaluated in terms of their objective benefits for speech perception and comprehension. For example, speech reception threshold (SRT) is a common metric which measures the percentage of sentences that listeners are able to listen to and repeat aloud. One such example of this is the Hearing in Noise Test (HINT) (Nilsson

et al., 1994), in which listeners are presented with simple sentences, in the presence of noise, and are asked to repeat each sentence as accurately as possible. The presentation level of the speech, relative to noise (SNR) is varied, in order to estimate the threshold at which the listener can adequately perform the task. The use of this and similar tasks, to evaluate different DSP options, is ubiquitous. For example, Klasen, den Bogaert, Moonen, and Wouters (2007) used HINT performance as a means to compare different binaural noise reduction algorithms for hearing aid users. Similarly, for CI users, performance with HINT sentences has been used extensively in the evaluation of novel DSP options (e.g. Zeng et al., 2005; Brockmeyer and Potts, 2011).

Although to a lesser extent than speech perception, music listening is very important to a large portion of HI patients. Kochkin (2012) showed that improved sound quality when listening to music is a factor that contributes to the overall uptake of hearing aids, while according to Stainsby et al. (1997), cochlear implant users rate music as the second most important listening activity, after speech.

However, perhaps because most hearing prostheses are concerned primarily with improving speech perception, and since the consideration of music as an input to either the HA or CI is comparably novel (Chasin & Russo, 2004), there is no equivalent objective paradigm for assessing music perception. In contrast to speech, different DSP approaches for music perception in aided listening have tended to be evaluated using subjective, preference-based measures. This method involves presenting musical stimuli that have been processed according to some experimental manipulation, and asking listeners which stimuli they prefer to listen to. For example, Croghan et al. (2014) investigated HA users' preferences for various different compression limiting and wide dynamic-range compression (WDRC) processing conditions (release time,

number of channels etc.), as applied to brief (∼13 secs) excerpts of classical and rock music. A forced-choice paired-comparison paradigm was used to assess preference, in which a single stimulus was presented twice, with different processing applied each time, and listeners indicated which version they preferred.

Where psychoacoustical methods have been applied to assess music perception in aided listening, the focus has been predominantly on constituent attributes of music perception, such as pitch perception and timbre discrimination (Drennan and Rubinstein, 2008; McDermott, 2004). Additionally, these assessments have tended to rely on monophonic, synthetic or otherwise impoverished musical stimuli (e.g. Rahne, Bhme, and Gtze, 2011), which might limit their application to real-world listening. As such, these types of assessment are likely unsuitable for clinical application (i.e. for evaluating an HA or CI fitting for music perception) or for facilitating general comparison between different types of DSP.

8.1.4 How could a more objective assessment be designed, and what would be benefits of this?

While individuals' subjective enjoyment of musical stimuli is of course an important consideration when discussing DSP options for music perception in aided listening, there are drawbacks to relying upon this method exclusively. Firstly, this approach makes it difficult to compare, with precision, the effects of different types of DSP when applied to music, since paired-comparison-based assessments of preference tend to record only a preference for A versus B etc., rather than the precise degree to which A is preferred. Even where degree of preference is recorded, there is likely to be inter-individual variability in the way that listeners interpret these kinds of judgements.

As noted by Roy et al. (2012b), preference judgements may be influenced by myriad extraneous variables, e.g. musical background, age, personality factors, and so on. Therefore, there tends to be relatively high variability associated with these judgements, as a consequence of individuals' different criteria used to make the judgements, and/ or differing interpretations of adjectives used as the bounds of perceptual scales (Gfeller et al., 2000). Additionally, preference ratings may be unstable over time subjective assessment at fitting may change after several months, whereas an objective metric should in principle be more robust. In other words, subjective measurements of music perception lack a quantifiable figure that is invariant between participants, as exists for speech perception.

In principle, music perception could be measured objectively in a similar way to speech perception, by using an adaptive (or 'staircase') procedure. This widespread psychophysical method is typically used in the case where one wishes to estimate a threshold at which some perceptual task can be performed. Within this paradigm, the difficulty of a task is manipulated according to the participant's performance i.e. the task becomes more difficult when the participant responds correctly, and becomes less difficult when the participant responds incorrectly (Levitt, 1971). Put another way, on each trial, physical characteristics of the experimental stimuli are determined by both their counterparts from the previous trial(s), and the participant's response(s), (Leek, 2001). For example, in the context of speech perception, a battery of sentences is presented to the listener, who must repeat each one aloud as accurately as possible. In this case, a 'correct' response denotes a completely accurate repetition of the sentence, and task difficulty is manipulated by adjusting the presentation level of the sentence, relative to speech-shaped noise. After each response, or after a number of responses, the difficulty of the task is adjusted, ac-

cording to the listener's performance correct responses lead to a decrease in SNR, while incorrect responses lead to an increase.

Building upon the adaptive procedure as described above, Levitt (1971) describes the 'transformed up-down procedure', in which task difficulty is not updated on every trial, but instead follows a pre-defined rule. For example, a 1-up-2-down adaptive procedure denotes that the difficulty of the task will be decreased following each incorrect response, but will only be increased following two consecutive correct responses. Additionally, the 'step size' (i.e. the increment by which task difficulty is adjusted) may be altered during the adaptive procedure. Typically, a larger step size is used to begin with, to minimise the amount of observations that occur far from the perceptual threshold of interest. As the listener approaches this threshold, the step size is reduced in order to more accurately estimate it. To estimate the threshold, a mean is usually calculated based on the task difficulty over a predetermined number of 'reversals' usually defined as the points at which the listener gives an incorrect response following a correct response, or vice versa.

The study described in this chapter represented a first attempt at adopting the adaptive procedure as described above, for the assessment of music perception. The next section provides a detailed outline of the assessment paradigm constructed, and the musical stimuli used.

8.2 Method

8.2.1 Participants

Nine HI listeners (mean = 69.09 years old, SD = 9.81) were recruited via Starkey Hearing Technologies' Subject Tracking and Retrieval volunteer database. All had

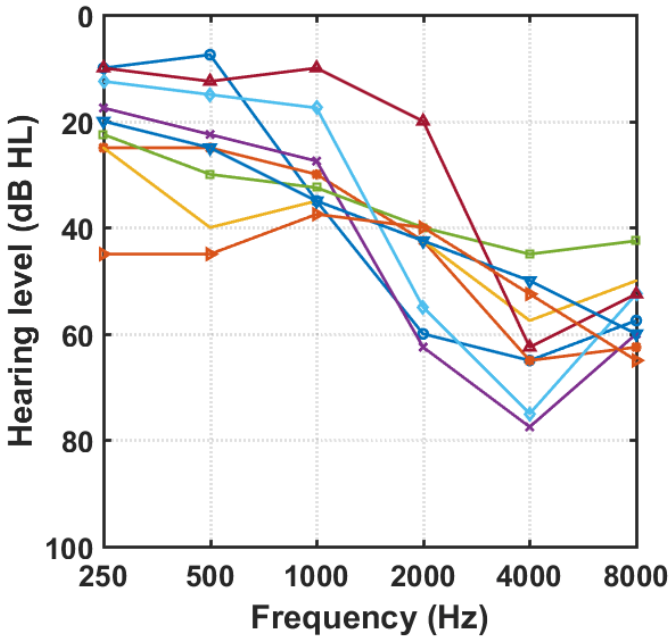


Figure 47: Audiograms for HI participants, averaged across right and left ears.

bilaterally symmetrical, normal or mild sloping to moderate sensorineural hearing loss, with air conduction thresholds at or below 70 dB HL at 8 kHz (Figure 47). All participants had previously expressed an appreciation of music and reported engaging in musical activities (e.g. playing an instrument, attending concerts).

Seven normally-hearing participants, all employees of Starkey Hearing Technologies, were also recruited (mean = 38.57 years old, SD = 11.19). These participants all had audiometric thresholds less than or equal to 25 dB HL across the 250 - 8000 Hz range. Participants in both conditions provided fully informed consent, and HI participants were reimbursed for their time.

8.2.2 Test stimuli

Participants listened to brief (mean = 6.36 seconds, SD = 1.19) musical excerpts, taken from eight pieces of four-part J. S. Bach chorales, from the Bach10 dataset

(Duan & Pardo, 2011). The original stimuli consisted of separate audio recordings of violin, clarinet, saxophone and bassoon, respectively playing the soprano, alto, tenor and bass parts of each piece. Using sample editing software (FL Studio), the individual instrumental tracks from each recording were split into an average of five excerpts, according to their constituent musical phrases.

For each of these excerpts, four different components were derived: the saxophone part alone, the clarinet part alone, all instrumental parts except for the saxophone, and all instrumental parts except for the clarinet. Hereafter, the isolated saxophone and clarinet components are referred to as 'target melodies', and the respective remaining instrumental parts are referred to as 'backgrounds' (i.e. clarinet alone = target melody, all instruments except for the clarinet = background. Because of the consideration that the instruments in the highest and lowest pitch ranges (violin and bassoon, respectively) may be more perceptually salient, these instruments were not used for the target melodies.

For all of the target melodies, along with the original version, a second 'note-omitted' version was created. In this version, a single note was removed from the melody, using sample editing software, with an approximately 5 msec linear fade-in and fade-out either side of the omission. To ensure approximately equivalent difficulty across different stimuli, the note removed was always a quarter-note (approximately 600 msec duration), always occurred on the off-beat, and never occurred at the start or end of a musical phrase (Figure 48).

In each trial, participants were first presented with a target melody, and then a 'mix' stimulus, consisting of a target mixed with its respective background (Table 26). For each mix, the relative level of the target melody compared to the background was varied, in increments of 1 dB, in order to adjust the difficulty of the task. Therefore,

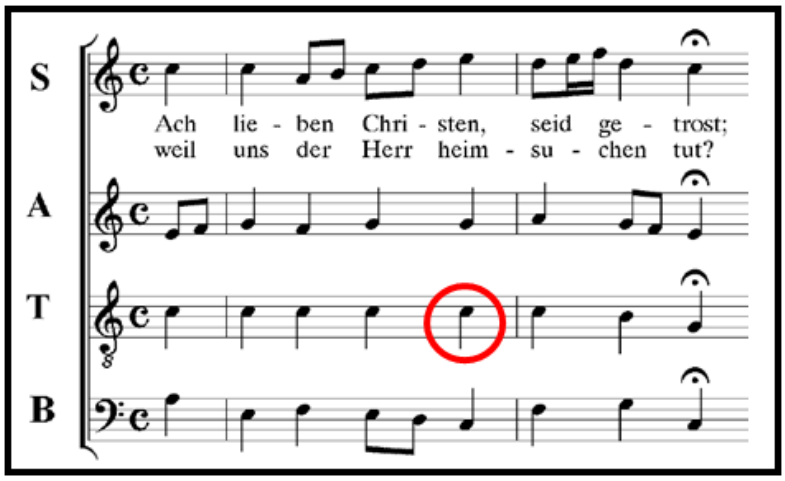


Figure 48: Example of a musical excerpt used. Red circle indicates the saxophone (Tenor) note omitted in the manipulated version of the stimulus.

task difficulty was inversely proportional to target-to-background ratio (TBR).

Table 26: Summary of the four target-mix stimuli pairs used.

Target	**Mix**
Clarinet (original)	All instrumental tracks (original) All instrumental tracks (note omitted from clarinet)
Saxophone (original)	All instrumental tracks (original) All instrumental tracks (note omitted from sax)

For the mixed stimuli, there were three different amplification conditions: dynamic range compression applied to the target melody and background individually (i.e. prior to mixing), dynamic range compression applied to the mix (i.e. after mixing the target and background together), and a control (Linear amplification no compression) condition. For the former two conditions, the NAL-NL2 adult prescriptive procedure (Keidser et al., 2011) was used to generate participant-specific insertion gains, based on individual audiograms (i.e. amounts of amplification, at different frequency bands, needed to bring the hearing level up to a pre-specified target). Since Starkey hearing aids use default 5 msec attack/ 300 release settings, the 'fast-acting' compression parameter within NAL-NL2 was used corresponding to approximately

5 msec attack / 70-120 msec release (Keidser, Dillon, Dyrlund, Carter, & Hartley, 2007). Insertion gains for each participant were interpolated for sixteen channels (each with a bandwidth of 625 Hz). These gains were used to derive the research fitting software (RFS) configuration files ultimately required to produce the necessary compression. Since NAL-NL2 does not provide insertion gains for frequencies above 8 kHz, the gains for these frequency bands were set to that prescribed for 8 kHz, following Sęk and Moore (2012).

For NH listeners, insertion gains were generated assuming a flat 50 dB hearing loss intended to provide perceptible compression ratios (ranging from 1.4 to 3 across the 156Hz - 10 kHz frequency spectrum for soft-moderate sounds, and 1.2 to 3 for moderate-loud sounds), such that participants would be able to notice the effects of the compressor (Neuman, Bakke, Mackersie, Hellman, & Levitt, 1998). However, stimuli presented to normally-hearing listeners were attenuated by a flat 10 dB RMS, such that the frequency shaping and compression from the NAL-NL2 prescription was retained, whilst the actual increase in broadband level was not, thereby reducing an effect of overall audibility. For the control condition, linear gain processing was applied to the stimuli by setting the gain prescribed by NAL-NL2 at the 'soft' and 'loud' listening levels as equivalent to the 'moderate' gain (65 dB long-term average speech spectrum (LTASS) – corresponding approximately to the average presentation level for background stimuli, before mixing in the target). In this condition, no compression was applied.

It was expected that the note omission manipulation would be most easily noticeable in the Independent condition, since the target instrument would receive relatively greater gain, compared to the accompanying instruments (see Figure 49). This was because, in the Independent condition, compressor activity was driven only by the

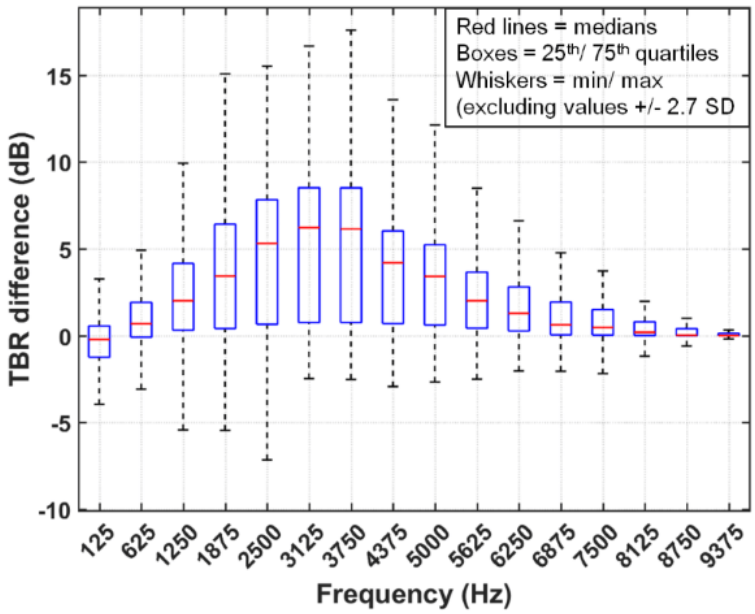


Figure 49: Amplified sub-band TBR differences between Independent and Mix conditions, averaged for all musical excerpts, at -6 dB TBR for clarinet stimuli. Amplification based on gains for flat 50 dB loss.

target instrument. Additionally, dynamic range compression applied to all of the instrumental tracks simultaneously (as in the Mix condition) increased 'across-signal modulation coherence' – i.e. the degree of similarity in amplitude modulations across the different instrumental tracks (Stone & Moore, 2007). In principle, this would be expected to increase the difficulty of identifying a missing note in any individual melody.

In total, there were 222 unique target-background pairs for each target instrument (saxophone, clarinet): eight musical pieces × average of five excerpts per piece × two manipulation conditions (original target, note-omitted from target) × three amplification conditions (Independent compression, Mix compression, Linear amplification).

Prior to testing, the headphone output was calibrated using a Brel & Kjr Type 4153 Artificial Ear coupler, Brel & Kjr 4192-B-001 condenser microphone and Larson Davis System 824 sound level meter. The headphones were clamped to the coupler,

and the microphone was attached to it. A 1 kHz sinusoidal tone was played at -25 dB RMS over the headphones, and the output level was recorded by the microphone and sound level meter. This level was measured to be (85 + 14) = 99 dB SPL, therefore the full-scale (0 dB FS) level was (99 + 25) = 124 dB.

Isolated target stimuli were always presented at 65 dB SPL, prior to amplification. To construct the mix stimuli, the backgrounds were also presented at 65 dB SPL, while the level of the target within the mix (and therefore the overall level of the mix stimuli) varied between trials, according to the TBR.

8.2.3 Procedure

Participants were seated in a sound-attenuated laboratory, and listened to the auditory stimuli via Sennheiser HD 600 headphones, powered by a Lavry DA10 digital to analog converter and RME Hammerfall DSP Multiface II audio interface.

Stimulus presentation, including both the adaptive logic and the interface used by participants was coded in MATLAB. A 1-up-2-down adaptive procedure was used to estimate the psychometric threshold (for 70.7 percentile performance) for the perception of note emissions in the different amplification conditions (Independent, Mix, Linear) (Levitt, 1971). Three adaptive tracks ran in parallel during the experiment, one for each amplification condition, following Micheyl, Divis, Wrobleski, and Oxenham (2010). The adaptive tracks were interleaved according to the amplification condition in each trial therefore, TBR was adjusted, and thresholds were tracked, independently for each amplification condition.

For each trial, participants were first presented with the target in isolation. After a 1 sec delay, the mix was then presented. A 2-AFC paradigm was used, wherein par-

Figure 50: Illustration of the test procedure for one trial. Prior to amplification, isolated target stimuli were presented at 65 dB SPL, and backgrounds were presented at 65 dB SPL. Level of targets presented with backgrounds varied, according to the TBR used.

ticipants judged whether the target instrumental track was the same as, or different to, the corresponding instrumental track within the mix (Figure 50).

On each trial, the manipulation condition (original versus note-omitted) was chosen at random. The presentation order of the excerpts from each piece was also randomized, although it was ensured that the same excerpt did not occur consecutively, and that all excerpts were used before any were repeated. Similarly, amplification condition order was pseudo-random, such that each condition occurred exactly once every three trials (though a condition no longer occurred once the threshold associated with it had been estimated). Participants responded using a graphical user interface (GUI) (Figure 51), presented on a computer monitor, by clicking on one of two buttons marked 'same' and 'different'. To prevent input errors, participants were then required to click a 'confirm' button in order to save the response, and advance to the next trial. During the main experiment, participants did not receive any feedback regarding their responses.

At the start of the staircase procedure, the TBR was set to 10 dB. Following each incorrect response, the TBR was increased (making the task less difficult), and following two consecutive correct responses, the TBR was decreased (making the task more difficult). To begin with, TBR was adjusted by increments of 4 dB, until the

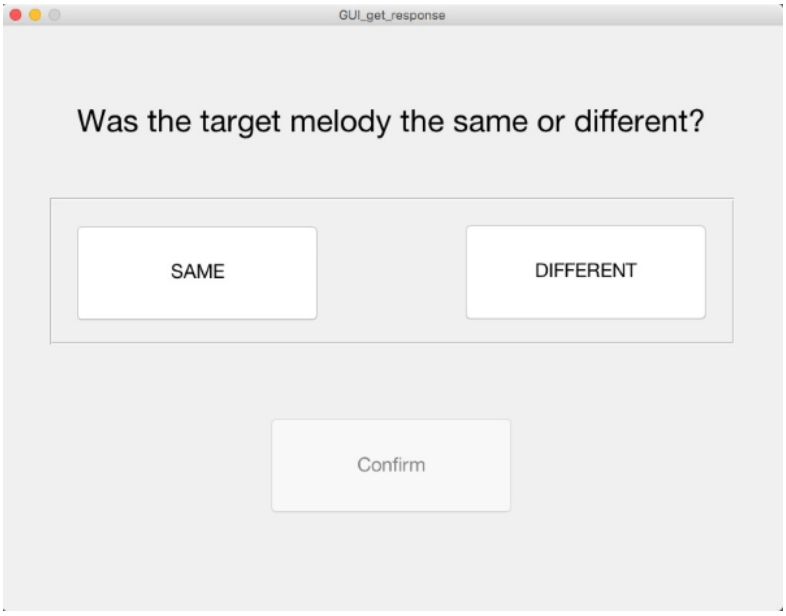


Figure 51: Screenshot of the GUI used for the 2AFC task.

first reversal from increasing to decreasing TBR occurred. At the second point, the TBR adjustment increment was reduced to 2 dB. After the second reversal from increasing to decreasing TBR, it was reduced to 1 dB. TBR was not limited to a particular range, though in practice typically varied from approximately -20 dB to 10 dB. Once a total of four reversals (either direction) occurred, using the 1 dB step size, the threshold was calculated as the mean TBR across these reversals. This entire process lasted approximately thirty-five minutes per target instrument, and was carried out two times per participant, with one block estimating the thresholds for the three amplification conditions with the clarinet playing the target melody, and the other block estimating thresholds for the saxophone playing the target melody. The order of these blocks was counterbalanced, such that half of the participants listened to the saxophone target melodies first, and half heard the clarinet first.

Prior to the staircase procedure, participants completed a practice block of twelve trials, in order to familiarize them with the experimental paradigm. Stimuli for the

practice phase were constructed using a piece from the Bach10 dataset (Duan & Pardo, 2011) that was not used in the data collection trials. Therefore, these were very similar to those in the main experiment. During this phase, the experiment ran as normal, except that no adaptive procedure was running, and instead stimuli drawn at random were presented at a constant TBR of 10 dB. Additionally, participants received text-based feedback (i.e. 'correct' or 'incorrect') via the GUI during the practice session. Including the practice trials, the experiment lasted for approximately two hours in total. All participants were tested individually, and all testing for each participant occurred within a single session.

