The processing of Dutch prosody with cochlear implants and vocoder simulations

Velde, D.J. van de

The processing of Dutch prosody with cochlear implants and vocoder simulations

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111

e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration: an excerpt from part 3 of Arnold Schoenberg's *Gurrelieder*, cantata for soloists, choruses and orchestra (1900-1911)

ISBN: 978-94-6093-245-8
NUR 616

The processing of Dutch prosody with cochlear implants and vocoder simulations

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 5 juli 2017
klokke 13.45 uur

door

Daan Johan van de Velde

geboren te Amsterdam
in 1984

Promotores:	Prof. dr. Niels O. Schiller (Universiteit Leiden) Prof. dr. ir. Johan H.M. Frijns (Leids Universitair Medisch Centrum)
Promotiecommissie:	Prof. dr. Lisa Cheng (Universiteit Leiden) Prof. dr. Carolien Rieffe (Universiteit Leiden) Prof. dr. Astrid van Wieringen (KU Leuven) Dr. Yiya Chen (Universiteit Leiden)

Contents

Acknowledgments

The completion of this thesis would not have been possible without the support from a great number of people and institutes. First of all, I owe my gratitude to three institutes: Leiden University Centre for Linguistics (LUCL), Leiden University Medical Center (LUMC) and Leiden Institute for Brain and Cognition (LIBC). They provided me with the funding, equipment and contacts that I needed. My principal host institute LUCL has been especially generous in its confidence in the completion of this thesis.

Of all the people who helped me, my expressions of gratitude first go to my supervisors. Niels Schiller was gracious enough to allow me the freedom I wanted to take in choosing topics and methodologies for experiments as well as providing me with ample opportunity to teach courses. Johan Frijns always kept me connected to the world of cochlear implants by inviting me to watch a surgery in the operating room, introducing me to other influential researchers in the field or showing the importance of collaborating with students to help performing experiments. With Vincent van Heuven I spent numerous hours on data analysis and interpretation in the first half of my PhD period. Mieke Beers, Claartje Levelt and Jeroen Briaire assisted me on all scientific stages such as methodology, stimulus creation, and acquiring medical-ethical approval for the final study. To all my supervisors I am extremely grateful.

Jos Pacilly, engineer at the Leiden University phonetics laboratory, cannot be thanked enough. With his invaluable and tireless

support he made every study possible by, among other things, writing *Praat* scripts, thinking methodologies through, building experimental setups and selecting optimal hardware and software for them, and showing interest at every stage of the project.

Special thanks are due to my colleague Zohreh Shiamizadeh. Throughout the years we discussed countless issues about phonetics, academic writing and publishing, Dutch habits, international politics and Iranian food. Thank you so much!

I am also indebted to the many other colleagues who directly or indirectly helped me with theoretical or practical issues: Kate Bellamy, Monique Bisschop, Johanneke Caspers, Yiya Chen, Lisa Cheng, Rongjia Cui, Elly Dutton, Nanda Ernanda, Andreea Geambasu, Aliza Glasbergen-Plas, Rob Goedemans, Margarita Gulian, Anne van der Kant, Elena Karvovskaya, Olga Kepinska, Viktorija Kostadinova, Saskia Lensink, Qian Li, Marieke Meelen, Marie-Catherine Michaux, Gareth O'Neill, Amanda Post da Silveira, Bobby Ruijgrok, Franziska Scholz, David Shakouri, Heleen Smits, Benjamin Suchard, Mulugeta Tsegaye, Daniil Umanski, Maaike van Naerssen, Marijn van 't Veer, Xander Vertegaal, Cesko Voeten, Mang Wang, Jurriaan Witteman, Junru Wu, Yang Yang, Bahar Zhalehgooyan, Sima Zolfaghari, Ting Zou, and any colleagues that I have forgotten to mention.

Outside LUCL many people have helped me. I thank Peter-Paul Boermans, Esther Scholink, and Walter Verlaan of LUMC's ENT department for sharing their equipment and audiological perspective on the setup of experiments and for helping me to recruit participants. I also thank a diverse group of researchers from many different institutes, including Aoju Chen, David Dekker, Paula Fikkert, Paul Govaerts, Lauren Harris, Lizet Ketelaar, Sylvia Mozziconacci, Carolien Rieffe, Anna Safar, Dirk Jan Vet, and researchers from the Oorgroep in Antwerp. Thanks to my academic friends David Morris, Richard Penninger, Stefano Cosentino, Carina Pals, David Greenberg, and Cherith Webb corners of the global academic world became much more reachable.

I received much support for statistical analysis. I am very grateful to Vincent Buurman, Joost van Ginkel, Willem Heiser, Pieter Kroonenberg, Kees Verduin, and Ron Wolterbeek.

I was lucky enough to collaborate with students in the minor program Brain and Cognition and in the Pre-University program. Without the admirable commitment of Stijn Frima, Linda ter Beek, Arian Khoshchin, Annemarie Tol, Sjors Bootsman, Maaike Voorhoeve and Brendan Analikwu three experiments would not have been performed. In particular, I would like to mention Giorgos Dritsakis, who at the time excelled as a master's student and in the meantime has earned his PhD degree on the experience of music by cochlear implant users at the University of Southampton. Thank you very much.

Equally indispensable were the more than 200 participants – among whom adults, parents and children with and without cochlear implants – for taking part, sometimes for sessions up to two hours, in one of the studies that make up this thesis. My sincere appreciation goes out to all these people.

Finally, I thank friends and family for the continuous interest they showed in my thesis.

Chapter **1**

Introduction

An estimated 360 million people (over 5% of the population) suffer from hearing loss worldwide, according to an estimate by the World Health Organization ("Deafness and hearing loss," 2015). The prevalence of deafness in the Netherlands is approximately 0.7% as of 2016 (Lamoré, 2016). Hearing loss (presumed equivalent to the impairment associated with being 'hard of hearing') is defined as a "hearing disorder, whether fluctuating or permanent, which adversely affects an individual's ability to communicate" and deafness as "a hearing disorder that limits an individual's aural/oral communication performance to the extent that the primary sensory input for communication may be other than the auditory channel" (American Speech-Language-Hearing Association, 1993). Possible causes of hearing loss are hereditary and acquired. Among hereditary causes, the most prevalent is connexin-26 deficiency (DFNB1) in the GJB2 gene. Possible syndromal hereditary causes are Waardenburg's syndrome and Usher's syndrome. Among the acquired causes are meningitis and reactions to ototoxic drugs. One to two in every thousand children is born with bilateral sensorineural hearing loss (Gravel & Tocci, 1998).

Hearing loss can have different repercussions for individual listeners. Following the WHO's International Classification of Functioning, Disability and Health (Stephens & Kerr, 2000), they experience problems detecting, recognizing and identifying sounds, appreciating sound quality, tolerating loud sounds, understanding speech in silence and noise, understanding spoken emotions, and localizing sound sources. Moreover, their education and career opportunities are compromised (Lang, 2002). Neurocognitive effects of (untreated) auditory deprivation have also been reported, such as problems with working memory (Marschark, Lang, & Albertini, 2002) and socio-emotional control (such as psychopathology; Theunissen, 2013), cognitive decline in older listeners and degradation of auditory cortex and its takeover by the visual modality (Glick & Sharma, 2016).

The observations above demonstrate the severity of the problem of hearing loss, both at the level of the individual listener and at the level of global socio-economic functioning. A variety of medical interventions are available to treat hearing loss, such as conventional hearing aids (sound amplification), bone-anchored hearing aids (BAHA; sound conduction through bones) and cochlear (CI) and auditory brainstem implants (ABI), the suitability of which depends, among other factors, on the severity and type of an individual's hearing loss. This thesis focuses on the cochlear implant.

The remainder of this chapter consists of sections introducing core aspects of this thesis. Section 1.1 discusses the goal and history of cochlear implantation and the mechanism behind cochlear implant hearing. Section 1.2 describes the phenomenon of prosody, the aspect of speech which forms the linguistic focus of this dissertation. Section 1.3 covers the distinction between perception and production of speech, both of which are investigated in this dissertation. Section 1.4 focuses on the acquisition of language by children with CIs, as a subset of the studies reported in this dissertation involve that population. Section 1.5 briefly discusses the usage of vocoders for research into CI hearing, a method that was adopted in three of the

studies in this thesis. Finally, an overview of the chapters of this thesis and the corresponding perspectives and hypotheses regarding the processing of prosody with CIs is provided.

1.1 Cochlear implants

Cochlear implants are prostheses of the inner ear partially restoring hearing for severely to profoundly deaf children and adults by providing an electrical reconstruction of sound directly to the auditory nerve. The basic functioning of a CI is based on the vocoder technique (see section 1.5). The functioning involves capturing of sound by a microphone attached near the outer ear, signal analysis by a speech processor, transmission of the processed signal to a transmitter attached to the scalp and subsequent electromagnetic transcutaneous transmission to a receiver on the inside of the skull, and finally to a set of between 12 and 22 electrodes inserted into the cochlea. The array of electrodes mimics the tonotopic organization of the basilar membrane by presenting lower frequencies with electrodes situated at the apical end of the cochlea and higher frequencies at the basal end of the cochlea. Of the many design options that exist, some of the more important ones concern the number of channels (electrodes), the shape of the analysis and synthesis filters, the rate and configuration of stimulation, and the position of the array in the cochlea. Detailed descriptions of CI design and functioning have been provided elsewhere (Wilson & Dorman, 2009).

Cochlear implantation has first been performed by Parisian electrophysiologist André Djourno and otolaryngologist Charles Eyriès in 1957 on a deaf patient (Djourno & Eyries, 1957; Eisen, 2009). With their single-channel implant, the recipient was able to discriminate lower from higher frequencies and environmental sounds but had no speech understanding beyond a small number of words. Otologists William House in Los Angeles, Blair Simmons at Stanford University, and Robin Michelson at the University of California-San

Francisco (UCSF) independently pursued this work with single-channel implants in the 1960s, allowing useful hearing sensations to deaf patients but also encountering issues with biocompatibility of the device. Concerns were raised by scientists regarding the feasibility of electrically reconstructing a signal as complex as that of speech (Jongkees, 1978; Lawrence, 1964; Simmons, 1966). However, in 1975, the National Institutes of Health (NIH) acknowledged the benefits of CI by showing improvements in speech production, lip reading and quality of life, spurring further research and its financial support (Bilger, 1977). Scientists at the UCSF, as well as Graham Clark at the University of Melbourne in Australia developed multichannel CIs, which later became the now commonly used Advanced Bionics Clarion and Cochlear Corporation's Nucleus devices, respectively. In the 1980s, Food and Drug Administration (FDA) approvals were granted for adult CI recipients and children as young as two years of age, allowing research to shift from safety to outcome issues. In 1991, the now common continuous interleaved sampling (CIS) strategy, a design whereby electrodes are never activated simultaneously to reduce channel interactions, was shown to further improve speech understanding (Wilson et al., 1991). Since then, a large variety of implant designs and speech coding strategies have been developed and the scientific and social acceptance of CI have grown considerably (Blume, 1999; Christiansen & Leigh, 2004; Wilson & Dorman, 2008).

The primary aim of CIs is to allow speech understanding. The candidacy criteria for cochlear implantation are multifaceted, evaluated on a case by case basis and differ per country, but in general some of the important eligibility criteria for CI are (i) that individuals and their relatives have realistic expectations of its benefits; (ii) that they are motivated to undergo the surgical procedure and persevere the ensuing rehabilitation; (iii) that they benefit less from conventional hearing aids; and (iv) that there is an absence of medical contra-indications, such as inner ear malformations. On the basis of these criteria, as much as 40% of cases presented led to specialists' decision

not to proceed to implantation in the United Kingdom between 1990 and 1994 (Summerfield & Marshall, 1995). However, due to improved implant technologies and benefits, candidacy criteria have become less stringent over the last decades (Niparko, Lingua, & Carpenter, 2009). The number of CI recipients have grown exponentially since these developments, with over 300,000 users worldwide as of 2014 (Wilson, 2014) and over 6,500 in the Netherlands as of 2015 ("Aantal implantaties in Nederland," 2016). Based on research showing that postlingually deafened adults improved their hearing scores after implantation as they showed up to 80% preimplantation phoneme perception scores, the Leiden University Medical Center's ENT department decided to adopt this preimplantation score as an upper limit for CI indication, as even higher scores provided no benefit of cochlear implantation (Snel-Bongers, Netten, Boermans, Briaire, & Frijns, submitted).

CIs have proven successful in allowing recipients to develop or process spoken language more efficiently than deaf children with a conventional hearing aid (Knoors, 2008; Lenden & Flipsen, 2007). This was shown by a number of outcomes: (i) a vocabulary growth at about 60% of normally hearing (NH) children's rate (Blamey et al., 2001; Geers, 2003); (ii) the production of longer sentences (Geers, 2003); (iii) improved sentence understanding (Geers & Moog, 1994); (iv) improved phoneme production (Geers & Moog, 1994); (v) speech perception abilities in quiet conditions within the norms of normally hearing individuals and communication over the telephone (Beadle et al., 2005); (vi) improved reading skills (Johnson & Goswami, 2010); and (vii) improved production of narratives (Boons et al., 2013; Crosson & Geers, 2001). Implantation can also allow participation in mainstream education and favorable career opportunities (Spencer, Gantz, & Knutson, 2004); however, those results are inconclusive and particularly mixed due to individual variation (Marschark, Rhoten, & Fabich, 2007; Punch & Hyde, 2011; Stacey, Fortnum, Barton, & Summerfield, 2006; Thoutenhoofd, 2006). The effects of CI on quality of life have so far also been inconclusive due to theoretical and

methodological inconsistencies and results between studies (Knoors, 2008). Nevertheless, with the above facts and figures about the device's psychophysical merits taken together, the CI could count as the most successful artificial sensory prosthesis.

Despite these merits, CI hearing faces a number of challenges. The input is degraded relative to normal hearing as a result of, among other factors, a limited number of effective electrodes, channel interactions, the single-sided character of the hearing (in case of unilateral implantation), possible cochlear malformations and dead regions of the auditory nerve, malfunctioning electrodes, and frequency shifts due to shallow electrode insertion depths (Wilson & Dorman, 2009). Of the three main dimensions that the auditory signal is composed of – the temporal, the dynamic and the pitch dimension – variations in the pitch dimension and, to a lesser degree, in the dynamic dimension are difficult to discriminate for CI recipients (Meister, 2011; Shannon, 2002). In the perception of speech, NH listeners rely on some dimensions more than others, depending on the listening task. Reliance means that when a dimension is unavailable for whatever reason, this compromises the recognition of the linguistic information in the speech signal. When a dimension provides information about speech, it is referred to as a 'cue' and the relative reliance by listeners on the dimensions as 'cue weighting'.

Due to CI users' perception difficulties, the voice's pitch (fundamental frequency or F0) and, to a lesser extent, the intensity dimensions pose notorious problems for them, prompting them to weight cues differently than normally hearing people do by balancing their reliance from F0 cues (partly) towards temporal and dynamic cues. These input and sound processing issues compromise their music perception, speech perception, spectral resolution, sound source localization, hearing in noise, the perception of acoustically less prominent morphosyntactic endings in languages such as Dutch and English, such as the suffix *-t* in *werkt* (third person singular of 'to work') which is non-syllabic and short (Hammer, 2010; Nikolopoulos, Dyar, Archbold, & O'Donoghue, 2004; Svirsky, Stallings, Lento,

Ying, & Leonard, 2002), and more general capacities such as verbal working memory and serial data recall (Nittrouer, Caldwell-Tarr, & Lowenstein, 2013; Pisoni, Kronenberger, Roman, & Geers, 2011). In view of these possible consequences, cue weighting is further studied in this thesis.

Linguistic performance by CI users notoriously shows much individual variation (Kane, Schopmeyer, Mellon, Wang, & Niparko, 2004; Peterson, Pisoni, & Miyamoto, 2010), begging the question what factors underlie those differences. For instance, performance on recognition of monosyllables ranges between almost zero percent correct to ceiling level after two years of implant experience, with standard deviations up to 30% (Wilson, 2006). The factors underlying this variation can be divided into demographic factors, psychosocial factors, device factors and neurocognitive factors. Demographic factors are factors such as the duration of hearing loss before implantation, the age at implantation (whereby children implanted at two years or younger tend to outperform the later-implanted children), the duration of implant usage and the family's socio-economic status and size (Anderson et al., 2004; Boons et al., 2012; Colletti, Mandalà, Zoccante, Shannon, & Colletti, 2011; Geers, Nicholas, & Sedey, 2003; Harrison, Gordon, & Mount, 2005; Leigh, Dettman, Dowell, & Briggs, 2013; McConkey Robbins, Green, & Waltzman, 2004; Niparko et al., 2010; Sharma, Dorman, & Kral, 2005; Sharma et al., 2004). Psychosocial factors include the presence of additional disabilities such as mental, emotional and social problems (Edwards, 2007; Shin et al., 2015). Device factors are factors such as the number of electrodes, the analysis and synthesis filter's shape, and the array's insertion position (Geers, Brenner, & Davidson, 2003). Finally, among the neurocognitive factors are (verbal) working memory, and intra- (auditory) and cross-modal (visual) neural reorganization due to auditory deprivation (AuBuchon, Pisoni, & Kronenberger, 2014; de Hoog et al., 2016; Finke, Buchner, Ruigendijk, Meyer, & Sandmann, 2016; Nittrouer et al., 2013; Pisoni, 2000). Of these, the duration of hearing loss, age at implantation, and the duration of implant usage, as

well as socio-economic status tend to surface as some of the main predictors of language performance outcome after implantation (Blamey et al., 2013; Holden et al., 2013; Moon et al., 2014). Although many factors have been identified, the individual variation is still not fully understood.

1.2 Prosody

Prosody is speech content that cannot be predicted from the information of individual segments or the coarticulation of subsequent segments (Lehiste, 1970; Rietveld & van Heuven, 2009). It is primarily conveyed by means of variations in F0, intensity and durations of any structural level of an utterance. The functions of prosody can be divided into linguistic, on the one hand, and emotional and indexical functions, on the other (Rietveld & van Heuven, 2009; Witteman, van IJzendoorn, van de Velde, van Heuven, & Schiller, 2011). Linguistic prosody pertains to information about the meaning of an utterance, such as phrasing by means of pauses, lengthening and intonation, word stress, information structure by means of pitch accents (the marking of new vs. known information in sentences) and sentence type (statement vs. question). Emotional and indexical prosody convey information about the emotion or attitude (e.g., irony) and demographics, such as identity, gender, age, dialect and health, of the speaker. The importance of emotion understanding in speech has been highlighted by research pointing to a correlation between emotional identification capacities, but not word identification scores, and quality of life (Schorr, Roth, & Fox, 2009).

A third type of prosody, which is not usually acknowledged independently in the literature, could be called basic prosody. Basic prosodic measures have no linguistic, emotional or indexical function. If anything, they could have an emotional or indexical function, but that is only relevant when it has been shown that changes in the parameters correlate with emotion or speaker identification scores in a

listening task. Without such demonstrated function, between-speaker and between-utterance variations could be considered ‘basic’, possibly stochastic prosodic variations. For instance, utterance duration or F0 declination could serve to infer emotion or speaker characteristics, but when such a link is not established, those measures would still count as basic. In Chapter 2 of this thesis, such basic prosodic measures were compared between speech of CI users and NH peers. Measures that would appear to distinguish between the two groups, could then be considered indexical prosodic measures.

Given this central role of prosody in development and usage of language together with CI users’ perceptual problems, it becomes clear that by missing out on important prosodic information such as information structure and indexical (speaker) information (Gilbers et al., 2015; Massida et al., 2011; Meister, Fursen, Streicher, Lang-Roth, & Walger, 2016), this group of language users is at risk of late and/or deviant language acquisition (Chatterjee & Peng, 2008; Giezen, Escudero, & Baker, 2010; Kong, Cruz, Jones, & Zeng, 2004). This warrants further research into the questions of what types of information are available to CI users, what the mechanism behind their capabilities and limitations is, how children acquire prosody, and if a limitations in perception have repercussions for production. This thesis intends to fill in some of these gaps. The last of these issues is discussed in the next section.

1.3 Speech perception and production

The relationship between speech perception and production can be approached from at least two different angles, that of its development influenced by a speaker’s hearing history (this could be called the ‘diachronic’ perspective) and that of its functioning during speech processing (the ‘synchronic’ perspective). First of all, the development of the relationship between speech perception and production seems in part to depend on an individual’s hearing history. For instance, both

congenitally deaf speakers (Osberger & McGarr, 1982) and speakers with acquired deafness (Waldstein, 1990) produce deviant speech, showing that deficient input has ongoing consequences for the output, even after the supposed establishment of an articulation routine. Speakers with acquired deafness, however, continue to produce normal speech for some time following the onset of deafness, which indicates that the acquired articulatory goals are robust enough to support proper production for some time without direct auditory feedback (Guenther, Ghosh, & Tourville, 2006).

Second, the functioning of the relationship between speech perception and production has been modeled by the Directions Into Velocities of Articulators model (DIVA; Guenther, 2006). In this model, which is based on neurolinguistic evidence, articulatory actions are viewed as motor programs for sound, syllables or sequences of syllables. These actions feedforwardly project system-internal abstract predictions of the structure to be produced, against which the auditory feedback provided by the actual output is checked for adequacy. In case of an inadequate output, an error is detected and the feedforward commands are updated. The output can for instance be inadequate as a result of disruption or feedback delay during articulation (Burnett, Freedland, Larson, & Hain, 1998; Perkell et al., 2007; Purcell & Munhall, 2006), because the speaker is still acquiring speech, because the speaker's articulators are still maturing, or because of deafness. The adequacy of the output is based on speech input provided by ambient speech. Deafness, therefore, may result in deviant speech because inappropriate sound structure representations have been established. The process of the evaluation of speech output against internal representations has been labeled 'monitoring' by other researchers (Levelt, 1983).

Together, the above observations suggest that proper speech perception is required for proper speech production and possibly vice versa as well. The 'diachronic' and 'synchronic' aspects of the relationship between perception and production capabilities are both relevant to this thesis, because children with cochlear implants by

definition have an abnormal hearing history and because their auditory input, even if stable since implantation, is degraded in relation to that of normally hearing individuals. Given the possible relationship between language perception and production capabilities, in combination with CI users' deviant perception performance and life history, it is therefore plausible to assume that CI recipients' production performance is also deviant. This hypothesis is tested in Chapter 6 of this thesis.

A number of studies have probed the possible correlation between perception and production in CI users. Peng (2005) tested school-aged children on the perception and production of the intonation of sentence type (declaratives vs. statements) and found that children with a good tone production also showed a good tone perception, but not necessarily vice versa, suggesting that for CI children good perception precedes good production. According to Peng (2005), the observations might reflect an indirect relationship between perception and production, in that other factors, such as age at implantation, might differentially underlie perception and production. In a series of experiments, O'Halpin (2010) tested prosody perception and production performance of school-aged children with and without cochlear implants. The participants indicated (a) whether utterances were pronounced as compounds or phrases (e.g., *greenhouse* vs. *green house*), (b) which of two words in a sentence carried focus (*It's a GREEN door* vs. *It's a green DOOR*, where capitals mark focus) or (c) which of three words carried focus (*The DOG is eating a bone* vs. *The dog is EATING a bone* vs. *The dog is eating a BONE*). In another experiment, the participants' production of these phrases was evaluated for appropriateness by a panel of NH listeners. The author reported no correlations between most of the perception and production scores. In a study on 47 primary-school-aged children with cochlear implants and 40 peers with hearing aids, Blamey et al. (2001) found a correlation between word and sentence comprehension performance, on the one hand, and intelligibility measures of spontaneous utterances, on the other hand. Speech

intelligibility scores in prelingually deafened CI users predicted post-implantation speech perception scores, whereas preimplantation speech perception scores with hearing aids constituted a weaker predictor (van Dijkhuizen, Beers, Boermans, Briaire, & Frijns, 2011; van Dijkhuizen, Boermans, Briaire, & Frijns, 2016). Other studies have shown mixed results regarding the correlation between perception and production by CI children, such as a lack of correlation between the Beginner's Intelligibility Test (Osberger, 1994) and the Prosodic Utterance Production test (Bergeson & Chin, 2008) or a correlation between emotion imitation and recognition (Lyxell et al., 2009; also see, Spencer et al., 2004).

These studies together demonstrate that it is at present unclear to what extent perception and production of speech are correlated in children with cochlear implants, as was also concluded in a recent review (Cysneiros, Leal, Lucena, & Muniz, 2016). This thesis joins this debate by studying perception and production of two types of prosody (linguistic and emotional) by CI children controlling for general linguistic and emotional maturation. The next section discusses the general background for this thesis regarding language acquisition by implanted children.

1.4 Language acquisition by children with cochlear implants

Language acquisition is thought to start as early as approximately three months before birth, when the fetus perceives mainly relatively loud and low-frequency (under 1000 Hz) environmental, bodily and some speech sounds from the mother (Graven & Browne, 2008). This is evidenced by newborns' preference for the maternal language over other languages (Mehler et al., 1988; Moon, Lagercrantz, & Kuhl, 2013). Auditory experience further shapes the very early stages of language acquisition by means of infant-directed speech, perceptual tuning in the first 6 months of life (Kuhl et al., 2006; Werker & Tees,

1984), and by guiding the perception of focus, syntactic information and phrase boundaries (Soderstrom, Seidl, Nelson, & Jusczyk, 2003).

Prosody plays a special role in acquisition. As a result of prenatal imprinting, newborns show a preference for native over non-native prosody, showing that the speech information has been processed (Moon, Cooper, & Fifer, 1993). Due to the intrauterine frequency selectivity, the speech sounds that penetrate are mainly prosodic, i.e., rhythmic and intonational. After birth, 'motherese' (prosodically exaggerated child-directed speech by caregivers) draws infants' attention to important components in speech (Liu, Kuhl, & Tsao, 2003; Thiessen, Hill, & Saffran, 2005). Prosody continues to play a pivotal role in language acquisition in the following months and years. At the age of approximately seven months, infants use prosodic patterns to segment the speech stream. Prosody thus paves the way for word learning (Johnson & Jusczyk, 2001). We can therefore conclude that the development of prosody starts early, probably forming the first stage in language acquisition, and proceeds to play an essential role in children's language acquisition until the young-adolescent age.

Given the importance of hearing experience for early language acquisition, it is not surprising that language acquisition develops differently in children with hearing loss. Most deaf children have two hearing parents (Mitchell & Karchmer, 2004), and consequently do not receive native sign language input. Deaf children can have delayed canonical onset and a restricted repertoire of babbling (Kuhl & Meltzoff, 1996; Oller & Eilers, 1988). They possibly do not catch up with NH peers (Vaccari & Marschark, 1997). This inability to catch up after a delay despite intensive efforts is thought to be due to a sensitive period in acquisition, i.e., an age window during which acquisition has to start in order to be able to reach a normal level as the end stage (Lenneberg, 1967; Werker & Hensch, 2015).

Congenitally deaf children with cochlear implants present an interesting case of atypical language development, since they experience a clear-cut delayed onset of spoken language acquisition, while enjoying – in most cases – a normal upbringing. For

congenitally deaf implanted children, the onset of spoken language acquisition coincides with the activation of the implant (Connor, Craig, Raudenbush, Heavner, & Zwolan, 2006; Tye-Murray, Spencher, & Woodworth, 1995). The study of pediatric CI recipients therefore allows the investigation of the effect of a delayed onset on language acquisition and the role of early non-linguistic maturation. CI children's language acquisition is delayed and can also be deviant relative to that of NH peers (Geers, Nicholas, Tobey, & Davidson, 2016; Robinson, 1998). Cochlear implantation improves speech production but after several years of implant usage, in some recipients, it still deviates from that of NH peers (Geers, Tobey, Moog, & Brenner, 2008).

Despite these differences, several studies observed a similar prosodic development in CI and typically developing (TD) children (Snow & Ertmer, 2009, 2012; Vogel & Raimy, 2002; Wells, Peppé, & Goulandris, 2004). Snow and colleagues (Snow & Ertmer, 2009, 2012) modeled children's intonational development until 24 months of age in terms of stages in F0 range on word accents. They found that CI children matched TD children's alternation between stages of increased and decreased pitch range. However, the CI recipients' development shows an interaction between implantation age and duration of implant usage, whereby children implanted after 24 months of age showed a development that was more advanced than would be expected based on their hearing age (i.e., the time since implantation) and whereby children implanted before 24 months of age showed a delay in their development. This suggests that maturation plays a role in prosody development in that some components of it continue without auditory input.