In order to assess the repeatability of the threshold estimating paradigm used, a subset of five NH participants were asked to attend a second experimental session. These second sessions followed the same procedure as described previously, except that the order of instrumental blocks was reversed (i.e. participants that initially heard the saxophone stimuli first, now heard the clarinet first). For all participants, there were at least two days between the experimental sessions.

8.2.4 Hypotheses

Firstly, it was expected that the test procedure would be reliable, and therefore that those participants taking part in two sessions would register similar thresholds for each session.

Importantly, it was expected that the test procedure would be sensitive enough to detect differences in thresholds, between the DSP conditions intended to make the task easier or more difficult. Specifically, it was hypothesised that the omission of notes from musical stimuli would be most noticeable in the Independent condition, since the target instrument would receive more amplification, compared to the ac-

companying instruments, than in either the Mix or Linear conditions. This pattern of results was predicted for both NH and HA participants.

In terms of overall thresholds, no significant differences were predicted between NH and HA participants, since HA participants listened with amplification equivalent to that of their HA, while NH participants listened with amplification prescribed for a flat 50 dB hearing loss. Therefore, stimuli were presented at approximately the same hearing level for both groups of participants, and frequency shaping as a consequence of amplification was roughly equivalent for both.

Lastly, although the lowest- and highest-register instruments (bassoon and violin, respectively) were not used as target melodies, it was considered that the two instruments used for target melodies (saxophone and clarinet) might differ in terms of their salience within the musical ensemble, thereby affecting task difficulty. During piloting, slightly lower thresholds tended to be achieved using the saxophone-as-target-melody stimuli, and so it was expected that a small effect of musical instrument upon overall thresholds might be found.

8.3 Results

Firstly, to estimate the error inherent in the adaptive procedure's threshold estimation, a Monte Carlo simulation was used to generate 10,000 'runs' of the experiment. The error distribution of the maximum a posteriori (MAP) estimation is shown in Figure 52. Mean error was -0.6 dB, and 95% of probability mass was between -3.4 to 1.3dB, indicating that the adaptive procedure produced reliable threshold estimates, albeit with a slight tendency to underestimate.

For the subset of NH participants who completed two testing sessions, threshold

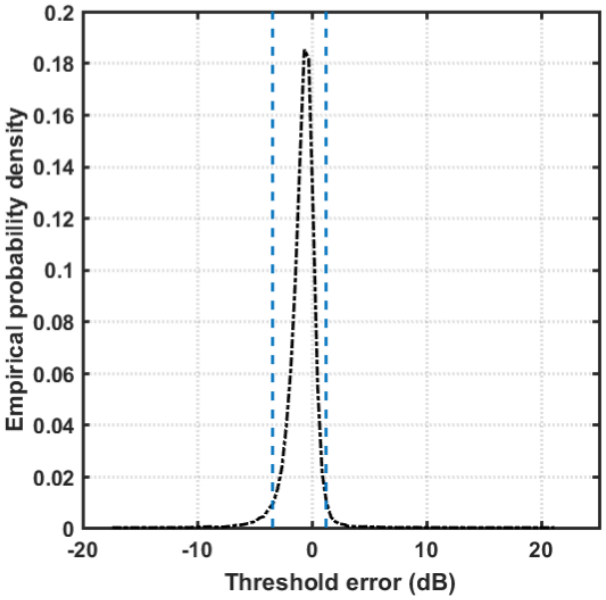


Figure 52: Threshold estimate error (MAP) for the adaptive procedure used, based on a Monte Carlo simulation of 10,000 runs. The area enclosed by the blue lines denotes 95% of probability mass.

estimates were relatively consistent from Session 1 to Session 2, despite one participant demonstrating markedly poorer performance in the second session (Figure 53). Pearson correlation showed that threshold estimates in the two sessions were strongly positively correlated, $r = .72$, $p < .001$.

To investigate the relative difficulty of the different musical excerpts used, percentages of correct responses for each excerpt, from each musical piece, were calculated (Figure 54). On the whole, stimulus difficulty was relatively uniform. To examine this further, logistic regression was used to predict correct responses, based on the interaction of musical piece $\times$ excerpt, after taking into consideration the effects of TBR, amplification condition, and manipulation (original versus note omission), in order to identify stimuli that had a significant influence upon the prediction made. Using this approach, in both the NH and HI groups, three stimuli were identified as being potentially problematic and, in particular, two of these were consistent across

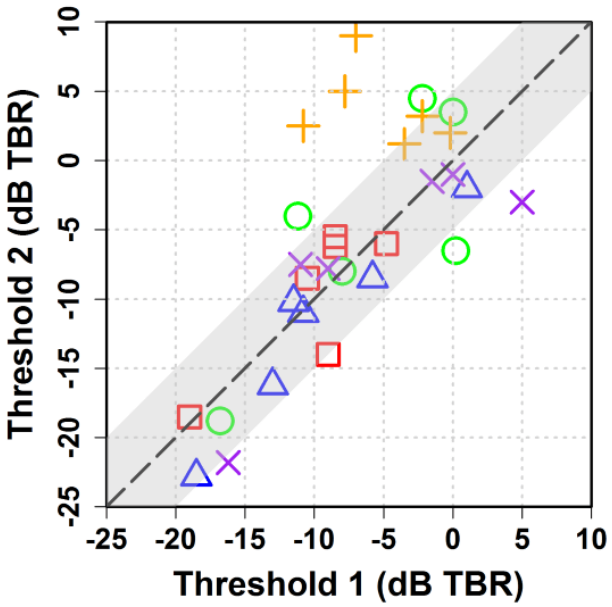


Figure 53: Relationship between NH participants' scores (indicated by the different symbols) in two different experimental sessions. Shaded area indicates +/- 5 dB, dotted line indicates +/- 5 dB.

all participants.

Individuals' thresholds for each amplification condition and target instrument are shown in Figure 55. In general, individuals showed better performance for the Independent stimuli (as indicated by lower perceptual threshold estimates), relative to the Mix stimuli, although this effect was more pronounced for NH listeners than HI listeners. Normal-hearing participants also appeared to perform better with Independent stimuli relative to Linear, although this effect was much less consistent for HI participants. In almost every condition, both NH and HI groups tended to achieve lower thresholds when listening to the saxophone target melodies.

To assess the impact of hearing impairment, target melody instrument, and amplification condition upon music perception, a 2 × 2 × 3 mixed ANOVA was computed, with hearing status (NH, HI), target melody instrument (clarinet, saxophone) and

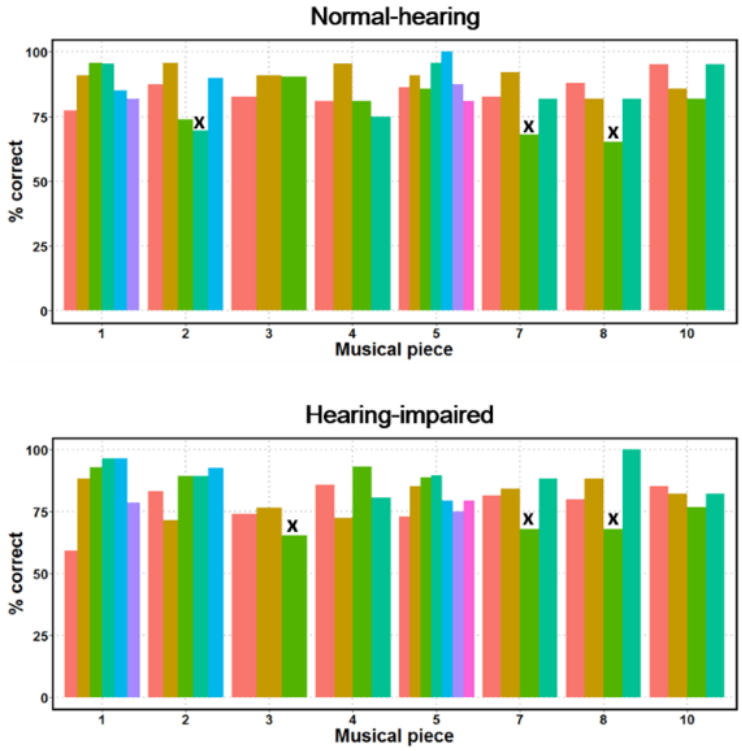


Figure 54: Total percentages of correct responses for each of the different stimuli used, across the different musical pieces and excerpts, for saxophone target melodies. × symbols indicate potentially problematic stimuli, as identified by the logistic regression analysis.

amplification condition (Individual, Mix, Linear) as independent variables and perceptual threshold as the dependent variable. ANOVA revealed significant main effects of both target melody instrument and amplification condition upon threshold estimates, $F(1, 70) = 88.82$, $p < .001$, $\eta^2 = .34$ and $F(2, 70) = 3.63$, $p = .032$, $\eta^2 = .03$, respectively. The effect of hearing status (NH, HI) was not significant, $F(1, 14), 4.13$, $p = .062$, $\eta^2 = .08$, potentially because of the reduced statistical power associated with the between-subjects comparison. None of the interaction effects were statistically significant.

For the main effect of amplification condition, a Bonferroni-adjusted post hoc test, showed a significant difference only between the independent and mix conditions,

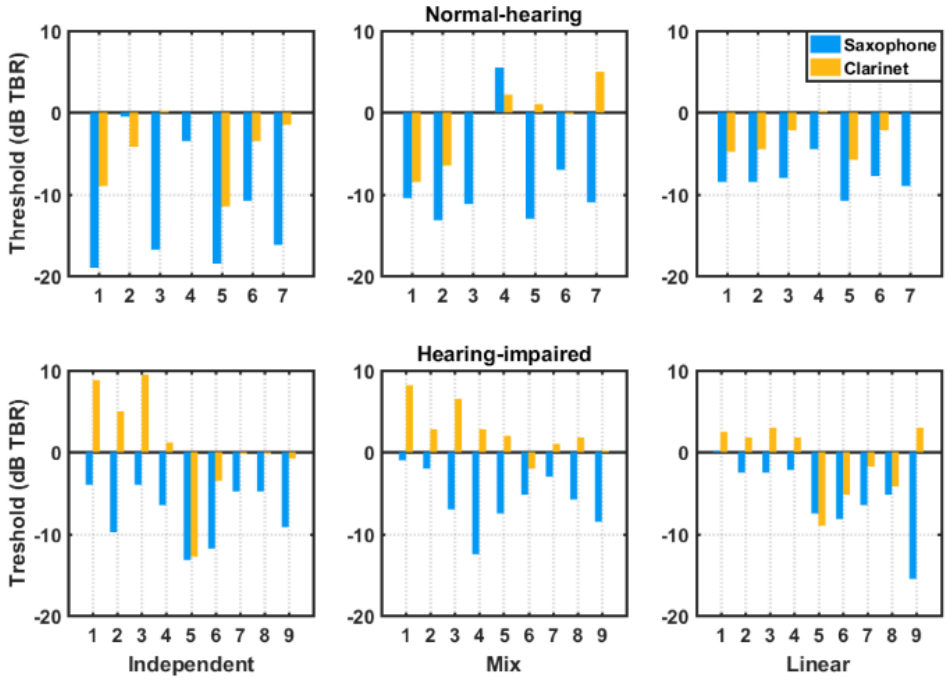


Figure 55: Individuals' estimated thresholds in the three amplification conditions, for both saxophone and clarinet target melodies. Lower thresholds indicate better task performance.

$p = .040$, $d = .66$. Figure 56 shows the mean differences in thresholds, across all participants, for the different amplification conditions.

Across all participants (both HI and NH, pooled), a linear robust regression model was able to predict the improvement in thresholds for the Independent condition, compared to the mix, based on individuals' performance in the former, $F(1, 28) = 12.53$, $p = .002$, $R^2 = .30$. This indicates that there was a greater beneficial effect of Independent amplification when listening at lower TBRs (Figure 57).

Lastly, to explore the difference observed in threshold estimates as a function of target musical instrument, measurements of timbre were calculated for all stimuli using MIRtoolbox for MATLAB (Lartillot & Toiviainen, 2007) (Table 27).

Unsurprisingly, the timbre values observed for the two musical instruments were dif-

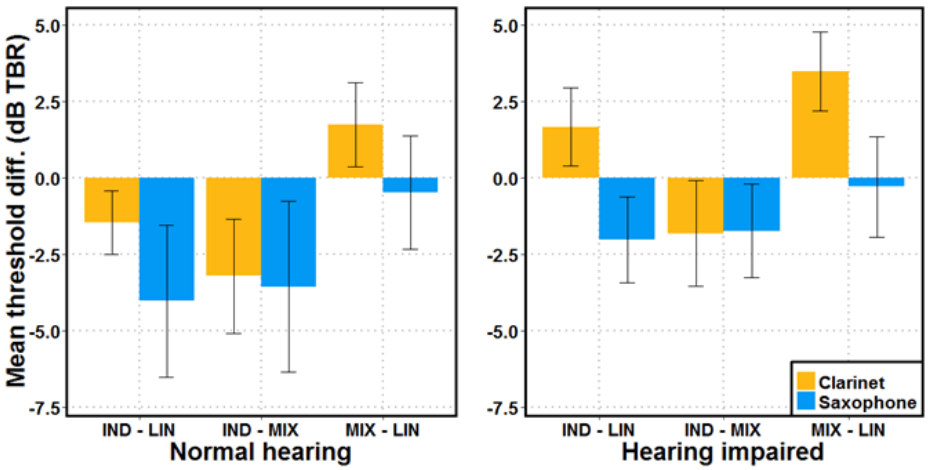


Figure 56: Mean threshold differences for each pairwise comparison of the three amplification conditions, for the two target instruments. For each comparison, a negative value indicates better performance with the first amplification condition. Error bars indicate +/- 1 SE. IND = Independent, LIN = Linear, MIX = Mix.

Table 27: Mean measurements of different timbre components for the Clarinet and Saxophone target melody stimuli.

Timbre measure	**Instrument**	**Mean value**
Brightness	Clarinet	0.19 (SD = 0.04)
	Saxophone	0.33 (SD = 0.04)
Roughness	Clarinet	0.11 (SD = 0.07)
	Saxophone	1.69 (SD = 0.94)
Spectral centroid	Clarinet	967.45 (SD = 113.79)
	Saxophone	1378.89 (SD = 118.30)

ferent (differences for all measures were significant at p <.001), denoting that the saxophone stimuli were on average brighter, rougher and with a higher-frequency spectral centroid. The larger amount of high-frequency energy present in the Saxophone signals is illustrated in Figure 58. The spectrograms clearly show that the Saxophone stimuli were more salient than the Clarinet stimuli, by comparison to their respective musical 'backgrounds'.

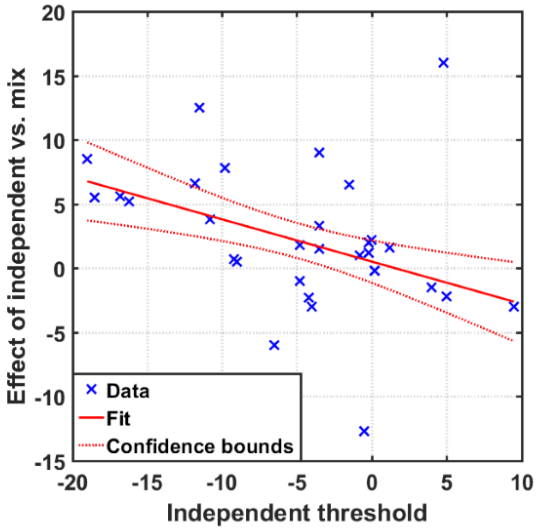


Figure 57: Linear regression model, predicting effect of Independent versus Mix, based on all participants' thresholds (NH and HI) in the Independent condition.

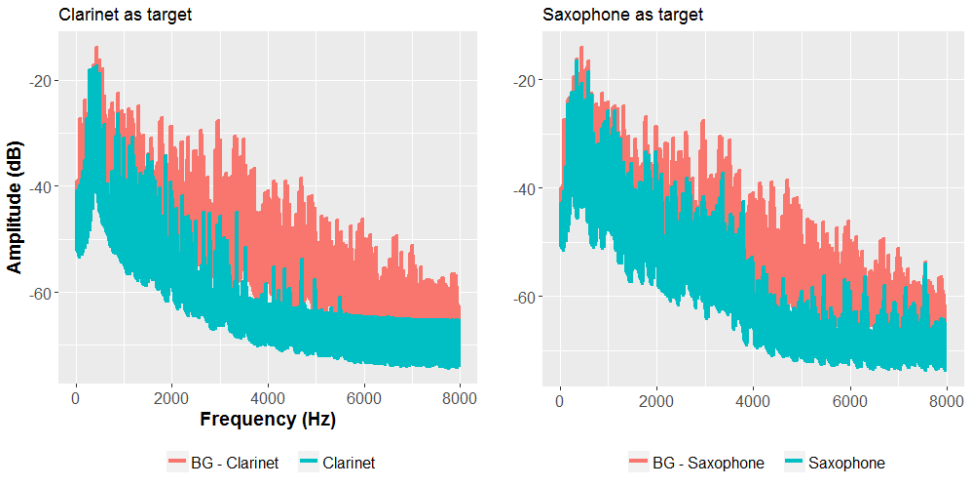


Figure 58: Frequency/ amplitude spectrograms for all of the target melodies played on the Clarinet and Saxophone, compared with their respective musical backgrounds.

8.4 Discussion

8.4.1 Summary of results

Firstly, the assessment procedure developed appeared to be sufficiently reliable. Analysis showed that the adaptive procedure made only very small errors across many simulated trials, tending to estimate thresholds very accurately. Additionally, participants registered very similar scores across different sessions – approximately 80% attained thresholds in the second session which were within ±5 dB of their initial thresholds. However, reliability was only assessed for a sub-sample of NH participants, and therefore the possibility exists that reliability may be lower with HA users.

Encouragingly, percentages of correct responses across each of the individual musical stimuli were relatively uniform, indicating that the different stimuli were of approximately equal difficulty. There were, however, two stimuli that appeared to be disproportionately difficult for both NH and HI participants. Upon closer inspection, it was not clear why this should have been the case: there were no obvious compositional differences in these excerpts, nor any noticeable artefacts created by the compression manipulation or otherwise. Nonetheless, the data indicated that these particular stimuli were problematic, and hence should be removed from the assessment procedure.

Thresholds were significantly affected by the amplification manipulation, as predicted on the basis of sub-band TBR differences between conditions. More precisely, only the Independent and Mix condition produced significantly different thresholds. The most likely reason for this result was that, in terms of both 'effective TBR' and across-signal modulation coherence (Stone & Moore, 2007), the contrast between

these two amplification conditions was the greatest. In other words, the Mix condition decreased the extent to which target melodies were amplified, relative to the background, while simultaneously increasing similarity between the amplitude envelopes of each of the instrumental parts. By contrast, the Independent condition had the opposite effect. Thus, there was a large difference created between the Mix and Independent conditions, whereas Linear amplification had an effect somewhere between the two.

This conclusion was supported by the regression analysis, which was able to predict the improvement in thresholds for the Independent condition, compared to Mix, based on individuals' performance in the former. This indicated greater improvement for Independent amplification when listening at lower TBRs. Thus, the increase in 'effective TBR' was most noticeable when the actual TBR was lower. The overall effect of amplification was indicative of the validity of the test procedure, since it was able to detect differences in listeners' thresholds when DSP manipulations intentionally altered the difficulty of the task.

As hypothesised, the effect of hearing status was not significant, although there was a trend towards NH participants performing better than HA participants. It is likely that this difference was driven, at least partially, by NH participants' relatively greater engagement with music, although unfortunately no data were collected to support this suggestion. Additionally, the sample of NH participants included a number of professional audiologists, who may have been more familiar with this type of experimental task.

Lastly, thresholds were significantly higher for clarinet target melodies, compared to saxophone. Although some difference was expected, the magnitude of this effect was larger than anticipated. Timbre analyses of the respective instrumental melodies

showed that saxophone stimuli were significantly brighter and rougher, with a greater proportion of high-frequency energy. Therefore, the likely explanation for this result is that saxophone target melodies were substantially more salient compared to their respective backgrounds. In any case, this finding highlighted the importance of carefully choice of musical instruments, when constructing this type of task.

8.4.2 Limitations

This was a relatively small-scale pilot experiment, and as such the number of participants recruited, and in particular HI participants, was quite low. Considering the relatively high variability in participants' TBR thresholds for the note detection task, it would be informative to have a larger sample size. This would facilitate the systematic examination of factors that potentially contributed to inter-individual differences – for example, musical abilities and/or experience. Unfortunately, due to time constraints during participant recruitment, it was not possible to test a large enough sample size to facilitate a musicianship comparison. Instead, all of the participants had a self-professed interest in music, but potentially varied considerably in their degree of formal instrumental tuition, and related indices of musical experience.

In cases where participants' performance became erratic – for example, due to fatigue – it is possible that the adaptive procedure used produced less meaningful results. When estimating a threshold using a given number of reversals, it is generally assumed that these reversals should cluster somewhat closely together. However, in some cases, reversals were more sparsely clustered, making it difficult to infer the 'true' threshold. In other words, erratic performance (caused, for example, by listener fatigue) can lead to less meaningful threshold estimates (see Figure 59). This is a problem inherent when estimating thresholds using an adaptive procedure in

this way, since an inference is made about the subject's psychophysical threshold is based upon only a subset of the available observations.

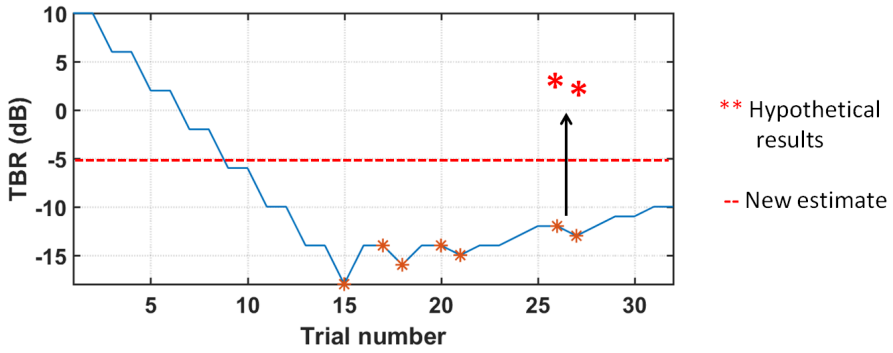


Figure 59: Illustration of how erratic performance within an adaptive procedure can lead to potentially misleading psychoacoustic threshold estimates.

There are at least two potential solutions that could circumvent this issue. The first option is the addition of an additional 'stopping condition' – that is, an extra criterion that must be met before the experimental 'run' is terminated and a threshold estimated. For example, as well as requiring that four reversals occur at the smallest step-size, it could also be required that these reversals fall within a specified TBR range. In principle, this would ensure that any 'outlying' reversals would be excluded from the threshold estimation, which should lead to more robust results. However, an obvious caveat to such an approach is that the duration of the experiment would be indeterminate, and could vary wildly between participants, depending on the consistency of their performance on the task. Therefore, some participants may complete many more trials than others, and be affected by fatigue to a much greater extent. Theoretically, such a paradigm could continue indefinitely, given that increasing fatigue might exponentially increase the prevalence of erratic results.

A second way to improve threshold estimation in the adaptive procedure, is by the use of a Bayesian modelling approach. This would essentially entail incremental

estimation of a participant's threshold after each trial, so that the final estimate would take into account all of the available data (i.e. the participant's performance at each presented TBR level). The primary advantage of this approach is that no observations are 'thrown away', which should in principle lead to more robust threshold estimation. However, a thorough comparison, investigating the advantages of using Bayesian modelling for psychoacoustic threshold estimation, is not within the scope of this thesis.

Lastly, it should be kept in mind that objectively 'accurate' perception does not always equal a good listening experience. In this study, perceptual 'verisimilitude' was assumed to be desirable and indicative of 'better' music perception, but this approach neglected aesthetic enjoyment. As Limb (2016) has pointed out, music may engage a somewhat different 'mode' of listening, which may be inherently less amendable to objective assessment. For this reason, the approach outlined in this chapter was intended to supplement rather than replace subjective methods of assessment. That is, thorough examination of music perception ability, either for clinical audiometry or for evaluation of novel DSP strategies, should include both subjective and objective measurements.

8.4.3 Conclusions

The 'music audiometry' paradigm trialled here appears to be both a reliable and valid way to examine perception of polyphonic music by HA users. In this experiment, the note omission task yielded similar performance across different test sessions, and was sensitive enough such that differences in the amplification of polyphonic music led to significantly different estimated psychometric thresholds. There may be some minor, inherent issues with the adaptive tracking paradigm, although the optimisation of

this threshold estimation procedure is outside the scope of this thesis. Additionally, the discrepancies revealed between threshold estimates with saxophone compared to clarinet target melodies highlight the importance of careful choice of experimental stimuli, in order to ensure equivalent task difficulty. In this case, saxophone target melodies led to less variable threshold estimates generally, and so future versions of this experiment could be streamlined by excluding clarinet target melodies.

The experiment must be interpreted with caution, however, since relatively few HA users were tested, although these initial results are promising and suggest that the paradigm developed may be a useful tool for the examination of music perception in HI listeners. As already alluded to, however, this approach is intended to supplement, rather than replace, preference-based assessments.

The development of an objective paradigm for music perception assessment in aided listening constitutes an important step towards evaluating (and hopefully thereby improving) hearing aid DSP for music. Eventually, it is hoped that a refined version of the paradigm piloted here could be used in much the same way that conventional speech audiometry, in order to assess the extent to which HA users are able to perceive music, and the extent to which this is improved by different fittings and/ or DSP approaches. More specifically, this 'music audiometry' approach could be used in conjunction with existing tests of sound quality and subjective preference, in order to evaluate the success of music programs.

8.5 Extension of this test procedure to cochlear implant users

In the next chapter, the feasibility of using this paradigm for other HI populations is investigated. Specifically, the procedure is trialled with simulated CI participants. The chapter first discusses the potential benefits of adopting such an assessment for

CI users, along with some potential difficulties, followed by a systematic evaluation of the paradigm with a CI-simulated NH sample.

9 Study 8: Objective assessment of music perception in cochlear implant-simulated listeners

9.1 Introduction

9.1.1 Overview

The central aim of this chapter was to build upon the objective music perception assessment procedure developed in the previous chapter, by assessing its suitability for cochlear implant (CI) users. Music perception in CI users, especially with respect to polyphony, is typically much worse than for HA users (Donnelly, Guo, and Limb, 2009; Penninger, Limb, Vermeire, Leman, and Dhooge, 2013) and therefore it was unclear whether the use of the same paradigm to measure psychoacoustic thresholds would be a plausible approach. Therefore, Study 8 sought to investigate this possibility. Unlike the previous study, the aim of this experiment was not to move towards a 'definitive' assessment of music perception, but instead to explore whether the use of same approach might or might not be valid when applied to the case of CI listening.

In this chapter, current techniques for assessment of music perception assessment specifically in CI users are described and evaluated. As in the previous chapter, the drawbacks of currently-used approaches are discussed, and the objective assessment procedure is offered as a potential, supplementary methodology.

9.1.2 How is music perception currently assessed for CI users?

The history of music perception assessment in CI users has followed a similar trajectory to music perception assessment with hearing aids (HAs). That is, the use of speech audiometry with CI users is a long-established tradition, whereas the measurement of music perception is a relatively novel emergence. As has been the case in HA research, most efforts to assess music perception for CI users have either: A) focussed on subjective, preference-based measures; or B) concentrated on objective measurement of constituent musical features, such as timbre or pitch.

Considering the former case first, some researchers have probed CI users' music perception abilities via measures that are not only subjective, but also survey-based. For example, in an investigation into the benefits of bilateral implantation for music perception, Veekmans, Ressel, Mueller, Vischer, and Brockmeier (2009) compared unilateral and bilateral CI users' responses to the Munich Music Questionnaire, which includes various questions about individuals' enjoyment of music and everyday listening habits. This approach is of course valuable in attaining an overview, but is far from a comprehensive assessment of music perception, and produces fairly broad data, which are unlikely to be useful, for example, in driving refinements made to CI DSP.

A recent example of a more sophisticated, subjective measure for music perception in CI users is the Multiple Stimulus with Hidden Reference and Anchor paradigm (MUSHRA), which was adapted for this population by Roy et al. (2012a). In the CI-MUSHRA, participants are presented with several different versions of a given stimulus, along with an original, unaltered version (i.e. 'Hidden Reference') and a heavily altered version ('Anchor'). Participants are blind to which stimulus is being presented during any given trial, and are asked to indicate their perception

of the sound quality of each musical excerpt, using a rating scale from 0 to 100. The rationale underlying this approach is that participants' responses for any given stimulus version are logically contextualised with respect to the Hidden Reference and Anchor stimuli (Roy et al., 2012a).

A similar approach, designed by Looi, Winter, Anderson, and Sucher (2011), involves the presentation of multiple musical stimuli, to be rated according to six bipolar visual analogue scales, using relevant adjectives such as 'rough' vs. 'smooth', 'natural', vs. 'unnatural', etc. Unlike the CI-MUSHRA paradigm, the emphasis here is not on the comparison of different stimuli, but on overall pre-to-post-test differences in participants' ratings. For example, Looi, Wong, and Loo (2016) used this method to evaluate potential differences in CI users' appreciation of music, following a training-based intervention.

Demonstrably, subjective paradigms like this have valuable applications for the assessment of music perception in CI users. In addition to being relatively quick and straightforward to administer, the methods described above clearly represent an improvement over simple 'paired comparison' measures, in which participants are simply asked to indicate a preference for stimulus A or stimulus B. In particular, subjective paradigms will always be relevant wherever researchers are concerned with CI users' enjoyment or 'appreciation' of music – particularly since these do not necessarily correspond to improvements in objective aspects of music perception. As discussed in the previous chapter however, subjective measures of music perception are not without their drawbacks. Briefly, individuals' preferences are subject to change over time, as a function of myriad factors, and therefore it may be problematic to use such measures as a basis for programming adjustments to the CI. Additionally, there is inevitable inter-individual variance in the criteria used for subjective

judgements of preference (e.g. some degree of sensory dissonance might be appraised as positive by one person and negative by another), and in the interpretation of any rating scales that are used.

In addition to the measures outlined above, several researchers have attempted to measure music perception in CI users via more objective methodologies, most of which have concentrated upon the perception of isolated musical attributes. For example Leal et al. (2003) conducted separate assessments to measure CI users' perception of: timbre – requiring participants to identify the musical instruments used to produce various brief solo instrumental melodies; pitch – requiring participants to make a same-different judgement between a target melody and various transpositions thereof; and rhythm – in which participants again made a same-different judgement, this time between pairs of monotone sequences played with various different rhythms. Several other researchers have utilised a very similar approach to this, essentially decomposing music into a handful of constituent auditory attributes, and examining individuals' sensitivity to these, thereby achieving a measurement of music perception by proxy. For example, as a clinical assessment of music perception by CI users, Nimmons et al. (2008) assessed pitch and timbre discrimination in essentially the same way as outlined above, but chose to also include a melodic, rather than rhythm, discrimination task. Indeed, most studies aimed at objectively assessing the music perception abilities of this population have examined some permutation of the above-mentioned musical features – e.g. Jung et al. (2010) also tested the perception of pitch, timbre and melody, whereas Gfeller and Lansing (1991) examined melody, timbre and rhythm.

However, there are some drawbacks to the broadly-related set of approaches outlined above. Even though there has been substantial research effort focussed on assessing

the auditory components underlying music perception (Cooper et al., 2008), there has been little consensus as to how best these should be measured, and exactly which sub-components are most relevant etc. Additionally, research has prevalently utilised familiar melody recognition tasks, which present various problems via the introduction of a significant memory component. Considering the above, Cooper et al. (2008) has criticised the lack of standardised objective procedures for the assessment of music perception by CI users. Therefore, the authors suggested co-opting the Montreal Battery for Evaluation of Amusia (MBEA) for this purpose. This assessment measures music perception in terms of six constituent components: Scale, Contour, Interval, Meter, Rhythm, and Musical Memory) (Peretz & Hyde, 2003).

In recent years, researchers have increasingly moved towards the use of standardised, objective testing procedures such as these. For example Cullington and Zeng (2011) used the MBEA for assessment of music perception in bilateral and bimodal (i.e. unilateral CI and contralateral HA) CI users. Wright and Uchanski (2012) set out to test whether simulated CI users would perform similarly on these measures, and whether they had any relationship with participants' enjoyment of music. The researchers used not only the MBEA, but also the University of Washington Clinical Assessment of Music Perception, in addition to the subjective Appreciation of Music in Cochlear Implantees measurement. For the objective measures, simulated CI users provided a good approximation of real CI users' performance. For the latter group, however, the correlation between these measures and subjective assessment of music enjoyment was significantly weaker. Therefore, the consensus of Wright and Uchanski (2012) was that enjoyment should not be taken for granted, and that a conclusive assessment of music perception in CI users should include both objective

and subjective components.

9.1.3 An alternative, objective assessment paradigm

Although the use of standardised objective assessments has undoubtedly been an improvement in the assessment of CI users' music perception, such measurements are still limited in that they primarily assess individual musical components (e.g. isolated tasks for the perception of rhythm, melody, etc.). Therefore, these types of tasks lack ecological validity, and thereby risk misrepresenting the real-world performance of CI users.

Accordingly, it might be useful to develop a paradigm that involves presenting participants with more 'complete', polyphonic pieces of music. The principle underlying this approach would be to probe the extent to which users can accurately perceive real music. As already noted, perceptual abilities relating to specific musical components do not always translate to performance with actual music (Bruns, Mürbe, and Hahne, 2016; Wright and Uchanski, 2012). Similarly, individuals' performance on more generic psychoacoustical measurements may be a poor predictor of the adequacy of music perception (Drennan & Rubinstein, 2008).

Considering the above, the assessment procedure developed in Study 7 was posited as a potentially useful tool for the measurement of music perception in CI users. As in the previous chapter, this measure was not intended as a definitive replacement for existing approaches, but rather as a supplement – an addition to a wider 'toolbox' of procedures that together may constitute a more thorough assessment of music perception.

9.1.4 Is this approach viable for CI users?

The principle risk of the proposed approach was that the CI, and therefore also the CI simulation, render it much more difficult to perceive individual instruments within a polyphonic stimulus (Donnelly et al., 2009). In Study 7, participants most likely approached the task by cognitively 'separating' out the individual instruments from the polyphonic mix, on the basis of timbre, and then attending to the relevant 'stream', i.e. the instrument playing the target melody. However, the CI simulation (and indeed the CI itself), complicates this task, since differences in timbre are much less pronounced due to more limited spectral resolution. In Study 7, participants were able to listen to the target melody, and then deliberately attend to the relevant timbre in the polyphonic stimulus, and compare the two melodies. Conversely, with CI simulation, it is likely that participants will listen to the target melody, and then focus on 'listening out' for this melody in the polyphonic stimulus. In the latter case, the difficulty of the task was expected to be drastically increased when the background instruments are the same level as, or louder than, the target melody, since there is very little timbre information available with which to disambiguate the competing musical streams. It should be noted, however, that the different instrumental parts were played in different pitch registers (as was also the case in Study 7), thereby providing another cue by which participants could tell them apart. Nonetheless, overall performance was expected to be much worse for normal-hearing participants listening with CI simulation, compared to HA simulation.

In any case, this procedure was still considered potentially informative in assessing music perception in CI users. Even if performance was worse than with HA users, and/ or was achieved via a different listening strategy, the results may nonetheless provide insight into the abilities of CI users to interpret and make judgements about

'real' music. That is, participants made judgements about recordings of existing polyphonic music, rather than impoverished excerpts contrived for experimentation. Therefore, whatever the approach taken to the task, it may be informative as to the real-world music perceptual abilities of CI users/ CI-simulated listeners.

9.1.5 Adaptations made to the assessment procedure

Before evaluating the assessment procedure described in Study 7 with (simulated) CI users, several refinements were made. Most importantly, for this study, all of the stimuli presented (as described in the previous chapter) were processed with a CI simulation. However, three more general refinements were also made to the procedure.

Based on the results from Study 7, only stimuli in which the target melodies were played on the saxophone were included in this study. Compared to the clarinet stimuli, performance was on average less variable with the saxophone stimuli, and overall performance was significantly better. Naturally, this meant that the experimental task became slightly easier on the whole, although this was not considered problematic, particularly since the CI simulation was expected to substantially increase overall difficulty. Additionally, the use of only one instrument made the assessment procedure shorter, and was expected to reduce some of the variation in participants' responses.

Secondly, since Study 7 identified two musical stimuli that were significantly more difficult than the rest, these stimuli were omitted. This was expected to improve the reliability of the adaptive procedure, by eliminating unwanted variability in task difficulty (i.e. variability unrelated to the manipulation of target-to-background ratio (TBR)).

Lastly, in Study 7, inter-individual variability in estimated thresholds was quite high. Musical expertise and engagement were not formally measured, and might have been able to explain some of this variance. Therefore, in Study 8, the same validated inventory of musical engagement as used in Studies 3 and 4 was administered to participants (Chin & Rickard, 2012).

The next section provides a recap of the assessment procedure outlined in the previous chapter, describing how it was adapted and evaluated for use with CI-simulated listeners.

9.2 Methods

9.2.1 Participants

Ten participants – a combination of undergraduate and postgraduate students at the University of Sheffield – were recruited via opportunity sampling (60% female, mean age = 27.3, SD = 5.6). All participants reported having normal hearing and either normal or corrected-to-normal vision. Prior to participation, each participant provided fully-informed consent. No financial incentive or reimbursement was offered to participants.

Unfortunately, it was not possible to recruit a sample of CI users for this study, due to time constraints associated with the process of obtaining ethical approval via the NHS. However, CI simulation has previously been demonstrated to be a good model for the CI, with respect specifically to music perception (Wright & Uchanski, 2012).

9.2.2 Procedure

The main experimental procedure was almost identical to that outlined in Study 7, with the key exception that all stimuli were processed with a CI simulation, as opposed to an HA simulation. Otherwise, the same musical stimuli were used (phrases excerpted from J. S. Bach chorales, excepting the two stimuli identified as disproportionately difficult in Study 7), and the same MATLAB program was used for stimulus presentation, data collection and threshold estimation. The CI simulation was achieved using the same noise-band-vocoding (NBV) approach described in Studies 1 to 4.

Briefly, the experimental task involved participants listening to an isolated musical melody (i.e. the 'target' melody), then listening to the same melody with accompanying instruments (i.e. the 'background'), in the context of a polyphonic musical mix. During the second presentation, a single note was sometimes (decided at random) omitted from the target melody, and participants were asked to make a 2-AFC judgement about whether the melody was the same or different in each instance. As in Study 7, there were three experimental conditions: 'Individual', in which dynamic range compression was applied separately to the target and background; 'Mix', in which the target and background were first mixed and then compressed together; and 'Linear', in which the target and background were linearly amplified, i.e. with no compression. These conditions was intermixed in presentation, with each consecutive trial containing stimuli from a pseudo-randomly-chosen condition, such that each condition occurred exactly once per three trials (though stimuli from each condition were no longer presented once the threshold associated with that condition had been estimated). For each condition, task difficulty was controlled separately with a 1-up-2-down adaptive procedure (estimating the 70.7% psychometric threshold, as in

Study 7). To adjust task difficulty, the relative intensity level of the target to background in the mixed presentation was varied. Isolated targets were always presented at 65 dB – in the mixed presentation, backgrounds were always presented at 65 dB, and targets were presented either quieter or louder, to make the task respectively more or less difficult.

Despite the difficulty of the note omission task being potentially increased by the use of CI simulation, the initial target-to-background ratio (TBR) was set to 10 dB: the same level used in Study 7. As in Study 7, following each incorrect response, TBR was increased, and after two consecutive correct responses, TBR was decreased. At first, TBR was adjusted in increments of 4 dB, until the first reversal from an increasing to decreasing TBR occurred. At this point, the adjustment increment was reduced to 2 dB. After the second reversal from an increasing to decreasing TBR, it was reduced to 1 dB. Once four reversals (either direction) had occurred at the 1 dB step size, the threshold was calculated as the mean TBR across these reversals. This process lasted approximately thirty-five minutes per target instrument. Unlike Study 7, this was carried out only once, since only one musical instrument (saxophone) was used to play the target melodies.

All participants first completed a practice block of six trials (half as many as Study 7, since only one 'target' musical instrument was used), in order to familiarise them with the procedure. This practice phase was identical to the main experimental procedure, excepting that the TBR remained constant, and participants received feedback ('Correct' or 'Incorrect') immediately after each response.

Since a primary aim of this study was to assess the repeatability of this paradigm for CI-simulated listeners, all ten participants attended a second experimental session. These follow-up sessions used exactly the same procedure as described previously,

and occurred two days after each participant's initial experimental session. To re-familiarise participants with the experimental procedure, and the nature of the CI simulation, the practice phase was also completed a second time, prior to the follow-up experiment proper.

Upon completion of the main test procedure, participants were asked to answer the Music Use questionnaire (Chin & Rickard, 2012), in order to assess whether musical training and/ or experience had any influence on performance in the main experimental procedure.

All participants were tested independently, and all testing took place in a quiet laboratory at the University of Sheffield. Stimuli were presented over Sennheiser HD 600 headphones, connected directly to a 15.6" ASUS N56VM laptop computer.

9.3 Hypotheses

It was hypothesised that the test-retest reliability of the paradigm would be similar to that observed in the previous study. More specifically, it was predicted that the majority of threshold estimates would be within a ±5 dB TBR range across the two experimental sessions, and that the estimates for each session would be strongly positively correlated.

Secondly, as in Study 7, it was hypothesised that participants' threshold estimates would be lower in the Independent amplification condition, compared to the Mix and Linear conditions. As described previously, WDRC applied independently to the target and background essentially rendered the TBR larger than it would otherwise be, meaning that the note omission task should have been easiest in this condition.

However, it was also considered that participants might approach the task differently,

owing to the different characteristics of the task when listening with CI simulation compared to HA simulation. Given that the CI simulation renders the different instrumental voices much less discriminable, it was predicted that participants might effectively perceive the stimulus as one melody, as opposed to four separate melodies. In this case, the benefit of independently-applied compression might be negated somewhat.

Lastly, it was expected that individuals with more extensive musical training might be more accustomed to the kind of 'analytical' listening necessary to detect missing notes in polyphonic music. Therefore, it was hypothesised that these individuals might register lower TBR thresholds. In particular, it was anticipated that more musically-trained individuals might be better able to cognitively 'partition' the polyphonic stream into its constituent instrumental parts, thereby showing a stronger advantage in the Independent condition.

9.4 Results

Similarly to the previous study, threshold estimates were relatively consistent between session one and session two, albeit with a greater trend towards improvement in the latter session, implying a stronger effect of practice than that observed in Study 7. Pearson correlation showed that threshold estimates in the two sessions were strongly positively correlated, even more so than in Study Six, $r(8) = .86$, $p < .001$ (Figure 60). In total, 76.67% of threshold estimates in the second session were within ±5 dB TBR of the corresponding estimates in the first session. There was also a small effect of practice, in that participants' estimated thresholds tended to be lower, on average, in session two compared to session one (overall means = 6.53 and 9.72, respectively).