In one of the experiments in this study, long-term effects of cochlear implantation on emotional and linguistic prosody perception and production are investigated by comparing school-age CI with NH children. Apart from probing possible deviations or delays in the acquisition of these four quadrants of prosody processing (linguistic prosody production, linguistic prosody perception, emotional prosody

production, and emotional prosody production) and the correlations between them, we test the hypothesis that emotional prosody is less delayed than linguistic prosody because the former is supposedly less dependent on rule-learning derived from input than the latter.

1.5 Vocoders

Sound processing in cochlear implants is based on the channel vocoding technique. Channel vocoders (short for voice encoder) are signal processing algorithms designed to reconstruct a sound signal in a parametrized way. The signal processing procedure follows two basic steps: analysis and resynthesis. In the analysis step, incoming sound is band-pass filtered into a number of contiguous frequency bands (channels). In the resynthesis step, the signal is resynthesized (with a reduced information load) by multiplying the dynamic envelope of each channel with a chosen source signal, band-pass filtering the resulting channels by the same filters as for the analysis part, and finally adding those channels together. The signal source can either consist of noise (noise vocoder) or of a sinewave (tone vocoder) (Loizou, 2006).

In CI models, variation exists in the settings that the vocoding technique allows to manipulate. Most importantly, the number of channels is typically between 12 and 22 and the source signal consists of a constant train of pulses delivered to the electrodes with a rate of several hundreds to several thousands of pulses per second per electrode. Moreover, the shape of the analysis and synthesis filters influences the amount of spectral smearing between filters. Steeper filter slopes cause less overlap than shallower filter slopes, improving discriminability of frequencies coded in different bands (Friesen, Shannon, Baskent & Wang, 2001).

Researchers use vocoder simulations of CIs to study CI hearing. This allows them to recruit participants with normal hearing, who are more numerous and form an audiologically more uniform

group than CI users. Moreover, it allows researchers to manipulate and study signal processing parameters that cannot be manipulated in CI users, since the settings in their devices are fixed. Results from studies using vocoders could, however, inspire the design of implants with improved settings. In this thesis, for the above reasons, vocoders were used to test the effect of filter slope on the discriminability of intonational and rhythmic variants of spoken sentences and musical fragments.

Limitations of vocoders as CI simulations should, however, be taken into account. The details of the signal processing procedure, the functioning of the ear, and the audiological background of the participants all differ between hearing and implanted individuals. Results from vocoder simulations cannot therefore be generalized to the population of CI users without caution. Ideally, tests with vocoders are followed up by tests with actual CI users in order to elucidate which vocoder settings most closely model the performance by the clinical population. These limitations of vocoder simulations will be dealt with in more detail in the respective chapters.

1.6 Overview of this thesis

This thesis investigates the processing of prosody by CI users from a number of perspectives, covering the mechanism and development of perception and production of the major types of prosody. These perspectives are covered by a number of broadly stated hypotheses of which more specific formulations are tested throughout different chapters. The motivations for these hypotheses will be stated in the chapters in which they are tested.

First of all, we investigate prosody by making a distinction between three major types, namely linguistic, emotional, and basic prosody, and studying one of them separately (basic prosody, in one study, Chapter 3) or comparing linguistic and emotional prosody (in three studies, Chapters 4, 5, and 6). There are fundamental differences

between linguistic and emotional prosody; e.g., knowledge of emotional prosody is possibly innate and universal, its cerebral processing right-lateralized and its realization of a gradient nature, whereas linguistic prosody is probably learned, less lateralized (Witteman et al., 2011) and its realization more discreet and rule-based. They might therefore be perceived and produced differently. A third type, basic prosody, is postulated as a rest category of prosodic measures that are performed without linking them to a linguistic or emotional function and is separately tested. We hypothesize that emotional prosody is differently recognized (**Hypothesis 1a**) and realized (**Hypothesis 1b**) than linguistic prosody. Second, emotional prosody perception and linguistic prosody perception are compared to music perception (elaborated below). It is predicted that emotional prosody is less correlated to music than linguistic prosody (**Hypothesis 1c**). Finally, we hypothesize that emotional prosody perception and production are less correlated than linguistic prosody perception and production (**Hypothesis 1d**).

The second perspective entails the distinction and relationship between speech perception and production. Perception (in three studies, Chapters 3, 4, and 5) and production (in one study, Chapter 1) are studied separately or in direct comparison (in one study, Chapter 6). We hypothesize that both perception (**Hypothesis 2a**) and production (**Hypothesis 2b**) are deviant in CI users, because they develop as an integrated system, which surfaces as a within-participant correlation between perception and production scores (**Hypothesis 2c**).

The third perspective is that of the relationship between prosody perception and music perception, two disciplines in which the acoustical dimensions of rhythm and melody are fundamental. In one study (Chapter 4), the hypothesis that NH listeners can be cue-specifically trained with musical materials to recognize musical melodies based on either melody or rhythm cues is tested (**Hypothesis 3**). Further, this training effect could transfer to reliance on the non-trained cue in melody perception (cross-cue transfer), on the trained

cues in prosody perception (cross-domain transfer) or to prosody perception for both cues (cross-cue plus cross-domain transfer) (as this does not involve a directional hypothesis, this issue is referred to as the **Transfer Issue**).

The fourth perspective is that of the mechanism of CI prosody hearing. CI users weight the cues they use to process prosody differently than NH listeners do. In this thesis, we compare prosody perception with the availability of temporal and F0 related cues by these two groups. Based on previous literature, **Hypothesis 4a** holds that of these two cues, CI users rely relatively heavily on temporal cues, as compared to their NH peers. **Hypothesis 4b** states that this cue weighting is reflected in speakers' speech output in that F0 related basic prosodic measures of CI users will deviate more than temporal prosodic measures. Within perception, it is hypothesized (**Hypothesis 4c**) that reduced channel interaction, as manipulated by steepening of channel filter slopes in vocoder simulations of CI hearing, will improve F0 perception, but not temporal perception.

The final perspective is that of the development of prosody in children. Two of the studies in this thesis were (retrospectively) performed with children with and without CIs (Chapters 2 and 6). We conjecture that language acquisition of CI children is delayed relative to that of NH peers by as much as the time until implantation (**Hypothesis 5a**), but that this delay is longer for prosody perception than for prosody production (**Hypothesis 5b**) and longer for linguistic prosody than for emotional prosody (**Hypothesis 5c**), and that CI children (partially) catch up with increasing experience with their device (**Hypothesis 5d**).

Chapter 2 reports a retrospective study of basic prosodic measures of prosody in spontaneous speech recordings of control children without and hearing-aged matched children with cochlear implants. The prosodic measures are categorized, from 'easy' to 'difficult' for CI users, as temporal, intensity related and F0 related and measured at 18, 24 and 36 months after implantation (for CI recipients) or birth (for

NH children). This study combines the perspectives of production, mechanism and development and tests **Hypotheses 2b**, **4b**, **5a**, and **5d**. It is predicted that production differs most for F0 related, less for intensity related and least for temporal measures and that any delay that exists with hearing-aged matched controls will be (partially) caught up after 36 months of CI experience, but more so for 'easier' measures.

Chapter 3 uses vocoder simulations of cochlear implant hearing to test the role of spectral smearing for intonation perception by normally hearing Dutch adults. Spectral smearing is the effect whereby the activation in a channel overlaps the area of a neighboring channels resulting in mixed (frequency) percepts. Sharper channel filters (i.e., with a steeper filter slope, expressed in dB/octave) reduce overlap and guarantee better F0 and intonation perception. Noise vocoder simulations are used instead of actual CI users, because they allow the manipulation of sound processing parameters (such as filter slopes) that could play a role in CI hearing but that the device of a given user does not allow to be manipulated (they could, however, be manipulated by redesigning a device). This study combines the perspectives of perception and mechanism and tests **Hypotheses 2a** and **4c**. Participants decide if naturally recorded but manipulated utterances that differ only in their F0 contour sound as a surprise, as news or as a predictable utterance. This setup, in which participants are asked to pay attention to the interpretation of the utterance, maximizes the likelihood that they listen to the stimuli as linguistic (intonational) and not just as acoustic (frequency varying) stimuli. It is hypothesized that intonation identification will be more accurate with a 40 dB/octave than with a 20 dB/octave condition, but that for both conditions it will be less accurate than in a control condition without vocoding.

Chapter 4 uses the same setup as the experiment described in Chapter 3 but extends its scope by using more different filter slopes (ranging

between 5 and 160 dB/octave), by making a distinction between emotional and linguistic prosody, and by making either temporal, F0 related or both cues available. This study combines the perspectives of the distinction between the two major types of prosody (emotional and linguistic), that of the perception and that of the mechanism and tests **Hypotheses 1a**, **2a**, **4a**, and **4c**. In this pair of experiments, NH Dutch adults decide (focus test) which of two words in a phrase carries sentential focus, or (emotion test) which of two emotions (happy or sad) is expressed in a phrase, whereby the phrases are highly similar to those in the focus test in order to justify a comparison between results of those two tests. These tests are repeated with and (as a control condition) without noise vocoding. It is hypothesized that intonation discrimination will improve with increasing filter slope and that this effect is smaller when temporal cues are available than when only F0 cues are available. The pattern of results might or might not differ between emotional and linguistic prosody. This experiment also functions as a validation for the stimuli, which are also used in several experiments in Chapter 6. Near-ceiling performance with the non-vocoded condition shows which of the stimuli appropriately convey focus position and emotions, thereby validating them for usage in further experiments.

Chapter 5 compares music perception to prosody perception. For the musical task, NH Dutch adults receive a short training to enhance their perception of either temporal (one group) or frequency (second group) perception of tone-vocoded stimuli and subsequently decide which of four possible well-known melodies was heard in conditions with only the rhythm of the melody available, only the tonal changes (but with all notes having the same duration) or both. They are also tested on emotional and linguistic prosody perception with the same cue conditions. The linguistic tasks are similar to those performed in the experiments in Chapter 4. This study combines the perspectives of the distinction between emotional and linguistic prosody, perception, the mechanism and music, and tests **Hypotheses 1a**, **1c**, **3**, **4a** and the

Transfer Issue. It is hypothesized that NH participants' perception in post-training tests is selectively enhanced for the trained cue. Further, this training effect could either transfer to non-trained cues in the same domain (i.e., within music; cross-cue transfer), in another domain but only for the same cue (i.e., to language; cross-domain transfer) or to another domain and another cue (cross-domain and cross-cue transfer).

Chapter 6 reports a set of experiments performed with young school-age children with and without CIs. They performed four core tests gauging their capabilities in the perception and production of both emotional and linguistic prosody. In the perception tests, temporal and F0 cues or both cues were made available. Additionally, participants performed three control tests aimed at probing their baseline level of non-verbal emotional development, of general linguistic development, and of basic picture identification and naming skills. Parents or caregivers completed a questionnaire about their children's language and medical background and the parents' socio-economic status. This set of experiments combines most of the perspectives of this thesis, viz. the distinction between linguistic and emotional perspectives, perception and production, the mechanism, and the development. It tests **Hypotheses 1a**,**b**,**d**; **2a**,**b**,**c**; **4a**,**b**; and **5a**,**b**,**c**,**d**.

Chapter **2**

Basic measures of prosody in spontaneous speech of children with early and late cochlear implantation

Abstract

Research on prosody in speech produced by children with cochlear implants (CI) has revealed deviations from the speech of normally hearing (NH) peers, such as a high fundamental frequency (F0), elevated jitter and shimmer, and inadequate intonation. However, three important dimensions of prosody (temporal, intensity, and spectral) have not been systematically investigated or compared in production research. Given that in general the resolution in CI hearing is best for the temporal, followed by the intensity, and worst for the spectral dimension, we may expect that this hierarchy is also present in the speech production.

9 Dutch Early Implanted (EI), 9 Late Implanted (LI; division at 2 years of age) children and 12 hearing age matched NH controls were tested at 18, 24, and 30 months after implantation (CI) or birth (NH). We expected that (1) there would be differences between CI recipients and controls on prosodic speech measures, (2) they would be smallest for temporal measures, followed by intensity measures and largest for

spectral measures, (3) they would be larger for later than for earlier implanted children (4) and they would diminish with increasing device experience.

From spontaneous speech data, 1,937 utterances were extracted. Of these utterances, nine outcome measures along the spectral, intensity and temporal dimensions were subjected to Principle Component Analysis (PCA) and, using Linear Mixed Modelling, compared between Group, Session, and Gender, as well as their interactions.

PCA combined three measures into one, leaving three temporal and three spectral measures. On most measures, interactions of Group and/or Gender with Session were significant. For CI recipients as compared to controls, performance on temporal measures was not in general more deviant than spectral measures, although differences were found for individual measures. LI had a tendency to be closer to NH than EI. Groups converged over time.

The hypothesis regarding differential deviations for the different phonetic dimensions was not supported. This suggests that the appropriateness of the production of basic prosodic measures does not depend on auditory resolution. Rather, it seems to depend on the amount of control necessary for speech production. Chronological age, hearing status and gender of the speaker influence the development of the measures.

2.1 Introduction

Most people who suffer from severe or profound hearing loss are nowadays treated with cochlear implantation (CI), which partly restores their hearing. Despite major advantages in spoken communication relative to pre-implantation, the CI recipients' hearing situation is not like that of normally-hearing (NH) people. Characteristics of the device and the CI recipient's auditory history limit, in particular, the perception of speech prosody (Meister et al., 2007), music (Looi, Gfeller, & Driscoll, 2012) and hearing in noise (Friesen, Shannon, Baskent, & Wang, 2001). This hearing situation does not only affect perception of speech, but is expected to result in deviant speech output as well, since there is a link between hearing capacity and speech production performance, i.e., self-monitoring of speech (Guenther, 2006; Levelt, 1983).

The speech of CI recipients has been investigated by at least two different types of studies. The first type (which can be called the 'normative' type) is to compare CI recipients' voices at one or more moments in time after implantation to their pre-implantation voices and/or to the voices of normally hearing peers, as part of the same study or as normative data from previous research (Evans & Deliyski, 2007; Goffman, Ertmer, & Erdle, 2002; Lane et al., 1998; Perrin, Berger-Vachon, Topouzkhanian, Truy, & Morgon, 1999; Seifert et al., 2002; Ubrig et al., 2011; Uchanski & Geers, 2003; Valero Garcia, Rovira, & Sanvicens, 2010). The second type of research (the 'on/off' type) involves a comparison between the performance of (more or less experienced) CI users in a condition in which their implant is temporarily turned off and one in which it is turned on again (Higgins, McCleary, & Schulte, 2001; Poissant, Peters, & Robb, 2006; Tye-Murray, Spencer, Bedia, & Woodworth, 1996).

Outcomes across studies of both types vary considerably, both in the direction and the amount of deviations (if any) from the norm. This variability has been attributed to the divergence in the following methodological factors: speech material (sustained vowels, syllables,

read-aloud continuous speech or spontaneous speech), assessment techniques (aerodynamic/physiologic, standard acoustic analysis, custom-made acoustic analysis or perceptual evaluation), age of the participants, speech-processing strategy of the implant and age of implant activation (Baudonck, van Lierde, Dhooge, & Corthals, 2011). The lack of convergence in the results so far is substantiated by a review of 27 articles about the voice quality of CI users (Coelho, Brasolotto, & Bevilacqua, 2012), which concluded that the number of effective studies is too small to draw clear conclusions.

Nevertheless, a number of impressionistic generalizations about voice and speech measures can be made from the pooled investigations on CI users with varying hearing histories so far. The fundamental frequency (F0) is high before implantation, on normative type studies (Oster, 1987; Perkell, Lane, Svirsky, & Webster, 1992; Szyfter et al., 1996; Ubrig et al., 2011) or when the implant is turned off, i.e., in on/off type studies (Monini, Banci, Barbara, Argiro, & Filipo, 1997; Poissant et al., 2006; Svirsky, Lane, Perkell, & Wozniak, 1992), and drops gradually after implantation. Variability of F0, or vF0 (Ball & Ison, 1984; Holler et al., 2010; Ubrig et al., 2011), and jitter (Fourcin, Abberton, Richardson, & Shaw, 2011; Hocevar-Boltezar et al., 2006) decrease after implantation. The nasal resonance of the speech is in general either too low (Monini et al., 1997; van Lierde, Vinck, Baudonck, De Vel, & Dhooge, 2005) or too high (Hassan et al., 2011a; Nguyen, Allegro, Low, Papsin, & Campisi, 2008; Svirsky, Jones, Osberger, & Miyamoto, 1998; Ubrig et al., 2011), but interacts with the principal resonance cavity of the sound (Baudonck, van Lierde, D'Haeseleer, & Dhooge, 2015). On a more global level, speech rate is low (Evans & Deliyski, 2007; Lane et al., 1998; Leder et al., 1987; Perrin et al., 1999) but increases with implant experience (Oster, 1987; Perkell et al., 1992). Correspondingly, the duration of speech elements is long at different linguistic levels, such as syllables (Lane, Matthies, Perkell, Vick, & Zandipour, 2001; Menard et al., 2007; Neumeyer, Harrington, & Draxler, 2010; Uchanski & Geers, 2003), words (Kishon-Rabin, Taitelbaum, Tobin,

& Hildesheimer, 1999; Uchanski & Geers, 2003; Waters, 1986), sentences (Leder et al., 1987; Uchanski & Geers, 2003), and paragraphs (Leder et al., 1987). Perceptually, the voice of CI users is rated to some degree as strained, rough, breathy, asthenic, unstable and hoarse (Baudonck, D'Haeseleer, Dhooge, & van Lierde, 2011; Horga & Liker, 2006; van Lierde et al., 2005).

It could be argued that even within the population of CI users differences in hearing history have differential effects on voice and speech measures. For instance, postlingually deafened adults might benefit from feedforward articulatory commands established during the period as hearing individuals, whereas speakers with prelingual hearing loss or children with postlingual hearing loss had no or little opportunity to establish those commands (Perkell et al., 1992; Perkell et al., 1997). However, speaker groups with different onsets of hearing loss have been rarely tested in a single study. Hassan et al. (2011b) found higher nasality values relative to a NH control group for adults with more than six years of hearing loss than for adults with less than three years of hearing loss. Richardson, Busby, Blamey, Dowell, and Clark (1993) measured vowel formants in two adults and three children, but the sample size was too small to draw firm conclusions. The question to what extent voice and speech measures differ between adult and pediatric CI recipients therefore largely remains an open question. The current study focused on children.

Despite its broad range, the research on CI speech has failed to fully consider a number of important theoretical and methodological aspects. First of all, some prosodic measures have not been investigated phonetically, such as the natural declination of F0 during an utterance or the ratio of voiced and unvoiced frames. These specific measures are potentially interesting because they could reflect CI recipients' difficulty with perceiving F0. Second, basic measures of prosody, i.e., prosodic measures that have not been linked to a linguistic or emotional function, have, to our knowledge, not been systematically compared across phonetic dimensions within a single study. A comparison between the temporal, intensity, and spectral

dimensions may allow connecting problematic phonetic aspects to auditory resolutions along those same dimensions. O'Halpin (2009) investigated accuracy of perception and production of duration, intensity and F0 cues of focused words, but this involved only one measure per dimension and was performed on laboratory instead of spontaneous speech. Third, measures were usually not compared at several points in time before and/or after implantation and/or for children with different ages at implantation. And finally, spontaneous speech has been neglected, even though voice differences can be expected between spontaneous speech and task-related speech (Vorperian & Kent, 2007). The use of spontaneous speech is important because it is the natural daily speaking mode. For instance, it could be argued that asking CI recipients to describe a picture, as in Evans and Deliyski (Evans & Deliyski, 2007), elicits a type of speech that is only spontaneous to a limited degree since the recipient is confronted not only with a specific semantic register but also with an experimental setting.

The present study aims to complement the body of research on CI users' speech characteristics by comparing a number of basic prosodic characteristics along three different phonetic dimensions in the spontaneous speech of young children: 'temporal', 'intensity', and 'spectral'. These dimensions were selected to reflect three important phonetic and acoustic parameters for which CI users have been found to have differential auditory resolutions and effectiveness (Cooper, Tobey, & Loizou, 2008; Moore, 2003; Shannon, 2002). This allows us to investigate to what extent perceptual competences are reflected in speech production. Measurements were repeated at three points in time after the onset of hearing and compared between children implanted before, or after the age of two years and a control group of normally hearing (NH) children of the same hearing age (Boons et al., 2012; Hayes, Geers, Treiman, & Moog, 2009; Holt & Svirsky, 2008). We conjectured that (1) the CI recipients' measures differed from those of the controls because they had less successful auditory feedback to control their laryngeal and articulatory output; (2) CI

recipients were least deviant on the temporal dimension, followed by the amplitude dimension and most deviant on the spectral dimension; (3) the late implanted group had more deviant outcomes than the early implanted group; and (4) that the differences between CI recipients and controls decreased with increasing experience with the device and that this decrease was faster for early implanted than for late implanted children.

2.2 Methods

2.2.1 Participants

The study included three groups. There were two experimental groups, consisting of nine children implanted before and nine after the age of two, respectively (Early/Late Implanted, EI/LI; both 6 boys and 3 girls) with mean chronological ages of two years and ten months (henceforth, '2;10'; SD: 0;7) and 6;8 (SD: 2;5) at the time of testing. These participants were profoundly deaf and received a CI at Leiden University Medical Center (LUMC). The third (control) group consisted of 12 normally hearing children (4 boys, 8 girls) with a mean age of 2;1 (SD: 0;4; NH group). Eleven of them were children of the CLPF (Clara Levelt – Paula Fikkert) corpus (Fikkert, 1994; Levelt, 1994), available through the CHILDES database (MacWhinney, 2000) and through personal communication. One was from a corpus compiled by Beers (1995).

Demographic, audiometric and implant characteristics for individual CI recipients and for groups, as well as results of one-way Analyses of Variance of group mean differences can be found in Table 1. Some variables require an explanation. Age at onset of hearing loss diagnosis reports the age at which hearing loss was first diagnosed, with *0* for presumed congenital deafness. The estimated duration of deafness is the time between the estimated onset of deafness and age at CI activation. The mean age over recordings is the arithmetic mean chronological age of all recordings of a recipient that were used for

analysis. This statistic was preferred over the age at first recording because not all sessions were available for all CI recipients (see the Data analysis section).

Groups were matched for hearing age, which is defined as the time since the onset of stable spoken language acquisition, i.e., without a changing hearing situation. For the CI group, this equals the time between CI activation and the time of recording; for the NH group, this equals the time between birth and the time of recording (i.e., chronological age). Matching for hearing age is a common procedure in CI language acquisition research, as language development of children with CIs has been found to match the development of NH children better by hearing age than by chronological age (Dornan, Hickson, Murdoch, & Houston, 2009; Fagan & Pisoni, 2010). This suggests that spoken language development starts with the onset of hearing and not necessarily at birth. Since in our study we were not interested in language development in general, but in phonetic development, we kept the amount of experience with stable spoken language input (i.e., hearing age) constant across participant groups.

Inclusion criteria for CI recipients were pediatric chronological age (under 11 years), bilateral pre- or postlingual severe-to-profound hearing loss, and a monolingual Dutch home environment. Exclusion criteria were reported additional social, cognitive or physiological disorders. All CI recipients were enrolled in the LUMC rehabilitation program for pediatric CI recipients, involving frequent speech training and six-monthly communication and social behavior follow-ups. The dividing line between Early and Late age of implantation was set at two years because differences in language outcomes have been observed between children implanted before or after this age, likely due to a boundary of one of the sensitive periods of language acquisition (Boons et al., 2012; Hayes et al., 2009; Holt & Svirsky, 2008; Werker & Hensch, 2015).

Matching groups for hearing age, combined with the selection by differential activation ages for different recipient groups

Table 1. Demographic and implant characteristics of CI recipients and the mean age of the control group. 'AB' is the Advanced Bionics HiRes 90k implant; 'Nucleus' is the Nucleus Freedom Contour Advance implant. BERA thresholds refer to the highest loudness levels in the left (L) and right (R) ear, respectively, that no BERA response was reported for. The group CI is the Early and Late Implanted groups taken together. SDs were rounded to whole months. Note that the (chronological) age and the hearing age are, by definition, the same for the NH group. Abbreviations: x;y.z – years;months.days. Numbers in parentheses indicate standard deviations, unless indicated otherwise. For Mean age over recordings and Mean hearing age over recordings, 2-way comparisons are Bonferroni corrected post-hoc analyses.

Group	**Subject number (gender)**	**Age at onset of hearing loss diagnosis (months)**	**Estimated duration of deafness (months)**	**Age at CI activation**	**Mean age over recordings**	**Mean hearing age over recordings**
	1 (M)	3	12	1;2.24	2;8.24	2;0.22
	2 (M)	0	13	1;1.20	2;8.28	2;1.18
	3 (M)	0	17	1;4.26	2;7.15	2;0.24
	4 (M)	0	12	0;11.26	2;7.08	2;1.26
EI	5 (F)	4	15	1;7.09	3;2.16	2;3.29
	6 (F)	2	16	1;5.23	3;1.28	1;10.7
	7 (M)	1	13	1;2.00	2;7.19	1;5.20
	8 (F)	4	10	1;1.26	2;6.23	1;8.15
	9 (M)	7	11	1;6.12	3;0.08	1;11.29
	MEAN	***2.3 (2.4)***	***13.2 (2.3)***	***1;3.19 (0;2.16)***	***2;10.9 (0;6.18)***	***1;11.18 (0;3.4)***
	1 (M)	0	49	4;1.08	5;4.05	1;10.12
	2 (F)	16	27	3;6.23	5;3.04	2;1.1
	3 (F)	30	16	3;9.17	5;3.04	2;0.18
	4 (M)	0	96	8;0.00	9;6.28	2;1.1
LI	5 (M)	16	86	8;5.28	10;2.02	2;0.24
	6 (M)	9	64	6;0.19	7;6.16	2;0.1
	7 (M)	12	47	4;10.22	6;4.08	1;5.20
	8 (M)	2	81	6;10.16	8;4.27	1;10.11
	9 (F)	0	25	2;1.27	3;7.18	2;0.7
	MEAN	***9.4 (10.2)***	***54.6 (28.9)***	***5;3.28 (2;1.27)***	***6;8.12 (2;4.22)***	***1;11.18 (0;2.12)***
CI	***OVERALL***	***5.9 (8.1)***	***33.9 (29.1)***	***3;3.23 (2;6.18)***	***4;9.11 (2;7.4)***	***1;11.13 (0;2.22)***
NH	***MEAN***				***2;0.15 (0;3.29)***	***2;0.15 (0;3.29)***
3-way	ANOVA *p* (*F*)				**<.001** (32.9)	.69 (.37)
EI-LI	ANOVA *p* (*F*)	0.059 (4.1)	**.001** (18.0)	**<.001** (31.0)	**<.001**	1
EI-NH	ANOVA *p* (*F*)				.54	1
LI-NH	ANOVA *p* (*F*)				**<.001**	1
CI-NH	ANOVA *p* (*F*)				**.002** (11.8)	.39 (.77)

Notes: [a] Calculations were based on available cases and on means of both ears where applicable

Table 1 (cont.)

Group	Subject number (gender)	Etiology	BERA threshold L/R (dB)	Implan-ted ear(s)	Implant type	Speech process-sor	Insertio depth (degrees
EI	1 (M)	unknown	92/90	bilateral	AB	PSP	467.99/483.1
	2 (M)	hereditary	95/100	right	AB	PSP	480.4
	3 (M)	unknown	108/103	right	AB	PSP	461.3
	4 (M)	hereditary	unknown	bilateral	AB	PSP	405.16/447.7
	5 (F)	unknown	103/103	bilateral	AB	PSP	465.53/425.1
	6 (F)	unknown	100/100	right	AB	PSP	547.7
	7 (M)	unknown	100/100	bilateral	AB	PSP	455.03/506.9
	8 (F)	unknown	105/105	right	AB	PSP	498.5
	9 (M)	unknown	100/100	bilateral	AB	PSP	437.05/560.5
	MEAN		***100.3 (4.6)[a]***				***479.47 (34.86)***
LI	1 (M)	unknown	100/100	left	AB	PSP	482.6
	2 (F)	meningitis	90/100	left	AB	Auria	575.6
	3 (F)	unknown	97/97	right	AB	Harmony	504.9
	4 (M)	unknown	100/85	left	AB	Harmony	
	5 (M)	unknown	90/90	left	Nucleus	Freedom	
	6 (M)	unknown	no response[b]	left	AB	PSP	
	7 (M)	unknown	100/80	left	AB	PSP	463.5
	8 (M)	meningitis	100/100	left	AB	PSP	512.9
	9 (F)	unknown	97/97	right	AB	Harmony	632.4
	MEAN		***95.2 (4.0)[a]***				***528.69 (63.46)***
CI NH	***OVERALL MEAN***		***97.7 (4.9)***				***499.16 (52.49***
3-way	ANOVA *p* (*F*)		**0.035** (5.42)				
EI-LI	ANOVA *p* (*F*)						0.073 (3.8

Note : [b] BERA performed in another medical center

unavoidably introduced a confound with chronological age. As can be seen in Table 1, therefore, measures relating to chronological age were statistically different between groups (except for EI vs. NH for chronological age), but not those relating to hearing age. The Spearman rank correlation between Group and Chronological age was 0.922. When fitting both Group and Chronological age into the statistical model (multilevel linear regression model), standard errors were highly inflated and parameter estimation became highly

unstable. We therefore only considered the variable Group in the statistical model, without chronological age. We will return to this complication in the Discussion section.