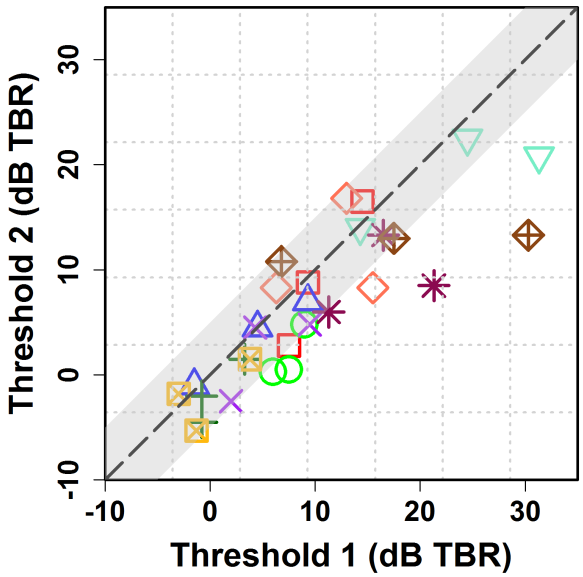


Figure 60: Relationship between participants' scores (indicated by the different symbols) in the two different experimental sessions. Shaded area indicates +/- 5 dB, dotted line indicates +/- 0 dB.

Individuals' thresholds for each amplification condition and experimental session are shown in Figure 61. Inter-individual variability was high, with estimated thresholds varying from -5 dB TBR to >30 dB TBR. Generally, performance was quite similar across the two sessions, with most participants registering a small improvement (i.e. a decrease) in the second session. The Independent condition appeared to result in the lowest thresholds overall, although this was not consistent for all participants.

Considering thresholds averaged over all participants, individuals' thresholds were clearly lowest in the Independent amplification condition, as hypothesised. Thresholds were also slightly lower in the Mix condition relative to Linear, though this difference was much less pronounced (Figure 62).

Corroborating the observed effect of practice, scores were also lower on average for all amplification conditions in session one, compared to session two. However, ANOVA

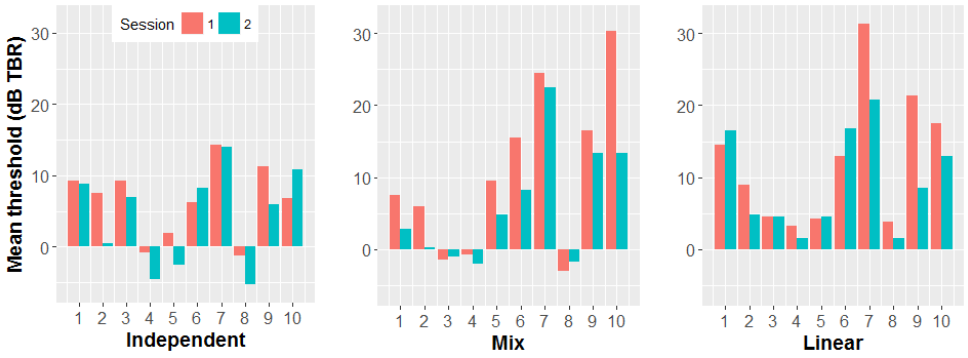


Figure 61: Individuals' estimated thresholds in the three amplification conditions, for sessions one and two. Lower thresholds indicate better task performance. X-axes represent participant ID.

revealed that neither the effects of Session nor Amplification condition, nor the interaction thereof, were significant (Table 28).

Table 28: Outcome of a factorial mixed ANOVA, examining how participants' estimated psychometric thresholds varied according to test session, amplification condition, and the interaction between these variables.

Effect	***F***	**df**	***p***	η^2
Session	0.07	1, 48	.797	<.01
Amplification	1.39	2, 48	.258	.05
Session × Amplification	0.18	2, 48	.839	<.01

Interestingly, despite the overall trend towards thresholds being lower in the Independent condition than the Mix condition (as was observed for HA and simulated HA listeners in the previous study), there were a subset of participants that displayed the opposite pattern. To explore this discrepancy, data from the MUSE questionnaire regarding participants' musical backgrounds were considered. For those participants with lower estimated thresholds in the Independent condition (N = 4), the average Index of Musical Training (IMT) score was 4.17, whereas participants with lower thresholds in the Mix condition had an average IMT score of 8.25. A Wilcoxon-Mann-Whitney test revealed that this difference was not significant, $W = 3$, $p = .067$, $r = .61$.

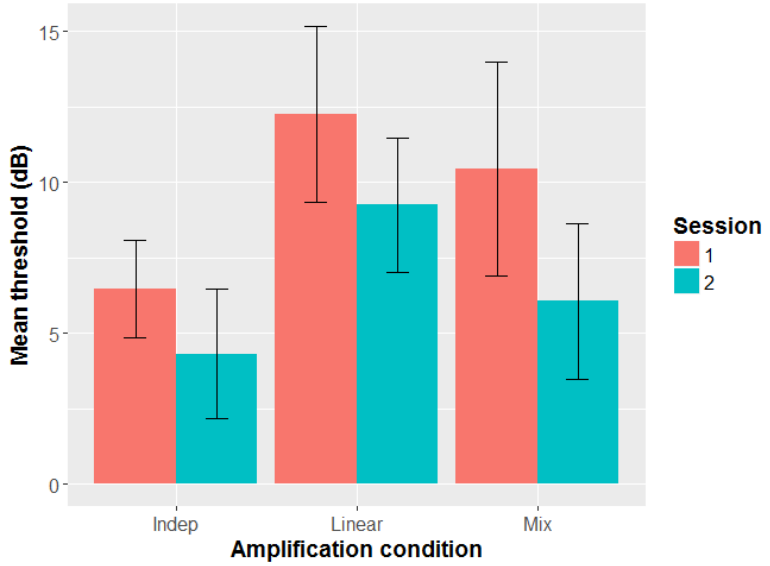


Figure 62: Individuals' estimated thresholds in the three amplification conditions, for sessions one and two. Lower thresholds indicate better task performance.

Lastly, no significant correlation was found between participants' overall thresholds and any of the subcomponents of the MUSE.

9.5 Discussion

9.5.1 Summary of results

Participants demonstrated relatively consistent test results across the two experimental sessions, implying that the paradigm is capable of detecting repeatable psychometric thresholds for CI-simulated listeners. Although a handful of participants demonstrated rather drastic improvement due to practice, the overall correlation between the two sets of threshold estimates was greater than for the previous study with HA users. Additionally, there was not a significant main effect of test session upon participants' thresholds, although the data showed a trend in the predicted direction. Taken together, the findings suggest that the test-retest reliability of this

paradigm is sufficient to warrant further testing with real CI users.

As anticipated, a stronger effect of practice was found for the CI-simulated version of this paradigm, compared to the HA-simulated version from Study 7. This most likely occurred because the stimuli were more unusual for participants to begin with, and therefore the effect of acclimatisation (as observed in Study 4) was more prominent.

Contrary to the hypothesis, the effect of amplification type was not significant, though the general pattern of results was very similar to the previous study, with Independent amplification associated with the lowest estimated thresholds, on average. There are three potential reasons for this lack of statistical significance. Firstly, the overall sample size was smaller in this study compared to Study 7 (N = 10 compared to N = 16, respectively). Secondly, inter-individual (i.e. non-systematic) variability, in terms of overall performance and improvement due to practice, was greater in this study. Thirdly, and perhaps most importantly, the amplification manipulation might not have had quite the same effect, because of its interaction with the CI simulation. Because the CI simulation made it much more difficult to attend selectively to different musical instruments, the effect of independently applying compression to one of the instruments may have been less impactful. That is, CI simulation reduced the spectral resolution of the stimuli, rendering the different instrumental tracks less distinct. Because of this, participants may have perceived the four instrumental melodies as one combined melody, regardless of the relative volume of the different instruments. Therefore, greater statistical power would have been required to detect a smaller net effect of amplification type, but instead this study had fewer participants than Study 7. Nonetheless, the fact that the results obtained were similar to the previous study is encouraging, and points towards the potential of this assessment paradigm.

Although there was an overall trend towards better performance in the Independent amplification condition, four of the ten participants achieved higher scores in the Mix condition. The most straightforward explanation for this is that these participants perceived the four instrumental voices present in the stimulus as just one combined melody. Therefore, either the Independent condition was less helpful, as alluded to above, or the Mix compression reinforced this 'combined melody' percept, in a way that was somehow conducive to detection of the note omission.

Interestingly, this tendency appeared to be associated with extent of musical training, with these four participants achieving higher scores on average for the Index of Musical Training. This pattern was contrary to the hypothesis, which suggested that musicians might be more adept at attending to individual musical instruments, thereby experiencing a relatively larger advantage in the Independent condition. Instead, it appeared that more highly musically-trained individuals were more likely to perform best in the Mix condition.

However, this difference was non-significant, likely because of the very small overall sample size and unequal 'group' sizes. In any case, it appeared that those participants achieving their lowest thresholds in the Mix amplification condition approached the task in a fundamentally different manner, and that this strategy was somehow related to musical training and/ or experience. Assuming this interpretation of the results to be correct, it was not immediately clear why more musical training should lead to a reduced tendency to attend to the constituent instrumental parts of polyphonic stimuli. Given the small sample sizes involved, however, and the lack of statistical significance, this apparent finding was most likely spurious. Nonetheless, this result highlights that, by comparison to the previous study, there was less of a de facto 'optimum' method of listening.

In any case, there was no overall performance advantage associated with musical training and/ or engagement. As in earlier studies involving CI-simulated musical excerpts (Studies 2 and 4), this might be due to the CI simulation being too disruptive.

9.5.2 Limitations

The most important limitation of this experiment was that the paradigm was trialled only with CI-simulated participants, as opposed to CI users. Therefore, without further testing, it is not possible to recommend this procedure for use in clinical settings. Specifically, it is difficult to know to what extent the test-retest reliability observed here would be maintained if CI users were tested. However, prior research, both within this thesis and elsewhere (e.g. Wright and Uchanski, 2012) has indicated that CI-simulated listeners perform similarly to real CI users in music perception tasks.

In fact, this study perhaps underestimated the reliability of the paradigm somewhat. Of the 23.33% of threshold estimates that were more than 5 dB TBR apart across the two sessions, every one of these occurred because of drastic improvement (mean = 9.36 dB TBR) in the second session. This large practice effect is likely due to two factors: unfamiliarity with the test paradigm, and unfamiliarity with the nature of CI simulated stimuli. While the former factor is likely to lead to a practice effect for CI users, the latter is not, since this population should already be accustomed to the sound of music via the CI, to some extent. Therefore, as with Study 7, although the results appear promising, a necessary next step is to expand the study to incorporate a larger sample of hearing-impaired listeners.

As in Study 7, this assessment procedure provides, by nature, only a partial picture

of CI users' music perception. That is, 'accurate' performance within this paradigm offers little insight into participants' aesthetic appreciation or subjective enjoyment of music. Much like the previous study, it was assumed that, at least to some extent, the two would be correlated, although previous research has shown that this is not always the case (Bruns et al., 2016). As with HAs, any comprehensive assessment of CI users' music perception ability must therefore include both objective and subjective elements.

9.5.3 Conclusions

The paradigm developed in Study 7 was applied to CI users, and the initial results were promising, although with a few reservations. It appears to be a reliable way to obtain information about the abilities of CI users to accurately perceive polyphonic music. Most importantly, participants' estimated psychoacoustic thresholds were very similar across different testing sessions, indicating that these scores captured were relatively stable over time.

However, the assessment procedure failed to detect significant differences in participants' thresholds as a function of the amplification manipulation. This did not appear to denote a fundamental flaw in the procedure per se, but instead was most likely resultant from the relatively small sample size studied, in addition to the particular manipulation used being less salient when listening via the CI than the HA. Given that perception of polyphony is particularly problematic for CI users, a more overt manipulation of the stimuli may be necessary in order to reliably influence participants' thresholds. This might be more so a reflection of the current state of music perception with the CI, than the assessment itself. In fact, leaving aside statistical significance, the observed results were very similar to those in Study 7, indicating

that the paradigm was able to elicit the performance differences expected for the different signal processing conditions.

Lastly, as in Study 7, inter-individual variability on this task was quite high. The extent of participants' musical training and engagement had little impact upon the thresholds obtained, therefore it is likely that variability in participants' performance was caused by some other factor(s). Identifying substantial contributors to variability in participants' thresholds would therefore be a sensible goal for future research.

In summary, the assessment procedure trialled here is of potential utility for the assessment of music perception in CI users. However, before the approach can be definitively recommended, more data must be gathered to support its effectiveness. Chiefly, it must be demonstrated that the procedure is sensitive enough to be able to detect predictable differences in thresholds caused by deliberate DSP manipulations. Although the pattern of results implied that this was the case, a larger sample of listeners should be tested in order to provide confirmation. Additionally, the procedure should of course be examined with a sample incorporating real CI users. As in Study 7, the essential aim of this research was to develop a paradigm comparable in spirit to speech audiometry, which may be used to evaluate music perception in CI users, and to evaluate any improvements made, for example due to the adoption of different DSP strategies. In this aim the study was successful, and has provided a good basis for additional investigation and refinement of the testing procedure.

9.6 Consolidating the insights offered by this thesis

The final section of this thesis presents a consolidation of the research conducted in this chapter and each of the preceding empirical chapters. Considering all of the studies carried out thus far, Chapter 10 presents some overarching conclusions which may

be drawn, along with implications for both research and clinical practice. Finally, owing to the tendency of research to produce at least as many questions as it answers, the chapter outlines some potentially fruitful avenues for further exploration.

10 Conclusions, implications and suggestions for future research

10.1 Overview: Asking questions

This thesis was concerned with the perception of music and speech, and the identification of emotions expressed in each, by hearing-impaired listeners. In order to evaluate the success of the thesis, this chapter considers what advances the research carried out here has made with respect to the state of knowledge in these areas.

It is said that 'science begins by asking questions and then seeking answers' (Vale, 2013, p. 680). So, too, did this thesis, and so shall the evaluation thereof. This chapter recaps the most pertinent research questions that were addressed herein. The thesis is evaluated by considering both the answers to these questions, and the novel questions that have emerged via the pursuit of these. Summarily, this chapter asks three overarching questions about the research carried out:

A) What were the main research questions addressed in the thesis?

B) On the evidence gathered, what are the answers to these research questions?

C) What questions are yet to be unanswered, and what new questions have emerged?

By framing the evaluation of the thesis in these terms, the chapter acknowledges that the formulation of good questions is at least as important to the scientific process

as the documentation of 'facts'. More importantly, it situates the thesis, not as an endpoint of a scientific investigation, per se, but simultaneously as various possible beginnings.

10.2 Summary of the research questions addressed

Throughout this thesis, there were several interrelated yet distinct 'threads' of scientific enquiry. These were characterised, broadly, by the following research questions:

1) Is emotion identification in speech and music possible when cochlear implant (CI) users are examined under more challenging test conditions (e.g. with acoustic feature attenuation)?

2) What listening strategies are responsible for CI users' above-chance decoding of emotion in speech and music?

3) To what extent can emotion identification by CI users be improved via training?

4) How effectively can emotion identification in CI users be modelled by normal-hearing (NH) participants listening with simulation?

5) Can the assessment of music perception in hearing-impaired (HI) individuals be improved, by using a more objective methodology?

In the sections to follow, these questions serve as a structure by which the impact of this thesis is evaluated. Specifically, questions A, B and C (from the previous section) are applied in turn to each of the primary research questions, assessing the extents to which they have been satisfactorily answered, and to which complementary lines of investigation have emerged.

10.3 Is emotion identification in speech and music possible when cochlear implant (CI) users are examined under more challenging test conditions?

10.3.1 Overview of the question

In previous studies, researchers have sometimes made assumptions about the emotion identification abilities of CI users, and accordingly have made decisions about experimental design that made this task more easily achievable. For example, Chatterjee et al. (2015) presented child-directed speech (to a sample including adults) with deliberately exaggerated expressive cues. Previous research has also decreased the difficulty of emotion identification by restricting participants' response options (e.g. Volkova et al., 2013).

On the basis of results such as these, it has been concluded that CI users can decode emotion in both speech and music with accuracy exceeding chance performance. This is a hugely important result, but one that leaves open a question about the limits of emotion identification abilities in this population. In particular, 'is above-chance performance still achievable if the experimental task is made more difficult?'.

Researchers have begun to answer this question by utilising more difficult test procedures. For example Luo et al. has documented above-chance performance in emotion identification in both speech and music, using 4- and 5-alternative forced-choice (AFC) paradigms, respectively (Luo et al., 2007; Luo, 2016). In this thesis, Studies 1 through to 6 built upon this approach, using a 5-AFC discrimination task for both speech and music, and additionally including a feature attenuation manipulation for CI-simulated participants.

With respect to music specifically, previous research has focussed upon the perception of emotions conveyed by musical composition, but has not specifically investigated CI users' sensitivity to musicians' expressive intentions. That is, musical emotions conveyed purely via *performance* as opposed to compositional cues (e.g. major/ minor mode). By contrast the stimuli presented here, throughout Studies 1 through to 6, varied only in terms of performance cues.

Taking the above factors into account, the emotion identification task employed in this thesis was, in principle at least, substantially more difficult than those used previously. Therefore, the studies carried out were able to provide valuable insight into CI users' proficiency in emotion identification, under more challenging test conditions.

10.3.2 What is the answer to this question?

In short, the answer to this question is a resounding 'yes'. That is, users *can* decode emotion at above-chance level, even when the difficulty of the identification task is increased, compared to previous studies. The longer, more conservative answer, however, adds the caveat that emotion identification in speech and music by CI users is *possible* under more challenging circumstances, but not necessarily *guaranteed*.

This answer is predicated upon two key observations. Firstly, the degree of variability between individual CI users is enormous. Myriad factors including (but not limited to): age, duration of deafness, experience with the CI, degree of residual hearing and device type are all likely to contribute to individuals' proficiency in emotion identification. Indeed, these factors almost certainly contributed to the variation observed in the studies conducted. Secondly, huge potential variability exists in terms of the emotional stimuli presented. A key example in the case of this thesis is the type of stimuli presented, i.e. speech compared to music – some listeners were

unable to reliably decode emotions expressed by music at above-chance level.

Another answer to this question might be that it depends on the specific emotion being expressed, or more accurately on a complex interaction between emotion, stimulus and listener. As an illustrative example, CI users largely struggled to identify happiness when expressed via speech, but not via music – at the same time, however, some participants performed substantially better than others.

As one would expect, the likelihood of attaining above-chance performance also depends on how difficult the task becomes. When listening with processing to attenuate different acoustic features, CI-simulated listeners were not always capable of accurately identifying emotions. Further to the point made above, rather than one manipulation uniformly disrupting perception, what emerged was a more complicated interaction, such that emotion identification was disrupted differentially for specific combinations of the experimental conditions studied.

Nonetheless, the evidence presented in this thesis is sufficient to conclude, with confidence, that emotion identification in speech and music is indeed possible when cochlear implant (CI) users are examined under more challenging test conditions. More importantly, the results obtained suggested that doing so is a worthwhile approach, in terms of the additional insights afforded.

10.3.3 What is there left to discover?

In principle, the emotion identification task could be made more difficult ad infinitum, although in practice there is probably little insight to be gained from doing so.

With this said, for speech perception in particular, a potentially interesting direction for further research may be to investigate how the listening strategies of CI users

vary with the addition of noise. This listening context typically presents a significant challenge for CI users (Fu and Nogaki, 2005; Nelson, Jin, Carney, and Nelson, 2003), though research in this area has tended to focus on speech intelligibility rather than recognition of emotional expression. Within this context, however, a more realistic approximation of everyday listening might generate more ecologically valid results, and also may have interesting effects upon the salience of the auditory feature cues studied.

Another area for further investigation could be the use of a different *type* of experimental task. For example, if emotions were appraised in reference to dimensional continua, instead of the popularly-used AFC method, the results obtained might be quite different. Of course, adopting this paradigm would prohibit the measurement of decoding proficiency by a metric as simple as percentage 'correct' responses, meaning that establishing what constitutes chance-level performance is slightly more complicated. Nonetheless, the degree of correspondence between CI and NH participants, in ratings along emotional dimensions like arousal and valence etc., could be illuminating. Specifically, the proximity (or lack thereof) of each group's ratings, along each emotional dimension, may reveal more subtle differences in the ways that CI and NH users approach this task, and/ or respond to training. Potentially, such an approach could uncover more fine-grained detail about the nature of emotion identification in each group, which would be an excellent supplement to the results reported already.

10.4 What listening strategies are responsible for CI users' above-chance decoding of emotion in speech and music?

10.4.1 Overview of the question

As discussed above, it has been relatively well-established that CI users are capable of above-chance emotion identification in both speech and music. What is less evident, however, is how exactly this level of performance is achieved. That is to say, the underlying mechanisms responsible are not well understood.

The CI impacts the acoustic signal quite drastically, producing a representation of sound that is impoverished in terms of spectral content (McDermott & McKay, 1994), and to a lesser extent intensity (Javel & Shepherd, 2000) and fine-grained temporal information (Duarte et al., 2016). Considering these factors, there is no reason to assume that emotion identification should necessarily proceed in the same way as for NH listeners. Indeed, several researchers have argued that different listening strategies are utilised by CI users, when decoding emotion via the sense of hearing.

For example, it has been suggested that CI users might compensate particularly for deteriorated frequency information by instead prioritising either temporal or intensity information. To this end, various studies showed, for example, that CI users were more likely to base emotion judgements on tempo rather than musical mode, relative to NH listeners (Caldwell et al., 2015). However, there is also the possibility that residual frequency information might be informative for some emotional stimuli, or that loudness information might play an important role. Thus far, there has been a lack of more comprehensive research into the precise extents that different acoustic features are utilised by CI users.

Accordingly, a key aim of this thesis was to investigate in greater depth the strategies involved for emotion identification in CI users. Obtaining knowledge about the cognitive mechanisms underlying this task is an important endeavour for two main reasons. Firstly, knowing which listening strategies are being used by participants, along with whether or not these strategies are optimal, has immediate consequences for rehabilitation, which would attempt to bring the two into alignment. Secondly, and relatedly, listening strategies indirectly provide information about the efficacy of the CI, and in which areas there may be shortcomings. For example, if listeners are resorting to less-than-optimal acoustic cues because these are better-preserved, then there is clearly room for improvement, either in digital signal processing (DSP) or in the CI itself. In short, understanding the listening strategies employed by CI users is informative about A) how people learn to adapt to impoverished sensory input, and B) the particular scenarios in which input is most severely disrupted by the CI.

10.4.2 What is the answer to this question?

In line with evidence presented previously, the studies carried out here provided some evidence of CI users' attendance to better-preserved auditory cues. However, to claim that CI users *primarily* attended to relatively-preserved features would be reductive and likely inaccurate. In fact, the results obtained with both CI and CI-simulated participants were much more complex.

The listening strategies prevalently used by participants may be best explained by the model posited by Peng et al. (2012). That is, when presented with impoverished sensory input, listeners' strategies will emerge as a compromise between two general motivations: A) attending to features most salient for the expression of emotion in the particular stimulus presented, and B) attending to those features least disrupted

by the sensory perturbation. For the studies carried out here, the relevance of this is that the type of stimulus (music vs. speech), the emotion conveyed, and of course any feature attenuation processing, all had a bearing upon which auditory features were most important for emotional expression. Not only that, but these factors appeared to influence the listening strategies that were adopted by human listeners – both CI users and CI-simulated listeners.

Without training, there appeared to be some tendency for CI-simulated participants to focus on frequency information, although this disappeared with training. In terms of the explanation offered above, frequency cues have typically been considered the most informative feature in emotional speech (Petrushin, 1999), and therefore it is natural that participants should place greater emphasis on this feature during emotion identification. However, with training, listeners learned that this was not the case – at least not to nearly the same degree – with CI-simulated stimuli.

However, even after training, CI-simulated strategies did not correspond perfectly to those of real CI users. The latter group were more likely to utilise timbre or voice quality cues – potentially because these cues are better-preserved by the implant than by the simulation, or alternatively because additional listening experience and/ or acclimatisation with the CI facilitates greater sensitivity to these cues. In other words, users may 'habituate', so to speak, to the sound of the CI, and therefore become more attuned to variation in timbral properties of sound.

The most straightforward answer to the question of which listening strategies are responsible for CI users' above-chance decoding of emotion is that there is not one prevailing strategy or set of strategies. Frequency information does not seem to be actively prioritised, nor does it appear to be systematically ignored. In fact, the different types of features seem to be utilised in approximately equal proportion,

on the whole. There were some small patterns – for example, temporal information appeared to be slightly more prioritised when participants listened to music compared with speech. Without further, more thorough study, however, it is difficult to know how to interpret these types of trends. In any case, the available evidence implies, in principle, that good performance should depend on one's ability to switch between listening strategies, in order to best decode many different emotions, across many different stimuli. Unsurprisingly, and consistent with the existent literature, the studies here did not uncover any 'quick-fix' solution or simple, unequivocally 'optimal' listening strategy.

To this last point, listening strategies were also more difficult to model for CI users, compared to NH listeners, most likely because of increased inter-individual variance in this population. Within the context of listening strategies, the important implication of this is that there may not be one meaningful 'average' strategy. Indeed, it may be important for future research to explore the nature of individual differences in listening strategies, investigating both the causes and the effects of these.

10.4.3 What is there left to discover?

With respect to this question, there are three logical goals that emerge for future research. The first of these is better demarcation of the specific strategies utilised by CI users, and how these might be predictable on the basis of individual differences. The studies reported here found that some participants performed drastically better than others. Elsewhere in the literature too, exceptionally able listeners have been documented – for example, Maarefvand, Marozeau, and Blamey (2013) reported the case study of a CI user with particularly excellent music perception skills. A logical question to ask, with respect to these high-achieving individuals, should be:

'is there something tangible that distinguishes these individuals, in terms of how they approach the emotion identification task? A related question would of course be: to what extent are the differences in listening strategies *responsible* for improved performance, rather than merely symptomatic of some larger underlying factor (e.g. greater residual acoustic hearing). Although the sample studied here was too small to support substantive conclusions, a subset of subjects performed substantially better, and therefore future research should ask: 'Can variation in listening strategies be part of the explanation?' Understanding this is central to a comprehensive understanding of the mechanisms underpinning emotion identification in CI users.

A second goal for future research should be to establish whether the apparently greater utilisation of timbral cues by real CI users arises from a longer period of effective 'training', or whether this instead constitutes a difference between CI and simulation. It was suggested previously that CI users might be more attentive towards subtle differences in timbre because they are more accustomed to the overarching disruptions to timbre caused by the CI. This proposition could be tested empirically, using a longitudinal experiment, by comparing CI users' listening strategies with those of CI-simulated listeners, after varying intervals of training. Such an experiment could provide valuable additional information about the process of rehabilitation, and its ability (or indeed inability) to alter the ways that people listen.

Lastly, an important goal for future research should be more qualitative confirmation of the inferences made here about listening strategies, which were on the basis of computational modelling and the distributions of participants' responses. Supplementary, qualitative research would be useful, potentially, in identifying subtler components of participants' listening strategies, which may not be immediately obvious from inspection of their performance. This might also be informative with

respect to the degree of similarity between CI and CI-simulated participants. Generally speaking, corroborative qualitative data could help to fortify and perhaps qualify the various conclusions made in this chapter.

10.5 To what extent can emotion identification by CI users be improved via training?

10.5.1 Overview of the question

Of course, a snapshot assessment of emotion identification by CI or simulated-CI participants is well and good, but a major component of the cochlear implantation is rehabilitation. This process can have an incredibly potent, positive effect on patients' performance in perceptual tasks, like emotion identification.

Taking account of this, various perceptual learning paradigms have been implemented with CI or CI-simulated listeners, and have reported practice effects for various tasks. The efficacy of this methodology has been demonstrated for tasks as diverse as non-word repetition (Burkholder et al., 2004) and musical instrument recognition (Driscoll et al., 2009). Using either single-session or lengthier, longitudinal designs, studies such as these have attempted to demonstrate the real-world value of deliberate practice for improving perceptual proficiency in CI users. However, prior to this thesis, no study had applied this paradigm to the problem of emotion identification.

Accordingly, several studies were conducted to evaluate the efficacy of such a paradigm for the improvement of emotion identification in both speech and music by CI users and CI-simulated listeners.

10.5.2 What is the answer to this question?

Much like the first question, regarding above-chance performance, the answer to this question is 'the exact extent varies, but emotion identification can definitely be improved'. As before, however, there are also a few caveats.

There are two related but separable effects that training could be expected to have upon emotion task performance in these studies. Firstly, a generic effect of acclimatisation to the sound of the CI simulation, and secondly, more specific effects of practice in terms of learning which (and how) auditory features are involved in the expression of different emotions. Both of these effects were observed within the studies reported in this thesis. In general, training was capable of improving overall emotion identification accuracy, in reducing overall biases (e.g. overestimation of anger) and in reducing the prevalence of previously common confusions, for example between anger and happiness.

However, the extent to which emotion identification was improved by training appeared to depend on both the individual undertaking the training, and the medium via which emotion is expressed (i.e. speech or music). On the basis of the evidence presented in this thesis, one can be confident that the training paradigm was effective for the case of speech perception, however this was not clearly evident for music. Nonetheless, as alluded to as part of the previous question, there were so-called 'star performers' for whom training with musical stimuli appeared to be quite beneficial. Unfortunately, there were also participants for whom the training produced little to no measurable benefit.

At present, it is difficult to know which factors affected participants 'responsiveness' to the training paradigm, and thereby to predict which individuals would benefit

most. Musicality was considered a potential explanatory factor, but the results obtained in this regard were inconclusive. Part of the reason for this is that the sample of participants was small, and characterised by a consistent lack of engagement with music, and therefore the sample lacked the variation sufficient to see a reliable effect.

10.5.3 What is there left to discover?

Importantly, the extent to which CI or CI-simulated participants' emotion identification abilities are improved by training might be affected by the experimental methodology. That is, alterations to the training procedure could improve its effectiveness, especially with musical stimuli. For example, presenting a different and/ or more diverse set of stimuli, or extending the training period to last several weeks could potentially enhance the efficacy of the 'intervention'.

In addition to this, the nature of the training itself might be adjusted for improved impact. More precisely, it has been suggested that providing explicit instruction during the perceptual learning paradigm might be more effective than simply providing feedback about correctness (Driscoll et al., 2009). For the paradigm considered here, this would entail describing to participants the specific correspondences between acoustic features and emotional states – e.g. 'the answer was Anger, which is often communicated by a more dissonant timbre and greater intensity' – as opposed to assuming that these will be inferred by participants. This method might be more successful in helping participants to adjust their listening strategies, if the features highlighted by the explicit feedback were not previously being strongly attended to. In fact, previous research has demonstrated the efficacy of a similar approach applied to the production of emotional expression in music (Juslin, Karlsson, Lindstrm, Friberg, & Schoonderwaldt, 2006).

Both the suggestions outlined above could potentially increase the overall benefits of training upon emotion identification, which in turn may have important implications for rehabilitation with CI users. Therefore, refinement of the training procedure would be a worthwhile pursuit for future research.

To a similar end, it would be sensible for future research to evaluate the extent to which the advantages attained via training are maintained at a follow-up evaluation. Of course, with respect to the implications for rehabilitation, it will be necessary to demonstrate that the benefits of training denote longer-lasting changes in the way that emotion identification is approached, as opposed to temporary improvements confined to the particular task.

10.6 How effectively can emotion identification in CI users be modelled by normal-hearing (NH) participants listening with simulation?

10.6.1 Overview of the question

Various studies have used a noise band vocoding (NBV) approach to simulate the perceptual effects of the CI in studies of both speech and music perception (e.g.Friesen et al., 2001; Giannantonio et al.; 2015. Although more sophisticated options for simulation exist (involving, for example, constructing a model of the human auditory system so that residual acoustic hearing can be accounted for (Stadler & Leijon, 2009)), it has been argued that NBV-based simulation is not only a good model of overall emotion identification performance, but also leads to patterns of responses by NH participants which are quite similar to those of CI users (Chatterjee et al., 2015).

However, it was unclear whether this approach to CI simulation would be as effective during the studies reported here for two reasons, both of which have already been alluded to. Firstly, the experimental paradigm was more difficult than perhaps any previous study, which might have had implications for the effectiveness of the simulation. Secondly, a training component was included, and the ability of the simulation to adequately capture the effects of this upon emotion identification was relatively unknown.

Therefore, this thesis investigated whether similar results would be obtained for CI and CI-simulated participants, whether these groups would be affected similarly by training, and whether – in line with previous research – simulation would lead to comparable distributions of errors and correct responses.

10.6.2 What is the answer to this question?

Essentially, the answer to this question is that emotion identification in CI users can be modelled quite well by using a simulation – broadly speaking, CI and CI-simulated listeners performed similarly well in the experimental tasks, and showed overlapping patterns of errors and correct responses. This observation is largely in agreement with the previous literature, and extends the conclusions made so far by finding that the efficacy of NBV-based simulation was robust to both increased task difficulty and the incorporation of a perceptual learning paradigm. Although the NBV approach provides by no means a perfect simulation of the CI (primarily because it does not account for many longer-term consequences of hearing loss, e.g. reorganisation of the auditory cortex), the results reported here support the suggestion that his technique provides a reasonable model of the CIs main effects.

Therefore, this approach appears to be a valid tool for investigating the underlying

listening strategies of CI users during emotion identification. The importance of this should not be understated – it is of course hugely important to understand how exactly CI users go about making emotional judgements. Understanding of this process has obvious consequences for patient rehabilitation, and may also provide wider insights in relation to the overall performance of the CI. However, recruitment of large samples of CI users is a lengthy, resource-heavy process, owing largely to the scarcity of volunteers. Further, the intensive listening studies inherent in this research can be quite taxing for this population, who are often elderly and already bombarded with all manner of listening tests. Therefore, knowing for which scenarios testing with CI-simulated listeners will or will not suffice is enormously important in allocating resources properly and maximising the efficiency of research effort.

Naturally, there were also several points of departure between simulated and real CI users, although the reasons for these were difficult to ascertain precisely. For example, CI users seemed to be less likely to overestimate anger than simulated listeners. They were also more inclined to utilise voice quality cues when listening to speech, and prioritised slightly different temporal cues for music stimuli. However, it is hard to say for sure whether these subtle differences reflect true discrepancies between the CI and simulation, or merely result from differences between the two sets of participants. In the studies carried out, for example, the NH listeners studied were both younger and much more engaged with musical activities, compared to the CI users. Even discounting demographic differences, there are inescapable dissimilarities between the two populations – most notably that the sound of the CI is a relative novelty for simulated listeners.

To summarise, in a broad sense, NBV-based simulation provides a very good approximation of the effects of the CI, and appears to be of great utility for empirical

research. The few points of divergence between the two groups likely reflect both shortcomings of the simulation, and inherent differences between CI and NH participants, though it is difficult at present to say in what proportion.

10.6.3 What is there left to discover?

It would be interesting to investigate which discrepancies between CI-simulated and real CI listeners arose strictly from relevant differences between these populations, as opposed to differences in factors like age and musicality. Although certain, strong inter-group differences, e.g. in engagement with music, are known to exist, it would nonetheless be useful to know whether better subject matching of NH and CI volunteers could lead to more congruous listening strategies.

Another potentially interesting direction for future research, albeit one slightly outside the scope of this thesis, would be to extend the simulation to encapsulate other elements of CI users' experience, and thereby provide a more comprehensive simulation of the effects of the CI. One point of discrepancy between the simulation and the CI itself, is that modern implants tend to have extensive and adaptable DSP, designed to enhance user experience in various ways. For example, most recently-developed implants have a dedicated program with DSP to facilitate the perception of music, which is not taken into account by simple NBV-based simulation. If this music-specific processing could be simulated, then the overall simulation should more closely mirror the real experience of CI users. More generally, extending the simulation to include these elements would provide a valuable means for NH volunteers to be utilised in provisional assessment of updates to CI DSP.

10.7 Can the assessment of music perception in hearing-impaired (HI) individuals be improved, by using a more objective methodology?

10.7.1 Overview of the question

In all of the studies reported so far, participants found the listening tasks more difficult with music than with speech. Since the inception of hearing prostheses this has been the case, largely because of greater research effort directed at speech perception, owing to its greater functional relevance in everyday life. More recently, however, manufacturers of both hearing aids (HAs) and CIs have begun to incorporate dedicated programs with DSP tailored specifically for music perception. This has already led to some improvement in terms of user satisfaction for music listening, but there is a great deal more work to be done before 'perfect hearing' is achieved (Limb, 2016).

Compared to the advances made in music DSP, behavioural testing has lagged behind somewhat, and music perception in hearing-impaired listeners is still primarily assessed using subjective methods. For example, various studies have utilised a paired comparison paradigm, in which participants compare variously-processed pieces of music, such that researchers may establish an overall hierarchy of listener preference (e.g. Croghan et al., 2014). However, listener preferences may be unstable over time, are inherently idiosyncratic, and the criteria responsible for their formation can be difficult to ascertain.

By contrast, speech has long been assessed by objective, psychophysical methodologies. For example, the Hearing in Noise Test (Nilsson et al., 1994) quantifies the exact signal to noise ratio at which presented speech is intelligible to listeners. Measures such as this are effective for both patient assessment, and for quantifying

the benefits of different DSP algorithms. The objective nature of these paradigms facilitates better comparisons, both across individuals, across devices.

A natural question to ask, therefore, is 'can this approach be adapted for the assessment of music perception?'. More specifically, the studies reported in this thesis aimed to first develop such a paradigm, and then evaluate its reliability and validity as a method of assessment.

10.7.2 What is the answer to this question?

On the basis of the studies conducted here, the answer is 'yes' – that is, objective assessment of music perception, following a similar protocol to that used in speech testing, has great potential. A novel assessment procedure based on detection of omitted notes in polyphonic musical stimuli was developed and separately piloted with groups of HA users and simulated CI listeners. For the most part, the results obtained were highly encouraging.

With HA users, the assessment paradigm was found to be both reliable and valid – it led to similar thresholds when participants were retested at a later interval, and the thresholds obtained demonstrated sensitivity to a DSP manipulation designed to elicit a difference in task performance. However, re-test reliability was only confirmed using NH participants (listening with HA simulation). Therefore, the only obstacle remaining is to demonstrate the reliability of the procedure with a larger sample of HA users. Even considering the small sample recruited, however, the procedure was successful in generating an objective and meaningful measurement of music perception. More precisely, the thresholds estimated by the adaptive procedure used seemed to be a good index of individuals' ability to accurately perceive each of the instrumental tracks in a polyphonic musical piece.

When considering the merit of applying this approach to CI users, the chief concern was that the CI is far less effective than the HA in clearly conveying polyphony (Penninger et al., 2013). Nonetheless, for the sample of NH, CI-simulated listeners tested, very good re-test reliability was demonstrated. That is, despite the increased task difficulty associated with the CI simulation, subjects achieved very similar scores when tested on two different occasions. However, the dynamic range compression manipulation – designed to elicit a difference in participants' thresholds – did not have a significant effect. Therefore, the validity of the task for CI users was not confirmed. The most likely explanation for this was that the particular compression manipulation was less salient in the context of CI simulation. In other words, increased difficulty in perceiving polyphony meant that a manipulation of the relative audibility of different musical elements was less impactful. In future, for use with CI patients, the task could potentially be adapted to assess other aspects of music perception that do not rely as heavily on polyphony.

Nonetheless, both studies clearly highlighted that the development of an objective, psychophysical measure of music perception is a worthwhile endeavour. In all, the results were promising, although they need not necessarily represent an endpoint. In fact, returning to the main question proposed, what the studies clearly demonstrated was that objective assessment of music perception in HI listeners was relatively easily implementable, and mostly produced the results expected. It will be down to further research to demonstrate that this constitutes an *improvement* in the way that music perception is assessed.

10.7.3 What is there left to discover?

In response to these findings, there are three main areas of enquiry that should be addressed by future research. Firstly, it will be important to gauge the reliability of the task using a reasonably large sample of HA users. Even though sufficient reliability was found for HA-simulated listeners, there is always the possibility of greater variability in real HA users. Therefore, a convincing demonstration of re-test reliability with this population is required before a paradigm like this could be considered for clinical applications. The same is true for CI users except, as mentioned above, that a slightly different task or auditory processing manipulation might have to be developed, in order to demonstrate the validity of the task.

It is also important to examine whether the measure developed is sensitive enough to detect real-world differences, for example between different DSP strategies for music. The pilot testing used a rather contrived manipulation, intended specifically to affect task difficulty. However, it would represent much greater practical utility, if the assessment were able to detect subtle, relevant differences between different devices and/ or DSP. As an example, future research could investigate the value of this assessment procedure in establishing the benefits of HA and CI manufacturers' specific 'music programs'.

A final, important question to be answered is: how well do the perceptual thresholds estimated by this procedure correspond to subjective enjoyment of music? Since aesthetic enjoyment is not always positively correlated with perceptual fidelity (Limb et al., 2010), further research should clarify what additional value is gained via objective testing of music perception. By comparing results using existing subjective measures against the procedure developed here, a much clearer picture should emerge, of the potential benefits of the latter.

10.8 Concluding remarks

In all, this thesis was successful in both providing answers to the stated research questions, and also in stimulating the formation of fruitful questions for further research. As a body of work, the thesis has produced at least five important, overarching insights, corresponding to the questions addressed above. However, consideration should also be given to what this body of work offers when considered as a cohesive unit.

In this respect, the thesis succeeded in two major areas. Chiefly, it has drawn attention to novel topics of enquiry, and helped to re-frame and re-focus existing ones. With respect to the study of emotion identification, greater emphasis was placed on the underlying cognitive and perceptual mechanisms responsible, and important steps were made in improving understanding thereof. Consideration was also given to the expressive cues present in musical performance, whereas much of the previous literature has focussed on musical composition. With respect to music perception more generally, for HI individuals, the thesis emphasised that different methods of behavioural assessment might be required, and accordingly proposed a novel, objective procedure.

Secondly, and equally importantly, the thesis has demonstrated that the use of innovative or lesser-used analytical and experimental approaches may be able to add important insights to what is known in the areas studied. Estimating listening strategies by training computational models to emulate human response patterns in auditory emotion discrimination, was an approach taken here that does not seem to have been utilised elsewhere in the literature. This is of course a rather 'indirect' way of accessing this information, yet the results obtained suggested that real value was added via this approach. Both the feature attenuation processing (in combination

with perceptual learning), and the various modelling techniques explored, provided an important basis for inference about the importance of different auditory features for emotion identification. It would be an additional success of this thesis, if the work carried out were to influence not only the topics of future research effort, but also the manner in which such studies are conducted.