EI recipients were implanted in the right ear (N = 4) or bilaterally (N = 5), whereas 7 out of 9 of the LI recipients were implanted in the left ear. All but one recipient received the Advanced Bionics HiRes 90k with a HiFocus 1j electrode and a PSP (including all the EI recipients), an Auria or a Harmony speech processor (Advanced Bionics, Sylmar, CA, USA); one recipient in the LI group was fitted with the Nucleus Freedom Contour Advance (Cochlear Corp, Sydney, Australia). Etiologies were unknown in most cases, except for hereditary causes and meningitis in two cases each. Insertion depth in degrees (computed as the mean between both ears if applicable) was not different between groups, but Brainstem Evoked Response Audiometric (BERA) thresholds were higher for EI than for LI.

2.2.2 Procedure

Speech recordings of the experimental participants were performed in playrooms at the department of pediatrics in LUMC. The setup consisted of a table, chairs, games and toys (such as cars and a kitchen) for children. A researcher observed and videotaped the session. Audio was recorded through the camera's integrated high-quality microphone or one attached to children and parents' clothing just below the head. Both in the recordings of the experimental and those of the control group, the child played with (a) parent(s) or a therapist/experimenter and sometimes also siblings. A child's speech was elicited when he/she did not speak much spontaneously. A recording session typically lasted between 20 and 30 minutes.

2.2.3 Data analysis

Audio channels were digitized with a 16-bit resolution and at a 48 kHz sampling frequency. Speech segmentation and phonetic analyses were performed by a trained linguist and phonetician (DV) using *Praat*

Table 2. List of prosodic measures performed for the analysis of the speech data, each listed under the phonetic dimension (temporal, intensity, spectral) that it is classified under for the current purpose. Abbreviation is the code by which it is referred to in the text (if unspecified, the full name is used). Unit is the mathematical unit used to describe an outcome of the measure. σ stands for syllable. Definitions are explained in the text.

Dimension	Measure (abbreviation)	Definition	Unit
Temporal	Articulation rate (ArtRate)	Number of syllables pronounced per second speech without pauses	σ/s
	Duration of the utterance (log) (DurUtt)	Base-*e* logarithm of the difference between final and initial time point of the utterance	s
	Voicing Ratio	Portion of voiced frames of an utterance as a percentage of the total number of analysis frames in the utterance	%
Intensity	Amplitude Perturbation Quotient (APQ)	(5-point scale). "The average absolute difference between the amplitude of a period and the average of the amplitude of its and its four closest neighbors, divided by the average amplitude."	%
	Harmonics-to-Noise Ratio (HNR)	The ratio between the energy that is in the periodic part and the energy that is in the aperiodic part of the voiced stretches of the signal	dB
Spectral	Declination	Global trend of F0 from beginning to the end of an utterance	Hz/s
	Mean F0	Mean of all pitch points (i.e., F0) of an utterance	Hz
	F0 standard deviation (SD F0)	Standard deviation of the mean of all pitch points (i.e., F0) of an utterance	Hz
	Pitch Perturbation Quotient (PPQ)	(5-point scale). "The average absolute difference between a period and the average of its and its four closest neighbors, divided by the average period."	%

software, *Version 5* (Boersma & Weenink, 2014). NH and CI recordings were matched for hearing age with a five-day margin per session (18, 24, 30 months). This yielded twenty recordings per group divided over hearing age sessions at 18, 24 and 30 months. Due to restricted data availability at source in combination with the strict matching criteria, this design suffered from missing data (see the

section Statistical Analysis). All recordings were subjected to the same data processing procedure. Nine phonetic prosody parameters were measured (Table 2). We will call them 'basic' measures because they do not involve linguistic or subjective judgements about the (un)naturalness, function or meaning of the prosody. They cover three fundamental acoustic dimensions of prosody: the temporal, the intensity and the spectral dimensions (Lehiste, 1970). The temporal measures were articulation rate (ArtRate), duration of the utterance (DurUtt) and Voicing Ratio. ArtRate is defined as the number of syllables pronounced per second speech without pauses (Goldman-Eisler, 1968). Numbers of syllables per utterance were determined from the recordings, on the basis of the realized, not the targeted, form of words. The duration of the utterance (DurUtt) was based on prosodic and syntactic integrity. The exact starting and end points were based on visual inspection of the waveform. Voicing Ratio refers to the percentage of frames of an utterance that are voiced. This was based on a pitch analysis whereby the time-step for frames was 75 ms and the pitch range of analysis was 100-600 Hz. The reason we consider this a temporal measure is that correct production of voicing specifically requires that the timing of the onsetand offset of vocal fold vibration is synchronized with the sequence of vowels and consonants.

The intensity measures are the five-point amplitude perturbation quotient (APQ) and Harmonics-to-Noise Ratio (HNR). APQ is "[t]he average absolute difference between the amplitude of a period and the average of the amplitude of it and its four closest neighbors, divided by the average amplitude."[1] This is a measure of local variability of the amplitude of an F0 period. HNR represents the ratio (expressed in dB) between the energy in the harmonics vs. the energy in the parts between the harmonics of the voiced stretches of the signal. Periodicity was detected using the cross-correlation method with a time-step of 10 ms, a pitch floor of 100 Hz, a silence threshold of 0.1 times the global maximum amplitude and 1 period per time window.[2] Despite the fact that HNR carries both spectral (absence or

presence of periodicity) and intensity-related signal information, we regard the intensity-related information as primary, since HNR is defined as a ratio of intensities, and is therefore an intensity measure itself. These intensity measures could count as prosodic measures because they involve voice quality measured over a full utterance.

The spectral measures are declination of F0, standard deviation of F0, the mean of F0 and the pitch perturbation quotient. Declination is the natural global downtrend of F0 from beginning to the end of an utterance (Strik, 1994). To our knowledge, declination has never been estimated in CI users' speech. Because its realization depends not only on physiological effort but also on linguistic choices for which good control of F0 is needed, we expect that CI recipients will relatively often disrupt the baseline deviation such that values will become less negative (shallower downtrends). Mean F0 was calculated as the mean of all pitch points (i.e., F0) of an utterance. Following previous research, we expect to find elevated values of mean F0 for CI users (Oster, 1987; Perkell et al., 1992; Szyfter et al., 1996; Ubrig et al., 2011). The standard deviation of F0 (SD F0) is computed as the deviation of the mean of all pitch points of an utterance. It could be taken as a proxy for the global variability of F0 over an utterance. Based on research on a comparable measure, vF0, the coefficient of long-term F0 variation (the relative standard deviation of the period-to-period F0) (Deliyski, 1993; Hocevar-Boltezar et al., 2006; Holler et al., 2010; Ubrig et al., 2011), we hypothesize higher values for the CI recipients than for the controls. Finally, the five-point PPQ is "[t]he average absolute difference between a period and the average of its and its four closest neighbors, divided by the average period."[3] This is a measure of local pitch variability.

The utterance was used as the unit of the measurements, as this counts as a unit for many aspects of prosody. It is the highest prosodic unit under discourse-level units where intonational boundaries and temporal organization coincide (Rietveld & van Heuven, 2016). Utterances that were inaudible and/or interrupted by other speakers were left out because their phonetic realization and/or analysis would

be unreliable. This yielded 1,973 utterances. From this set, in order to avoid improbable values due to pitch detection errors, utterances were removed from the analysis if the declination was more than two standard deviations away from the mean (1.8%), resulting in 1,937 utterances for analysis. Different participants provided different raw and net numbers of utterances, but all measures were performed for every available utterance.

A risk of using a corpus of spontaneous speech is that the speech material is not equal between groups. It is especially important for Voicing Ratio and, to a lesser extent, for ArtRate that the realized segmental material be phonetically balanced. We therefore obtained an approximation of the number of tokens per phoneme used in the whole data set of each Group. Figure 1 displays the token occurrence per phoneme as a percentage of the total number of tokens in the group. The graph shows that the distributions of allophone tokens are highly comparable between groups. A second possible pitfall in corpus research is the number of syllables. However, according to an ANOVA, there was no effect of Group on the mean number of syllables per participant ($F(2,27) = 1.25$, $p = .30$).

2.2.4 Statistical analysis

Statistical tests were carried out using *IBM SPSS Statistics, Version 21* (IBM, Armonk, NY). Each participant was measured at three planned occasions and each occasion provided multiple (unique) utterances. The statistical model took into account that utterances were correlated within participants. For each of the seven dependent variables separately, a multilevel linear regression model was used to describe the differences between the groups and between time points of measurement, with within-subject correlation being modelled by introducing a random subject intercept. This was done by modelling the correlation structure before the fixed structure (Fizmaurice, Laird, & Ware, 2011). The procedure started by applying a very complex and well-fitting model and subsequently reduced it using Restricted Maximum Likelihood and Maximum Likelihood Ratio tests. When a

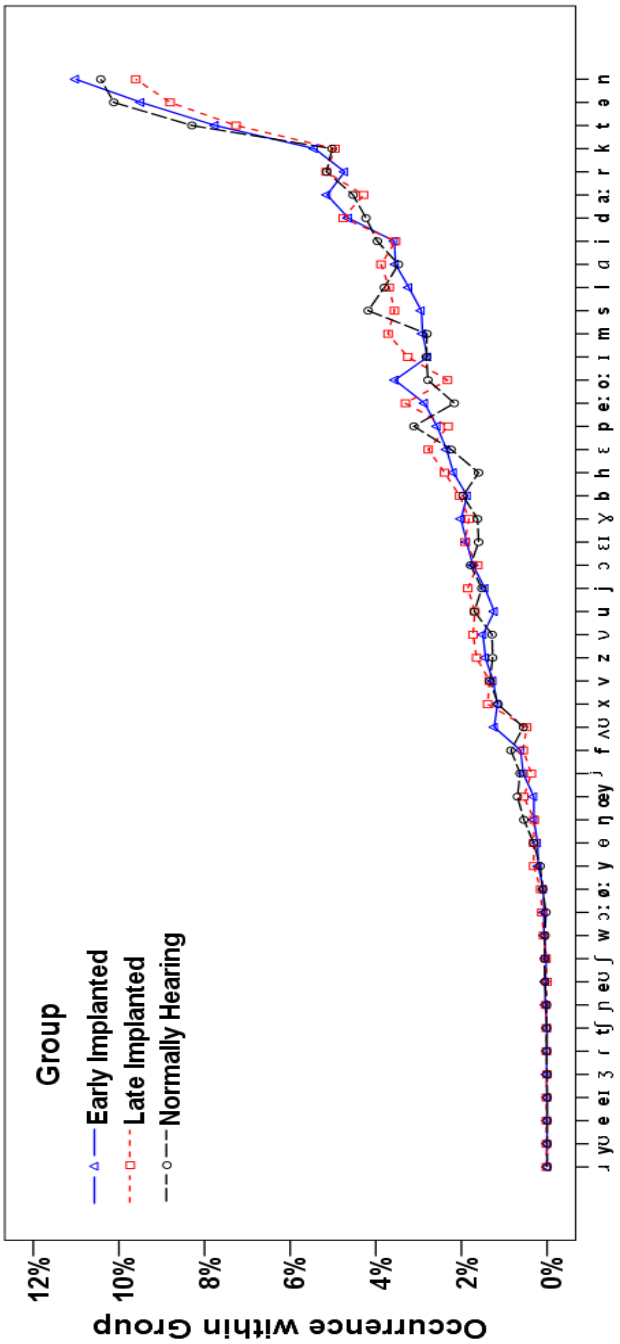

Figure 1. Token frequencies of 48 vowel and consonant phonemes of Dutch (N = 22,106) for three groups of speakers (Early Implanted, Late Implanted, Normally Hearing), as a percentage of all tokens within each respective group. Allophone, on the abscissa, refers to the allophone types found in the combined speech material of all participants. The transcriptions into these allophones were were left unaltered as they were created *Praat*. Type and token frequencies were obtained by automatically converting an orthographic transcription of all utterances to a broad phonetic transcription.

decision could not be based clearly solely on Likelihood Ratio tests, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were considered to decide on the most appropriate model (Fox, 2008). Models were fit using the Linear Mixed Model procedure in *SPSS*. A significance threshold of $p = 0.05$ was adopted.
In order to explore possible correlations among the nine dependent variables obtained for the analysis (see Table 2), an exploratory factor analysis using a principal component extraction method and a varimax rotation was conducted using heuristics and steps taken from Meyers, Gamst, and Guarino (2006). All correlation coefficients are shown in the correlation matrix in Table 3. The data were screened by considering both univariate and multivariate descriptive measures. All variables were interval variables and, except for DurUtt, approximately normally distributed. DurUtt was logarithmically transformed (with base *e*). Using these variables, all variable pairs appeared to be bivariate normally distributed with the exception of the pair ArtRate - DurUtt. The Kaiser-Meyer-Olkin measure of sampling adequacy for this pair was 0.612, which is not considered adequate given a criterion of 0.7. However, a factor analysis showed that three variables were correlated to a medium to high degree, viz. HNR, PPQ and APQ. Considering only these three relatively strongly correlated variables, the Kaiser-Meyer-Olkin measure was adequate (0.707). Bartlett's test of sphericity was, however, significant both when including and excluding the three non-highly correlated variables ($\chi^2(36) = 4032.65$, $p < .001$; $\chi^2(36) = 2919.03$, $p < .001$). We concluded that the dataset was appropriate for factor analysis. In the factor analysis considering all nine dependent variables, four eigenvalues greater than 1 were found (2.553, 1.404, 1.078, and 1.044).

Given the preference for interpretable dependent variables, and also taking into consideration that the second principal component consisted of two variables with only a small correlation (0.280), only the first component was constructed. The factor (henceforth, Factor 1) was constructed by standardizing and summating the three dependent

variables that were involved in the component (HNR, PPQ and APQ). Further analysis was thus done using the seven (almost uncorrelated) dependent variables.

Table 3. Correlation matrix with coefficients of the Pearson correlations between the nine dependent variables, plus two-tailed significance indications and *p-value* (between parentheses). Definitions of the measures can be found in Table 2.

Measure	HNR	PPQ	APQ	Mean F0	SD F0	Voicing Ratio	DurUtt ($\log_e$)	ArtRate
PPQ	**−.598**							
APQ	**−.763**	**.674**						
Mean F0		**−.146**	**−.263**					
SD F0	**−.112**	**.111**	0.028	**.280**				
Voicing Ratio	**.274**	−0.044	**.045**	0.008	**−.106**			
DurUtt ($\log_e$)	**.118**	**−.128**	**−.174**	0.026	**.201**	**−.111**		
ArtRate	**.090**	**−.177**	**−.101**	**.090**	**.048**	0.034	**.163**	
Declination	−0.011	**.050**	0.037	**.049**	0.021	0.006	0.013	0.038

Notes: Correlations in boldface were significant. In this table, correlation coefficients >.045 were significant at the $p < .05$ level, and correlation coefficients >.090 were significant at the $p < .01$ level.

As explained in the section Data analysis, recordings were missing on one or two sessions for some participants. There were a number of causes: 1) the recording contained no or hardly any analyzable child utterances (1 case, EI); 2) the recording did not exist because the child had been implanted too recently (3 cases, EI); 3) the recording at that session was not performed because that was not deemed necessary by the speech therapist given his/her development or because some other test was performed during that visit (2 cases, LI); 4) technical problems (2 cases, LI); 5) the session fell outside the range ever recorded by an LUMC speech therapist for a participant (16 cases, NH). Recording selections were based on the chronological age during recording and not on the quality of their content. We therefore believe our data are Missing Completely At Random or

perhaps Missing At Random (Fizmaurice et al., 2011) which allowed us to use a linear mixed model that uses the likelihood function to estimate the parameters in an unbiased way. For a recent review on the problem of and solutions for missing data in otorhinolaryngological research, see Netten et al. (2016).

In sum, seven independent linear mixed model (LMM) analyses were run, each for one of the dependent variables (one of which, Factor 1, is a combination of three of the original variables). We were interested in the effect of the independent variables Group (EI, LI or NH) and Session (a hearing age of 18, 24 or 30 months). Though its effect was not a focus in itself, the variable Gender of the participant was added as well, viz. in order to account for a possible confounding effect because genders were not equally divided across groups (see Table 1).

2.3 Results

Mean values and standard deviations (in parentheses) of all nine dependent variables and Factor 1 are listed in Table 4. This includes the values aggregated over one, two, and three independent variables (Group, Session and Gender). APQ, HNR, and PPQ will not be discussed separately, as they have been merged into Factor 1. Means and confidence intervals of the seven dependent variables left after factor analysis are shown in Figure 2. The development in hearing age in months (Session) was plotted on the abscissa. This was split by Group and Gender (left panels), and separately, for clarity, split by only Group (right panels).

The grouping of APQ, HNR, and PPQ into Factor 1 eliminated one of the phonetic dimensions under investigation, viz. the intensity dimension, as the two intensity measures were both part of that procedure. Results of the remaining seven variables will now be discussed in turn. Following the Principle of Marginality, main effects were not interpreted when more complex terms present in the model

were significant (Fox, 2008). Further, individual regression coefficients were not interpreTable 1n those cases either, because they cannot be considered separately from the interactions. Table 5 lists the best-fit models and statistics of the component effects for all seven dependent variables. Best-fit models refer to the combination of terms

Table 4. Mean values and standard deviations (right sides of columns) of all nine dependent measures and Factor 1, divided over Group (EI: Early Implanted, LI: Late Implanted, NH: Normally Hearing), Gender, and Session (hearing ages of 18, 24, and 30 months). Factor 1 is the sum of z-transformed values of HNR, APQ, and PPQ. Definitions of the measures can be found in Table 2. 'Syll': syllable; 'mos': months.

Group	Session (mos.)	Measure											
		ArtRate (syll/s)		DurUtt (log_e, s)		Voicing Ratio (%)		APQ (%)		HNR (dB)		Declination (Hz/s)	
EI	18	2.27	.67	0.55	.09	0.68	.16	6.62	2.98	12.76	4.9	–8.16	101.33
	24	2.78	.77	0.58	.09	0.6	.14	4.62	1.55	14.3	3.62	–16.08	92.12
	30	2.86	.98	0.56	.11	0.64	.16	6.56	3.14	12.37	4.3	3.82	91.84
LI	18	2.94	1.21	0.51	.1	0.63	.18	7.55	4.23	10.47	6.11	–32.94	116.36
	24	3.3	1.1	0.54	.1	0.65	.17	6.04	2.97	13.13	4.51	–32.73	91.13
	30	2.78	.77	0.57	.13	0.64	.14	5.09	2.26	13.92	3.73	0.43	84.79
NH	18	2.22	.69	0.47	.05	0.75	.18	7.64	4.22	11.89	4.68	–56.57	110.78
	24	2.5	.81	0.52	.08	0.63	.15	5.69	2.14	13.38	3.79	–14.45	127.42
	30	2.78	.77	0.57	.09	0.62	.14	5.42	2.05	14.89	3.72	–4.7	66.52
Total	18	2.44	.83	0.51	.09	0.68	.18	7.25	3.85	11.7	5.37	–31.26	111.04
	24	2.7	.88	0.54	.09	0.63	.16	5.52	2.32	13.54	3.96	–19.34	111.77
	30	2.78	.85	0.57	.11	0.63	.14	5.66	2.51	13.96	4.03	–1.12	78.6
EI		2.63	.83	0.57	.1	0.63	.16	5.74	2.7	13.31	4.28	–8	94.76
LI		2.94	1.4	0.54	.11	0.64	.17	6.2	3.34	12.61	5	–24.08	98.05
NH		2.5	.75	0.53	.09	0.65	.16	5.86	2.58	13.65	4.01	–16.99	110.88
Total		2.63	.83	0.54	.1	0.64	.16	5.92	2.84	13.28	4.38	–16.43	103.48

Table 4 (cont.)

Group	Session (months)	Measure									
		Declination (Hz/s)		Mean F0 (Hz)		SD F0 (Hz)		PPQ (Hz)		Factor 1 (*z*)	
EI	18	–8.16	101.33	321.25	49.84	56.63	29.62	1.07	.53	0.35	3.01
	24	–16.08	92.12	325.16	53.92	61.11	24.51	0.97	.33	–0.89	1.82
	30	3.82	91.84	321.46	54.02	56.29	27.35	1.22	.54	0.94	3.02
LI	18	–32.94	116.36	310.73	63.24	53.95	31.49	1.33	.74	1.67	3.97
	24	–32.73	91.13	306.65	58.71	53.31	26.38	1.1	.47	0.13	2.77
	30	0.43	84.79	291.03	41.12	50.5	24.08	1.01	.37	–0.58	2.01
NH	18	–56.57	110.78	304	102.64	43.29	27.67	1.29	.65	1.35	3.47
	24	–14.45	127.42	330.08	48.2	51.83	23.04	1	.38	–0.28	1.95
	30	–4.7	66.52	304.46	33.49	48.17	21.93	0.98	.37	–0.74	2.05
	18	–31.26	111.04	312.42	73.69	51.72	30.19	1.23	.65	1.11	3.54
	24	–19.34	111.77	323.15	53.15	54.38	24.52	1.02	.4	–0.33	2.18
	30	–1.12	78.6	306	43.22	50.97	24.25	1.05	.44	–0.21	2.48
EI		–8	94.76	323.01	52.83	58.47	26.85	1.07	.47	–0.01	2.68
LI		–24.08	98.05	303.66	56.43	52.74	27.29	1.14	.55	0.37	3.1
NH		–16.99	110.88	318.72	56.44	49.57	23.53	1.03	.44	–0.21	2.33
Total		–16.43	103.48	315.9	55.96	52.83	25.74	1.07	.48	0	2.66

listed in the column Terms of the best-fit model in Table 5. Unless stated otherwise, the focus of the interpretation will be on Group and Session (the right panels of Figure 2), because Gender was considered a confounding variable. The left panels of Figure 2 are shown for the sake of completeness.

Table 5. Best-fit models and statistics of component effects for all seven measures left after factor analysis. The best-fit model refers to the combination of factors (Group, Gender, Session and all their interactions) that was found to be the best Linear Mixed Model for the data of each measure. It consists of the combined terms for that measure. See the text for the criteria used for finding the best-fit model. The statistics of component effects refer to the *F*-value, degrees of freedom and *p*-value found for each term in the best-fit model. df: degrees of freedom; significant differences (at $p = .05$) are in boldface. Degrees of freedom were rounded off to the nearest integer value.

Measure	Terms of the best-fit model	Statistics of the term			
		F	df1	df2	*p*
ArtRate	Group	1.97	2	24	.16
	Gender	6.42	4	186	**<.001**
	Session	10.05	2	217	**<.001**
	Group × Gender	2.11	2	24	.14
	Group × Session	6.60	2	217	**.002**
	Gender × Session	1.51	4	186	.20
	Group × Gender × Session	6.42	4	186	**<.001**
DurUtt	Group	.00	1	26	1.0
	Gender	.88	2	26	.43
	Session	57.23	2	1864	**<.001**
	Group × Session	12.16	4	1670	**<.001**
	Gender × Session	8.14	2	1780	**<.001**
	Group	.82	2	20	.45
	Gender	1.71	1	20	.21
	Session	7.55	2	182	**.001**
	Group × Gender	.48	2	20	.62
	Group × Session	7.82	4	156	**<.001**
	Gender × Session	5.34	2	182	**.006**
	Group × Gender × Session	2.05	4	156	.090
Decli-nation	Session	7.29	2	1402	**.001**
Mean F0	Group	.98		26	.39
	Gender	.094	1	26	.76
	Session	19.53	2	1897	**<.001**
	Group × Session	11.86	4	1880	**<.001**

SD F0	Group	4.95	2	23	**.016**
	Gender	.076	1	23	.79
	Session	5.76	2	1759	**.003**
	Group × Gender	2.44	2	23	.11
	Gender × Session	4.25	2	1759	**.014**
Factor 1	Group	.33	2	25	.72
	Gender	1.26	1	25	.27
	Session	30.11	2	1913	**<.001**
	Gender × Session	19.12	2	1888	**<.001**
	Group × Session	13.06	2	1828	**<.001**

The best-fit for ArtRate was with all separate (Group, Gender, Session) and combined independent variables together. Given that the three-way interaction is the most complex significant term, all other effects must be interpreted with caution. Articulation rates were on average 2.63 syllables/s (syll/s) for the EI group, 2.94 syll/s for the LI group, and 2.50 syll/s for the NH group. Panel 1b in Figure 2 shows that from 18 to 30 months, the EI and the NH children experienced a rise in ArtRate, with the EI being ahead of the LI, and that the LI children converged with NH starting from higher values. The EI were therefore closer to the NH than the LI on only one of the three sessions. To our knowledge, the only previous study comparing speech or articulation rates in children with and without CIs is by Perrin et al. (1999). They found lower rates for the clinical group than for the typically developing group. However, their participants were older (9 to 14 years) than ours and the researchers did not report absolute outcome values. The values of all groups in the current study were on the lower side but within the range reported in studies on 3- to 5-year-olds discussed in (Flipsen, 2002). Rates tended to increase with age (e.g., Amir & Grinfeld, 2011) and to be lower in atypically developing populations including (adult) CI users (Evans & Deliyski, 2007; Lane et al., 1998; Smith, Roberts, Smith, Locke, & Bennett,

2006). Recipients in the studies on CI were all implanted as adults. In the current study, groups were confounded by chronological age and groups with a higher mean age had faster rates. This suggests that pediatric cochlear implantation does not prevent the typical increase in articulation rate with age.

DurUtt was best fit with Group, Gender, Session, Gender × Session, and Group × Session. Interpretable are differences in development between Groups (our focus) and, separately, between Gender. Figure 2, Panel 2b shows that at 18 months the NH had the shortest utterances, the LI had longer utterances, and EI the longest, but there was a convergence over time towards high values, with LI showing a straighter development than EI. The LI, with 1.72 s (transformed back from the logarithmic value) were further away from the controls (1.70 s) than the EI were (1.77 s). Utterance or sentence lengths (measured in syllables, phones or seconds) of typically and atypically developing populations tended to increase with age (Flipsen, 2002 and references therein; however, see Kadi-Hanifi & Howell, 1992), but this was not currently reflected, as the oldest group (LI) did not show the longest duration. For our older participants (LI), the value was low in comparison to values mentioned in the literature. In one study on the unrestricted speech of three groups of 4-, 7-, and 11-yeard-old stutterers and age-matched non-stutterers (Kadi-Hanifi & Howell, 1992), the average durations of the first two control groups were both 5.15 s. This, together with the observation that values in the three groups of the current study, despite being significantly different, were in an absolute sense very close together, suggests that the utterance duration length depended not on the chronological age, but rather on the hearing age (which was matched between groups). The convergence over time could be due to differential mechanisms for the three groups, as suggested by a comparison between DurUtt and ArtRate. Because a higher articulation rate would, all else being equal, result in shorter utterances, the increase in DurUtt for the NH must be due to the number of syllables, the duration of silence within utterances, or both. To further investigate this possibility, mean

numbers of syllables were computed (see the Data analysis section for the procedure) split between groups and sessions. For the 18, 24, and 30 months sessions, respectively, numbers of syllables were 2.2, 3.4, and 5.0 in the NH group, 3.7, 5.1, and 5.0 in the EI group, and 4.0, 5.0, and 5.3 in the LI group. According to an ANOVA, the interaction between Group and Session for this measure was highly significant ($F(4,1929) = 5.26$, $p < .001$). ArtRate and number of syllables per utterance developing more synchronously for the controls than for the CI recipients, it is very probable that control participants' utterances were longer because of an increasing number of syllables. The CI recipients, on the other hand, would tend to articulate faster on longer utterances without adding syllables. This could point at a more limited verbal working memory (compare, e.g., Burkholder & Pisoni, 2003). In conclusion, CI recipients' utterance duration seems to develop with hearing (not chronological) age and to be restricted by a relatively limited verbal working memory.