Discounting the various methodological caveats that have already been discussed, there is one primary shortcoming of this thesis. In fact, it may be better phrased not as a weakness, but as an important goal of the wider research area, that has not yet been fully realised. The insights gained from this thesis – via additional exploration already delineated in this chapter – should be translated into concrete advice that can reasonably be disseminated in order to improve the lives of those suffering from hearing loss. The thesis has uncovered (or in some cases begun a process of uncovering) a variety of results with potentially impactful consequences for this population, and therefore it would be heartening to see these meaningfully integrated within clinical environments.

As Charles Limb has repeatedly argued, the end-goal of hearing loss restoration must be to provide patients with perfect hearing, rather than a merely 'adequate' substitute (Limb, 2016). Achieving this ambitious goal will require that any advances made in academic research are explored thoroughly, and are thereby translated to concomitant steps forward for clinical practice.

References

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, *1*(1), 21–62.

Adorno, T. W. (1992). *Quasi una fantasia: Essays on modern music (translated by Rodney Livingstone).* London, UK: Verso.

Agresti, A. (2007). *An introduction to categorical data analysis.* New York City, NY: Wiley.

Ahmed, D. G. (2017). *Processing of musical and vocal emotions through cochlear implants* (Unpublished master's thesis). McGill University.

Ahtisaari, M., & Karanam, K. (2015). *Music and emotion.* Retrieved 2017-06-28, from `http://syncproject.co/blog/2015/7/21/music-and-emotion`

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*(1), 3.

Ambert-Dahan, E., Giraud, A.-L., Sterkers, O., & Samson, S. (2015). Judgment of musical emotions after cochlear implantation in adults with progressive deafness. *Frontiers in Psychology*, *6*(181).

Anders H. Jessen, H. P. A., Lars Baekgaard. (2014). *What is good hearing aid sound quality, and does it really matter?* Retrieved 2017-08-28, from `http://www.audiologyonline.com/articles/what-good-hearing-aid-sound-12340`

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, *33*(4), 245–266.

Andersson, C., Campbell, D., Farquharson, A., Furner, S., Gill, J., Jackson, A., ... Whybray, M. (2006). *Assistive technology for the hearing-impaired, deaf and deafblind.* Berlin, Germany: Springer Science & Business Media.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms* (pp. 1027–1035).

ASA. (1960). *American standard acoustical terminology.* New York City, NY: American Standards Association.

ASHA. (2003). Cochlear implants: Working group on cochlear implants. *ASHA Technical Report*.

Bach, J. (2012). A framework for emergent emotions, based on motivation and cognitive modulators. *International Journal of Synthetic Emotions (IJSE)*, *3*(1), 43–63.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, *8*(2), 53–57.

Balkwill, L.-L., Thomspon, W. F., & Matsunaga, R. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. *Japanese Psychological Research*, *46*(4), 337–349.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, *12*(4), 579–599.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*(1), 28–58.

Barrett, L. F. (2009). Variety is the spice of life: A psychological construction approach to understanding variability in emotion. *Cognition and Emotion*,

23(7), 1284–1306.

Barrett, L. F., Gendron, M., & Huang, Y.-M. (2009). Do discrete emotions exist? *Philosophical Psychology*, *22*(4), 427–437.

BCIG. (2016). *Quality standards: Cochlear implant services for children and adults.* Retrieved 2017-05-10, from `http://www.bcig.org.uk/wp-content/uploads/2016/05/BCIG-Quality-Standard-2016.pdf`

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2007). *Neuroscience* (Vol. 2). Philadelphia, PA: Lippincott Williams & Wilkins.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Knowledge discovery and data mining workshop* (Vol. 10, pp. 359–370).

Bess, F. H., & Humes, L. (2008). *Audiology: The fundamentals.* Philadelphia, PA: Lippincott Williams & Wilkins,.

Beurg, M., Fettiplace, R., Nam, J.-H., & Ricci, A. J. (2009). Localization of inner hair cell mechanotransducer channels using high-speed calcium imaging. *Nature Neuroscience*, *12*(5), 553–558.

Bhatara, A., Laukka, P., & Levitin, D. J. (2014). Expression of emotion in music and vocal communication: Introduction to the research topic. *Frontiers in Psychology*, *5*(399).

Bhatara, A., Tirovolas, A. K., Duan, L. M., Levy, B., & Levitin, D. J. (2011). Perception of emotional expression in musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 921–934.

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion*, *19*(8), 1113–1139.

Bilger, R. C., Black, F. O., & Hopkinson, N. (1977). Research plan for evaluating subjects presently fitted with implanted auditory prostheses. *The Annals of Otology, Rhinology & Laryngology. Supplement*, *86*(3), 21–24.

Blair, R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, *14*(4), 698–718.

Blamey, P., Arndt, P., Bergeron, F., Bredberg, G., Brimacombe, J., Facer, G., ... others (1996). Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants. *Audiology and Neurotology*, *1*(5), 293–306.

Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (version 5.1. 05) [computer software]. amsterdam, the netherlands: Institute of phonetic sciences.* Retrieved from `http://www.praat.org/`

Boone, D. R., & Plante, E. (1993). *Human communication and its disorders*. Englewood Cliffs, NJ: Prentice Hall.

Bouhuys, A. L., Bloem, G. M., & Groothuis, T. G. (1995). Induction of depressed and elated mood by music influences the perception of facial emotional expressions in healthy subjects. *Journal of Affective Disorders*, *33*(4), 215–226.

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breitenstein, C., Lancker, D. V., & Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, *15*(1), 57-79.

Bricker, S. (2016). *Music the way it was meant to be heard.* Retrieved 2017-06-23, from `http://www.starkey.com/blog/2016/03/streaming-music-with-hearing-aids`

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software*, *25*(4), 1–22.

Brockmeyer, A. M., & Potts, L. G. (2011). Evaluation of different signal processing options in unilateral and bilateral cochlear freedom implant recipients using r-spacetm background noise. *Journal of the American Academy of Audiology*, *22*(2), 65-80.

Broekens, J. (2012). In defense of dominance: Pad usage in computational representations of affect. *International Journal of Synthetic Emotions (IJSE)*, *3*(1), 33–42.

Brown, S. (2000). The "musilanguage" model of music evolution. In B. M. N. L. Wallin & S. Brown (Eds.), *The origins of music* (pp. 271–300). Cambridge, MA: MIT Press.

Bruns, L., Mürbe, D., & Hahne, A. (2016). Understanding music with cochlear implants. *Scientific reports*, *6*, 32026.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Interspeech* (Vol. 5, pp. 1517–520).

Burkholder, R., Pisoni, D., & Svirsky, M. (2004). Perceptual learning and nonword repetition using a cochlear implant simulation. *International Congress Series*, *1273*, 208–211.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Busso, C., & Narayanan, S. S. (2008). The expression and perception of emotions: comparing assessments of self versus others. In *Interspeech* (pp. 257–260).

Butler, R. A., & Flannery, R. (1980). The spatial attributes of stimulus frequency and their role in monaural localization of sound in the horizontal plane. *Perception & Psychophysics*, *28*(5), 449–457.

Buyens, W., van Dijk, B., Moonen, M., & 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., & Moonen, M. (2015, Oct). A stereo music preprocessing scheme for cochlear implant users. *IEEE Transactions on Biomedical Engineering*, *62*(10), 2434–2442.

Byrd, S., Shuman, A. G., Kileny, S., & Kileny, P. R. (2011). The right not to hear: The ethics of parental refusal of hearing rehabilitation. *The Laryngoscope*, *121*(8), 1800–1804.

Cabanac, M. (2002). What is emotion? *Behavioural Processes*, *60*(2), 69–83.

Calcagno, V., & de Mazancourt, C. (2010). glmulti: An R package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software*, *34*(1), 1–29.

Caldwell, M., Rankin, S. K., Jiradejvong, P., Carver, C., & Limb, C. J. (2015). Cochlear implant users rely on tempo rather than on pitch information during perception of musical emotion. *Cochlear Implants International*, *16*(3), 114–120.

Cannam, C., Landone, C., & Sandler, M. (2010, Oct). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the acm multimedia 2010 international conference* (pp. 1467–1468). Firenze, Italy.

Carhart, R. (1946). Selection of hearing aids. *Archives of Otolaryngology*, *44*, 1–18.

Carney, A. E., & Nelson, D. A. (1983). An analysis of psychophysical tuning curves in normal and pathological ears. *The Journal of the Acoustical Society of America*, *73*(1), 268–278.

Cawley, G. C. (2006). Leave-one-out cross-validation based model selection criteria

for weighted ls-svms. In *The 2006 ieee international joint conference on neural network proceedings* (pp. 1661–1668).

Chasin, M. (2003a). *Five differences between speech and music for hearing aids.* Retrieved 2017-05-18, from `http://www.audiologyonline.com/articles/five-differences-between-speech-and-1116`

Chasin, M. (2003b). Music and hearing aids. *The Hearing Journal*, *56*(7), 36–38.

Chasin, M. (2014). *Programming hearing aids for listening and playing music, presented in partnership with the association of adult musicians with hearing loss (aamhl).* Retrieved 2017-06-12, from `http://www.audiologyonline.com/articles/programming-hearing-aids-for-listening-12915`

Chasin, M., & Russo, F. A. (2004). Hearing aids and music. *Trends in Amplification*, *8*(2), 35–47.

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., ... Christensen, J. A. (2015). Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Hearing Research*, *322*, 151–162.

Chernykh, V., Sterling, G., & Prihodko, P. (2017). Emotion recognition from speech with recurrent neural networks. *arXiv (preprint)*.

Chin, S. B., Bergeson, T. R., & Phan, J. (2012). Speech intelligibility and prosody production in children with cochlear implants. *Journal of Communication Disorders*, *45*(5), 355–366.

Chin, T.-C., & Rickard, N. S. (2012). The music USE (MUSE) questionnaire: An instrument to measure engagement in music. *Music Perception: An Interdisciplinary Journal*, *29*(4), 429–446.

Chittka, L., & Brockmann, A. (2005). Perception space – The final frontier. *PLoS Biology*, *3*(4), e137.

Clark, G. (2003). *Cochlear implants: Fundamentals and applications.* New York City, NY: Springer.

Clark, J. G. (1981). Uses and abuses of hearing loss classification. *Asha*, *23*(7), 493–500.

Cooper, W. B., Tobey, E., & Loizou, P. C. (2008). Music perception by cochlear implant and normal hearing listeners as measured by the Montreal Battery for Evaluation of Amusia. *Ear and Hearing*, *29*(4), 618-626.

Corretge, R. (2012). *Praat vocal toolkit: A praat plugin with automated scripts for voice processing.* Retrieved 2016-09-05, from `http://www.praatvocaltoolkit.com`

Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(6), 864–881.

Cousineau, M., Demany, L., Meyer, B., & Pressnitzer, D. (2010). What breaks a melody: Perceiving {F0} and intensity sequences with a cochlear implant. *Hearing Research*, *269*(1-2), 34–41.

Cousineau, M., Demany, L., & Pressnitzer, D. (2009). What makes a melody: The perceptual singularity of pitch sequences. *The Journal of the Acoustical Society of America*, *126*(6), 3179–3187.

Coutinho, E., & Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. *Cognition and Emotion*, *27*(4), 658–684.

Cox, R. M., & Alexander, G. C. (1995). The abbreviated profile of hearing aid benefit. *Ear and Hearing*, *16*(2), 176–186.

Cray, J. W., Allen, R. L., Stuart, A., Hudson, S., Layman, E., & Givens, G. D. (2004). An investigation of telephone use among cochlear implant recipients. *American Journal of Audiology*, *13*(2), 200–212.

Croghan, N. B. H., Arehart, K. H., & Kates, J. M. (2014). Music preferences with hearing aids: Effects of signal properties, compression settings, and listener characteristics. *Ear and Hearing*, *35*(5), 170-184.

Cullington, H. E., & Zeng, F.-G. (2011). Comparison of bimodal and bilateral cochlear implant users on speech recognition with competing talker, music perception, affective prosody discrimination and talker identification. *Ear and Hearing*, *32*(1), 16.

Curtis, M. E., & Bharucha, J. J. (2010). The minor third communicates sadness in speech, mirroring its use in music. *Emotion*, *10*(3), 335–348.

Dalla Bella, S., Berkowska, M., & Sowiński, J. (2011). Disorders of pitch production in tone deafness. *Frontiers in Psychology*, *2*(164).

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York City, NY: G. P. Putnam's Sons.

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness.* Boston, MA: Houghton Mifflin Harcourt.

Darwin, C. (1871). *The descent of man and selection in relation to sex.* London, UK: Murray.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222.

Davis, W. (2004). *What is frequency compression?* Retrieved 2017-09-02, from `https://www.audiologyonline.com/ask-the-experts/what-is-frequency-compression-543`

Davitz, J. R. (1969). *The language of emotion.* San Diego, CA: Academic Press.

de Cheveigne, A., & Kawahara, H. (2002). Yin, a fundamental frequency estimator

for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

Dellaert, F., Polzin, T., & Waibel, A. (1996, Oct). Recognizing emotion in speech. In *Proceedings of the fourth international conference on spoken language processing (icslp)* (Vol. 3, pp. 1970–1973).

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*. Upper Saddle River, NJ: Prentice Hall.

Digeser, F. M., Hast, A., Wesarg, T., Hessel, H., & Hoppe, U. (2012). Melody identification for cochlear implant users and normal hearers using expanded pitch contours. *European Archives of Oto-Rhino-Laryngology*, *269*(11), 2317–2326.

Donnelly, P. J., Guo, B. Z., & Limb, C. J. (2009). Perceptual fusion of polyphonic pitch in cochlear implant users. *The Journal of the Acoustical Society of America*, *126*(5), 128–133.

Drennan, W. R., Oleson, J. J., Gfeller, K., Crosson, J., Driscoll, V. D., Won, J. H., ... Rubinstein, J. T. (2015). Clinical evaluation of music perception, appraisal and experience in cochlear implant users. *International Journal of Audiology*, *54*(2), 114–123.

Drennan, W. R., & Rubinstein, J. T. (2008). Music perception in cochlear implant users and its relationship with psychophysical capabilities. *Journal of Rehabilitation Research and Development*, *45*(5), 779–789.

Driscoll, V. D., Oleson, J., Jiang, D., & Gfeller, K. (2009). Effects of training on recognition of musical instruments presented through cochlear implant simula-

tions. *Journal of the American Academy of Audiology*, *20*(1), 71–82.

Duan, Z., & Pardo, B. (2011, Oct). Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1205–1215.

Duarte, M., Gresele, A. D. P., & Pinheiro, M. M. C. (2016). Temporal processing in postlingual adult users of cochlear implant. *Brazilian Journal of Otorhinolaryngology*, *82*(3), 304–309.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria* (Tech. Rep.). The Methodology Center.

Eargle, J. M. (2012). *Handbook of recording engineering*. Berlin, Germany: Springer Science & Business Media.

Eickers, G., Loaiza, J. R., & Prinz, J. (2017). Embodiment, context-sensitivity, and discrete emotions: A response to moors. *Psychological Inquiry*, *28*(1), 31–38.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169–200.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129.

Elizondo, D. (2006). The linear separability problem: Some testing methods. *IEEE Transactions on Neural Networks*, *17*(2), 330–344.

Elzouki, A. Y., Stapleton, F. B., Harfi, H., Oh, W., Whitley, R. J., & Nazer, H. (2011). *Textbook of clinical pediatrics*. Berlin, Germany: Springer Science & Business Media.

Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4),

1491–1506.

Fearn, R., Carter, P., & Wolfe, J. (1999). The perception of pitch by users of cochlear implants: Possible significance for rate and place theories of pitch. *Acoustics Australia*, *27*(2-41).

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition*, *37*(1), 1–9.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Development*, *64*(3), 657–674.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r*. Thousand Oaks, CA: Sage.

Fitz, K., Burk, M., & McKinney, M. (2009). Multidimensional perceptual scaling of musical timbre by hearing-impaired listeners. In *Proceedings of meetings on acoustics 157asa* (Vol. 6).

Fitz, K., & McKinney, M. F. (2010). Music through hearing aids: Perception and modeling. *The Journal of the Acoustical Society of America*, *127*(3), 1951–1951.

Fletcher, H. (1934). Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *Journal of the Acoustical Society of America*, *6*, 59–69.

Florentine, M., Buus, S., Scharf, B., & Zwicker, E. (1980). Frequency selectivity in normally-hearing and hearing-impaired observers. *Journal of Speech, Language, and Hearing Research*, *23*(3), 646–669.

Fox, E. (2008). *Emotion science cognitive and neuroscientific approaches to understanding human emotions*. London, UK: Palgrave Macmillan.

Fox, S. I. (1996). *Human physiology (9th edition)*. Boston, MA: McGraw-Hill Higher

Education.

Fredelake, S., & Hohmann, V. (2012). Factors affecting predicted speech intelligibility with cochlear implants in an auditory model for electrical stimulation. *Hearing Research*, *287*(1-2), 76–90.

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, *110*(2), 1150–1163.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., ... Koelsch, S. (2009). Universal recognition of three basic emotions in music. *Current Biology*, *19*(7), 573–576.

Fu, Q.-J., & Galvin, J. J. (2008). Maximizing cochlear implant patients' performance with advanced speech training procedures. *Hearing Research*, *242*(1-2), 198–208.

Fu, Q.-J., & 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.

Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, *5*(1), 123–147.

Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music*, *24*(1), 68–91.

Gabrielsson, A., & Lindström, E. (2010). The role of structure in the musical expression of emotions. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 367–400). New York City, NY: Oxford University Press.

Galushkin, A. I. (2007). *Neural networks theory.* Berlin, Germany: Springer Science & Business Media.

Galvin, J. J., Fu, Q.-J., & Nogaki, G. (2007). Melodic contour identification by cochlear implant listeners. *Ear and Hearing*, *28*(3), 302–319.

Gantz, B. J., Woodworth, G. G., Knutson, J. F., Abbas, P. J., & Tyler, R. S. (1993). Multivariate predictors of audiological success with multichannel cochlear implants. *Annals of Otology, Rhinology & Laryngology*, *102*(12), 909–916.

Geers, A. E. (1986). Vibrotactile stimulation: Case study with a profoundly deaf child. *Journal of Rehabilitation Research and Development*, *23*(1), 111–117.

Geier, L., Barker, M., Fisher, L., & Opie, J. (1999). The effect of long-term deafness on speech recognition in postlingually deafened adult Clarion cochlear implant users. *Annals of Otology, Rhinology & Laryngology*, *1147*(1), 114.

George, J. M. (1996). Trait and state affect. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 145–171). San Francisco, CA: Jossey-Bass.

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

Gfeller, K., Knutson, J. F., Woodworth, G., Witt, S., & 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., & 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., Mehr, M., & Witt, S. (2001). Aural rehabilitation of music perception

and enjoyment of adult cochlear implant users. *Journal of the Academy of Rehabilitative Audiology*, *34*, 17–28.

Gfeller, K., Olszewski, C., Rychener, M., Sena, K., Knutson, J. F., Witt, S., & Macpherson, B. (2005). Recognition of "real-world" musical excerpts by cochlear implant recipients and normal-hearing adults. *Ear and Hearing*, *26*(3), 237–250.

Gfeller, K., Witt, S., Adamek, M., Mehr, M., Rogers, J., Stordahl, J., & Ringgenberg, S. (2002). 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.

Giannantonio, S., Polonenko, M. J., Papsin, B. C., Paludetti, G., & Gordon, K. A. (2015, 08). Experience changes how emotion in music is judged: Evidence from children listening with bilateral cochlear implants, bimodal devices, and normal hearing. *PLoS ONE*, *10*(8), e0136685.

Gibson, J. J. (1966). *The senses conisdered as perceptual systems.* Boston, MA: Houghton-Mifflin.

Gifford, R. H. (2011). Who is a cochlear implant candidate? *The Hearing Journal*, *64*(6), 16–18.

Gifford, R. H., Shallop, J. K., & Peterson, A. M. (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiology and Neurotology*, *13*(3), 193–205.

Glasberg, B. R., & Moore, B. C. J. (1986). Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *The Journal of the Acoustical Society of America*, *79*(4), 1020–1033.

Glasberg, B. R., & Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, *50*(5), 331–342.

Good, A., Gordon, K. A., Papsin, B. C., Nespoli, G., Hopyan, T., Peretz, I., & Russo, F. A. (2017). Benefits of music training for perception of emotional speech prosody in deaf children with cochlear implants. *Ear and Hearing*, *38*(4), 455–464.

Goy, H., Pichora-Fuller, M. K., Singh, G., & Russo, F. A. (2016). Perception of emotional speech by listeners with hearing aids. *Canadian Acoustics*, *44*(3).

Graham, M. C., Priddy, L., & Graham, S. (2014). *Facts of life: Ten issues of contentment.* Denver, CO: Outskirts Press.

Greasley, A. (2016). *Effects of advanced hearing aid settings on music perception.* Retrieved 2017-06-12, from `http://musicandhearingaids.org/effects-of-advanced-hearing-aid-settings-on-music-perception/`

Green, T., Faulkner, A., Rosen, S., & Macherey, O. (2005). Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification. *The Journal of the Acoustical Society of America*, *118*(1), 375–385.

Greenberg, S., & Kingsbury, B. E. (1997). The modulation spectrogram: In pursuit of an invariant representation of speech. In *Proceedings of the ieee international conference on acoustics, speech, and signal processing* (Vol. 3, pp. 1647–1650).

Greeno, J. G. (1994). Gibson's affordances. *Psychological Review*, *101*(2), 336–342.

Gregory, R. L. (1970). *The intelligent eye.* Washington, DC: Education Resources Information Center.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories.* Chicago, IL: University of Chicago Press.

Griffiths, P. E. (2004). Emotions as natural and normative kinds. *Philosophy of Science*, *71*(5), 901–911.