The best fit for Voicing Ratio was the one consisting of all separate and combined independent variables. The interpretable effects were Group × Session (this study's focus) and Gender × Session. In Figure 2, Panel 3b, it can be observed that CI recipients' Voicing Ratios started out lower than the controls' but converged towards comparable levels. The EI decreased in the first interval and were more variable, whereas the LI increased and were more constant. CI Recipients had a lower Ratio mainly at 18 months. EI children were not clearly more or less deviant than the LI children. It has been argued that children acquiring a first language pay attention to the distinction between voiced and voiceless intervals in the input in order to discover the rhythmic system of the language (Dellwo, Fourcin, & Abberton, 2007). Apparently, the implanted children did pay attention to this, but learned to time their voicing like NH peers 18 to 30 months after implantation.

The optimal fit for Declination was with only Session. Declinations became shallower over time, going from −31 to −1 Hz/s for all participants combined (Figure 2, Panel 4b; Table 4).

Declinations were less negative for the CI recipients, but mainly so at 18 months. EI participants were further from the NH values than LI at 18 months, closer at 24 months and about equally close at 30 months. These were only trends, however, since only the effect of Session was significant for Declination. 'T Hart, Collier, and Cohen (2006) summarized the declination *D* of utterances under 5 s. in semitones per time unit as $D = -11/(t + 1.5)$, with *t* in seconds (also see Rietveld & van Heuven, 2016). This formula was found to both predict spontaneous and read-aloud utterances fairly accurately, although for spontaneous speech a somewhat shallower declination was reported. Given the overall mean F0 of 316 Hz and an overall utterance duration of 1.72 s. in our study, declinations of around -92 Hz/s were expected, which is much steeper than what we found (-16 Hz/s). This may be due to the fact that our participants were children, as it has been claimed that in very young children some units of speech (i.e., short 'breath groups') show no declination (Lieberman, 1986).

Mean F0 was best fit with Group, Gender, Session, and Group × Session. Mean F0 developed differently among Groups (Figure 2, Panel 5b). The EI children showed hardly any changes, whereas the LI children's F0 dropped from 311 Hz at 18 months to 291 Hz at 30 months, and the NH children peaked in the middle session (from 304 to 330 Hz and back). With overall averages of 323, 304, and 319 Hz for the EI, LI, and NH groups, respectively. Mean F0 was, contrary to expectation, not higher in general in CI recipients, but only on two sessions for the EI and on one session for the LI. Further, EI were not clearly less deviant than the LI. The hypotheses regarding Mean F0 were therefore not confirmed. In one review of F0 values of children of different ages in 21 studies (Vorperian et al., 2005), the F0 value of one-and-a-half-year-old children (comparable to the mean age of the control group in the current study) was between 300 and 350 Hz, that of 3-year-old children (approximately the mean age of the Early Implanted group in the present study) ranged between 250 and 300 Hz and the value of the 7-year-old children (around the mean age of the Late Implanted group) ranged between around 240 and 280 Hz.

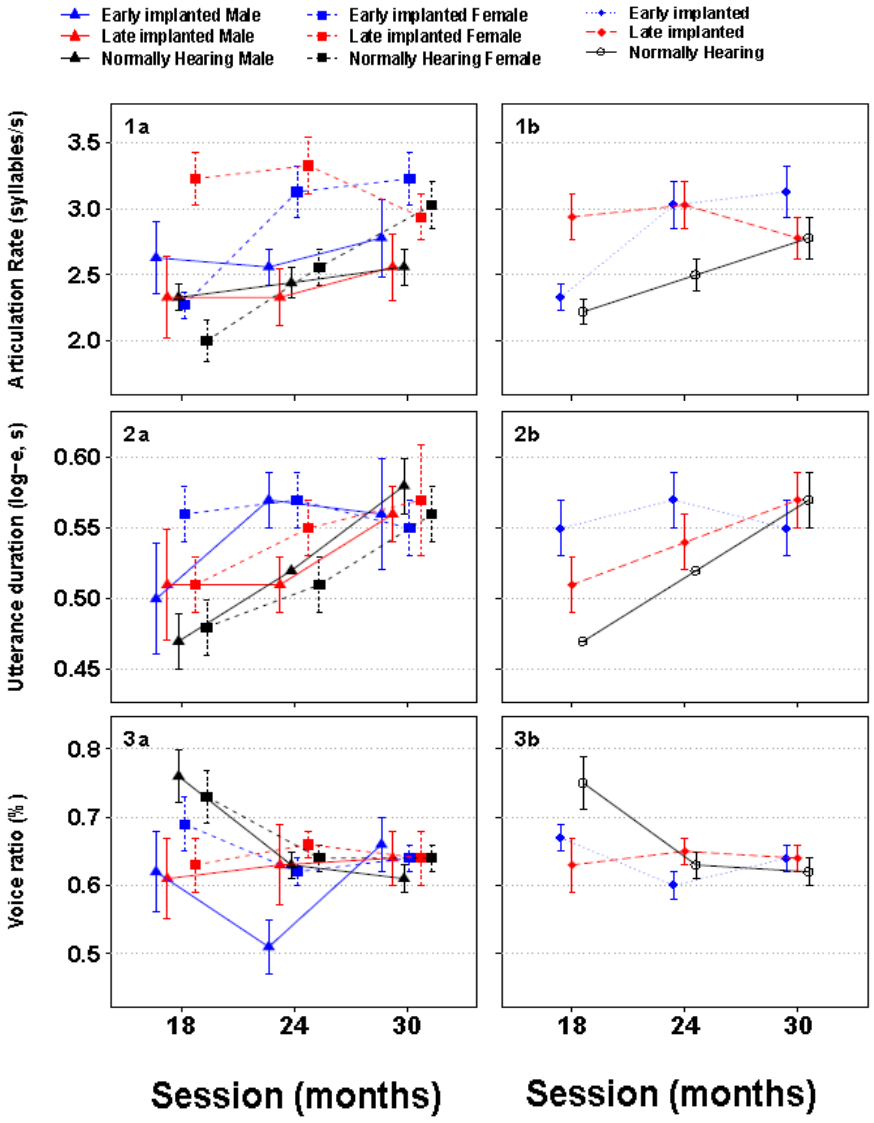

Figure 2. Plots of mean values of Articulation Rate, Utterance Duration (log-*e* transformed), Voicing Ratio, Declination, Mean F0, SD F0, and Factor 1. Factor 1 is the sum of *z*-scores of HNR, APQ, and PPQ. Hearing age in months (Session) is plotted on the abscissa. Left panels show results split by Gender, Group, and Session (Hearing Age in months). Right panels show the same results but aggregated over Gender. Error bars represent 95 % confidence intervals. The x-coordinates were jittered for the sake of clarity.

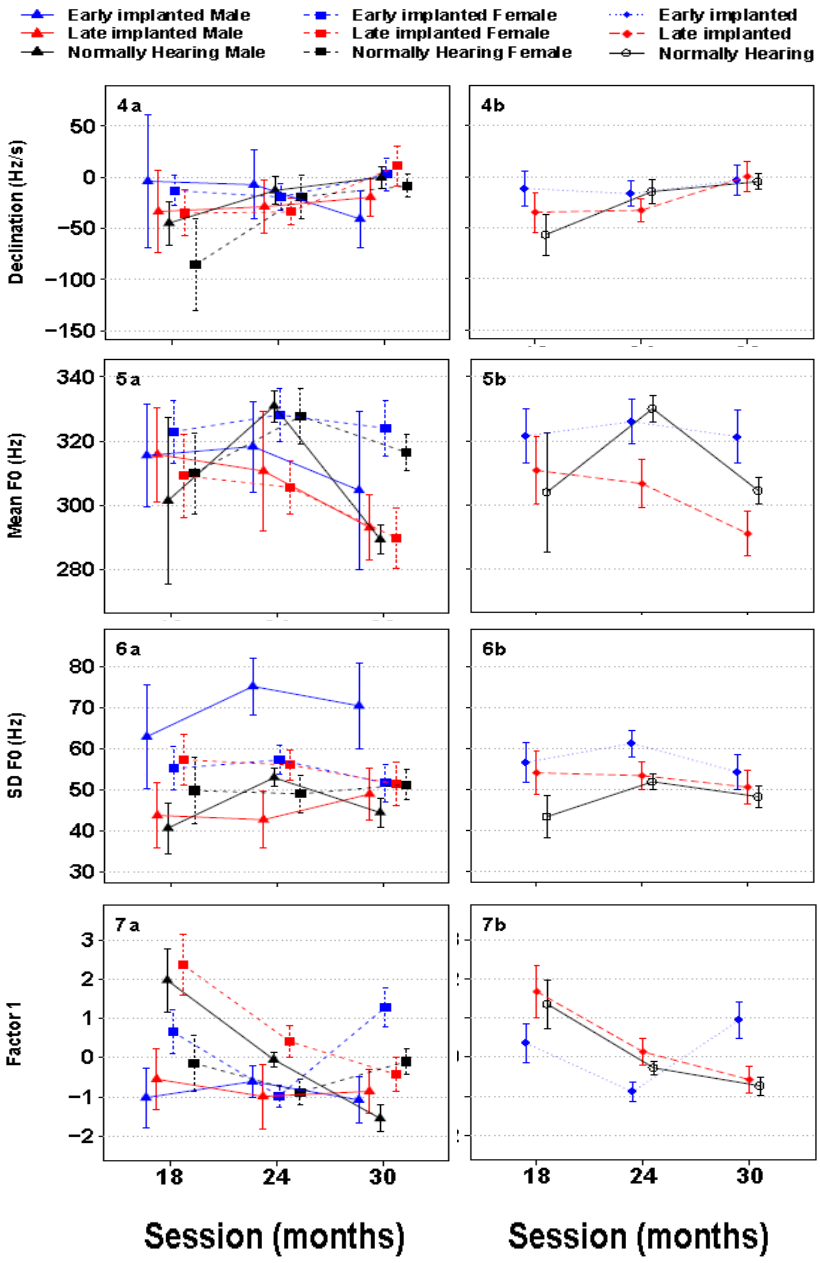

Figure 2 (cont.).

Interestingly, values of all our groups were in the range corresponding to the age of the youngest (NH) group, which suggest that hearing age, not chronological age, steered Mean F0.

SD F0 was best fit with Group, Gender, Session, Group × Gender, and Gender × Session. The Gender × Session interaction was the only interpretable effect. We can see in Figure 2, Panel 6a, that in general girls had extremer values and more variability than boys. There was, however, no overall difference in development between groups. The higher values for SD F0 for CI recipients (85.5 Hz for EI, 52.7 Hz for LI) as compared to controls (49.6 Hz) were in line with the predictions. The LI were, however, closer to the NH than the EI were. These values, especially those of the EI group, were considerably higher than those reported in an exploratory study on normative voice measurement values for younger and older adults (Goy, Fernandes, Pichora-Fuller, & van Lieshout, 2013), i.e., 26 Hz for males and 45 Hz for females. However, the participants in that study were much older (mean age 19.1 y. for the younger group) than those of the present study. This might explain the difference, as it has been suggested that with maturation children's voices become more stable (Kent, 1976). The literature shows mixed results concerning the effects of implantation age and implant experience on long-term frequency variability in implanted children. Holler et al. (2010) observed only an effect of time in sound (i.e., the sum of the time before the onset of deafness and the time since implant activation). Hsu et al. (2013) found an improvement (i.e., reduction of variability) as a function of experience, but no effect of implantation age. In a study by Campisi et al. (2005), there was no influence of implantation age nor of device experience. The current study is in agreement with results showing a convergence over time to normal values and more normal starting values for later implanted children.

Factor 1 was fit with Group, Gender, Session, Gender × Session, and Group × Session. Interpretable are the effects of Gender × Session and, our focus, Group × Session. Factor 1 was a combined factor. It therefore did not afford a prediction in the direction of

possible deviation nor for a direct comparison with previous research. The high correlation of the three variables of Factor 1 (APQ, HNR and PPQ) is in agreement with previous literature (Hillenbrand, 1987). The measures most likely all stem from glottal pulse irregularity. Higher PPQ relates to higher APQ, in part because the energy from one pulse interacts with the energy from the next, more variability in pulse duration resulting in more variability in inter-pulse intensity resonance. The correlation between HNR and perturbation measures is due to shifts in measured zero-crossings (PPQ), and contributions to the pitch-pulse amplitudes (APQ) as a result of added random fluctuations, respectively (Hillenbrand, 1987). Because of this mechanism underlying the correlation between its three measures, we consider Factor 1 as the laryngeal factor. As reflected in Figure 2, Panel 7b, the LI children developed in parallel with the control group, following a downward trend, whereas the EI children had their very own trajectory, starting lower and ending higher. This could entail that laryngeal control requires maturation more than speech experience.

To summarize, we predicted that prosodic measures would differ between participant groups, with larger deviations from the norm for the LI than for the EI children. No interpretable main effects of Group were found, but we did observe a significant three-way interaction (Group × Gender × Session) on ArtRate as well as significant interactions between Group and Session, indicating differential developments, on DurUtt, Voicing Ratio, Mean F0, and Factor 1. For the Group × Session interactions, the LI showed a more constant development (or lack of development) than the EI on DurUtt, Voicing Ratio, and Factor 1, but not on Mean F0, where the EI were very constant but where the LI's values decreased much more. The LI's values were closer to the NH's than the EI's value on DurUtt, two out of three sessions of Mean F0, and Factor 1, but not on Voicing Ratio, where the two recipient groups were about equally different from the controls. On Declination and SD F0, no main effect of or interaction with Group surfaced as significant.

2.4 Discussion

The aim of this study was to compare the development of two dimensions of phonetic measures of prosody in the spontaneous speech of children with early (EI) and late (LI) cochlear implantation with those of normally hearing (NH) peers. These dimensions were the temporal (Articulation Rate, Utterance Duration, Voicing Ratio) and the spectral (Declination, Mean F0, Standard Deviation of F0) dimensions. A separate factor (Factor 1) was constructed as an arithmetic combination of Amplitude Perturbation Quotient, Harmonic-to-Noise Ratio and Pitch Perturbation Quotient. On both dimensions, deviations for CI recipients have been observed in the literature, but they have not systematically been compared in spontaneous speech production across different measures. We predicted that (1) CI recipients and controls would differ from each other, (2) they would differ least on the temporal and most on the spectral measures, (3) EI children would differ less from controls than LI children and (4) differences from the norm would diminish with increasing implant experience.

First of all, there were two confounding factors in this study, viz. chronological age and gender. We will discuss these two issues. As outlined in the Statistical Analysis section and Table 1 (see the column 'Mean age over recordings'), the three participant groups had statistically different mean chronological ages. This was an unavoidable consequence of selecting for differential implantation ages while matching for hearing age. We have to take into consideration that any differences found between these groups could in principle also have been caused by age differences, or a combination of hearing age and chronological age. There are, however, two arguments to consider the age effect negligible. First, as an approximation of the effect of chronological age, we obtained Pearson correlations between chronological age and all of the dependent variables for all Group, separately. Out of 27 (i.e., 9 variables × 3 groups) cells, 13 correlations were below 0.1, 8 were

below 0.2, while the largest coefficient was 0.409. This suggests that chronological age does not greatly influence any of the dependent variables. Second, for some measures, the pattern of results is not consonant with what would be predicted on the basis of the groups' chronological age. DurUtt is expected to increase with age, but the oldest group (LI) had values in between those of the other groups. On Voicing Ratio, groups did not clearly differ (apart from their developmental path). For Declination, the Group effect was not significant, but a trend (shallower declinations for older children) contrary to hypothesis could be discerned for two out of three Sessions. The values of Mean F0 are anticipated to drop with age, but a clear difference (i.e., independent of Session) in that direction was only observed between the two recipient groups and, moreover, that difference was smaller than what was suggested by the literature given the age difference between the groups. On SD F0, the oldest group (LI) was below the middle group (EI) but they were both above the youngest group (NH). For these reasons we conclude that the role of chronological age is small at most and does not prevent us from drawing conclusions based on differences between groups. When there are no differences between groups, it can be argued that results are dependent on hearing age, not chronological age. When the CI recipients' values are too low or too high relative to the age of the NH group, this is a sign that their hearing status influences the prosodic parameters of their voice. When the same pattern of results anticipated based on age is shown for all groups, this can be interpreted as a sign that cochlear implantation does not prevent a normal age-based development for this measure.

The second confounding factor was Gender. Gender was involved in effects on most measures (all but F0 and Declination) and, given that proportions of Gender were not equal across groups, that factor could potentially explain (some of) the effects of Groups. But note, first, that the proportion of Gender was only different between controls on the one hand and CI recipients on the other hand (i.e., not between the two recipient groups). And second, whereas girls were

more variable in their development on DurUtt and Factor 1, the NH, despite their higher proportion of girls, were not more variable than the CI recipients. Likewise, the extremer and straighter development on Voicing Ratio and SD F0 for girls was not reflected in the trajectory of the NH group. We therefore feel safe to conclude that Gender is not responsible for differences in comparisons between recipient groups and the control group.

Our hypotheses were partly borne out. The first hypothesis (the CI recipients' measures differ from those of the controls) was supported for some, but not all, measures, although always in interaction with Gender and/or Session. This implies that hearing through a cochlear implant affects the development of speech due to the period(s) of atypical auditory sensations before and/or after implantation. This is in line with earlier literature reporting vocal deviations for CI children (e.g., Baudonck et al., 2015; Evans & Deliyski, 2007; Hocevar-Boltezar et al., 2006; Horga & Liker, 2006; Lane et al., 1998; Neumeyer et al., 2010; Oster, 1987; Poissant et al., 2006; Szyfter et al., 1996; Ubrig et al., 2011; van Lierde et al., 2005). This could imply that the atypical hearing situation of this population affects its vocal output in a general sense. It does not, however, specify to what level of perceptual detail this connection has an effect, i.e., if all acoustic parameters would be equally affected or if more problematic parameters would be more affected than relatively successful parameters. Our second hypothesis, the main focus of this study, was aimed at shedding light on that issue. We conjectured that CI users' voice deviances would be larger for the spectral measures, and smaller for temporal measures. This prediction was not in general supported by the results. The developments of Groups differed on three temporal measures (DurUtt, Voicing Ratio, and, in interaction with Gender, ArtRate), one spectral measure (Mean F0), and on the laryngeal factor (Factor 1). No effect was found for two spectral measures (Declination and SD F0). Importantly, this suggests that there is no clear correspondence between the degree of perceptual difficulty with a phonetic parameter and proficiency for that same

parameter in production, as the poorer resolution for the spectral as opposed to the temporal dimension of the auditory signal was not reflected in a pattern of more deviant spectral than temporal speech measures.

Several previous studies have addressed the question of the relationship between perception and production performance of pediatric CI recipients. Peng and colleagues investigated Mandarin tone recognition and production by means of picture selection and naming, respectively (Peng, Tomblin, Cheung, Lin, & Wang, 2004). Across their thirty participants, they found a significant (r = .44) inter-test correlation. It has to be noted, however, that the correlation became non-significant when the top three performers were removed from the analysis. In another study, they compared appropriateness of elicited utterances' intonation with question vs. statement discrimination, finding a correlation of r = .65 (Peng, Tomblin, & Turner, 2008). Children with and without CIs in a set of experiments by O'Halpin (2009) had to decide whether utterances were compounds or phrases (e.g., *bluebottle* vs. *blue bottle*) and to identify which word in a phrase carried a focal accent. Scores on those tasks were compared to the participants' difference limens for F0, intensity and duration of synthetically manipulated nonsense syllables. O'Halpin concluded that the implanted children payed least attention to F0 cues, more to amplitude cues and most to duration cues. In production, however, these dimensions did not clearly differ from each other in their level of appropriateness. Moreover, interestingly, no correlations between participants' appropriateness of production and reliance on the acoustic dimensions was found except that an appropriate production of amplitude and duration was more related to a good perception of duration than of amplitude or F0. The results of this study suggest that despite differential perceptual competence of acoustic dimensions, this is not generally reflected in differential competence of those dimensions in production.

Nakata, Trehub, and Kanda (2012), testing Japanese pediatric CI recipients and NH controls, found a correlation of r = .56 for scores

on prosody-based emotion recognition and rated appropriateness of imitated prosody. In a study on Mandarin-speaking children, Zhou, Huang, Chen, and Xu (2013) reported a significant correlation (r = .56) between accuracy for lexical tone identification on a picture selection task, and intelligibility of tones produced by picture naming. If broken up into individual tones, the correlation was significant for only two out of the four tones tested.

Taken together, studies about the perception and production of prosody in CI users, although not consistently, provide some evidence of a relationship in performance abilities between the two. There is, however, no evidence for a relationship per acoustic dimension, i.e., perceptual performance on a specific dimension does not predict the performance on that dimension in production. The present study is in agreement with the latter finding, since no clear advantage for a presumably better dimension (temporal over spectral) was observed. A number of explanations for the lack of correspondence between perception and production in the current study could be proposed. First of all, for speakers in general, the proficiencies in production and perception of speech could be independent of each other. This, however, appears not to be the case, given that the present study as well as previous work have demonstrated that there are discrepancies in the speech of individuals with hearing impairment with or without cochlear implants (e.g., Evans & Deliyski, 2007; Lane et al., 1998; Oster, 1987; Perkell et al., 1992; Perrin et al., 1999; Seifert et al., 2002; Szyfter et al., 1996; Ubrig et al., 2011) (Ball & Ison, 1984; Fourcin et al., 2011; Kishon-Rabin et al., 1999; Menard et al., 2007;

Nguyen et al., 2008; Svirsky et al., 1998). As a more direct indication, speech is altered soon after temporarily switching a CI off or back on (Higgins et al., 2001; Monini et al., 1997; Poissant et al., 2006; Svirsky et al., 1992; Tye-Murray et al., 1996). A second, more plausible account, therefore, would be that there is a relationship between production and perception, but that the difference in auditory resolution between the two dimensions currently studied is not large enough to result in a difference in production. This is also unlikely

since the spectral and temporal resolution for most CI users cover two extremes, from very good to very poor, respectively (Moore, 2003; Shannon, 2002; Vorperian & Kent, 2007). A third possibility is that, although the spectral dimension is poorly processed, it is produced successfully because it is an automatic by-product of speech, i.e., it does not involve conscious linguistic or paralinguistic choices but is a physiological consequence of choices in other dimensions that may be consciously controlled. For instance, increasing a syllable's intensity for emphasis might be automatically paired with elevated pitch due to accelerated vocal fold vibration. Indeed, the two spectral measures showing a good performance, declination and SD F0, could be considered relatively uncontrollable variables, whereas the worse performance of Mean F0 could reflect its controllable nature. On the other hand, Factor 1 was relatively deviant, but would count as a less consciously controllable variable. Moreover, deviations in the temporal dimension would not be expected even for controllable variables, but they were found. All temporal measures were, however, in fact deviant as well as controllable and therefore it could be hypothesized that controllability plays a more important role than auditory resolution. This account is supported by at least two other considerations. First, our finding that CI recipients articulated faster on longer utterances (more so than the controls) could point to a limited verbal working memory span (Burkholder & Pisoni, 2003). That same limitation would also be part of the origin of a lack of control in the cases of prosodic parameters that require pronunciation choices assuming that would also be relatively taxing for verbal working memory. Second, the account would be in line with the claim that a lack of auditory feedback affects long-term parameters more than short-term parameters (Hsu et al., 2013), as both distinctions contrast the more linguistic with the more physiological parameters. Taking the above considerations together and abstracting away from underlying causes, we conclude that the quality, or lack thereof, of the acoustic speech dimensions received by implanted children is not directly reflected in comparable quality in those dimensions in their

output, but that instead the controllability of prosodic voice parameters seems to be a more determining factor.

Our third hypothesis was that the LI would show more deviant outcomes than the EI group because they experienced a longer period without stable auditory input. LI's values were in general closer than the EI's to the NH's values, viz. on a temporal parameter (DurUtt), part of a spectral factor (Mean F0) and Factor 1, but not on another temporal measure (Voicing Ratio). Further, the LI children showed a less changeable development than the EI children on two temporal measures (DurUtt, Voicing Ratio) and the laryngeal factor, but it was the other way around for one spectral measure (Mean F0). Therefore, it seems that LI children did not deviate more than EI children; if anything, it was the other way around. This is in disagreement with most of the literature on the language development of CI users, where earlier implantation is associated with outcomes closer to the norm or with faster development. One possible cause for this is that four out of nine LI children had a late onset of hearing loss (between 12 and 30 months). This might have given them an advantage relative to the EI group, since in the time spent with relatively normal hearing prior to hearing loss they would have had some opportunity to establish speech goals from which they could still benefit after implantation. This could have partly compensated for the possible disadvantage from late implantation, resulting in less difference between the LI and EI groups.

Another possible cause is the fact that we focused on the more specific issue of voice and speech measures. Within the literature about age effects, few studies have done that. Advantages for earlier implantation or longer time in sound at various ages have been found regarding various segmental and suprasegmental variables (Tobey et al., 1991), glottal measures (Hocevar-Boltezar, Vatovec, Gros, & Zargi, 2005) and nasality (Hassan et al., 2011b), but not for formant values (Neumeyer et al., 2010). In one longitudinal study, prelingually deaf CI recipients showed a faster improvement but with more deviant starting values than postlingually deaf adults on a range of glottal

measures (Hocevar-Boltezar et al., 2006). The results of the present study add to this overview by supporting the studies showing no benefit of earlier implantation (at any age) for prosody production. Instead, it does for some measures but not for others, possibly reflecting a compensatory combination of factors relating to perceptual resolution, controllability, implantation age and duration of hearing loss of the CI recipients. Future research should address a greater variety of measures and participant groups within a single study to disentangle these factors.

The fourth hypothesis stated that the differences between CI recipients and controls would decrease with increasing experience with the device and that this decrease would be faster for the early implanted than for the late implanted children. Groups converged over time on ArtRate (in interaction with Gender), DurUtt, Voicing Ratio, to some extent on Factor 1 (only LI and NH), and as a tendency on Declination and SD F0, but there was no convergence on Mean F0. These findings suggest that experience with the implant brought most voice parameters closer to the norm. This effect was stronger for temporal than for spectral measures. It held irrespective of implantation age. Our results resonate with previous reports showing improvement of some voice measures with increasing implant experience (Hassan et al., 2011b; Hocevar-Boltezar et al., 2006; Lenden & Flipsen, 2007), and especially research showing improvement of temporal (Goffman et al., 2002) but not spectral (Campisi et al., 2005) measures. Taken together, our results underline the suggestion that implant experience has a positive effect on prosody production, but more consistently so for temporal than for spectral measures.

Conclusions and future directions

The current study suggests that the appropriateness of different phonetic dimensions of the basic prosody of an utterance did not

directly reflect the auditory resolution for the corresponding acoustic dimensions. The higher resolution for temporal structure than for spectral detail did not in general entail more successful production of temporal than spectral aspects of prosody in an utterance. Instead, it seemed that the parameters that required a relatively high level of articulatory and/or laryngeal control or planning (ArtRate, DurUtt, Voicing Ratio, Mean F0 and perhaps DurUtt) were somewhat more problematic than the parameters that were by-products of speaking (Declination, Factor 1, and SD F0). The data in this study did not shown an advantage of implantation before vs. after two years of age, but the outcomes improved with increasing implant experience.

The results of this study could be used as a recommendation for speech therapists to pay attention to the early development of basic prosodic measures of implanted children. I.e., using recordings of relatively spontaneous speech, they would have to monitor the measures that are at the risk of deviating and rehearse the necessary glottal and articulatory control and verbal working memory. It should be noted that the development of prosody can differ between parameters, between early and late implanted children and between genders. In future research, more different phonetic parameters should be compared in order to investigate more deeply the underlying cause of problems with some but not other parameters. It is also recommended that production results are directly compared with individuals' auditory resolutions on different dimensions, in an attempt to elucidate the possible correlation between perception and production in children with cochlear implants. Finally, in order to more clearly separate the effects of chronological age and hearing age, it would be advisable to orthogonally compare those two factors by testing early and late implanted children with the same chronological age, on the one hand, and with the same hearing age, on the other.

Acknowledgements

We are thankful to the Leiden Institute for Brain and Cognition (LIBC) for supporting this research. We are also grateful to ing. Jos Pacilly of the Leiden University phonetics laboratory for support involving signal processing. We thank Walter Verlaan of the LUMC, who helped with software and hardware issues related to recording and digitization. Statistician Vincent Buurman of the Faculty of Social Sciences helped us tremendously in analyzing the data.

[1]*Praat* manual, *Voice 3. Shimmer*.

[2]With these settings for analyzing HNR, analysis windows did not overlap, since with children's typical the analysis window is shorter than the time-step of 10 ms. With this procedure results are not based on the complete signal. In an informal comparison of the two procedures (non-overlapping vs. overlapping with 4.5 windows per period) the HNR values in the non-overlapping procedure were shown to be between 10% and 50% higher than with the overlapping method. It therefore has to be taken into account that with the overlapping method, lower HNR values would have been found.

[3]*Praat* manual, *Voice 3. Jitter*.

Chapter **3**

The effect of spectral smearing on the identification of pure F0 intonation contours in vocoder simulations of cochlear implants

This chapter is based on:

van de Velde, D. J., Dritsakis, G., Frijns, J. H., van Heuven, V. J., & Schiller, N. O. (2015). The effect of spectral smearing on the identification of pure F0 intonation contours in vocoder simulations of cochlear implants. *Cochlear Implants International, 16*, 77-87. Doi: 10.1179/1754762814Y.0000000086

Abstract

Objectives: Performance of cochlear implant (CI) users on linguistic intonation recognition is poorer than that of normally-hearing listeners, due to the limited spectral detail provided by the implant. A higher spectral resolution is provided by narrow rather than by broad filter slopes. The corresponding effect of the filter slope on the identification of linguistic intonation conveyed by pitch movements alone was tested using vocoder simulations.

Methods: Re-synthesized intonation variants of naturally produced phrases were processed by a 15-channel noise vocoder using a narrow (20 dB/octave) and a broad (40 dB/octave) filter slope. There were three different intonation patterns (rise/fall/rise–fall), differentiated purely by pitch and each associated to a different meaning. In both slope conditions as well as a condition with unprocessed stimuli, 24 normally hearing Dutch adults listened to a phrase, indicating which of two meanings was associated to it (i.e., a counterbalanced selection of two of the three contours).

Results: As expected, performance for the unprocessed stimuli was better than for the vocoded stimuli. No overall difference between the filter conditions, however, was found.

Discussion and conclusions: These results are taken to indicate that neither the narrow (20 dB/octave) nor the shallow (40 dB/octave) slope provide enough spectral detail to identify pure F0 intonation contours. For users of a certain class of CIs, results could imply that their intonation perception would not benefit from steeper slopes. For them, perception of pitch movements in language requires more extreme filter slopes, more electrodes, and/or additional (phonetic/contextual) cues.

3.1 Introduction

With the current implant technology, most users of cochlear implants (CI) can develop a good general understanding of speech in favorable listening circumstances. However, the average performance of implant users in the perception of speech intonation remains much poorer than that of normally-hearing (NH) listeners (Chatterjee and Peng, 2008; Peng et al., 2009; Souza et al., 2011). Intonation is a type of prosody. Prosody, or the 'melody of speech', refers to the combined phonetic aspects of an utterance that cannot be explained by effects of the (juxtaposition of) individual vowels and consonants. For instance, all vowel categories in a language have intrinsic fundamental frequencies, but this is not an example of prosody, since those frequencies are predictable from the type of the vowel (Rietveld and van Heuven, 2009). Some important acoustic parameters of prosody are pitch (F0) movements, intensity changes, and temporal structure. The problems that CI users have with perceiving intonation are associated with intonation being primarily conveyed by F0 movements.

CI users have problems with spectral perception for a number of reasons. First, the F0 is usually not directly transmitted because that frequency may be too low. Second, F0 cannot be reconstructed from higher harmonics, because those harmonics are not resolved. Third, in as far as pitches can be differentiated, the resolution is very low, because the spectral bands that the signal is analyzed into overlap (Faulkner et al., 2000; Green et al., 2004; Qin and Oxenham, 2005). Nevertheless, some degree of pitch perception in the F0 range has been shown to be possible. One of the mechanisms proposed to account for this is that the listeners are cued by the dynamic envelope of higher unresolved harmonics, because this envelope varies with the same frequency as F0 (Green et al., 2004).

There exists a large variability in speech performance between implant users, due to device- and patient-related factors such as the type of implant, duration of deafness, and age at implantation (Boons

et al., 2012; Geers et al., 2013; Lazard et al., 2012). In order to control for the effects of confounding parameters on speech perception of CI users, vocoder simulations have been widely used with NH listeners (Dorman and Loizou, 1998; Shannon et al., 1995; Crew et al., 2012). Vocoders process speech in a manner comparable to the implant processor. The signal is first analyzed into a number of frequency bands. Subsequently, the temporal envelope is extracted for each frequency band, for which the signal is low-pass filtered and used to modulate a noise or sine carrier (Loizou, 2006). Noise-vocoded speech has been shown to better model F0 perception by CI users than sine-vocoded speech. It is suggested this occurs because sine-wave vocoding provides spectral detail not available in a CI as opposed to noise-band vocoders which eliminate fine-structure cues (Whitmal et al., 2007; Souza and Rosen, 2009). Noise-band vocoders have been shown to produce speech intonation perception scores consistent with CI users' outcomes (Chatterjee and Peng, 2008; Peng et al., 2009).

As mentioned above, the main motivation for using vocoders instead of actual patients is the control of patient- and device-related parameters. Characteristics such as the duration of deafness, the age at implantation, the duration of implant use, and the etiology of deafness are inherent to hearing impaired but not to NH listeners. Vocoders also allow for manipulation of individual signal processing parameters that could affect perception. Most parameters in real patients' devices, including the speech processor algorithm, are individual to the patient and are not subject to adjustment. As a result, experimental investigation that requires the control over fine signal processing settings with larger groups of subjects would not be possible with CI patients. Other advantages, as mentioned by Laneau et al. (2006), are that the comparison between the acoustic model of NH listeners and the electric model of CI users provides theoretical insights into the mechanism of hearing and it also reveals potential causes for limitations by CI users. Finally, a much larger pool of NH subjects than of CI users is in general available, making investigations on CI perception considerably more feasible for researchers. If an accurate

model for CI perception can be found, research on CI performance has the potential to become larger in scale.

Previous studies have used noise-band vocoder simulation to assess some aspects of speech perception of CI recipients using at least two different approaches with regard to spectral resolution: by varying the number of analysis channels, or by producing different degrees of channel interaction. Different numbers of frequency channels are used to simulate the number of electrodes of the CI processor. Although increasing the number of frequency bands increases spectral detail and has been reported to improve speech perception (Fu et al., 2005; Henry and Turner, 2003; Qin and Oxenham, 2005; Stone et al., 2008), no significant differences in performance are generally observed when the number of channels increases beyond six to eight (Dorman et al., 1997, Xu et al., 2005). Thus, a relatively limited spectral resolution suffices for reasonable speech recognition. Remarkably, Shannon et al. (1995) demonstrated that a high level of speech recognition can be achieved with as few as four analysis channels. It has been suggested that the number of spectral channels transmitted by the vocoder cannot completely account for the poorer performance of CI users compared to NH listeners in speech recognition and discrimination tasks (Dorman et al., 1997; Friesen et al., 2001; Henry and Turner, 2003; Shannon et al., 1998). The degree of channel interaction simulated by using narrower and shallower filter slopes is an additional factor to consider.

Channel interaction refers to the overlap of spectral regions of adjacent electrodes as a result of spread of excitation, which is known to occur in CIs. Interaction between channels reduces the spectral detail of a signal (‘spectral smearing’) and, as a consequence, it compromises speech recognition performance (Crew et al., 2012; Henry and Turner, 2003). Previous studies have tested the effects of spectral smearing by varying the slope of the noise filters. Shannon et al. (1998) tested the effect of spectral smearing on speech recognition by comparing 3, 6, and 18 dB/octave filters with a standard condition with almost no channel overlap. They found that the performance on

the recognition of sentences, vowels, and consonants was still very high with the steepest slope, but that the performance with the 3 and 6 dB/octave filters, although above chance, was significantly below that of the standard condition. Fu and Nogaki (2005) showed that speech recognition is better with a 24 dB/octave filter than with a 6 dB/octave filter. In a study by Litvak et al. (2007), the number of channels (15) was kept constant and four different filter slopes were used (5, 10, 20, and 40 dB/octave). Recognition scores for synthetic vowels and consonants decreased with shallower filter slopes. Comparing the results to data from Saoji et al. (2005), they concluded that the slopes in the region from 4 to 30 dB/octave matched the CI performance well (i.e., there was no significant difference). Still, the CI patients performed slightly worse than the NH subjects, which was hypothesized to be due to the NH listeners benefiting from dynamic and temporal cues more than CI users do. More recently, the effects of channel interaction were investigated in a non-speech pitch task. Crew et al. (2012) used sinewave vocoder simulations with 16 band-pass filters and 3 filter slopes (24, 12, and 6 dB/octave) to test musical pitch perception. They replicated the findings of speech perception studies with steeper slopes producing more channel interaction and poorer performance in melodic contour identification. The authors suggested that the results were comparable to those of CI users in Zhu et al. (2011) and to those in Luo et al. (2007) who used sine-wave vocoder simulations as well. It should be noted here, however, that, as explained above, noise-band vocoder simulations might be more representative of a CI on pitch-related tasks due to the limited spectral information provided.

Although studies have been devoted separately to intonation perception and to channel interaction with vocoders, there is a paucity of research on the combination of the two. Therefore, the present study is concerned with the effect of channel interaction on intonation perception. In linguistics, intonation is analyzed as a series of pitch accents, i.e., connected F0 targets lending prominence to some of the syllables in an utterance (Ladd, 1996). Although different acoustical

parameters (cues) covary in the production of accents, pitch movements have been claimed to be by far the most important perceptual cue to the presence of an accent. This prominence of pitch movements for accents is in contrast with the perception of stress, for which durational and intensity cues are relatively important (van Heuven and Sluijter, 1996). This difference makes intonation more suitable than stress for the examination of linguistic pitch pattern perception. Moreover, we chose a type of pitch accent in Dutch, the variants of which are believed to be distinguished from each other by pitch movements only: accents with the pragmatic meanings of news, surprise, and predictability (Rietveld and van Heuven, 2009). The drawback of multi-cue phenomena would be that researchers can be less certain that stimuli in which only one of those cues is manipulated are processed as in natural language perception, since in natural language processing the cues would not be isolated. Pure pitch intonation is not uncommon in languages as, for instance, the distinction between questions and declarative sentences can also fall in this category (Rietveld and van Heuven, 2016). Because this type of intonation is expected to be especially difficult to perceive for CI users, the problem of intonation perception by CI users is a real issue. Because the perception of speech melody requires more spectral detail than the perception of the segmental (vowels and consonants) layer of speech (Smith et al., 2002), we used relatively steep filter slopes.

3.2 Methods

The identification of speech intonation contours was examined using a 15-channel noise-band vocoder with a 40 and a 20 dB/octave noise filter. The simulation algorithm used for the present study was the same as the one used in Litvak et al. (2007). Listeners performed an intonation identification task listening to vocoded and unprocessed speech. The stimuli were three basic Dutch melodic shapes, which are thought to be conveyed solely by F0. The intensity and the duration of

the stimuli were kept constant by using the same recorded phrase as a basis for superposition of all three intonational variants, whereas the filter slope was systematically varied in order to manipulate the amount of spectral detail available to the participants. Our first hypothesis was that NH listeners would have more difficulty in discriminating melodies in the vocoded than in the unprocessed condition. Our second hypothesis was that a steeper slope (40 dB/octave) should induce better recognition than a shallower slope owing to the smaller amount of channel interaction (Litvak et al., 2007; Shannon et al., 1995; Souza et al., 2011). If participants would be unable to perform the task with these settings, this would imply that more extreme filter slopes, possibly a higher number of electrodes and/or additional (non-pitch related) cues were required for the identification of intonation.

3.2.1 Participants

Twenty-four (14 female, 10 male) native Dutch speakers with NH, aged between 18 and 27 (mean age = 22.5 years) agreed to participate in this experiment as volunteers. Although no formal tests were performed for this, participants did not report any hearing or cognitive problems; all were (graduate or undergraduate) students of Leiden University. Participants were naive to the scientific goal of the experiment. Prior to the experiment, they were given ample time to read an information form which explained the setup of the experiment and the tasks they would be asked to perform. The study was approved by the Ethical Committee Social Sciences and Humanities at Leiden University. Participants had the right to withdraw from participation at any time during the procedure without any negative consequences for them.

3.2.2 Stimuli

Seven different short phrases were recorded by a male native Dutch speaker (DV) using a Sennheiser MKH416T condenser type microphone and Adobe Audition 1.5 (Adobe Systems, San Jose, CA,

USA): *een verbanddoos* (a first aid kit), *een cadeaubon* (a gift certificate), *morgenavond* (tomorrow evening), *over een uur al* (in an hour already), *naar de Veluwe* (to the Veluwe), *naar Leeuwarden* (to Leeuwarden), and *een agenda* (an agenda). This last phrase was only used as a practice stimulus. The phrases were selected because they were semantically and pragmatically logical utterances that could be produced with the three intonation types envisioned, consisted mainly of voiced segments and had a similar stress pattern (viz. main stress on the penultimate syllable, or the antepenultimate syllable if the final syllable was the reduction vowel 'schwa'). The recording sampling rate was 44,100 Hz, and the sampling resolution was 16 bit. The phrases were originally produced by the speaker with a rise–fall intonation to express the information as 'news'. The tokens were stylized using *Praat* software, *Version 5.3* (Boersma & Weenink, 2012) and then re-synthesized to obtain the three different F0 contours. For this manipulation, a rising or falling F0 contour was superimposed on the natural declination of the utterance. The contours were created as pitch accents, so they had the approximate duration of a syllable. The duration of the accent and of the whole phrase was the same for all three contour types per phrase. The phrase durations varied between 740 and 1000 ms. All manipulations had a nine semitone range between the high and low declination. In the rising contour, the range was twelve semitones at the end of the utterance. Since the upper line does not decline towards the end of the utterance, the distance between the lower declination line and the upper line became larger than nine semitones, yet the rise itself still covered only nine semitones. The dynamic range was scaled to 0.99 and the durations of all tokens of each phrase were equalized. This process yielded three versions of each of the seven phrases varying only in the shape of the pitch contour: a falling, a rising, and a rising–falling contour. The resulting set of re-synthesized stimuli comprised 18 stimuli. Importantly, the use of naturally produced re-synthesized stimuli, in comparison with previous studies where synthetic

phonemes were used (Litvak et al., 2007; Shannon et al., 1995; Souza et al., 2011), ensured that the stimuli were relatively realistic.

For the 18 stimuli, noise-band vocoder processing was implemented using *Matlab R2014a* (The MathWorks, Inc., Natick, MA, US), following the same algorithm as described in Litvak et al. (2007). The basic steps were as follows: the stimuli were digitally sampled at 17,400 Hz and then analyzed with a short-term Fourier transform. This output was grouped into 15 non-overlapping, logarithmically spaced analysis channels. The envelope of each band was extracted by averaging the square root of the total energy in the channel, implying a low-pass filter of 68 Hz. This output modulated a similarly synthesized noise band, which had the same center frequency as the analysis channel but the slope of which (the rate of the drop-off of the noise spectrum away from the center frequency) was either 20 or 40 dB/octave to simulate two different amounts of spread of excitation that may occur in an electrically stimulated cochlea (Litvak et al., 2007). This process yielded 54 stimuli (6 phrases × 3 contours × 3 processing conditions). Center and cut-off frequencies of all bands are given in Table 1.

Since the applied vocoder processing does not pass frequencies under 350 Hz, no direct cues for the F0 below that threshold were available, the highest F0 in the intonation contours of our stimuli being 200 Hz (the highest frequency in the falling contour version of *een cadeaubon*). Instead, we conjectured that judgements had to be based on spectral differentiation of higher harmonics and/or on the dynamic envelop of (resolved or unresolved) harmonics. For tonal contour re-synthesis, *Praat* applies Pitch-Synchronous Overlap and ADD (PSOLA) (Moulines and Charpentier, 1990), which creates the tones as glottal pulses and thus includes harmonics. The spectral smearing introduced by the vocoder processing most likely rendered the harmonics unresolved and thus the most probable cue, if present, would be the temporal envelope of the harmonics.

Table 1. Center and cut-off frequencies of the 15 non-overlapping bands produced by the vocoder algorithm.

Band number	Center frequency (Hz)	Lower cut-off frequency (Hz)	Higher cut-off frequency (Hz)
1	384	350	421
2	461	421	505
3	554	505	607
4	666	607	730
5	800	730	877
6	961	877	1053
7	1155	1053	1266
8	1387	1266	1521
9	1667	1521	1827
10	2003	1827	2196
11	2407	2196	2638
12	2892	2638	3170
13	3475	3170	3809
14	4176	3809	4577
15	5017	4577	5500

3.2.3 Procedure

The speech intonation identification task was conducted in a sound-treated booth. All sound stimuli were presented via Sennheiser HD414SL headphones. The subjects were seated 1 m from the computer screen and gave their answers by pressing buttons on a keyboard. Before the experiment, the participants were given detailed written instructions for the task. The participants first completed a training session of approximately 10 minutes, designed to familiarize them with the type of stimulus used and with the experimental task. The stimulus used in the training session was different from the test stimuli and it was the same phrase throughout the session. Correctness feedback was presented on every training trial.

In both the training and the experimental sessions, the target contours were associated with semantic labels: rise, fall, and rise–fall

were labelled as surprise, predictability, and news, respectively. These semantic labels were used in order to evoke linguistic instead of acoustic judgements, i.e., to make the participants less conscious of acoustic patterns as such but to listen to it as speech. Nevertheless, a genuine correct response could not be given (conscious or unconscious) without identification of the acoustic patterns. Based on performances observed during a pilot study, the task of identifying the meaning from the pitch contour (in the training session as well as in the experimental session) was made easier by reducing the number of patterns for the listeners to choose from three to two. The meaning of the target pattern, the written phrase, and a picture of the pitch contour shape were presented on the computer screen while the stimuli were played. This last addition was also based on the pilot study and was meant to help participants to recognize the contours. Ideally, participants would perform the identification based on the meaning of the utterance and use the pictures of the contour shapes as a support for their judgements.

In the experimental phase, the entire stimulus set was presented twice and in pairs. Each trial contained one pair of stimuli. This yielded 54 pairs (1 pair for every 1 of 6 phrases × 3 contours × 3 processing conditions) × 2 repetitions = 108 trials. The total set was divided into three blocks of 36 pairs each. In each block, there was a different target pattern to identify. The order of the six blocks was counterbalanced across listeners (four participants for each block). In each trial, two stimuli of the same vocoder condition but a different pitch contour were presented sequentially. The order of the presentation of the two stimuli (target and non-target) was counterbalanced over contour pair within blocks. Conditions within each block were randomized. Each stimulus had a duration of 1000 ms, and the silent interval between trials, during which responses were collected, was 4000 ms. Stimuli were presented at loudness levels of normal speech listening (around 65 dB SPL). The listener had to indicate which of the two stimuli expressed the target pattern in a two-way alternative forced choice task by pressing number 1 for the first

or number 2 for the second sound, on the numerical part of the keyboard, with two different fingers. The response accuracy and reaction time were collected. Reaction times were measured from the onset of the second of the pair of phrases played instead of at the end because decision making could in principle start during (not after) the second phrase. Although the durations of the phrases varied, reaction times were not corrected for this because all stimuli were equal between processing conditions.

An experimental session lasted fifteen minutes. No feedback was provided during testing. The experiment was set up and controlled by *E-prime 2.0* software (Psychology Software Tools, Pittsburgh, PA, USA; Schneider, Eschman, & Zuccolotto, 2012). The sessions took place at the Leiden University phonetics laboratory, over a period of three weeks depending on the availability of the participants. Statistical analysis was performed using *IBM SPSS, Version 21*; a significance threshold of $p = 0.05$ was adopted.

3.3 Results

Null responses (1.0% of the cases) and responses with a reaction time of more than two standard deviations (608 ms) from the original mean, i.e., reaction times below 426 ms and above 2588 ms, were excluded from further analysis. They were considered either unreliably fast (responding without full processing of the stimulus) or unreliably slow (responding with too much interference from higher-order cognitive functions) outliers, as is usual in psycholinguistic research (Baayen and Milin, 2010). This omission represented 8.0% of the non-null-response cases (0.5% too fast, 7.5% too slow). No further data were eliminated. In the discussion that follows, the dataset in which null responses were excluded but not the cases with extreme reaction times will be referred to as the ‘reduced dataset’, whereas the dataset in which only the null responses were excluded will be referred to as the ‘larger dataset’.

All cases that were too fast (1.6% within that condition) were in the unprocessed conditions. The unprocessed condition had the smallest percentage of too slow cases (4.4, 10.6, and 7.7% in the unprocessed, 20 dB/octave, and 40 dB/octave conditions, respectively). The number of cases eliminated differed significantly across filter slope and target contour conditions, as revealed by a Pearson Chi-square test (for filter slope, $\chi^2(2) = 21.32$, $p < 0.001$; for target contour, $\chi^2(2) = 8.02$, $p = 0.018$). However, this was due to the relatively low number of eliminated cases in the unprocessed condition. When the unprocessed condition was left out of the comparison, the Chi-square test was no longer significant (for filter slope, $p = 0.067$; for target contour, $p = 0.32$). This last Chi-square test was considered more meaningful than the Chi-square test with the unprocessed condition included. This is for two reasons. First, the number of too slow cases in the unprocessed condition is expected to be lower than in the two processed conditions to begin with. And at the same time, a difference in eliminated cases between precisely the two processed conditions is critical. Second, the results for the main effects (analyzed in the same way as the main analysis) were the same for all comparisons as with the reduced dataset, i.e., the same presence or absence and direction of effects. The results were also the same for all post-hoc comparisons, except for two that did reach significance with the reduced dataset but not with the larger dataset. After Bonferroni correction for the nine pairwise comparisons of target contours, i.e., three for each filter slope condition ($p = 0.05/10 = 0.005$ was adopted), the difference in accuracy between the rise and rise–fall contour in the unprocessed condition was only marginally significant ($F(1,311) = 8.64$, $p = 0.007$), and the difference in reaction times between the rise and rise–fall contour in the 40 dB/octave condition was not significant ($F(1,281) = 5.45$, $p = 0.027$). Taken together, however, on the basis of the lack of differences in main effects between the full and reduced dataset and the fact that the number of eliminated cases was not significantly different between the processed conditions, it was assumed that there is no reason to believe that the

eliminated cases influenced the current dataset in a meaningful way. Further analyses are based on the reduced dataset because the remaining trimmed reaction time values were believed to more reliably reflect task processing than the reaction times with the outliers included.

A repeated-measure analysis of variance (ANOVA) using the Huynh–Feldt adjustment for degrees of freedom was run on the remaining data. It revealed a significant effect of the vocoder condition both on the percentage of correct responses ($F(2,46) = 135.77$, $p < 0.001$) and on the response latencies ($F(1.385,46) = 22.15$, $p < 0.001$). Subsequent tests indicated that performance in the unprocessed condition (90.3%) was significantly better than in the 20 dB/octave slope both for the response accuracy ($F(2,46) = 201.27$, $p < 0.001$) and for the reaction times ($p < 0.001$) and the 40 dB/octave slope both for the accuracy ($F(2,46) = 258.22$, $p < 0.001$) and for the reaction times ($p < 0.001$). However, there was no difference between the 20 and the 40 dB/ octave conditions.

Accuracy for the unprocessed conditions, as tested with a binomial test over the frequencies of correct and incorrect responses per condition, was significantly above chance in the unprocessed condition ($p < 0.001$) and for the 20 dB/octave condition ($p < 0.001$), but not for the 40 dB/octave condition (the test proportion was defined as 0.50 because although there were three levels in the target contour condition, subjects only chose between two of them per trial). Out of 24 subjects, 18 scored better in the 20 dB/octave than in the 40 dB/octave condition.

Figs. 1 and 2 show the effects of the vocoder condition and the pitch contour on the percentage of correct responses and on the response latencies, respectively. The graphs demonstrate a better performance for the unprocessed (90% correct responses) than the stimuli of the two processed conditions (57% and 52% correct responses for the 20 and 40 dB/octave conditions, respectively). As for the target contour, no effect was observed, but there was a significant interaction between contour and slope conditions ($F(8,16)$

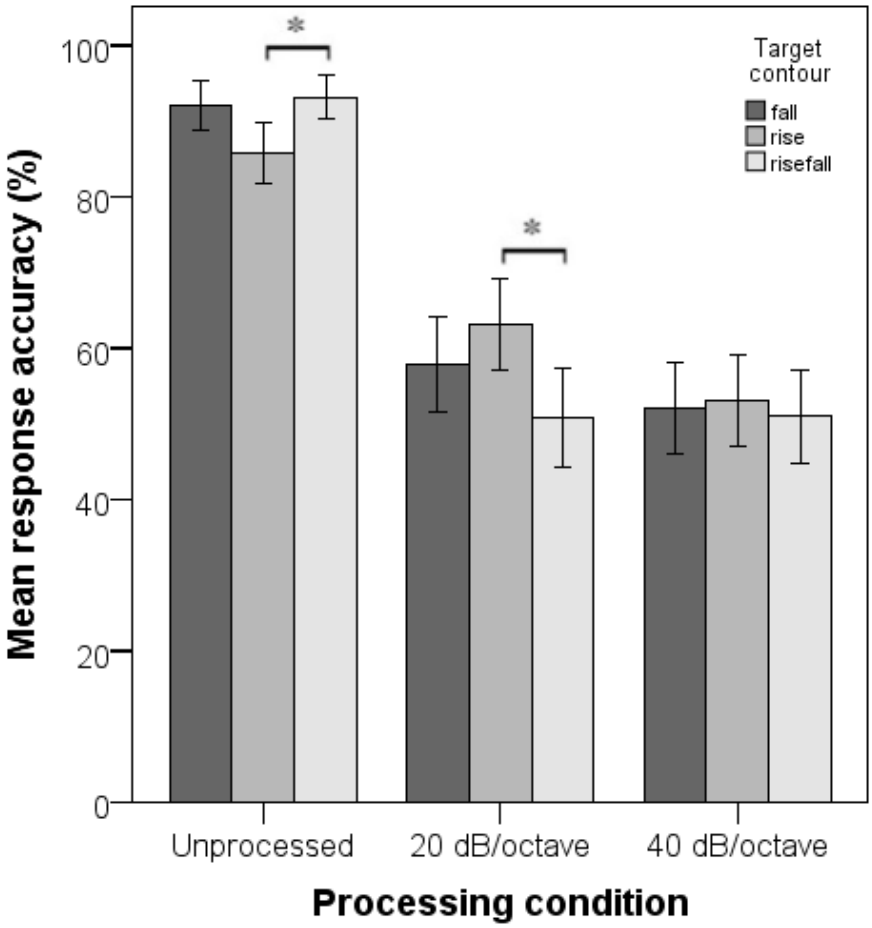

Figure 1. Response accuracy (%, percentage correct) for three intonation contours as a function of processing condition. Error bars display 95% confidence intervals. The overall effect of processing condition between the unprocessed and the 20 dB/octave condition and between the unprocessed and the 40 dB/octave condition, respectively, was significant ($p < .001$). *Significant ($p \leq .05$).