Gulya, A. J., Minor, L. B., Glasscock, M. E., & Poe, D. (2010). *Glasscock-shambaugh*

surgery of the ear. Shelton, CT: People's Medical Publishing House.

Gunes, H. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, *1*(1), 68–99.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hackney, C. M., & Furness, D. N. (2013). The composition and role of cross links in mechanoelectrical transduction in vertebrate sensory hair cells. *Journal of Cell Science*, *126*(8), 1721–1731.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, *16*(9), 458–466.

Handel, S. (1993). *Listening: An introduction to the perception of auditory events.* Boston, MA: MIT Press.

Hanley, B., & Goolsby, T. W. (2002). *Musical understanding: Perspectives in theory and practice*. Ontario, Canada: Canadian Music Educators' Association.

Helms, J., Müller, J., Schön, F., Winkler, F., Moser, L., Shehata-Dieler, W., ... others (2001). Comparison of the TEMPO+ ear-level speech processor and the CIS PRO+ body-worn processor in adult MED-EL cochlear implant users. *ORL: Journal for Oto-Rhino-Laryngology, Head and Neck Surgery*, *63*(1), 31–40.

Heng, J., Cantarero, G., Elhilali, M., & Limb, C. J. (2011). Impaired perception of temporal fine structure and musical timbre in cochlear implant users. *Hearing Research*, *280*(1-2), 192–200.

Hermansky, H. (1997). The modulation spectrum in the automatic recognition of speech. In *Proceedings of the ieee workshop on automatic speech recognition and understanding* (pp. 140–147).

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric

preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.

Hockenbury, D. H., & Hockenbury, S. E. (2007). *Discovering psychology.* New York City, NY: Worth.

Holmes, M., & Cole, J. (1983). Pseudo-resonance in the cochlea. In E. de Boer & M. Viergever (Eds.), *Mechanics of hearing* (pp. 45–52). Dordrecht, Netherlands: Springer.

Hopyan, T., Gordon, K. A., & Papsin, B. C. (2011). Identifying emotions in music through electrical hearing in deaf children using cochlear implants. *Cochlear Implants International*, *12*(1), 21–26.

Hopyan, T., Manno III, F. A., Papsin, B. C., & Gordon, K. A. (2015). Sad and happy emotion discrimination in music by children with cochlear implants. *Child Neuropsychology*, *6*, 1–15.

Hopyan, T., Peretz, I., Chan, L. P., Papsin, B. C., & Gordon, K. A. (2012). Children using cochlear implants capitalize on acoustical hearing for music perception. *Frontiers in Psychology*, *3*(425).

Hopyan-Misakyan, T. M., Gordon, K. A., Dennis, M., & Papsin, B. C. (2009). Recognition of affective speech prosody and facial affect in deaf children with unilateral right cochlear implants. *Child Neuropsychology*, *15*(2), 136–146.

Hornsby, B. W. Y. (2012). Understanding the problems with aided speech understanding. *Innovations*, *2*(1), 1–5.

House, D. (1994). Perception and production of mood in speech by cochlear implant users. In *Proceedings of the international conference on spoken language processing* (pp. 2051–2054). Yokohama, Japan.

Hughes, M. L. (2010). *Fundamentals of clinical ECAP measures in cochlear implants: Part 1: Use of the ECAP in speech processor programming (2nd ed.).*

Retrieved 2017-05-10, from http://www.audiologyonline.com/articles/fundamentals-clinical-ecap-measures-in-846

Hume, D. (2012). Emotions and moods. In S. P. Robbins, T. A. Judge, & T. C. Campbell (Eds.), *Organizational behavior* (pp. 258–297). London, UK: Pearson.

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*(4), 319–330.

Javel, E., & Shepherd, R. K. (2000). Electrical stimulation of the auditory nerve: III. Response initiation sites and temporal fine structure. *Hearing Research*, *140*(1-2), 45–76.

Johannsmeier, S., Heeger, P., Terakawa, M., Heisterkamp, A., Ripken, T., & Heinemann, D. (2017). *Optical cell stimulation for neuronal excitation.* Presentation at SPIE BiOS.

Johnstone, T., & Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proceedings of the XIVth international congress of phonetic sciences* (pp. 2029–2032). San Francisco, CA.

Joris, P. X. (2009). Recruitment of neurons and loudness. *Journal of the Association for Research in Otolaryngology*, *10*(1), 1–4.

Jung, K. H., Cho, Y.-S., Cho, J. K., Park, G. Y., Kim, E. Y., Hong, S. H., ... Rubinstein, J. T. (2010). Clinical assessment of music perception in korean cochlear implant listeners. *Acta Oto-Laryngologica*, *130*(6), 716–723.

Juslin, P. N. (1997). Emotional communication in music performance: A functionalist perspective and some data. *Music Perception: An Interdisciplinary Journal*, *14*(4), 383-418.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(6), 1797-1813.

Juslin, P. N. (2001). Communicating emotion in music performance: A review and a theoretical framework. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 309–337). New York: Oxford University Press.

Juslin, P. N., Karlsson, J., Lindstrm, E., Friberg, A., & Schoonderwaldt, E. (2006, 07). Play it again with feeling: Computer feedback in musical communication of emotions. *Journal of Experimental Psychology: Applied*, *12*, 79–95.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, *1*(4), 381–412.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814.

Juslin, P. N., & Sloboda, J. E. (2010). *Handbook of music and emotion: Theory, research, applications.* Oxford: Oxford University Press.

Kalathottukaren, R. T., Purdy, S. C., & Ballard, E. (2017). Prosody perception and production in children with hearing loss and age-and gender-matched controls. *Journal of the American Academy of Audiology*, *28*(4), 283–294.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). New York City, NY: Wiley.

Keidser, G., Dillon, H., Dyrlund, O., Carter, L., & Hartley, D. (2007). Preferred low-and high-frequency compression ratios among hearing aid users with moderately severe to profound hearing loss. *Journal of the American Academy of Audiology*, *18*(1), 17–33.

Keidser, G., Dillon, H., Flax, M., Ching, T., & Brewer, S. (2011). The NAL-NL2 prescription procedure. *Audiology Research*, *1*(24), 88–90.

Khanna, P., & Sasikumar, M. (2011). Recognizing emotions from human speech. *Thinkquest˜2010*, 219–223.

Kirchberger, M., & Russo, F. A. (2016). Dynamic range across music genres and the perception of dynamic compression in hearing-impaired listeners. *Trends in Hearing*, *20*, 1–16.

Kirkwood, D. (2005). When it comes to hearing aids, "more" was the story in '04. *The Hearing Journal*, *58*(5), 28–37.

Klasen, T. J., den Bogaert, T. V., Moonen, M., & Wouters, J. (2007, April). Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues. *IEEE Transactions on Signal Processing*, *55*(4), 1579–1585.

Kochkin, S. (2000). MarkeTrak V: "Why my hearing aids are in the drawer": The consumers' perspective. *The Hearing Journal*, *53*(2), 34–36.

Kochkin, S. (2010). MarkeTrak VIII: Consumer satisfaction with hearing aids is slowly increasing. *The Hearing Journal*, *63*(1), 19–20.

Kochkin, S. (2012). MarkeTrak VIII: The key influencing factors in hearing aid purchase intent. *Hearing Review*, *19*(3), 12–25.

Koelsch, S., Wittfoth, M., Wolf, A., Müller, J., & Hahne, A. (2004). Music perception in cochlear implant users: An event-related potential study. *Clinical Neurophysiology*, *115*(4), 966–972.

Kong, Y.-Y., Mullangi, A., Marozeau, J., & Epstein, M. (2011). Temporal and spectral cues for musical timbre perception in electric hearing. *Journal of Speech, Language, and Hearing Research*, *54*(3), 981–994.

Kral, A., Hartmann, R., Mortazavi, D., & Klinke, R. (1998). Spatial resolution of cochlear implants: The electrical field and excitation of auditory afferents. *Hearing Research*, *121*(1-2), 11–28.

Kricos, P. B. (2017). *Tips for hearing in noise.* Retrieved 2017-

03-05, from http://www.betterhearing.org/hearingpedia/counseling-articles-tips/tips-hearing-noise

Krumhansl, C. (2010). Plink: "Thin slices" of music. *Music Perception*, *27*(5), 337-354.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Kuk, F., Keenan, D., Korhonen, P., & Lau, C.-c. (2009). Efficacy of linear frequency transposition on consonant identification in quiet and in noise. *Journal of the American Academy of Audiology*, *20*(8), 465–479.

Ladefoged, P., & McKinney, N. P. (1963). Loudness, sound pressure, and subglottal pressure in speech. *Journal of the Acoustical Society of America*, *35*, 454–460.

Landsberger, D. M., Padilla, M., & Srinivasan, A. G. (2012). Reducing current spread using current focusing in cochlear implant users. *Hearing Research*, *284*(1-2), 16-24.

Landsberger, D. M., Vermeire, K., Claes, A., Van Rompaey, V., & Van de Heyning, P. (2016). Qualities of single electrode stimulation as a function of rate and place of stimulation with a cochlear implant. *Ear and Hearing*, *37*(3), 149–159.

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Review of personality and social psychology: Emotion* (pp. 25–59). Newbury Park, CA: Sage.

Lartillot, O., & Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. In *Proceedings of the international conference on digital audio effects* (pp. 237–244). Bordeaux, France.

Lasak, J. M., Allen, P., McVay, T., & Lewis, D. (2014). Hearing loss: Diagnosis and management. *Primary Care: Clinics in Office Practice*, *41*(1), 19–31.

Lassaletta, L., Castro, A., Bastarrica, M., Prez-Mora, R., Madero, R., Sarri, J. D., & Gaviln, J. (2007). Does music perception have an impact on quality of life

following cochlear implantation? *Acta Oto-Laryngologica*, *127*(7), 682–686.

Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: A preliminary study. *Journal of Phonetics*, *24*(3), 313–335.

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLoS ONE*, *7*(12), e52508.

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004, 01). Measuring empathy: Reliability and validity of the empathy quotient. *Psychological Medicine*, *34*(5), 911–920.

Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ez` (R package version 4.4-0)

Leal, M. C., Shin, Y. J., Laborde, M.-l., Calmels, M.-n., Verges, S., Lugardon, S., ... Fraysse, B. (2003). Music perception in adult cochlear implant recipients. *Acta Oto-Laryngologica*, *123*(7), 826–835.

LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology*, *23*(4-5), 727–738.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292.

Leek, M. R., Molis, M. R., Kubli, L. R., & Tufts, J. B. (2008). Enjoyment of music by elderly hearing-impaired listeners. *Journal of the American Academy of Audiology*, *19*(6), 519–526.

Leff, J. P. (1973). Culture and the differentiation of emotional states. *The British Journal of Psychiatry*, *123*(574), 299–306.

Lennie, T. (2017). *Universality in the language of emotions revisited: Towards a re-*

vised methodology for interpreting acoustic cues in musical affect (Unpublished master's thesis). The University of Sheffield.

Levine, P. A., Miyamoto, R. T., Myres, W. A., Wagner, M., & Punch, J. L. (1987). Vibrotactile devices as sensory aids for the deaf. *Otolaryngology – Head and Neck Surgery*, *97*(1), 57–63.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Lim, D. J. (1986). Functional structure of the organ of Corti: A review. *Hearing Research*, *22*(1), 117–146.

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, *11*(5), 1021–1031.

Limb, C. J. (2011, October). *Charles limb: Building the musical muscle [video file].* Retrieved from `https://www.ted.com/talks/charles_limb_building_the_musical_muscle?language=en#t-574196`

Limb, C. J. (2016, October). *Cochlear implants and music: The quest for perfect hearing.* Presentation at music & cochlear implants symposium, Snekkersten, Denmark.

Limb, C. J., Molloy, A. T., Jiradejvong, P., & Braun, A. R. (2010). Auditory cortical activity during cochlear implant-mediated perception of spoken language, melody, and rhythm. *Journal of the Association for Research in Otolaryngology*, *11*(1), 133–143.

Lindsay, P. H., & Norman, D. A. (1977). *Human information processing: An introduction to psychology.* New York City, NY: Academic Press.

Lindström, E., Juslin, P. N., Bresin, R., & Williamon, A. (2003). expressivity comes from within your soul: A questionnaire study of music students' perspectives on expressivity. *Research Studies in Music Education*, *20*(1), 23–47.

Liu, X., Chen, Q., Wu, X., Liu, Y., & Liu, Y. (2017). CNN based music emotion classification. *arXiv (preprint)*.

Liu, Y.-W. (2012). *Hilbert transform and applications.* Rijeka, Croatia: InTech Open Access.

Livingstone, S. R., Thompson, W. F., Wanderley, M. M., & Palmer, C. (2015). Common cues to emotion in the dynamic facial expressions of speech and song. *The Quarterly Journal of Experimental Psychology*, *68*(5), 952-970.

Loebach, J. L., Pisoni, D. B., & Svirsky, M. A. (2009). Transfer of auditory perceptual learning with spectrally reduced speech to speech and nonspeech tasks: Implications for cochlear implants. *Ear and Hearing*, *30*(6), 662.

Looi, V., Gfeller, K., & Driscoll, V. D. (2012). Music appreciation and training for cochlear implant recipients: A review. In *Seminars in hearing* (Vol. 33, pp. 307–334).

Looi, V., McDermott, H., McKay, C., & Hickson, L. (2008). Music perception of cochlear implant users compared with that of hearing aid users. *Ear and Hearing*, *29*(3), 421–434.

Looi, V., Winter, P., Anderson, I., & Sucher, C. (2011). A music quality rating test battery for cochlear implant users to compare the FSP and HDCIS strategies for music appreciation. *International Journal of Audiology*, *50*(8), 503–518.

Looi, V., Wong, Y., & Loo, J. H. (2016). The effects of training on music perception and appreciation for cochlear implant recipients. *Advances in Otolaryngology*,

2016.

Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, *78*(2), 341–348.

Luo, X. (2016, Oct). *Musical emotion recognition by normal-hearing listeners and cochlear implant users.* Presentation at music & cochlear implants symposium, Snekkersten, Denmark.

Luo, X., Fu, Q.-J., & Galvin, J. J. (2007). Vocal emotion recognition by normal-hearing listeners and cochlear implant users. *Trends in Amplification*, *11*(4), 301–315.

Lupsakko, A., Taina, Kautiainen, H. J., & Sulkava, R. (2005). The non-use of hearing aids in people aged 75years and over in the city of Kuopio in Finland. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, *262*(3), 165–169.

Maarefvand, M., Marozeau, J., & Blamey, P. J. (2013). A cochlear implant user with exceptional musical hearing ability. *International Journal of Audiology*, *52*(6), 424–432.

Madsen, S. M., & Moore, B. C. (2014). Music and hearing aids. *Trends in Hearing*, *0*(0), 1–29.

Madsen, S. M., Stone, M. A., McKinney, M. F., Fitz, K., & Moore, B. C. (2015). Effects of wide dynamic-range compression on the perceived clarity of individual musical instruments. *The Journal of the Acoustical Society of America*, *137*(4), 1867–1876.

Mao, K. Z. (2004). Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(1), 629–634.

Mauger, S. J., Warren, C. D., Knight, M. R., Goorevich, M., & Nel, E. (2014). Clinical evaluation of the nucleus® 6 cochlear implant system: Performance

improvements with SmartSound iQ. *International Journal of Audiology*, *53*(8), 564–576.

McAdams, S. (1993). Recognition of sound sources and events. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human auditiion* (pp. 146–198). Oxford, UK: Oxford University Press.

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

McDermott, H. J., & McKay, C. M. (1994). Pitch ranking with nonsimultaneous dual-electrode electrical stimulation of the cochlea. *The Journal of the Acoustical Society of America*, *96*(1), 155–162.

McFarland, W. (2000). Speech perception and hearing aids. In R. E. Sandlin (Ed.), *Textbook of hearing aid amplification* (pp. 37–53). London, UK: Singular Publishing Group.

Mehrabian, A. (1996). Pleasure-arousal. dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology: Developmental Learning Personality Social*, *14*(4), 261–292.

Meister, H., Lausberg, I., Kiessling, J., Walger, M., & von Wedel, H. (2002). Determining the importance of fundamental hearing aid attributes. *Otology & Neurotology*, *23*(4), 457–462.

Meng, Q., Zheng, N., & Li, X. (2016). Loudness contour can influence mandarin tone recognition: Vocoder simulation and cochlear implants. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*(99), 1-1.

Mershon, D. H., Desaulniers, D. H., Kiefer, S. A., Amerson, T. L., & Mills, J. T. (1981). Perceived loudness and visually-determined auditory distance. *Perception*, *10*(5), 531–543.

Mertens, P. (2004). The prosogram: Semi-automatic transcription of prosody based

on a tonal perception model. In *Proceedings of the international conference on speech prosody.*

Metselaar, M., Maat, B., Krijnen, P., Verschuure, H., Dreschler, W., & Feenstra, L. (2008). Comparison of speech intelligibility in quiet and in noise after hearing aid fitting according to a purely prescriptive and a comparative fitting procedure. *European Archives of Oto-Rhino-Laryngology*, *265*(9), 1113–1120.

Micheyl, C., Divis, K., Wrobleski, D. M., & Oxenham, A. J. (2010). Does fundamental-frequency discrimination measure virtual pitch discrimination? *The Journal of the Acoustical Society of America*, *128*(4), 1930–1942.

Mildner, V., & Koska, T. (2014). Recognition and production of emotions in children with cochlear implants. *Clinical Linguistics & Phonetics*, *28*(7-8), 543–554.

Mithen, S. (2009). The music instinct. *Annals of the New York Academy of Sciences*, *1169*(1), 3–12.

Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303.

Møller, A. R. (2000). *Hearing: Its physiology and pathophysiology.* London, UK: Academic Press.

Moon, I. J., & Hong, S. H. (2014). What is temporal fine structure and why is it important? *Korean Journal of Audiology*, *18*(1), 1–7.

Moore, B. C. J. (1996). Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear and hearing*, *17*(2), 133–161.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing.* London, UK: Academic Press.

Moore, B. C. J. (2016). Effects of sound-induced hearing loss and hearing aids on the perception of music. *Journal of the Audio Engineering Society*, *64*(3),

112–123.

Moore, B. C. J., & Carlyon, R. P. (2005). Perception of pitch by people with cochlear hearing loss and by cochlear implant users. In C. Plack, R. Fay, A. Oxenham, & A. Popper (Eds.), *Pitch* (Vol. 24, pp. 234–277). New York City, NY: Springer.

Moore, B. C. J., Peters, R. W., & Glasberg, B. R. (1992). Detection of temporal gaps in sinusoids by elderly subjects with and without hearing loss. *The Journal of the Acoustical Society of America*, *92*(4), 1923–1932.

Moore, D. R., & Shannon, R. V. (2009). Beyond cochlear implants: Awakening the deafened brain. *Nature Neuroscience*, *12*(6), 686–691.

Most, T., & Aviner, C. (2009). Auditory, visual, and auditory-visual perception of emotions by individuals with cochlear implants, hearing aids, and normal hearing. *Journal of Deaf Studies and Deaf Education*, *14*(4), 449–464.

Most, T., & Peled, M. (2007). Perception of suprasegmental features of speech by children with cochlear implants and children with hearing aids. *Journal of Deaf Studies and Deaf Education*, *12*(3), 350–361.

Most, T., Weisel, A., & Zaychik, A. (1993). Auditory, visual and auditoryvisual identification of emotions by hearing and hearing-impaired adolescents. *British Journal of Audiology*, *27*(4), 247–253.

Mozziconacci, S. J. L., & Hermes, D. J. (1999). Role of intonation patterns in conveying emotion in speech. *Journal of the International Phonetic Association*, *99*, 2001–2004.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLOS ONE*, *9*(2), e89642.

Müller, J., & Raine, C. (2013). Quality standards for adult cochlear implantation. *Cochlear Implants International*, *14*(2), 6–12.

Murray, D. J., & Hanson, J. V. (1992). Application of digital signal processing to hearing aids: A critical survey. *Journal of the American Academy of Audiology*, *3*(2), 145–152.

Myers, R. (1990). *Classical and modern regression with applications.* Boston, MA: Duxbury.

Nadol, J. B., Young, Y.-S., & Glynn, R. J. (1989). Survival of spiral ganglion cells in profound sensorineural hearing loss: Implications for cochlear implantation. *Annals of Otology, Rhinology & Laryngology*, *98*(6), 411–416.

Naimi, B., a.s. Hamm, N., Groen, T. A., Skidmore, A. K., & Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling. *Ecography*, *37*, 191–203.

Nair, D., Large, E. W., Steinberg, F., & Kelso, J. A. S. (2002). Expressive timing and perception of emotion in music: An fMRI study. In C. Stevens (Ed.), *Proceedings of the seventh international conference on music perception and cognition* (pp. 627–630). Adelaide, Australia: Causal Productions.

Nakata, T., Trehub, S. E., & Kanda, Y. (2012). Effect of cochlear implants on children's perception and production of speech prosody. *The Journal of the Acoustical Society of America*, *131*(2), 1307-1314.

Nelson, P. B., Jin, S.-H., Carney, A. E., & Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *113*(2), 961–968.

Nemer, J. S., Kohlberg, G. D., Mancuso, D. M., Griffin, B. M., Certo, M. V., Chen, S. Y., ... Lalwani, A. K. (2017). Reduction of the harmonic series influences musical enjoyment with cochlear implants. *Otology & Neurotology*, *38*(1), 31–37.

Neuman, A. C., Bakke, M. H., Mackersie, C., Hellman, S., & Levitt, H. (1998). The

effect of compression ratio and release time on the categorical rating of sound quality. *The Journal of the Acoustical Society of America*, *103*(5), 2273–2281.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085–1099.