= 3.93, p = 0.01). Performance for the rise–fall contour was significantly better than for the rise in the unprocessed condition for the accuracy (p = 0.013) but not for the reaction time. In the 20 dB/octave condition, the rise was significantly better than the rise–fall for the accuracy (p = 0.008) but not for the reaction times. No other significant interactions were observed. The difference in the reaction

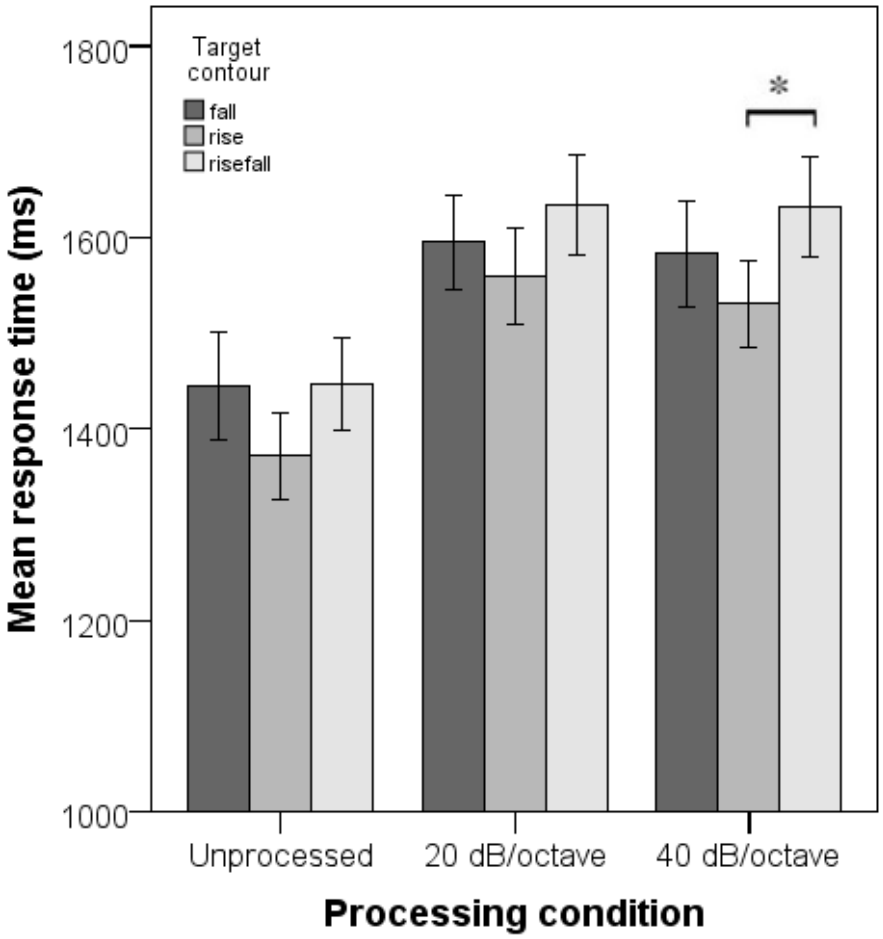

Figure 2. Reaction times (ms) for three intonation contours as a function of processing condition. Error bars display 95% confidence intervals. The overall effect of processing condition between the unprocessed and the 20 dB/octave condition and between the unprocessed and the 40 dB/octave condition, respectively, was significant ($p < .001$). *Significant ($p \leq .05$).

times between the rise–fall and the rise in the 40 dB/octave condition was marginally significant ($p = 0.050$; higher for the rise–fall than for the rise). Table 2 shows the mean and standard deviations of the response accuracy and reaction times of target intonation contours per

Table 2. Subjects' mean values and standard deviations (SD) of accuracy (%, percentage correct) and reaction times (RT) for intonation contours and processing conditions. Error bars display 95% confidence intervals. The overall effect of processing condition between the unprocessed and the 20 dB/octave condition and between the unprocessed and the 40 dB/octave condition, respectively, was significant ($p < .001$).

Processing condition	**Intonation contour**							
	fall		**rise**		**rise–fall**		**Total**	
	Acc. mean (SD) (%)		**RT mean (SD) (ms)**		**Acc. mean (SD) (%)**		**RT mean (SD) (ms)**	
Unprocessed	92.0 (27.2)	1445 (466)	85.8 (35.0)	1372 (385)	93.2 (25.3)	1447 (422)	90.3 (29.6)	1421 (426)
20 dB/octave	57.9 (49.5)	1595 (399)	63.1 (48.3)	1560 (406)	50.9 (50.1)	1634 (402)	57.5 (49.5)	1595 (403)
40 dB/octave	52.1 (50.1)	1583 (450)	53.2 (50.0)	1530 (373)	51.0 (50.1)	1631 (425)	52.1 (50.0)	1581 (418)
Total	67.5 (46.9)	1540 (445)	67.8 (46.8)	1484 (397)	66.8 (47.1)	1563 (426)	67.3 (46.9)	1528.6 (423.8)

* significant ($p < .05$)

processing condition, as well as the pooled means of intonation contours and processing conditions separately.

A test of the three-way interaction between the phrase, the contour, and the processing condition only revealed a significant effect of the phrase in the unprocessed condition, for response accuracy ($p < 0.001$) where the rise–fall was better than the rise, but not for reaction time. Pairwise comparisons showed that the performance for the phrase *een cadeaubon* was significantly poorer than the performance for all the other phrases. A repeated-measure ANOVA revealed that in the unprocessed condition, the performance on the phrase *een cadeaubon* was significantly lower in the rise (48.9%) than in the fall (81.8%) and the rise–fall (82.2%) ($F(2,134) = 4.60$, $p = 0.001$). When *een cadeaubon* was omitted from the analysis, the interaction between the phrase and the contour was no longer significant. In addition, there was no longer a significant interaction between slope condition and target contour. Apparently, the effect of

the phrase in the unprocessed condition is the same as the effect of the contour. Therefore, the interaction between the contour and the condition can well be interpreted as an interaction between the phrase and the condition. No interaction between the phrase and the contour was found in the 20 dB/octave filter slope.

3.4 Discussion

To our knowledge, this work is the first to test the ability of NH listeners to rely only on F0 for intonation identification of stimuli with varying amounts of spectral smearing. For the identification judgements, participants selected the pragmatic meaning (news, surprise, or predictability) associated to the phrase they heard. The results of the present study indicate that the performance of NH listeners with vocoded stimuli is significantly poorer than the performance with unprocessed stimuli for the three pitch contours (rise, fall, and rise–fall). Listeners' intonation identification was adversely affected under spectral degradation (as seen in the two vocoded conditions) but was high (92–93% correct responses) for the full-spectrum stimuli. These results are consistent with previous findings regarding the adverse effects of spectral degradation on pitch perception (Peng et al., 2009; Souza et al., 2011) and confirm our hypothesis that NH listeners would have more difficulty in discriminating between intonation patterns with the vocoded than with the unprocessed stimuli. The reaction times are long (around 1500 ms) but this latency includes (part of) the duration of the second stimulus phrase.

It is known that NH participants listening to vocoder simulations can take advantage of dynamic and temporal cues for intonation identification. However, since intensity and duration were kept constant in the present study, NH listeners had to exclusively rely on F0 information. When this information becomes obscure, intonation identification turns out to be a(n) (unsurmountable)

challenge for vocoder listeners. In contrast, the hypothesis that a steeper noise filter slope (40 dB/octave) would produce better intonation identification than a shallow slope (20 dB/octave) as a result of less channel overlap was not confirmed. This indicates that there were no cues available in the signal that could be effectively used by the listeners. Presuming that the vocoder processing did not eliminate all differences between the contours, any such differences were not sufficient for contour discrimination. Thus, the current type of processing eliminated any effective cues, be it spectral or temporal.

Performance could have been affected by at least two factors, however. First, because no formal tests were performed to assess the hearing status of the subjects, some subjects' hearing might have compromised their scores. Second, general difficulty to distinguish the contours based on the meaning could have played a role. Although high, the performance in the unprocessed condition reached well below a perfect score, with 86% correct in the case of the surprise contour and 92% and 93% in the predictability and news contours, respectively. In the processed conditions, this may have added to the difficulty of the discrimination in addition to the signal degradation.

The performance of NH listeners in intonation identification with the two filter slopes is not (directly) in agreement with the results of Litvak et al. (2007) who found that subjects could identify synthetic vowels and consonants with slopes as sharp as (20 and 40 dB/octave) or shallower (5 and 10 dB/ octave) than ours. Performance increased when slopes were steeper. These results suggest that identification of purely F0-related intonation was more difficult than segment recognition with the use of noise-band vocoders and with the current settings. This is not entirely surprising, given that listeners can use multiple cues for vowel and consonant identification; not only dynamic or temporal cues, but also additional cues in the frequency domain, e.g. the burst spectra of stop consonants or the noise spectra of fricative consonants (Dorman et al., 1997). This is confirmed in Shannon et al. (1995), who found that considerable reduction of spectral cues still allows for a surprisingly high level of phoneme

recognition. Possible reliance on non-spectral cues might explain the effect of sharpening the noise filters from 20 to 40 dB/octave found by Litvak et al. (2007) compared to the present study. The present results were also not in line with those of Crew et al. (2012), who tested musical pitch contour discrimination using a tone (sine-wave) vocoder simulation. The stimuli were relatively similar to those of the current study, as they were controlled for duration and amplitude and a comparable number of band-pass filters were used (16). They found that increasing the filter slopes from 6 through 12 to 24 dB/octave had a positive effect on the performance. This would suggest that a difference could also be found between slopes of 20 and 40 dB/octave, as in our study, since those slopes should enhance spectral differentiation even more. However, since that was not borne out, the discrepancy between the results of Crew et al. (2012) and the present study could be due either to the different vocoder type employed (sine wave vs. noise vocoder) or to the stimuli used (musical vs. linguistic stimuli) or both.

The present data show that NH listeners were not able to use the additional spectral detail provided in the 40 dB/octave condition effectively in intonation identification. Indeed, unsolicited comments by the majority of the subjects suggested that most of the vocoded stimuli given for pattern identification sounded identical. It is possible that more extreme filter slopes might have revealed a clearer effect. A slope of 20 dB/octave, used here as the relatively narrow filter, is still steep compared to (parts of) the settings adopted in some of the previous studies and also on the steep side compared to settings of actual CIs (Fu and Nogaki, 2005; Litvak et al., 2007; Shannon et al., 1998). It should be noted that a majority of subjects (75%) showed better performance on the 20 dB/octave than on the 40 dB/octave condition. Moreover, the overall performance in the 20 dB/octave condition was significantly above chance, whereas the performance on the 40 dB/ octave condition was not. However, because the difference between the significant result and the non-significant result (i.e., the comparison between slope conditions) was not itself significant, these

results are not interpreted as reflecting a real performance difference. The finding that there was an interaction between the performance on the target contour in the 20 dB/octave and the unprocessed condition, can be accounted for by assuming that contour identification could have been driven by general processing restrictions: the double pitch change that is involved in the rise–fall contour (one change for the rise and another for the fall) would be easier to recognize than a single rise or fall because there are two points for identification instead of one. However, two types of result speak against this hypothesis. First, the mean reaction time, although just a trend, for rise–fall was slower (1447 ms) than for the rise (1372 ms), suggesting that the rise–fall was more difficult to process (see below). And second, this contour effect in the unprocessed condition was much larger for the phrase *een cadeaubon* than for the other phrases. The effect disappeared when *een cadeaubon* was removed from the analysis, so either the double pitch change explanation does not hold or it explains only the effect found for *een cadeaubon*. The 20 dB/octave condition showed the opposite effect. The better performance for the rise than the rise–fall in this condition suggests that the double pitch change explanation cannot account for the data or that a different mechanism is responsible for processing in the 20 dB/octave condition than in the unprocessed condition. A possible explanation is in terms of processing time. Due to the poor spectral definition, the listener needed additional processing time. This time could have been available in the rise but not in the rise–fall, since the rise–fall accent was a relatively short part of the phrase (160–350 ms depending on the phrase), whereas the rise allowed the total time of the phrase (700–850 ms) for analysis. The trend in the reaction times in the 20 dB/octave condition were in line with this account, because their mean was longer for the rise–fall (1634 ms) than for the rise (1560 ms). Since the rise did not score uniformly better than the other contours, the bigger tone range at the end of the rise (as compared to the other contours) was not an important cue to identification. However, since participants could be using different strategies for different processing

conditions, it could be the case that the tone range was crucial in the 20 dB/octave condition but not in the unprocessed condition. Yet, it is still unlikely that it plays a role because the larger tone range is only virtual, in that it is only larger if the declination of the phrase is extrapolated.

The results of this study have implications for the understanding of intonation perception by CI users, in as far as our vocoder processing reflects CI processing. Litvak et al. (2007) compared their results on the identification of vowels and consonants of stimuli vocoded using the algorithm that was also adopted in the present study, with the results of Saoji et al. (2005) on actual CI users. Scores decreased with decreasing steepness of the filter slope. By comparing best-fit lines for vocoder and CI listeners, they concluded that the performances on the filter slopes between 4 and 30 dB/octave best matched those of the patients in Saoji et al. (2005). In the light of these findings, the results from our study could imply that even with filter slopes that are steep (as much as 40 dB/octave) relative to the condition that CI users' performance matches with (i.e., between 4 and 30 dB/octave), they could not achieve intonation identification. If, on the other hand, our shallower slope is closer to CI performance than our steeper slope, as suggested by tendencies to a better performance in the shallow slope condition, it is conceivable that CI users have more opportunities for hearing the F0 than our participants did. The tendency for better performance in the 20 dB/octave than in the 40 dB/octave condition could be indicative of an advantage of having more band-pass overlap (i.e., shallower slopes) as opposed to having steeper filter slopes. A cut-off of 350 Hz for the lowest band-pass filter removes direct cues to F0, leaving only temporal information (below 68 Hz due to the temporal envelope cut-off) as a cue. The settings in our study reflect a class of clinically used CIs. It has been found in previous research using that type of devices that shallow slopes facilitate temporal resolution (Drennan et al., 2010; Won et al., 2012). Although this account is disfavored by results from a pilot study (not reported here) in which the shallower slopes (5 and

10 dB/octave) did not show better performances than the conditions tested here, these more extreme conditions should be tested with additional subjects to rule out or confirm a possible advantage of the shallower slopes.

For the devices that our settings reflect, our study suggests that in order to support pure intonation perception, the filter slopes used in the current study are not recommended all else being equal, for the users cannot benefit from temporal nor from spectral cues. Either temporal cues should be exploited by using shallow slopes (5–10 dB/octave), or spectral cues should be exploited. The latter might be realized by increasing the number of electrodes, which in itself would necessitate the use of steep filter slopes. It is likely that if patients with CIs do achieve intonation identification, they do this to a relatively large extent, compared to NH listeners, based on cues other than direct or indirect cues to intonation, such as covarying phrase-level temporal or dynamic speech cues (i.e., in the linguistic sense of the stress and rhythm of an utterance), or, in the case of daily language comprehension, linguistic or extra-linguistic contextual cues.

Conclusions

The present results confirmed that the limited F0 resolution provided by noise-band vocoder simulations reduces the ability to identify intonation patterns by means of pitch alone. On average, no significant difference was found between a 20 dB/octave filter and a 40 dB/octave filter. These slopes were relatively steep compared to conditions from previous research that were found to suffice for discrimination of segmental speech information. This study therefore suggests that even relatively extreme filter slopes do not provide sufficient spectral resolution for the identification of intonation that is conveyed purely by F0 movements. No direct or indirect cues to F0, such as the temporal envelope of higher harmonics that could be in the signal were effective. There are CI devices that use comparable

settings, apart from probably shallower slopes. This has implications for explanations of the perception of intonation by the relevant users. Our message in essence is that if intonation perception by CI users succeeds, it is plausible that an extremely shallow slope is a benefit for pure intonation perception and/or that the users exploit not just pitch but additional bottom-up cues (such as phrase level temporal or dynamic cues) and/or top-down (contextual) cues. An alternative for increasing the spectral resolution worth exploring is to increase the number of electrodes and use accordingly steep filter slopes. We further suggest that future research address a broader range of slope conditions and compare them to conditions in which alternative or additional cues are present. Finally, it is recommended that tests also be carried out with actual CI users in order to examine which vocoder setting comes closest to their performance.

Acknowledgements

We are thankful to Willemijn Heeren and Jos Pacilly of the Phonetics Laboratory at Leiden University and to Jeroen Briaire, audiologist and clinical physicist at Leiden University Medical Center, for their technical assistance. Finally, Carl Verschuur from the Institute of Sound and Vibration Research at the University of Southampton was of great help concerning the interpretation of the results.

Chapter **4**

The perception of emotion and focus prosody with varying acoustic cues in cochlear implant simulations with varying filter slopes

This chapter is based on:

van de Velde, D. J., Schiller, N. O., van Heuven, V. J., Levelt, C. C., van Ginkel, J., Beers, M., Briaire, J. J., Frijns, J. H. M. (2017). The perception of emotion and focus prosody with different cues and filter slopes with. *Journal of the Acoustical Society of America*, *141*, 3349-3363. Doi: 10.1121/1.4982198

Abstract

This study aimed to find the optimal filter slope for cochlear implant simulations (vocoding) by testing the effect of a wide range of slopes on the discrimination of emotional and linguistic (focus) prosody, with varying availability of F0 and duration cues. Forty normally hearing participants judged if (non-)vocoded sentences were pronounced with happy or sad emotion, or with adjectival or nominal focus. Sentences were recorded as natural stimuli and manipulated to contain only emotion- or focus-relevant segmental duration or F0 information or both, and then noise-vocoded with 5, 20, 80, 120, and 160 dB/octave filter slopes. Performance increased with steeper slopes, but only up to 120 dB/octave, with bigger effects for emotion than for focus perception. For emotion, results with both cues most closely resembled results with F0, while for focus results with both cues most closely resembled those with duration, showing emotion perception relies primarily on F0, and focus perception on duration. This suggests that filter slopes affect focus perception less than emotion perception because for emotion, F0 is both more informative and more affected. The performance increase until extreme filter slope values suggests that much performance improvement in prosody perception is still to be gained for CI users.

4.1 Introduction

Current cochlear implants (CI) allow people suffering from severe to profound sensorineural hearing loss to attain a high level of speech understanding in favorable listening conditions (Wilson and Dorman, 2007). Some aspects of the acoustic signal, however, remain difficult to discern. Whereas the discrimination of rhythm and intensity is close to the performance by normally hearing (NH) people, discrimination of pitch is one of the most difficult tasks for CI users (Shannon, 2002; Limb and Roy, 2014). There are at least three major causes underlying this difficulty. First of all, although the incoming signal is usually analyzed into ten to twenty frequency bands (channels), the number of bands that the user can effectively benefit from is limited; i.e., in speech perception tasks CI users at best perform at a level comparable to that seen in CI simulations with about eight channels (Friesen et al., 2001). Second, pitch perception by means of temporal cues has an upper limit of around 300 Hz (Zeng, 2002). Finally, a less studied cause limiting spectral resolution is the slope of the analysis filters defining the frequency bands. Slopes with a shallow roll-off overlap each other more than those with a steep roll-off, resulting in more spectral smearing. Moreover, even with steep analysis filters, spectral smearing is also induced by overlapping neuron areas stimulated by adjacent electrodes (Tang et al., 2011), a factor represented by means of the synthesis filter in vocoder simulations. It remains unknown, however, what the theoretically optimal filter slope for frequency discrimination is given a certain number of channels. Using vocoder simulations of CIs, this study aims to find such an optimum for a specific aspect of speech in which pitch plays a central role (i.e., prosody).

Previous studies using vocoder simulations have shown that steeper filter slopes yield higher segmental speech perception scores but performance reaches an asymptote at some level of steepness. For example, recognition scores for sentences, consonants and vowels by normally-hearing listeners using four-channel CI simulated (vocoded)

stimuli, for which the slopes of the synthesis filters were varied between 3, 6, 18, and 24 dB/octave reached an asymptote at 18 dB/octave (Shannon et al., 1998). When 12, 36, and 48 dB/octave slopes were included, the asymptote was at 12 dB/octave (Fu and Shannon, 2002). Comparable slopes values where performance reached an asymptote were reported for vowel (12 channels) and consonant (8 to 12 channels) recognition in a study using five numbers of channels (2, 4, 8, 16, 32) and three slope conditions: 24 dB/octave for both the analysis and the synthesis slope, 24 dB/octave for the analysis and 6 dB/octave for the synthesis slope, and 6 dB/octave for both slopes (Baskent, 2006).

Other vocoder studies found that performance increased until higher slope values. Litvak et al. (2007) tested vowel and consonant perception with a 15-channel vocoder varying the synthesis filter slopes between 5, 10, 20, and 40 dB/octave. Scores improved with each increasing slope. Comparing their results with those from Fu and Nogaki (2005) of actual recipients, they concluded that CI users' performance corresponded most closely with the 5 dB/octave slope condition. Bingabr et al. (2008) tested vocoded sentence and monosyllabic word recognition with 4, 8, and 16 channels and synthesis filter slopes of 14, 50, and 110 dB/octave that modeled broad (monopolar) and narrow (bipolar) electrode configuration; they also took into account the difference in dynamic range between CI and NH listeners, defined as 50 dB/15 dB = 3⅓ times larger in NH listeners. The slope of the analysis filter was held constant at 36 dB/octave. In general, performance improved from 14 to 50 dB/octave, but leveled or decreased from 50 to 110 dB/octave. The effect of slope was stronger for higher numbers of channels. These studies show that the filter slope steepness beyond which performance stops improving can vary greatly, possibly depending on the task and vocoder parameters such as the number of channels.

The above studies, however, were concerned with segmental perception. Very few studies have addressed the effect of filter slope on the perception of musical melodies or of suprasegmental

components of speech, the topic of this study (i.e., prosody, relatively long signal types conveyed primarily by tonal, but also by dynamic and temporal shape). Crew et al. (2012) studied the effect of filter slope (24, 12, and 6 dB/octave) on melodic contour identification with a 16-channel sinewave vocoder. Melodic contours were nine combinations of flat, rising and falling intervals, each existing in variants with spacings of 1, 2, and 3 semitones. Participants selected the perceived contour on every trial. Performance deteriorated monotonically with widening filter slopes and with decreasing semitone spacing, showing that as with segmental perception, the steepening of filter slopes has a positive effect on prosody perception.

More extreme slopes were explored by van de Velde et al. (2015). They used a 15-channel vocoder to establish the discriminability of intonation contours in which pitch was varied (through resynthesis) to reflect the pragmatic meanings of surprise, expectedness and news. By asking the participants which meaning they thought was expressed, the researchers ensured that they listened to the stimuli in a functional way. Filter slopes were 20 and 40 dB/octave. Chance level performance was observed for both of these conditions, suggesting that for intonation discrimination even steeper slopes than 40 dB/octave are required, as these more extreme slopes are more likely to allow F0 discrimination than shallower slopes.

The literature reviewed above suggests that, similar to segmental perception, prosodic pitch (i.e., intonation) perception benefits from better frequency selectivity in the form of steeper filter slopes. However, whereas for segmental identification scores reached asymptote at 40 dB/octave (Litvak et al., 2007), performance for intonation perception was still at chance for 40 dB/octave (van de Velde et al., 2015), despite using the same number of channels (though some other vocoding parameters differed between the studies). Given the results of those studies, we hypothesize that, given comparable tasks, intonation perception requires greater channel independence, perhaps as realized by means of electrode configuration or steeper filter slopes, than segmental perception, because intonation

perception relies more heavily on spectral versus temporal information relative to segmental perception. An exploration of more extreme filter slopes seems therefore warranted, and was the aim of this study.

This exploration was done using noise vocoder simulations since, in contrast with actual CI perception, this allowed (1) manipulation of signal processing parameters, (2) inclusion of a uniform NH listener cohort, and (3) a comparison with previous studies using vocoders. Although these simulations have been shown to closely model actual CI perception (Dorman and Loizou, 1997; Dorman et al., 1997), a number of discrepancies between real and simulated CI hearing must be pointed out. First of all, as mentioned above, the effective number of channels is lower in real CIs than in simulations. Second, whereas filter slope, representing the amount of channel interaction, in principle can be indefinitely increased in simulations, it is likely limited to around 5 dB/octave for CI users. Third, CI recipients may have severe irregularities in patterns of neuronal survival affecting the regions activated by electrodes. Fourth, the (speech) amplitude range of CI hearing is only about a third as large as that of NH individuals (Bingabr et al., 2008), causing filter slope decay to reach the bottom of the dynamic range sooner in CI users. Fifth, steeper slopes may cause the electrical signal to reach fewer neurons, thus limiting the sound's amplitude in CI users. Finally, CI users' perception for all signal types is based on temporal information, whereas NH listeners also exploit F0, spectral, and intensity cues.

These discrepancies limit but do not preclude the representativeness of simulations for actual CI perception. As for the first two discrepancies (channel number and filter slope), despite results from the literature indicating an interaction between filter slope and channel number, we chose to keep the channel number constant, as that factor was not the focus of the study and would have made the task too long and burdensome for the participants. We used 15 channels for two reasons. First, extreme filter slopes are likely to be most (or even only) effective for higher numbers of channels (up to

certain limits), because channels are more difficult to segregate in a denser configuration (Stafford et al., 2014). Second, the studies by Litvak et al. (2007) and van de Velde et al. (2015) also used 15 channels, allowing a relatively straightforward comparison between their results and ours.

The selection of the exact range of filter slopes to be tested was based on pilot data, starting from findings in the literature that for higher channel numbers only the more extreme filter slopes are likely to show an effect since they are spaced closely together (Bingabr et al., 2008; Stafford et al., 2014). The pilot study explored several filter slopes to identify the range between chance and ceiling performance on a simple two-alternative forced choice (2AFC) prosody discrimination task, similar to the main experiment of this study. Using stimuli with the template '[ARTICLE] [ADJECTIVE] [NOUN],' participants judged if an emotionally intended phrase was pronounced as sad or happy (in one subtest), or if it carried sentential accent on the adjective or on the noun (in another subtest). The pilot results suggested that performance might show an asymptote only with values as extreme as 160 dB/octave and that chance-level performance might occur at 5 dB/octave. For these reasons, the slopes tested here ranged from 5 to 160 dB/octave. We hypothesized that performance on intonation discrimination would increase with increasing filter slope steepness. The third, fourth, and final discrepancies between CI and vocoder perception warrant additional caution in generalizing the results of this study to CI users, as these differences might prove any effect of filter slope found to be less pronounced in the clinical population.

To test if filter slope had the hypothesized effect particularly on F0-based prosody, the stimuli were divided over three conditions varying the availability of two possible types of cues, viz. rhythmic and pitch cues. We hypothesized that the cost of vocoding would be larger for pitch than for rhythmic cues, because filter slope affects the availability of pitch cues more than that of temporal cues. To investigate if different kinds of prosody would be influenced in a

different way or to a different degree by filter slope, we tested two types of prosody, namely linguistic and emotional prosody. This is a fundamental distinction in prosody types, as linguistic prosody conveys information about syntax or semantics while emotional prosody conveys information about the state of the speaker. The two prosody types have been found to be associated to different relative degrees with the two cerebral hemispheres (Witteman et al., 2011). Based on findings on the relative importance of F0 and duration parameters in vocal emotion expression (Williams and Stevens, 1972; Murray and Arnott, 1993) and sentential focus (Sityaev and House, 2003), we conjectured that linguistic prosody (in this case, sentential focus) would rely relatively heavily on temporal information but relatively little on F0 information as compared to emotional prosody. This would suggest that CI users would have more difficulty with emotion than with focus perception; if focus perception is indeed relatively unaffected by filter slope (because temporal information is relatively important), then that would facilitate focus perception for them.

To summarize the rationale of the study, using vocoder simulations of cochlear implants, we explored the influence of (synthesis) filter slope on the perception of prosody. The goal was to find the as yet understudied range of filter slopes between chance and ceiling performance and more particularly the optimal filter slope value within that range. The results are intended to represent the effect of spectral degradation on prosody perception for a specific group of CI users (those with 15 channel devices). We hypothesize that the strongest effect of filter slope would occur for a high number of channels and correspondingly (extremely) sharp filters. The results of this study could be meaningful to the future design of CIs, because a design goal for future implants is to reach higher numbers of channels.

4.2 Methods

In this study, we investigated the effect of filter slope (hereafter referred to as 'Filter slope' as a statistical condition) on the accuracy of focus and emotion discrimination (reflecting the two major types of prosody, i.e., linguistic and emotional prosody) in vocoder simulations of cochlear implants, when one or both of two cue types, namely F0 and temporal cues ('Cue' condition), were present in the signal. This was tested by means of a simple 2AFC task in which participants, in each trial, heard either an emotional or a focused variant of a phrase of the form 'ARTICLE] [ADJECTIVE] [NOUN]' and either judged the speaker's emotion (happy or sad) or identified the word that was focused (the adjective or the noun). The filter slopes were 5, 20, 80, 120, and 160 dB/octave, as well as a control condition without vocoder processing (but varying in availability of F0 and/or temporal cues). We hypothesized that filter slope would have a stronger effect when only F0 was present as a cue than when only duration was present, therefore influencing emotional prosody more strongly than linguistic prosody, because the former by hypothesis relies more on F0 cues than on duration cues relative to the latter. The inclusion of the condition with both cues simultaneously present allowed us to explore these relative forms of reliance. The availability of cues in the stimuli was realized by resynthetically replacing the F0 contour or the segmental durations of emotional or focus utterances, respectively, onto separately recorded emotion- or focus-neutral tokens of the same phrase. In this way, we assured that the emotion and focus positions could only be recognized based on the cues under investigation (F0 and duration) because all other components in the signal were identical between the two response options (i.e., they were both based on the exact same neutral token).