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

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden markov models. In *Proceedings of the european conference on speech communication and technology (eurospeech '01)* (pp. 2679–2682). Aalborg.

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, *18*(1), 9–35.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2017). vegan: Community ecology package [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=vegan` (R package version 2.4-3)

Orbelo, D. M., Grim, M. A., Talbott, R. E., & Ross, E. D. (2005). Impaired comprehension of affective prosody in elderly subjects is not predicted by age-related hearing loss or age-related cognitive decline. *Journal of Geriatric Psychiatry and Neurology*, *18*(1), 25–32.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical

analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911.

Oxenham, A. J. (2013). Revisiting place and temporal theories of pitch. *Acoustical Science and Technology*, *34*(6), 388–396.

Palmer, C. (1997). Music performance. *Annual Review of Psychology*, *48*(1), 115-138.

Parveen, F. (2017). *Plasmonic stimulation of electrically excitable cells* (Unpublished doctoral dissertation). University of South Florida.

Patel, A. D., Iversen, J. R., Wassenaar, M., & Hagoort, P. (2008). Musical syntactic processing in agrammatic Broca's aphasia. *Aphasiology*, *22*(7-8), 776-789.

Patel, A. D., Wong, M., Foxton, J., Lochy, A., & Peretz, I. (2008). Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Perception: An Interdisciplinary Journal*, *25*(4), 357–368.

Peng, S.-C., Chatterjee, M., & Lu, N. (2012). Acoustic cue integration in speech intonation recognition with cochlear implants. *Trends in Amplification*, *16*, 67–82.

Peng, S.-C., Lu, N., & Chatterjee, M. (2009). Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners. *Audiology and Neurotology*, *14*(5), 327–337.

Peng, S.-C., Tomblin, J. B., & Turner, C. W. (2008). Production and perception of speech intonation in pediatric cochlear implant recipients and individuals with normal hearing. *Ear and Hearing*, *29*(3), 336–351.

Penninger, R. T., Limb, C. J., Vermeire, K., Leman, M., & Dhooge, I. (2013). Experimental assessment of polyphonic tones with cochlear implants. *Otology & Neurotology*, *34*(7), 1267–1271.

Pentoś, K. (2016). The methods of extracting the contribution of variables in artificial

neural network models – Comparison of inherent instability. *Computers and Electronics in Agriculture*, *127*, 141–146.

Peppé, S., & McCann, J. (2003). Assessing intonation and prosody in children with atypical language development: The PEPS-C test and the revised version. *Clinical Linguistics & Phonetics*, *17*(4-5), 345–354.

Peretz, I., & Hyde, K. L. (2003). What is specific to music processing? insights from congenital amusia. *Trends in Cognitive Sciences*, *7*(8), 362–367.

Peretz, I., Vuvan, D., Lagrois, M.-É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140090.

Petersen, B., Hansen, M., Sørensen, S. D., Ovesen, T., & Vuust, P. (2014, Dec). Aspects of music with cochlear implants–music listening habits and appreciation in danish cochlear implant users. In *Proceedings of the international symposium on auditory and audiological research* (Vol. 4, pp. 461–466).

Petersen, B., Mortensen, M. V., Hansen, M., & Vuust, P. (2012). Singing in the key of life: A study on effects of musical ear training after cochlear implantation. *Psychomusicology: Music, Mind, and Brain*, *22*(2), 134.

Petersen, B., Sørensen, S. D., Pedersen, E. R., Parsons, C., & Vuust, P. (2015). *Wrap it in rap! Music making with adolescent CI users.*

Petersen, B., Weed, E., Sandmann, P., Brattico, E., Hansen, M., Sørensen, S. D., & Vuust, P. (2015). Brain responses to musical feature changes in adolescent cochlear implant users. *Frontiers in Human Neuroscience*, *9*(7).

Petrushin, V. (1999, Nov). Emotion in speech: Recognition and application to call centers. In *Proceedings of artificial neural networks in engineering* (pp. 7–10).

Philip, R. C. M., Whalley, H. C., Stanfield, A. C., Sprengelmeyer, R., Santos, I. M., Young, A. W., ... Hall, J. (2010, 11). Deficits in facial, body movement

and vocal emotional processing in autism spectrum disorders. *Psychological Medicine*, *40*, 1919-1929.

Pick, G., Evans, E. F., & Wilson, J. P. (1977). Frequency resolution in patients with hearing loss of cochlear origin. In E. F. Evans & J. P. WIlson (Eds.), *Psychophysics and physiology of hearing* (pp. 273–281). London: Academic Press.

Pinyon, J. L., Tadros, S. F., Froud, K. E., Wong, A. C., Tompson, I. T., Crawford, E. N., ... Housley, G. D. (2014). Close-field electroporation gene delivery using the cochlear implant electrode array enhances the bionic ear. *Science Translational Medicine*, *6*(233), 233–254.

Plack, C. J. (2013). *The sense of hearing.* Hove: Psychology Press.

Plomp, R., & Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, *18*(1), 43–52.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis.* New York City, NY: HarperCollins.

Plutchik, R. (2001). The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, *89*(4), 344–350.

Pocock, G., Richards, C. D., & Richards, D. A. (2013). *Human physiology.* Address: Oxford University Press.

Popelka, G. R., & Moore, B. C. (2016). Future directions for hearing aid development. In G. R.Popelka, B. C. J. Moore, R. R. Fay, & A. N. Popper (Eds.), *Hearing aids* (pp. 323–333). New York City, NY: Springer.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734.

Powers, T. A., & Fröhlich, M. (2014). Clinical results with a new wireless binaural directional hearing system. *Hearing Review*, *21*(11), 32–34.

Price, C., Thierry, G., & Griffiths, T. (2005). Speech-specific auditory processing: Where is it? *Trends in Cognitive Sciences*, *9*(6), 271 - 276.

Quinto, L., Thompson, W. F., & Taylor, A. (2014). The contributions of compositional structure and performance expression to the communication of emotion in music. *Psychology of Music*, *42*(4), 503-524.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Rahne, T., Bhme, L., & Gtze, G. (2011). Timbre discrimination in cochlear implant users and normal hearing subjects using cross-faded synthetic tones. *Journal of Neuroscience Methods*, *199*(2), 290–295.

Raine, C. (2013). Cochlear implants in the united kingdom: Awareness and utilization. *Cochlear Implants International*, *14*(1), 32–37.

Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, *16*(2), 143-160.

Ratanamahatana, A., & Keogh, E. (2004). Everything you know about dynamic time warping is wrong. In *3rd workshop on mining temporal and sequential data, in conjunction with 10th ACM SIGKDD International conference on knowledge discovery and data mining (KDD-2004).* Seattle, WA.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 532–538). New York City, NY: Springer.

Reiss, L. A., Turner, C. W., Erenberg, S. R., & Gantz, B. J. (2007). Changes in pitch with a cochlear implant over time. *Journal for the Association for Research in*

Otolaryngology, *8*(2), 241–257.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*(1), 129.

Rigo, T. G., & Lieberman, D. A. (1989). Nonverbal sensitivity of normal-hearing and hearing-impaired older adults. *Ear and Hearing*, *10*(3), 184–189.

Rosenblum, L. D. (2004). Perceiving articulatory events. In J. G. Neuhoff (Ed.), *Ecological psychoacoustics* (p. 219-248). Burlington, MA: Elsevier Academic Press.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Roy, A. T., Jiradejvong, P., Carver, C., & Limb, C. J. (2012a). Assessment of sound quality perception in cochlear implant users during music listening. *Otology & Neurotology*, *33*(3), 319–327.

Roy, A. T., Jiradejvong, P., Carver, C., & Limb, C. J. (2012b). Musical sound quality impairments in cochlear implant (CI) users as a function of limited high-frequency perception. *Trends in Amplification*, *16*(4), 191-200.

Rubinstein, J. T., Parkinson, W. S., Tyler, R. S., & Gantz, B. (1999). Residual speech recognition and cochlear implant performance: Effects of implantation criteria. *Otology & Neurotology*, *20*(4), 445–452.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin*, *110*(3), 426–450.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Russo, F. A., & Fanelli, D. (2016). Perception of emotion in music by hearing-impaired and hearing-aided listeners. *Canadian Acoustics*, *44*(3).

Saarikallio, S., Vuoskoski, J., & Luck, G. (2014). Adolescents' expression and perception of emotion in music reflects their broader abilities of emotional communication. *Psychology of Well-Being*, *4*(1), 21.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, *398*(6730), 760–760.

Saheer, L., & Potard, B. (2013). *Understanding factors in emotion perception* (Tech. Rep.). Idiap Research Institute.

Sandlin, R. E. (2000). *Textbook of hearing aid amplification.* Boston, MA: Cengage learning.

Sartre, J.-P. (2002). *Sketch for a theory of the emotions.* London: Routledge.

Sataloff, R. T., & Sataloff, J. (2006). Conductive hearing loss. In R. T. Sataloff & J. Sataloff (Eds.), *Occupational hearing loss* (pp. 129–194). Boca Raton, FL: CRC Press.

Sauter, D. A., Panattoni, C., & Happ, F. (2013). Children's recognition of emotions from vocal cues. *British Journal of Developmental Psychology*, *31*(1), 97–113.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143–165.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227–256.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, *44*(4), 695–729.

Scherer, K. R., Banse, R., Wallbott, H., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*(2), 123-148.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal

expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92.

Scherer, K. R., & Oshinsky, J. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, *1*(4), 331–346.

Schienle, A., Stark, R., Walter, B., Blecker, C., Ott, U., Kirsch, P., ... Vaitl, D. (2002). The insula is not specifically involved in disgust processing: an fMRI study. *Neuroreport*, *13*(16), 2023–2026.

Schmidt, J., Herzog, D., Scharenborg, O., & Janse, E. (2016). Do hearing aids improve affect perception? In *Physiology, psychoacoustics and cognition in normal and impaired hearing* (pp. 47–55). New York City, NY: Springer.

Schow, B., Friedland, D. R., Jensen, J., Burg, L., & Runge, C. L. (2012). Electrode failure and device failure in adult cochlear implantation. *Cochlear Implants International*, *13*(1), 35–40.

Schuknecht, H., Kimura, R., & Naufal, P. (1973). The pathology of sudden deafness. *Acta Oto-Laryngologica*, *76*(1-6), 75–97.

Sęk, A., & Moore, B. C. (2012). Implementation of two tests for measuring sensitivity to temporal fine structure. *International Journal of Audiology*, *51*(1), 58–63.

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*(3), 617–627.

Shannon, R. V. (1989). Detection of gaps in sinusoids and pulse trains by patients with cochlear implants. *The Journal of the Acoustical Society of America*, *85*(6), 2587-2592.

Shannon, R. V. (1992). Temporal modulation transfer functions in patients with cochlear implants. *The Journal of the Acoustical Society of America*, *91*(4),

2156-2164.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303-304.

Sharpe, D. (2015). Your chi-square test is statistically significant: Now what? *Practical Assessment, Research & Evaluation*, *20*(8), 1-10.

Shipp, D., & Nedzelski, J. (1995). Prognostic indicators of speech recognition performance in adult cochlear implant users: a prospective analysis. *The Annals of otology, rhinology & laryngology. Supplement*, *166*, 194–196.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*(23), 167–196.

Smith, J. D., & Kemler, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General*, *113*(1), 137.

Smith, J. D., & Shapiro, J. H. (1989). The occurrence of holistic categorization. *Journal of Memory and Language*, *28*(4), 386–399.

Smith, K. J. (2016). *Nature of mathematics*. Boston, MA: Cengage Learning.

Stadler, S., & Leijon, A. (2009). Prediction of speech recognition in cochlear implant users by adapting auditory models to psychophysical data. *EURASIP Journal on Advances in Signal Processing*, *2009*, 5:1-5:9.

Stainsby, T. H., McDermott, H. J., McKay, C. M., & Clark, G. M. (1997). Preliminary results on spectral shape perception and discrimination of musical sounds by normal hearing subjects and cochlear implantees. *Scientific Publications*, *10*(1104), 11-14.

Stevens, D., Charman, T., & Blair, R. J. R. (2001). Recognition of emotion in facial expressions and vocal tones in children with psychopathic tendencies. *The Journal of Genetic Psychology*, *162*(2), 201-211.

Stiles, D. J. (2013). Sarcasm recognition in children with hearing loss: The role of context and intonation. *Journal of Educational Audiology*, *19*, 3-11.

Stone, M. A., & Moore, B. C. J. (2007). Quantifying the effects of fact-acting compression on the envelope of speech. *The Journal of the Acoustical Society of America*, *121*, 1654-1664.

Strait, D. L., Kraus, N., Skoe, E., & Ashley, R. (2009). Musical experience and neural efficiency – Effects of training on subcortical processing of vocal expressions of emotion. *European Journal of Neuroscience*, *29*(3), 661-668.

Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., & Gwaltney, C. A. (1999). Monosyllabic word recognition at higher-than-normal speech and noise levels. *The Journal of the Acoustical Society of America*, *105*(4), 2431-2444.

Sundberg, J. (1999). The perception of singing. In D. Deutsch (Ed.), *The psychology of music* (pp. 171–214). San Diego, CA: Academic Press.

Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaikes information criterion. *Behavioral Ecology and Sociobiology*, *65*(1), 13–21.

Tao, D., Deng, R., Jiang, Y., Galvin III, J. J., Fu, Q.-J., & Chen, B. (2015). Melodic pitch perception and lexical tone perception in Mandarin-speaking cochlear implant users. *Ear and Hearing*, *36*(1), 102-110.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, *1*(2), 155-182.

Thayer, R. E. (1989). *The biopsychology of mood and arousal.* Oxford: Oxford University Press.

Thompson, W. F., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica*, *2006*(158), 407-424.

Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage

hypothesis. *Proceedings of the National Academy of Sciences*, *109*(46), 19027-19032.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion*, *4*(1), 46-64.

Tillmann, B., & Bigand, E. (2004). The relative importance of local and global structures in music perception. *The Journal of Aesthetics and Art Criticism*, *62*(2), 211-222.

Tobey, E. A., Shin, S., Prashant, M. S., & Geers, A. (2011). Spoken word recognition in adolescent cochlear implant users during quiet and multi-speaker babble conditions. *Otology & Neurotology*, *32*(3), 413.

Tomkins, S. S. (1962). *Affect, imagery, consciousness: Vol. I. The positive affects.* New York City, NY: Springer.

Townshend, B., Cotter, N., Van Compernolle, D., & White, R. L. (1987). Pitch perception by cochlear implant subjects. *The Journal of the Acoustical Society of America*, *82*(1), 106-115.

Trainor, L. J., & Corrigall, K. A. (2010). Music acquisition and effects of musical experience. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music perception* (pp. 89–127). New York City, NY: Springer.

Truax, B. (2001). *Acoustic communication.* Santa Barbara, CA: Greenwood Publishing Group.

Truax, B. (2016). Environmental sound and its relation to human emotion. *Canadian Acoustics*, *44*(3).

Truong, K. P., Leeuwen, D. A., Neerincx, M. A., & Jong, F. M. (2009). *Arousal and valence prediction in spontaneous emotional speech: Felt versus perceived emotion.* International Speech Communication Association.

Uys, M., Pottas, L., Vinck, B., & Van Dijk, C. (2012). The influence of non-linear

frequency compression on the perception of music by adults with a moderate to severe hearing loss: Subjective impressions. *South African Journal of Communication Disorders*, *59*(1), 53–67.

Vale, R. D. (2013). The value of asking questions. *Molecular Biology of the Cell*, *24*(6), 680–682.

Vandergrift, L., Goh, C. C. M., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning*, *56*(3), 431-462.

van Heugten, M., Volkova, A., Trehub, S. E., & Schellenberg, E. G. (2014). Childrens recognition of spectrally degraded cartoon voices. *Ear and Hearing*, *35*(1), 118-125.

Van Zijl, A. G. W., & Sloboda, J. (2011). Performers experienced emotions in the construction of expressive musical performance: An exploratory investigation. *Psychology of Music*, *39*(2), 196-219.

Varshney, L. R., & Sun, J. Z. (2013). Why do we perceive logarithmically? *Significance*, *10*(1), 28–31.

Veekmans, K., Ressel, L., Mueller, J., Vischer, M., & Brockmeier, S. (2009). Comparison of music perception in bilateral and unilateral cochlear implant users and normal-hearing subjects. *Audiology and Neurotology*, *14*(5), 315–326.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (4th ed.). New York City, NY: Springer.

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition and Emotion*, *22*(4), 720-752.

Volkova, A., Trehub, S. E., Schellenberg, E. G., Papsin, B. C., & Gordon, K. A. (2013). Children with bilateral cochlear implants identify emotion in speech

and music. *Cochlear Implants International*, *14*(2), 80-91.

Volkova, A., Trehub, S. E., Schellenberg, E. G., Papsin, B. C., & Gordon, K. A. (2014). Children's identification of familiar songs from pitch and timing cues. *Frontiers in Psychology*, 5: 863.

Von Békésy, G., & Wever, E. G. (1960). *Experiments in hearing* (Vol. 8). New York City, NY: McGraw-Hill.

Vos, P. G., & Troost, J. M. (1989). Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception: An Interdisciplinary Journal*, *6*(4), 383-396.

Vuust, P., Brattico, E., Glerean, E., Seppnen, M., Pakarinen, S., Tervaniemi, M., & Ntnen, R. (2011). New fast mismatch negativity paradigm for determining the neural prerequisites for musical ability. *Cortex*, *47*(9), 1091–1098.

Wakabayashi, A., Baron-Cohen, S., Wheelwright, S., Goldenfeld, N., Delaney, J., Fine, D., ... Weil, L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences*, *41*(5), 929 - 940.

Wang, D. (2017). Deep learning reinvents the hearing aid. *IEEE Spectrum*, *54*(3), 32–37.

Warnes, G. R., Bolker, B., Lumley, T., & Johnson, R. C. (2015). gmodels: Various R programming tools for model fitting [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=gmodels` (R package version 2.16.2)

Wayner, D. S. (1990). *The hearing aid handbook: User's guide for children.* Washington, DC: Gallaudet University Press.

Wei, W. I., Wong, R., Hui, Y., Au, D. K. K., Wong, B. Y. K., Ho, W. K., ... Chung, E. (2000). Chinese tonal language rehabilitation following cochlear

implantation in children. *Acta Oto-Laryngologica*, *120*(2), 218-221.

Welling, D. R., Ukstins, C. A., et al. (2013). *Fundamentals of audiology for the speech-language pathologist.* Jones & Bartlett Publishers.

Whitmer, W. M., Brennan-Jones, C. G., & Akeroyd, M. A. (2011). The speech intelligibility benefit of a unilateral wireless system for hearing-impaired adults. *International Journal of Audiology*, *50*(12), 905–911.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75fox*(5), 1182–1189.

WHO. (2012). *WHO global estimates on prevalence of hearing loss.*

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis.* New York City, NY: Springer-Verlag.

Wilson, B. S. (2004). Cochlear implants: Auditory prostheses and electric hearing. Springer handbook of auditory research, vol. 20. In F. G. Zeng, A. N. Popper, & R. R. Fay (Eds.), *Engineering design of cochlear implants* (pp. 14–52). New York City, NY: Springer.

Wilson, B. S., & Dorman, M. F. (2008). Cochlear implants: A remarkable past and a brilliant future. *Hearing Research*, *242*(1), 3–21.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2012). The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing. *The Journal of the Acoustical Society of America*, *131*(2), 1465–1479.

Witt, S., Murray, K. T., & Tyler, R. S. (2000). Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients. *Journal*

of the American Academy of Audiology, *11*, 390–406.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143–146).

Wong, A. C., & Ryan, A. F. (2015). Mechanisms of sensorineural cell damage, death and survival in the cochlea. *Aging, Neurogenesis and Neuroinflammation in Hearing Loss and Protection*, 7: 58.

Wright, P., He, G., Shapira, N. A., Goodman, W. K., & Liu, Y. (2004). Disgust and the insula: fMRI responses to pictures of mutilation and contamination. *Neuroreport*, *15*(15), 2347-2351.

Wright, R., & 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), 350–365.

Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing*, *90*(5), 1415–1423.

Yip, M., Jin, R., Nakajima, H. H., Stankovic, K. M., & Chandrakasan, A. P. (2015). A fully-implantable cochlear implant SoC with piezoelectric middle-ear sensor and arbitrary waveform neural stimulation. *IEEE Journal of Solid-state Circuits*, *50*(1), 214–229.

Yost, W. (2003). Audition. In R. W. P. Alice F. Healy (Ed.), *Handbook of psychology: Experimental psychology.* New York City, NY: Wiley.

Zeng, F.-G. (2002). Temporal pitch in electric hearing. *Hearing Research*, *174*(1), 101–106.

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., ... Cao, K. (2005). Speech recognition with amplitude and frequency modula-

tions. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2293-2298.

Zentner, M. R., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, *8*(4), 494-521.

Zentner, M. R., Meylan, S., & Scherer, K. R. (2000). *Exploring musical emotions across five genres of music. Presentation at 6th international conference of the society for music perception and cognition. August 5-10, Keele, UK.*

Zhang, T., & Kuo, C. C. J. (1999, Mar). Hierarchical classification of audio data for archiving and retrieving. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing* (Vol. 6, p. 3001-3004).

Zhang, T., & Kuo, C. C. J. (2001, May). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, *9*(4), 441-457.

Zwicker, E., & Fastl, H. (1990). *Psychoacoustics: Facts and models* (Vol. 22). Springer.

Zwicker, E., & Schorn, K. (1978). Psychoacoustical tuning curves in audiology. *Audiology*, *17*(2), 120-140.