4.2.1 Participants

Forty university students (29 women, 11 men) volunteered as participants and received credits if desired. Their mean age was 23.1

years, ranging between 18 and 35 years and with a standard deviation of 4.1 years. People with hearing problems, an age exceeding 60 years or without Dutch as their mother tongue were not recruited. Hearing was assessed by means of tone audiometry at the octave frequencies between 0.125 and 8 kHz (Audio Console 3.3.2, Inmedico A/S, Lystrup, Denmark). Candidates with a hearing loss of more than 20 dBHL above the lowest loudness tested (20 dBHL, the software's standard test), i.e., with a minimal loss of 40 dBHL, at any of the frequencies were excluded. This was the case for two people. All participants gave their written informed consent and filled in a short questionnaire about their education level and experience with sound manipulation and music (Appendix A). Most of them listened to music and engaged in music playing or singing for several hours a week, but most of them did not work with digital sound processing. This survey indicates that, on average, the cohort is used to active listening to audio material. The study was approved by the ethical committee of the Faculty of Humanities of Leiden University.

4.2.2 Stimuli

There were two different tests, an emotion recognition test and a focus recognition test, for which different phrases were recorded as natural stimuli in a sound-treated booth by a professional linguist (CL), at a sampling frequency of 44,100 kHz and a sampling depth of 32 bit. For the emotion test, the speaker was asked to pronounce twelve phrases following the template article-color-noun (e.g., *een rode stoel*, 'a red chair') in three variants: (1) without a specific emotion (neutral), (2) with a happy-sounding emotion and (3) with a sad-sounding emotion. The way the phrases were pronounced to convey the emotions was left to the speaker. However, she was asked to clearly distinguish them, keeping in mind that the same stimuli would also be used for a listening test with children in another study). Consequently, the prosody could have been realized as typically child-directed. The phrases were 1.5 to 2 seconds long.

The phrases for the focus test were twelve utterances of the template article-color-noun-*en een* (e.g., *een gele bloem en een*, 'a yellow flower and a'), highly comparable but not identical to those of the emotion test. The two trailing words were added to prevent phrase-final prosody on the noun. Three variants were produced for each phrase: (1) with neutral focus, i.e., the adjective and the noun carried focus as equally as possible; neutral), (2) with narrow focus on the color and (3) with narrow focus on the noun. For the neutral focus, the speaker was asked to speak relatively monotonously and to avoid sentential accents on any of the words. Since a phrase without focus is unlikely in practice, at least from the perceptual perspective, we aimed at equal prominence on the two words without requiring or claiming that the two words were either both focused or both unfocused, therefore calling the result 'neutral focus'. For both the emotion and the focus stimuli, the speaker was asked to keep the general speaking rate more or less constant across the variants, in order to avoid any large phrase-level temporal differences between variants that might result in ceiling-level performance in discrimination. This control of speaking rate was not believed to neutralize all duration information because it is not possible for a speaker to manipulate all phonemic and sub-phonemic temporal details in a phrase. Like the emotion phrases, the phrases were 1.5 to 2 seconds long.

As a next step, stimuli for both tests were all resynthesized into three variants with respect to the availability of the phonetic cues 'F0', 'Duration' and 'Both', using the *Praat* software, *Version 5* (Boersma & Weenink, 2014). The motivation for this step was to control for the availability of cues in the stimuli to be judged. It was done by importing the respective cues from the emotional or focused utterance onto the neutral variant of the same phrase per segment (i.e., maintaining the alignments with the vowels and consonants). This involved (1) the phrase's pitch contour (for the F0 condition), (2) the segment durations (Duration condition), (3) both the pitch contour and the durations (Both condition). We presumed that the two emotions, on the one hand, and the two focus positions, on the other hand, would

be acoustically systematically different such that there might exist a basis for participants' discrimination. As evidence of acoustic difference, however, gross acoustic measures were performed on the stimuli after resynthesis, using *Praat*.

Table 1 presents an overview of the mean F0 and the standard deviation (SD; reflecting phrase-level variability) and range of F0 as well as mean duration and intensity of phrases. Values were averaged over the twelve stimuli per emotion/focus condition and per cue condition. For the emotion stimuli, the F0 mean, SD and range were larger for happy than for sad variants in the conditions where pitch cues were present, whereas in the Duration condition those values were almost equal between the two emotions. In the conditions where duration cues were present (the Duration and Both conditions), however, sad stimuli were 9.2% longer than happy stimuli, whereas they were equal in the F0 condition.

As for the focus stimuli, in the F0 and Both conditions, F0 mean and range were lower for stimuli with nominal focus than for those with adjectival focus, but had a higher F0 SD. Durations were equal between focus positions in the F0 and Both conditions. In the Duration condition, all measures, including duration, were highly comparable between focus positions. As phrase-level durations in the Duration condition were found to be similar between focus positions, we investigated if the durations of the focused words were different. Table 1, part C shows that in the F0 condition, the difference in duration between the adjective and the noun is similar for the two focus conditions (reflecting the elimination of duration cues), but that the focused word was always longer than the non-focused word in the Duration and Both conditions. This shows that duration cue information other than phrase-level duration was present in the stimuli. For both the emotion and the focus stimuli, intensity values were similar in all conditions.

These results show that there were systematic acoustic differences between conditions and that the cues present in the signal

Table 1. Acoustic measurements of stimuli used in the emotion test (A) and in the focus test (B and C). Numbers represent the averages over the 12 stimuli (sentences) per cue condition and per emotion/focus condition. Mean F0, F0 SD, and F0 range refer to the mean, the standard deviation and the range of all pitch points in a stimulus, respectively. In panels A and B, the duration and intensity values refer to the respective measurements of the stimulus phrase as a whole. In panel C, duration values concern the adjective and the noun (i.e., as part of the complete phrases) of the stimuli of the focus test.

A. Emotion test

Cue	Emotion	Mean F0 (Hz)	F0 SD (Hz)	F0 range (Hz)	Duration (s)	Intensity (dB)
F0	**Happy**	324.1	113.9	377.0	1.67	71.71
	Sad	267.6	41.2	151.3	1.67	72.51
Duration	**Happy**	228.1	57.6	204.5	1.84	72.91
	Sad	232.6	57.8	205.4	2.01	73.08
Both	**Happy**	327.1	115.6	382.2	1.84	71.65
	Sad	269.4	41.4	173.7	2.01	72.63

B. Focus test

Cue	Focus position	Mean F0 (Hz)	F0 SD (Hz)	F0 range (Hz)	Duration (s)	Intensity (dB)
F0	**Adjective**	326.7	92.3	346.0	1.76	72.95
	Noun	265.1	105.7	320.9	1.76	72.41
Duration	**Adjective**	240.9	90.8	439.0	1.58	73.17
	Noun	238.5	91.7	422.5	1.56	73.13
Both	**Adjective**	321.6	93.6	367.4	1.55	72.81
	Noun	272.9	104.3	319.3	1.56	72.32

C. Durations of adjectives and nouns in the focus test

	Focus position	Duration of the adjective (s)	Duration of the noun (s)
F0	**Adjective**	0.52	0.47
	Noun	0.51	0.47
Duration	**Adjective**	0.52	0.44
	Noun	0.41	0.55
Both	**Adjective**	0.52	0.43
	Noun	0.39	0.56

corresponded to the conditions (i.e., the F0 condition had F0 cues and no duration cues, and vice versa), except for the total sentence duration measure in the focus test, which was similar for the two focus positions. Any duration or other temporal cue that participant might rely on to distinguish between focus positions must therefore be internal to the phrase, i.e., the relative durations of segments or syllables. The acoustic measurements further show that the speaker recording the stimuli was partly successful in controlling the general speaking rate because the overall durations of the two emotional variants of the stimuli in the emotion test differed by only 9.2%. She was more successful maintaining her speaking rate with the focus stimuli, where the difference was 1.3%. For the latter stimuli, however, focused words were longer than non-focused words, such that the speaking rate on the sub-phrasal level was not constant across stimuli. It is therefore plausible that overall phrasal durations provided a duration cue that listeners could rely on in the emotion test while relative word durations provided duration cue in the focus test.

The final stimulus processing step involved simulating cochlear implant hearing by means of vocoding. The 15-channel noise vocoder described in Litvak et al. (2007) was implemented in *Matlab R2015a* (The MathWorks, Inc., Natick, MA, US). The basic steps of this algorithm are as follows. First, it samples the signal at 17,400 Hz and divides it into 256 bins using a short-term Fourier transform. It then analyses the signal into fifteen non-overlapping, rectangularly-shaped, logarithmically spaced frequency bands, uses their amplitude envelopes to modulate similarly spaced noise bands, and finally sums the fifteen channels. There is an implicit low-pass envelope detector with a cut-off frequency of 68 Hz. Note that this cut-off frequency was too low to allow temporal perception of most of the F0 cues in the stimuli in the present study, since their mean F0 values were much higher than 68 Hz. This implies that if listeners were able to process F0 cues, it would be based on information other than temporal.

The slopes of the synthesis filters in the simulation can be varied to mimic greater or lesser spectral smearing. All stimuli in both

experiments were processed with each of the following five filter slopes: 5, 20, 80, 120 and 160 dB/octave. The selection of these slopes is based on a pilot study exploring the range from (near-)chance to (near-)ceiling level performance. The first three slope values differed by a factor of 4 but the final three were more closely spaced in to facilitate identification of a possible asymptote in that region. All stimuli were finally scaled to the same peak amplitude in order to neutralize any level differences between the various stimulus and filter slope conditions. The relatively high scores that were reached in the most favorable condition in the pilot tests ensured that the emotions and focus positions, were conveyed successfully enough to use these stimuli for the experiment.

In each experiment, participants heard both processed and unprocessed stimuli. The processed stimuli consisted of three of the five filter slope conditions, instead of all five, in each of the three phonetic cue conditions (per test: 12 phrases × 2 emotions/focus positions × 3 phonetic cues × 3 filter slopes = 216 items). The reason for selecting only three out of five filter slope conditions per participant was to limit the task burden. A Latin square design in which all participants received all conditions, but each a different (but balanced) subset of items was not considered a good alternative to relieve the task burden because in that case very few items would remain per participant. Instead the ten possible combinations of three out of five conditions were balanced across participants by creating ten subgroups of four participants. Missing data were therefore 'missing by design' (Schafer, 1997). The unprocessed stimuli included the neutral unprocessed phrases (12 items) and the non-vocoded stimuli in each of the three phonetic cue conditions (12 phrases × 2 emotions/focus positions × 3 phonetic cues = 72 items). Each of these triplets of non-neutral phonetic cue blocks was preceded by one warm-up trial.

4.2.3 Procedure

The emotion and focus tests were performed together in a single session in the same setting, on a computer with headphones in a sound-treated booth. The order of the two tests was counterbalanced across participants. The presentation level of the stimuli was determined by adjusting a dummy stimulus until the participant found the level comfortable. In practice, this was around 65 dB SPL. This level was maintained for all conditions of both tests in the session. Tests were preceded by a practice phase to familiarize participants with the procedure and with the type of stimuli. In both tests, practice stimuli consisted of eight vocoded and eight non-vocoded stimuli with varying filter slopes, forming a representative subset of the experimental stimuli. This was the only vocoded speech the participants were presented with before actual testing. The practice phase was followed, in this order, by a phase consisting of the block of neutral stimuli, a phase of three blocks of unprocessed stimuli (one block per phonetic cue) and finally a phase of nine blocks of processed stimuli (also blocked per cue). Per phase, the order of blocks as well as the order of stimuli within each block was randomized. However, in the processed phase, the three blocks of phonetic cues per filter slope condition, although randomized, were completed before continuing to the next filter slope. In all trials, participants were presented with one auditory stimulus and were asked to indicate by button-press which of two emotions (happy or sad) or focus positions (focus on the color or on the noun) they perceived, respectively (a 2AFC task). Participants had 5,000 ms to respond, starting from the onset of the sound file, but a trial jumped to the next when a response was given within that window. In the emotion test, a picture of the object mentioned in the phrase (e.g., a blue ball) was shown as well as a happy and a sad face with positions corresponding to the option buttons (left and right). The position of the faces was swapped halfway through the experiment. In the focus test, a picture of the object and printed words of the two critical elements of the phrase were shown (e.g., blue and ball in Dutch). The position of these

words was not swapped during the experiment because it would create a conflict if the first sounding element (the color) were shown to the right of the second sounding element (the noun). Response accuracy was registered for analysis, where a response counted as correct if the emotion or focus position intended by the speaker was identified as such and as incorrect if the unintended option was selected. For the unprocessed stimuli, for each trial, participants were also asked to indicate the certainty of their response on a five-point scale (1 for very uncertain, 5 for very certain). The goal of this was to find if there were response biases inherent to the basic stimuli, i.e., high certainty rates coupled with correct answers would be a sign of a lack of a response bias. An experimental session lasted around one hour.

4.2.4 Statistics

All statistical analyses involved *d'* or certainty as the dependent variable whereby *d'* is a transformation of accuracy scores per participant per cell of the design. This was done to account for possible response biases, which may be particularly influential in two-alternative response tasks. In this procedure, following signal detection theory, for any trial, the correct option is viewed as signal and the incorrect option as noise. Correctly choosing the signal counts as a hit (and the probability of doing so as the hit rate), and choosing the signal when it was noise counts as a false alarm (and the probability of doing so as the false-alarm rate). From this, *d'* is calculated by subtracting the *z* score of the false alarm rate from the *z* score of the hit rate (Stanislaw and Todorov, 1999), whereby a *d'* score of 0 corresponds to complete insensitivity (chance level performance) and a score of 2.5 corresponds to a percentage correct of around 90% (Macmillan and Creelman, 2004). Following a conventional solution (Macmillan and Kaplan, 1985), perfect scores in a cell, which are computationally unresolvable, were replaced by 100%/2*N*, where *N* is the number of items in the cell (24). Results are presented as *d'* scores.

A distinction was made in the analysis of the effect of Cue in the non-vocoded condition versus the effect of Cue and Filter slope in all accuracy data together (vocoded condition with the non-vocoded condition as a baseline). Recall that certainty data were collected only in the non-vocoded condition. The variances of *d'* and certainty scores over cue condition were tested for homogeneity using Mauchly's test and if necessary corrected for degrees of freedom using the Greenhouse-Geisser correction. Subsequently, the effect of Cue in the non-vocoded condition was tested with a Repeated Measures Analysis of Variance (RM ANOVA) because results were compared across levels of the condition Cue, which were completed by all participants. In order to account for the missing data in the design, Multilevel Modeling (Goldstein, 1987) was used, with filter slope and phonetic cue as independent variables and *d'* as the dependent variable (Stanislaw and Todorov, 1999; Macmillan and Creelman, 2004). In order to avoid computational problems of a multilevel model with an incomplete dataset (e.g., non-positive definite Hessian matrices), the multilevel models were restricted to the assumptions equal to RM ANOVA (compound symmetry). There were random intercepts for Filter slope and Cue but not for the interaction. These assumptions were not all met for all cells of the data structure. A more stringent interpretable model, however, was not believed to be available, and so no transformations or corrections were applied. Therefore, the results of the vocoded condition have to be approached with caution. All post-hoc tests were Bonferroni-corrected.

4.3 Results

We present the results of neutral stimuli, non-vocoded non-neutral stimuli, and vocoded stimuli (including non-vocoded non-neutral stimuli as a control condition) in turn. Only the non-null responses were taken into account in all of the analyses, i.e., the trials for which a response was detected with the available time window.

4.3.1 Neutral stimuli

The participants' task for the neutral stimuli was identical to that for all other stimuli, namely, to choose the emotion or focus position of the presented stimuli. Note that the stimuli, as per their neutral status, were not recorded with a specific emotion or focus position and that there were therefore only incorrect response options available for the participants. The neutral stimuli were analyzed to find out if there was a bias in the perception of emotion or focus position, respectively, and the analysis therefore consists only of percentages per response option and the certainty results. This bias analysis was performed to complement the *d'* analysis of all other stimuli because a bias in the neutral stimuli would reflect a bias inherent to the segmental basis of the stimuli, whereas a bias in the other stimuli would be a bias involving the prosody (since non-neutral stimuli were composed of the segmental layer of the neutral stimuli and the prosody of the non-neutral stimuli). Non-null responses covered 96.0% of the data in the emotion test and 94.7% in the focus test and only those were further analyzed. In the emotion test, sad responses represented 64.4% of cases and happy responses 35.6%. In the focus test, 81.3% of responses were with focus on the noun and 18.7% with focus on the adjective (color). In both tests, the mean certainty was 3.2 points with an SD of 1.3 on a scale of 1 (very uncertain) to 5 (very certain), indicating that people were not very certain of their responses, but that there was a bias towards perceiving the non-manipulated prosody as sad over happy and a strong bias of perceiving them as focused on the noun as opposed to the adjective. Alternatively, the sad and noun-focused responses could be seen as functioning more as defaults than the happy and adjective-focused responses, respectively. These results will be further discussed in the section Non-vocoded stimuli.

4.3.2 Non-vocoded stimuli

The non-neutral non-vocoded stimuli served as a control condition for the vocoded stimuli, differing from them only in the absence of vocoding. These non-vocoded stimuli involved those that were

pronounced with a specific emotion or focus and of which four variants were presented to the participants: unprocessed and with F0, duration, or both cues available. The goal of this part of the analysis was to find out if the emotions and focus positions intended by the speaker were successfully conveyed, i.e., if the participants were able to recognize them as such with a high level of accuracy. If so, this would indicate that the emotions and focus positions were in principle well conveyed and that a possible lack of an effect in the vocoder simulation condition would not be due to unsuccessful production of the raw stimuli. This analysis further allowed us to investigate which cues participants relied on without the intervention of vocoding.

Of all responses, 1.2% were null-responses (i.e., no response detected in the allotted time window) and not analyzed. In the emotion test, the percentages of null responses were 0.1% in the unprocessed condition (all cues present), 0.7% in the F0 condition, 2.5% in the Duration condition, and 0.6% in the Both condition. In the focus test, these percentages were 0.1%, 2.3%, 2.8%, and 0.3%, respectively. Results of *d'* scores and response certainty per phonetic cue and per test are shown in Table 2 and in Figure 1. They show that *d'* scores vary between 0.3 (corresponding to just above chance level performance) and 3.9 (a very high sensitivity corresponding to near-ceiling level performance) and that certainty scores are on a par with them. These patterns suggest differences in difficulty between Cue conditions in both tests. In order to test if there was an effect of phonetic cue (Cue) on *d'* scores as well as on certainty of the response, means were subjected to a RM ANOVA per Test (emotion or focus test). In both the emotion and the focus test, Mauchly's test indicated that the assumption of sphericity was violated both for *d'* (emotion test: χ^2 (5) = 195.93, $p < .001$; focus test: χ^2 (5) = 38.27, $p < .001$) and for Certainty (emotion test: χ^2 (5) = 51.13, $p < .001$; focus test: χ^2 (5) = 35.32, $p < .001$), leading us to use the Greenhouse-Geisser correction for degrees of freedom. Post-hoc tests for levels within the Cue condition were Bonferroni-corrected.

Table 2. Certainty and *d'* scores per test (emotion test and focus test) and per cue condition for non-vocoded stimuli. In the F0 condition, F0 information was available for the listeners, in the Duration condition segmental durations and in the Both condition both cues were available simultaneously. In the Unprocessed condition, the stimuli were natural.

Test	Cue	Certainty (SD)	*d'* (SD)
Emotion	Unprocessed	4.7 (0.7)	3.98 (0.25)
	F0	4.4 (0.9)	3.76 (0.47)
	Duration	3.3 (0.9)	0.25 (0.44)
	Both	4.6 (0.8)	3.94 (0.29)
	Total	**4.2 (1.0)**	**2.98 (1.63)**
Focus	Unprocessed	4.6 (0.8)	3.78 (0.42)
	F0	3.8 (1.0)	2.72 (1.25)
	Duration	3.0 (0.9)	1.12 (0.74)
	Both	4.2 (1.0)	3.25 (0.93)
	Total	**3.9 (1.1)**	**2.72 (1.33)**
Total	Unprocessed	4.7 (0.8)	3.88 (0.36)
	F0	4.1 (1.0)	3.24 (1.07)
	Duration	3.1 (0.9)	0.68 (0.75)
	Both	4.4 (0.9)	3.6 (0.77)
	Total	**4.1 (1.1)**	**2.85 (1.49)**

In the emotion test, the effect of Cue was significant both for *d'* ($F(1.06,41.41) = 225.41$, $p < .001$) and for Certainty ($F(1.70,75.25) = 89.48$, $p < .001$). Bonferroni post-hoc tests revealed that for *d'*, all pairwise comparisons with Duration were highly significant ($p < .001$) while all other comparisons were not significant (p at least .68). For Certainty, all pairwise comparisons with Duration as well as Unprocessed vs. F0 were highly significant ($p < .001$), F0 vs. Both was significant ($p = .002$) and Unprocessed vs. Both was just

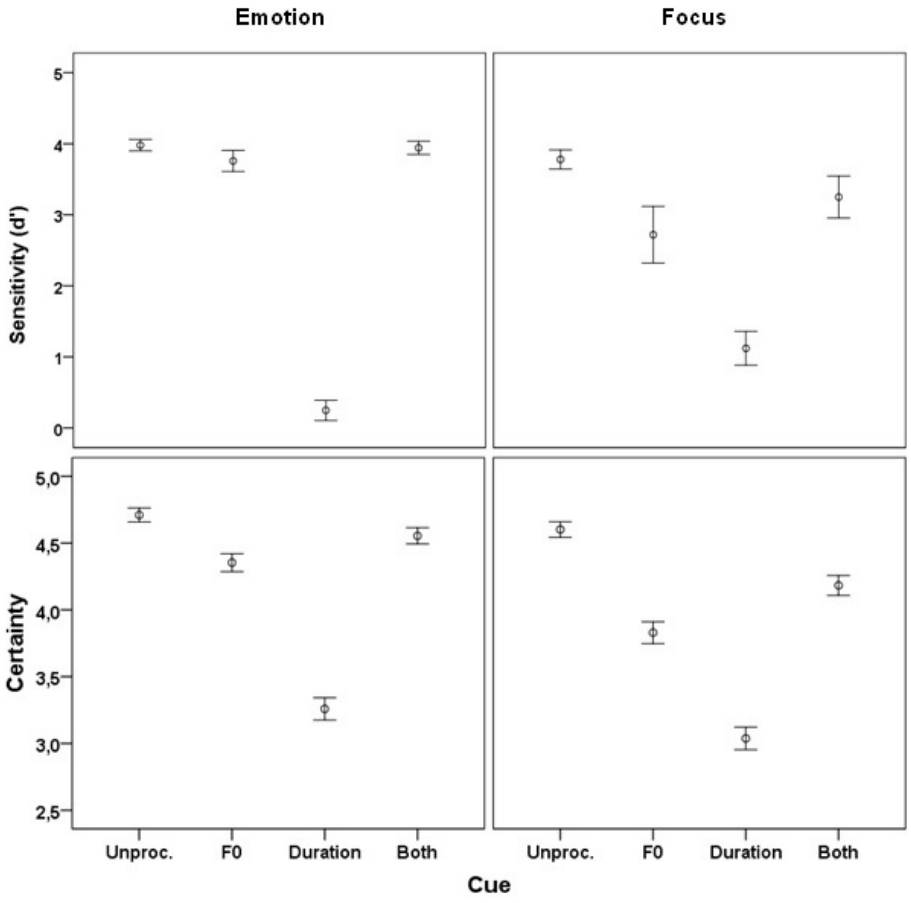

Figure 1. *d'* scores (top panels) and Certainty (bottom panels) scores per Cue (abscissa) and per Test (columns) for the non-vocoded stimuli. Error bars represent 95% confidence intervals. Unproc (Unprocessed) refers to non-resynthesized stimuli. In the F0 condition, F0 information was available for the listeners, in the Duration condition segmental durations and in the Both condition both cues were available simultaneously.

significant (p = .049). In the focus test, the effect of Cue was significant for *d'* [$F(1.77, 69.14) = 72.36$, $p < .001$] as it was for Certainty [$F(1.99, 77.40) = 50.48$, $p < .001$]. Bonferroni-corrected post-hoc tests revealed that for *d'*, all comparisons were highly significant ($p < .001$) except Unprocessed vs. Both, which was

significant (p = .022) and F0 vs. Both, which was not significant (p = .12). For Certainty, all pairwise comparisons were highly significant (p < .001).

Together, these results show that both the Emotions and the Focus positions intended by the speaker were well conveyed, since near-ceiling level accuracy was achieved in some conditions. For Emotion, participants relied mostly and heavily on F0 as opposed to Duration, given that scores for the F0 and Both condition were near-ceiling level while scores for the Duration condition were near-chance level. For Focus, there was less information in the F0 than for Emotion given the lower score on F0 and Both than in the Emotion test; it was, however still the cue that listeners relied on most given that F0 performance was closer to Both performance than Duration performance was). For Focus, Duration information was more useful than for Emotion, but still did not provide much information. These scores parallel the percentages of null responses in the different conditions.

4.3.3 Vocoded stimuli

The analysis of the vocoded condition involved the investigation of the main effects, interactions and post-hoc effects of the Cue and Filter slope conditions on *d'* scores (there were no certainty data). Data were analyzed per test (emotion or focus test) with Multilevel modeling because they suffered from missing data, as explained in the section Statistics. Non-vocoded data were re-included in the analysis as a baseline for comparison with the filter slope conditions. In other words, whereas in the previous analysis they were analyzed within the non-vocoded condition across cues, they were now analyzed as one of the filter slope conditions. Descriptive statistics in the form of mean *d'* scores of cells and overall means are presented in Table 3.

In the emotion test, the effects of Filter slope (F(5,241.66) = 187.60, p < .001) and Cue (F(2,149.32) = 268.55, p < .001) on accuracy, as well as their interaction (F(10,266.36) = 73.07, p < .001) were highly significant. All three post-hoc comparisons between the

Table 3. Means and standard deviations of Accuracy scores, and, where applicable, split by Test, Cue, and Filter slope, for vocoded stimuli. In the F0 condition, F0 information was available for the listeners, in the Duration condition segmental durations and in the Both condition both cues were available simultaneously.

Test	**Cue**	Sensitivity (*d'*)					
		5 dB/ octave	20 dB/ octave	80 dB/ octave	120 dB/ octave	160 dB/ octave	**Total**
Emotion	F0	0.05 (0.44)	0.09 (0.71)	1.29 (0.91)	3.32 (0.75)	1.54 (0.93)	**1.88 (1.65)**
	Duration	0.24 (0.69)	0.69 (0.46)	0.4 (0.55)	0.23 (0.51)	0.42 (0.82)	**0.36 (0.59)**
	Both	0.37 (0.67)	0.64 (0.52)	2.01 (0.96)	3.48 (0.59)	1.89 (0.98)	**2.24 (1.53)**
	Total	**0.22 (0.62)**	**0.47 (0.63)**	**1.23 (1.05)**	**2.34 (1.63)**	**1.28 (1.1)**	**1.69 (1.65)**
Focus	F0	0.27 (0.45)	0.08 (0.53)	0.11 (0.51)	1.19 (1.08)	0.34 (0.46)	**0.98 (1.35)**
	Duration	0.72 (0.68)	1.96 (0.98)	1.73 (1.13)	1.77 (1.08)	1.72 (0.81)	**1.47 (0.99)**
	Both	1.14 (1.03)	2.16 (1.25)	2.15 (1.18)	2.66 (1.17)	2.15 (0.94)	**2.35 (1.26)**
	Total	**0.71 (0.83)**	**1.4 1.34)**	**1.33 (1.32)**	**1.87 (1.25)**	**1.4 (1.08)**	**1.77 (1.41)**
Total	F0	0.16 (0.46)	0.08 (0.62)	0.7 (0.94)	2.25 (1.42)	0.94 (0.95)	**1.43 (1.58)**
	Duration	0.48 (0.72)	1.32 (1)	1.06 (1.11)	1 (1.14)	1.07 (1.04)	**0.91 (0.99)**
	Both	0.75 (0.94)	1.4 (1.22)	2.08 (1.07)	3.07 (1.01)	2.02 (0.96)	**2.3 (1.4)**
	Total	**0.46 (0.77)**	**0.93 (1.14)**	**1.28 (1.19)**	**2.11 (1.47)**	**1.34 (1.09)**	**1.73 (1.54)**

levels of Cue were highly significant at a Bonferroni-corrected significance of $p = .015$ (all three $p < .001$). The post-hoc comparisons between the six Filter slope conditions (that is, the actual five slopes of the vocoded condition plus the non-vocoded condition) were all highly significant at the corrected threshold of $p = .003$ ($p \leq .0001$), except for the ones between 5 dB/octave and 20 dB/octave ($p = .068$), and between 80 dB/octave and 160 dB/octave ($p = .44$). Figure 2

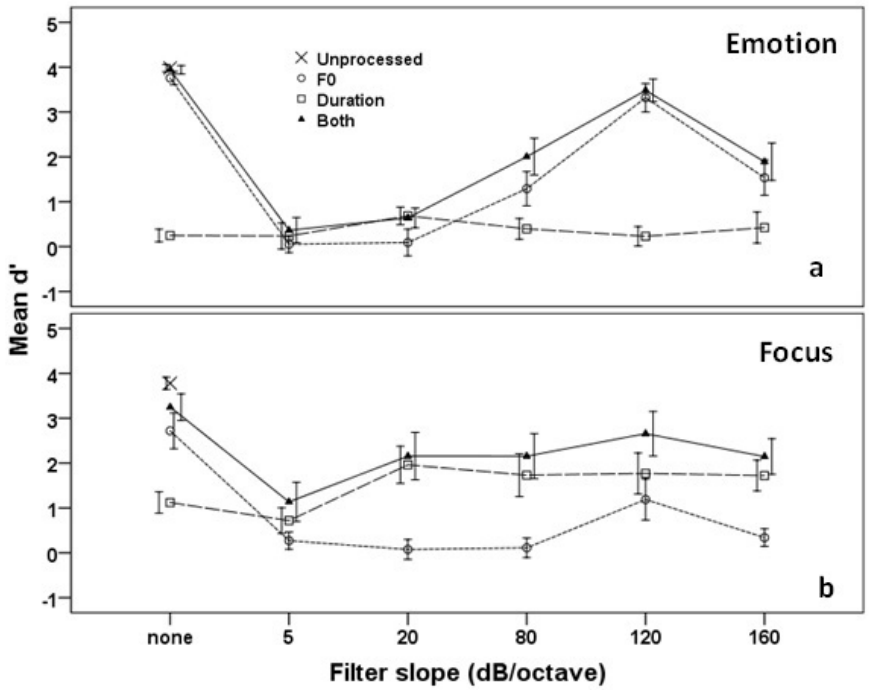

Figure 2. *d'* scores per Filter slope (abscissa) and Cue (line types), for each Emotion discrimination (a) and Focus discrimination (b) tests in the vocoded conditions. Included are the results for the unprocessed condition (crosses) which is only relevant for the 'none' filter slope (non-vocoded condition), in the top left of each panel. Error bars represent 95% confidence intervals.

(panel a) shows that this effect of Filter slope differs per Cue condition. Whereas for the conditions including F0 (i.e., the F0 and Both conditions) *d'* scores increase from 5 dB/octave to 120 dB/octave, approximating ceiling level performance, and drop again above 120 dB/octave, for the Duration condition there is overall much less differentiation and scores are only slightly above chance level. This pattern of results shows emotion perception is based on the F0 and not the Duration cue (given the comparable patterns for the F0 and Both condition) and that filter slope has a large effect always and

only when the F0 cue is present (as performance on the Duration condition was near chance level for all slope conditions). This cue weighting corresponds to that observed in the non-vocoded condition, suggesting that listeners did not adapt their listening strategy to the unnaturalness of the vocoded stimuli. The results therefore seem to reflect a relatively natural listening strategy.

In the focus test, the effects of Cue ($F(5,247.68) = 38.76$, $p < .001$), Filter slope ($F(2,164.92) = 164.14$, $p < .001$), and the interaction ($F(10,283.34) = 36.75$, $p < .001$) on accuracy scores were highly significant. Post-hoc comparisons for Cue were all highly significant at $p = .015$ ($p < .001$). Post-hoc comparisons for Filter slope were significant at $p = .003$, except those between Non-vocoded and 120 dB/octave and those between 20 dB/octave, 80 dB/octave, and 160 dB/octave. The comparison between 120 dB/octave and 160 dB/octave was marginally significant ($p = .004$). Figure 2b shows that filter slope differentially affects the respective cues. The pattern in the Duration condition mimics the Both condition more closely than the F0 condition does, indicating that Duration is weighted more heavily than F0. This result contrasts with the cue weighting in the non-vocoded condition, as in that condition Duration was weighted less heavily than F0. Figure 2 further shows that there is no performance improvement with increasing filter slope beyond 20 dB/octave, except for a peak at 120 dB/octave for the F0 and Both conditions, which suggests that for (certain) extreme filter slopes only F0 provides additional information. The effect of filter slope is not as large as in the emotion test, as there is less variation in scores per Cue condition. This could be due to Duration being at the same time the most important cue and the cue that is least affected by filter slope.

In summary, these results show, first of all, that increasing filter slope facilitates prosody perception. In the emotion test, performances ranged between near chance level for 5 dB/octave to near ceiling level performance for 120 dB/octave. The effect was, however, less strong in the focus test, where Cue conditions with a higher peak performance also had a higher performance for the most

difficult slope condition, possibly due to a greater reliance on Duration, which is less affected by filter slope than F0 is. Second, in both tests, the 120 dB/octave condition, and not the sharpest filter (160 dB/octave), shows the performance that is closest to that of the non-vocoded condition. We will return to this paradoxical result in the Discussion section. Finally, the results demonstrate that both for emotion and focus discrimination, F0 and Duration are used differently. In the emotion test, the patterns of F0 and Both were closest together, whereas in the focus test, those of Duration and Both were closest together. This suggests a reliance mostly on F0 cues in the emotion test and on Duration cues in the focus test.

4.4 Discussion

This study aimed to find how extreme (as well as intermediate) filter slopes influenced the discriminability of emotional and linguistic prosody in a 15-channel cochlear implant simulation. We conjectured that increasing filter slope would have a facilitating effect on performance due to reduced channel interaction. A second question was how this function would differ depending on the availability of F0 vs. durational cues. This was investigated by superposing the two respective cues, individually or together, from utterances with the specific prosody onto variants of those utterances pronounced with neutral emotion and focus. The hypothesis was that F0 would be more affected than Duration, but, due to difference in cue weighting, this could have different implications for emotion and for focus perception.

4.4.1 The effect of filter slope on the discrimination of emotional and linguistic prosody

The effect of filter slope was explored with values ranging from 5 through 20, 80, and 120 to 160 dB/octave, as well as an unprocessed control condition. In the unprocessed condition, scores approached

ceiling, assuring that intended emotions and focus positions were successfully conveyed. As expected, steeper slopes yielded higher scores than shallower slopes. As shown by bias-neutral *d'* scores, performance increased monotonically from chance or near-chance level at 5 dB/octave to performance approaching ceiling level (Emotion) or around 90% (Focus) at 120 dB/octave in the most informative (Both) condition. Importantly, however, performance dropped again significantly to levels similar to those of the 80 dB/octave condition at 160 dB/octave. These results indicate that, up to a certain point, speech perception benefits from increasing the steepness of the slopes. This supports results from earlier studies on the effect of filter slope on vowel and consonant recognition (Shannon *et al.*, 1998; Fu and Shannon, 2002; Fu and Nogaki, 2005; Baskent, 2006; Litvak *et al.*, 2007; Bingabr *et al.*, 2008), as well as on prosody and music perception (Laneau *et al.*, 2006; Crew *et al.*, 2012). Further, it extends, but does not contradict, the findings of van de Velde *et al.* (2015), whose filter slopes (20 and 40 dB/octave) form a subset within the range of the present study. Performance on segmental perception has been found to reach a plateau around 12 or 18 dB/octave (Shannon *et al.*, 1998; Fu and Shannon, 2002), or, in one study, at 40 dB/octave (Litvak *et al.*, 2007). Sentence and word recognition showed asymptotic performance between 50 and 110 dB/octave, but since no intermediate values between 14 (the shallowest slope tested) and 50 dB/octave were included, the slope value where performance actually saturates might also be lower (Bingabr *et al.*, 2008). The present results, nevertheless, found much steeper optimal slopes, namely at 120 dB/octave. A margin of around 20 dB/octave has to be taken into account because of the spacing of the filter slope values included, so the actual optimum slope might lie between 100 and 140 dB/octave. Galvin *et al.* (2009) reviewed studies on frequency selectivity in the form of number of channels required to reach at least 80% correct performance for different types of signals by NH listeners using vocoders. Understanding of easy and difficult speech in quiet required less than five and less than ten channels, respectively; emotional and

linguistic (Mandarin tone) prosody recognition necessitated around 15 channels; identification of musical melodies without rhythmic cues demanded over 20 channels; and musical melody recognition required as many as 40 channels, possibly suggesting that higher frequency resolution requirements (due to its importance for the task or due to it being more difficult to segregate from the rest of the signal) correspond to increased task difficulty. We therefore submit that the higher filter slope saturation level that we found compared to studies on segmental perception occurred because perception of prosody requires greater frequency selectivity, possibly enhanced by increased channel independence, than segmental perception (cf., for instance, Laneau *et al.*, 2006).

The demonstrated effect of filter slope begs the question of what mechanism underlies it. The discrimination of F0 patterns, which was the most demanding task for the participants, could in principle be sustained by at least two mechanisms: spectral encoding (resolving F0 based on harmonics represented in respective filters) and temporal encoding (finding F0 based on the dynamic temporal envelope). Spectral encoding, however, is unlikely to have played a role, since the filter bandwidths, each spanning at least a quarter of an octave, are too broad to resolve harmonics. Further, as the envelope detector's cut-off frequency of 68 Hz was lower than most of the F0 values in the stimuli, temporal encoding must have been minimally effective or occurred only indirectly.

This raises the question how the manipulated filter slope influenced the accuracy of the perception of F0 cues, as was found in this study. Anderson *et al.* (2012) tested spectral ripple detection (discriminating logarithmic amplitude modulation from flat spectra) at different amplitude modulation depths (AMD) and ripple frequencies by CI users and found that detection of higher ripple frequencies required greater modulation depths. AMD therefore acts as a low-pass filter, with low AMDs lowering the cut-off frequency of the broadband noise more than high AMDs do. In NH participants listening through the same vocoder as in the current study, Litvak *et*

al. (2007) showed a negative correlation between amplitude modulation thresholds (the minimal detected AMD) and filter slopes varying from 5 dB/octave to 40 dB/octave, indicating that, as for the CI users in Anderson *et al.* (2012), spectral contrast detection in CI simulations with shallower slopes requires deeper amplitude modulations than with steeper slopes. We therefore contend that AMD might explain our results, i.e., that the filter slope effectively changed the AMD of the signal, since steeper slopes of neighboring filters cross each other at a lower amplitude than shallower slopes do. Through the suggested coupling of AMD with a broad cut-off frequency (Anderson *et al.*, 2012), filter slope indirectly introduced a broad low-pass filter. This could have influenced temporal processing of (low-frequency) periodicity cues. The exact mechanism behind the perception of F0 cues with the current signal processing settings is an interesting issue that is recommended for future research.

Interestingly, participants in our study performed optimally at 120 dB/octave but poorer at the steepest filter slope, 160 dB/octave, despite a monotonic improvement from 5 dB/octave up. Apparently, there is a functional limit to the steepness of the filter. This echoes results in Bingabr *et al.* (2008), where NH participants showed a performance decrement in some conditions with 4 or 8 channels on monosyllabic word recognition and sentence-in-noise tests from 50 to 110 dB/octave. These results could be related to the observation from previous studies that speech perception does not benefit from a narrower (e.g., bipolar) electrode configuration, but that, instead, a wider (e.g., monopolar) configuration might be equally or even more beneficial (Zwolan *et al.*, 1996; Pfingst *et al.*, 1997; Kwon and van den Honert, 2006; Zhu *et al.*, 2012). As with the results from the present study, this is counterintuitive because a narrower configuration, or, correspondingly, steeper filter slopes, is (are) expected to produce less channel interaction. It has been suggested that this is either (1) because a narrower configuration activates fewer neurons or (2) because the location of activated neurons is not optimal in that configuration (Pfingst *et al.*, 1997; Pfingst *et al.*, 2001; Kwon

and van den Honert, 2006; Zhu *et al.*, 2012). As for the first account, when fewer neurons are activated, a higher stimulation amplitude is required to achieve the same loudness, resulting in a disadvantage for the narrower configuration if this is not controlled for experimentally. In our case, however, channels were so close together (approximately a quarter of an octave) that they overlap even with the steepest filter slope, such that all neurons encompassed by neighboring channels would still be activated. As for the second account, a suboptimal location of recruited neurons can be due to dead regions along a recipient's cochlea or to incomplete frequency range coverage due to a shallow insertion depth. As we tested normally-hearing people, this is unlikely to have been a factor. We submit, therefore, that both accounts are relevant for actual CI users, but not for simulations, and that another explanation is in order. One possibility, tentatively suggested by Stafford *et al.* (2014), who found a performance plateau for slopes between 10 dB/mm and 17 dB/mm, is that the inherent filtering limits of the cochlea had been (almost) reached. Although we cannot disprove this account, it remains an open question why performance would decline between 120 dB/octave and 160 dB/octave.

4.4.2 The effect of phonetic cue on the discrimination of emotional and linguistic prosody

Acoustic measurements of the stimuli with transplanted prosody (but without vocoding) showed that the respective transplanted cues (F0, Duration, or Both) were available in the intended cue conditions, i.e., the response options in each test (sad vs. happy or noun vs. adjective focus) differed exactly and only with respect to the transplanted cue(s). This assured that responses and results were based on those cues. Note that in the focus test, the response options in the Duration condition differed not with respect to overall duration (as they did in the emotion test), but with respect to the duration of the focus word. Although other duration cues could have been available, focus word duration was assumed to provide at least one of the cues.

Cue reliance differed between emotional and linguistic (focus) prosody perception. In the case of emotional prosody, participants relied almost exclusively on F0, as witnessed by the fact that for slope conditions above 20 dB/octave, scores in the F0 and Both conditions were close together while those of the Duration condition were much lower. In the 20 dB/octave condition, however, listeners relied entirely on Duration. Most likely, this was because very little spectral information was preserved by the process of vocoding in that condition, leaving only duration information to exploit. By that reasoning, with 5 dB/octave slopes, the condition that even more rigorously affected F0 perception, the reliance on Duration would have had to be even more pronounced. In that condition, however, reliance on the two cues was balanced. It is possible that the distortion of the signal was so great that onsets and offsets of segments and syllables were not perceived, compromising the use of duration cues for segment, syllable, and/or word identification.

This explanation is supported by a study finding a negative effect of channel interaction on segment and word identification by CI users, an effect which was accounted for by assuming that channel interaction obscures boundaries between formant peaks and disrupts, among other phenomena, the amplitude envelope, resulting in compromised voicing distinctions and syllabic patterns (Stickney *et al.*, 2006). This would lead listeners to rely equally on all available prosodic cues, since Duration and F0 might be equally unhelpful. In line with this, participants' informal comments regarding the intelligibility of the phrases in all slope conditions suggested that segments, syllables, and words were considerably more difficult to identify in the shallowest filter slope condition than in the steepest filter slope condition. Note that this perceived intelligibility is not a confound explaining the overall pattern of results across filter slope conditions, as it does not explain why there were different patterns for different cues. Moreover, in the 5 dB/octave and 20 dB/octave conditions, performance was so close to chance that the pattern of results regarding cue weighting can be viewed as a tendency at most.

In contrast with emotion perception, for focus perception, participants relied predominantly on Duration, as Duration scores were almost as high as Both scores, whereas F0 scores were considerably lower. Exceptions to this pattern were found in the 5 dB/octave and 120 dB/octave condition. With 120 dB/octave slopes, Duration was still dominant, but not any more dominant than with the 20, 80, and 160 dB/octave slopes, whereas F0 showed a prominent peak. At 120 dB/octave, therefore, F0 was relatively important. This shows that F0 information is relevant for focus perception but is a less salient cue for focus perception than for emotion perception. F0 can and will be exploited only when vocoding optimally (within the limits allowed by the types of processing) preserves it. Duration information, however, can compensate for a lack of F0 information. In the 5 dB/octave condition, both cues were used, but Duration was dominant (although less in this condition than with 20, 80, or 160 dB/octave slopes). As with emotion perception, we conjecture that the sound quality is compromised to such a degree that alignment of segments with prosody is unreliable. Still, however, duration information was more usable with focus perception than with emotion perception because scores with duration are higher than those with F0. This might be because duration information for focus perception is prominent and segmentally independent (i.e., more aligned with complete words than with individual segments) enough to survive the distortion. This salience of duration information might also in part explain why it is dominant and sufficient in other slope conditions.

Our results are compatible with previous research on the way cue availability affects linguistic prosody perception with (simulated) cochlear implants. Pediatric recipients and NH peers in O'Halpin, (2009) judged if natural utterances were pronounced as compounds or phrases (e.g., *greenhouse* vs. *green house*) and which of two or three words in a sentence carried focus (e.g., *The DOG is eating a bone* vs. *The dog is EATING a bone* vs. *The dog is eating a BONE*). Participant-level comparison of performances on these tests with separately-assessed difference limens for F0, intensity, and duration in

prosody showed that whereas the controls made use of all available cues, the CI recipients in general relied primarily on duration and amplitude cues and less on F0 cues. A similar cue weighting strategy was found for CI users and vocoder listeners in Peng *et al.* (2009). In a task where participants decided if natural sentences and one-word stimuli in which F0, intensity and duration cues were incrementally resynthesized sounded as a question or as a statement, CI and vocoder listeners, compared to the full-spectrum (natural) situation, partially traded F0 cues for duration and intensity. In a similar paradigm for NH, CI-only, and CI users with amplified residual hearing, Marx et al. (2014) showed that for the CI-only group, question/statement discrimination was affected by neutralization of amplitude and temporal cues but not by neutralization of F0 cues, whereas the other groups showed the opposite pattern of results, suggesting that F0 is an important cue but is not available to or used by CI users.

Cue weighting in emotional prosody is less studied. Vocal emotion recognition was more affected by amplitude normalization for CI users than for NH listeners (Luo *et al.*, 2007). In another test, subgroups of these listener groups performed better with an increasing number of channels (tested on 1, 2, 4, and 8 channels) and, orthogonally, with a higher cut-off temporal frequency (400 vs. 50 Hz), showing, according to the authors, use of both F0 (channel number) and temporal (cut-off frequency) cues. However, performance did not improve beyond 2 channels.

From this literature, a pattern of results emerges in which under conditions of (simulated) CI hearing, perception of prosody is based primarily on temporal and intensity cues and much less on spectral (F0) cues. The present research is, to our knowledge, the first to compare emotional and linguistic prosody on this issue. Our results support findings showing a dominance of non-F0 cues. However, this is only the case for linguistic prosody. Emotional prosody, which is less studied, shows a reliance on F0 cues. We therefore submit that the cue weighting found in research so far is relevant for linguistic prosody, but not for emotional prosody.

4.4.3 Implications for CI users

Speech perception performance by CI users corresponds to that of NH listeners using vocoded speech with a maximum of around eight channels (Friesen *et al.*, 2001; Baskent, 2006) and filter slopes of around 12 dB/octave or less (Shannon *et al.*, 1998; Shannon, 2002; Fu and Nogaki, 2005). If we interpolate values with that filter slope from our results and translate *d'* scores to percent correct, values of around 60% for emotion discrimination and 75% for focus discrimination could be obtained in the condition involving all available cues. Although in our experiment this was above chance (50%), it has to be taken into account that in real life, emotion perception entails open-set recognition instead of closed-set discrimination, and therefore actual vocal emotion recognition performance is most likely lower than in the experiment. This difficulty may reflect the observation that CI users have more difficulty perceiving emotions than people with normal hearing do, and that they rely relatively heavily on visual instead of vocal information (Winn *et al.*, 2013; Strelnikov *et al.*, 2015; however, see Most and Michaelis).

The generalizability of the current results to actual CI perception has to be viewed in light of the numerous technical and physiological differences between CI and vocoder listening mentioned in the section Introduction. The results hold for CIs with the current number of channels (15; see also the section Limitations below). Further, to translate filter slope values to current spread along the basilar membrane in CI users, a correction would need to be made for the difference in dynamic range (Bingabr *et al.*, 2008, suggest dividing vocoder values by 3.3). Note that results of our study do not require this correction, as they are intended only to model (not equal) CI perception. Finally, the effect of filter slope that we observed might be weaker in CI users because channel interactions can be aggravated by dead regions in the auditory neuron population and because higher filter slopes will activate fewer neurons and thus might convey the signal less effectively. Despite these nuances, the vocoder applied in the current study was shown to reliably model CI segmental

perception in a study using the same algorithm, albeit with shallower filter slopes (Litvak *et al.*, 2007). With the slopes that Litvak *et al.* (2007) found to correspond to those of CI users (5 to 30 dB/octave), our results show that the F0 and Duration cues are weighted equally up to around at least 20 dB/octave for emotion perception, duration is given much more weight than F0 beginning at 20 dB/octave and onwards for focus perception. These results therefore extend the findings by Litvak *et al.* (2007) by differentiating phonetic cues in prosody perception at realistic filter slopes.

Another way in which the present investigation extends Litvak *et al.* (2007) is by its exploration of more extreme slope values. We found that with the current parameters, the theoretical target filter slope for prosody perception is between 100 and 140 dB/octave. Although this may not currently be technologically and physiologically feasible, it is important to view the realistic values and performance into the perspective of this theoretical filter slope optimum. That is, for emotion perception, the realistic values are about 35% lower than the performance that would be obtained if filter slope were not a limiting factor, and for focus perception this is about 10% (that is, the percentage correct difference between the optimal filter slope of 120 dB/octave and the scores for the realistic slopes of between 5 and 10 dB/octave). The optimal filter slope value that we have identified marks a functional limit to filter steepness. In other words, making the slopes steeper improves prosody perception but only up to a certain point (around 120 dB/octave). This result is in contrast with research showing that for segment recognition, an asymptote is reached at much lower levels, even with more complex tasks (Litvak *et al.*, 2007). The current study therefore complements the literature by showing that for optimal prosody perception, even with a simple 2AFC choice task in acoustically optimal conditions (no background noise), much better spatial selectivity is required than for segmental identification.

Our results further suggest that the difficulty CI users have perceiving emotion may differ from the difficulty they have

perceiving focus. Depending on the filter slope, performance ranged between 56% and 95% for emotion discrimination and between 68% and 87% for focus discrimination. This suggests that for shallower (more realistic) slopes, focus perception is easier than emotion perception while for hypothetically steeper slopes, emotion perception is more successful. The reason for this is that focus perception is based more on temporal cues, which are less affected by vocoding, than spectral cues. In contrast, for emotion perception, F0 provides even more information than temporal cues do for focus, but it is only effectively available for steeper slopes. It has to be noted that while these results are valid for the current vocoding algorithm and the current stimuli, they cannot be generalized without caution to other vocoding techniques, cochlear implant speech processors, or stimuli. Performance is dependent on the exact audiological history and abilities of the listener, the paradigm in which prosody needs to be perceived (e.g., discrimination vs. identification) and the way the stimuli are pronounced. However, since linguistic and emotional prosody were presented to the same participants under equal circumstances, the difference in performance is likely to reflect inherent differences between those two types of signals, and merits further research (e.g., Witteman *et al.*, 2011). Because an extension with additional speakers, thus multiplying the number of stimuli, would have made the task too arduous for participants, this is left as a follow-up for future research in which, based on our results, only pivotal filter slope values can be included.

4.4.4 Limitations

A number of drawbacks of this study apart from those addressed in separate sections have to be taken into account. First of all, there was only one speaker involved. As individual speakers are known to vary in their realization of emotional (Scherer, Banse, Wallbott, & Goldbeck, 1991) and linguistic (Kraayeveld, 1997) prosody, the results of this study may not be generalized to other speakers. This is despite the fact that the near-ceiling level discrimination scores in the

unprocessed condition showed that the emotions and focus positions were successfully conveyed. In future research, paradigms might be considered in which emotions and focused elements are realized more naturally, e.g. by means of role playing or reading lists of items with contrastive constituents (Krahmer & Swerts, 2001; Velten, 1968). It has to be noted, however, that in our study, the use of stimuli from multiple speakers would have rendered the experiment too long and burdensome for the listeners. Moreover, as we asked the speaker to keep the speaking rate across variants of each phrase more or less constant as well as to produce (unnatural) emotionally and focus-neutral variants, more natural elicitations were not feasible.

A second limitation of this study concerns the control of speaking rate by the speaker recording the stimuli. This was done to remove gross temporal differences between emotional or focus variants because they would hypothetically not tax the reliance on durational nuances within phrases but any effect of duration could instead reflect, for instance, overall listening time per stimulus, which is not a phonetic measure. This control of speaking rate, however, did make the stimuli less natural, since the speaker had to suppress a difference that she might have realized otherwise. Given that this procedure made two response options (two emotions or two focus positions, respectively) more similar to each other, it cannot explain results by itself, but its consequence was in fact an underestimation of the differentiability of emotion or focus variants based on durational cues. The results apply mainly to phrase-internal duration differences. The control of speaking rate, as shown by the acoustical measures, was more successful for the focus than for the emotion stimuli, as for the latter the difference in average phrase duration between variants was much higher than for the former. This entails that the result that focus perception weighted duration cues more heavily than F0 cues, while this was the other way around for emotion perception, was underestimated because even with the additional duration cues that were available for the emotion relative to the focus stimuli, they were

not relied on, whereas the fewer duration cues that were available for the focus stimuli were relied on.

A fourth limitation is that (for practical reasons) we only tested one channel number. The channel number we have chosen is believed to theoretically represent a type of CI (currently an Advanced Bionics device) that makes use of current steering and that in future developments might benefit from techniques, such as multipole algorithms, that allow channel interactions that are much more reduced than currently achieved. A lower channel number (as was also suggested by our pilot test) was less likely to show an effect of filter slope for a wide range of slope values (Stafford et al., 2014). Nevertheless, in order to gain a completer image of the effect of filter slope on prosody perception, it is mandatory that in future studies other channel numbers are investigated.

A final limitation is that we investigated only two cues, F0 and duration. This was done to unravel the relative weighting of these two types of information, which would have been impossible or greatly complicated if other cues were available as well. These alternative cues did not play a role in the present experiment because only F0 and duration cues were made available to the listeners, namely by transplanting those aspects of the prosody onto the same segmental basis for both variants of the phrases per test. Other types of information, such as intensity and spectral information, could, however, also support emotion and focus discrimination (Scherer et al., 1991; van Heuven & Sluijter, 1996). The lack of alternative cue availability in our study nevertheless underestimates the discriminability of the emotions and focus positions. It is likely that the weighting of the cues currently investigated would be different if other cues were available as well because other cues might be more reliable. It has to be noted, however, that the cues studied allowed very high sensitivity when combined (the Both condition), implying that they were sufficient for successful discrimination and that the task did not require other cues to be present.

Conclusions

The purpose of this study was to investigate the effect of filter slope on the perception of emotion and focus prosody with different available cues (only F0, only duration, and both). A number of conclusions can be drawn from the results.

1) Emotion and focus discrimination improve with steeper filter slopes. This improvement is more pronounced for emotion perception than for focus perception, i.e., emotion perception performance starts from lower levels at shallow slopes and increases to higher levels at steep slopes than focus perception.
2) At 5 dB/octave, the shallowest slope tested, performance is close to chance level, but higher for focus than for emotion perception; at 120 dB/octave, where performance was optimal, scores were around 90% correct, but higher for emotion than for focus perception.
3) The optimal filter slope for both emotion and focus perception is between 100 and 140 dB/octave, which can be considered a theoretical target value. At 160 dB/octave, the steepest slope tested, performance is poorer than at 120 dB/octave.
4) In emotion perception, the F0 cue is weighted more heavily than duration cues, whereas in focus perception, duration cues are weighted more heavily than F0 cues. In emotion perception, F0 is more informative but only becomes available with steep slopes. In focus perception, on the other hand, duration cues, although less informative than F0 cues in emotion perception, are less compromised by vocoding such that they are relatively well preserved with shallow slopes.
5) Cochlear implant users hypothetically score around 35% lower than the performance observed at the optimum filter slope for emotion perception and around 10% for focus perception. It is worthwhile further reducing channel interactions in CI users,

because there is much room for improvement in the area of prosody perception.

Acknowledgements

Leiden University Centre for Linguistics (LUCL), Leiden University Medical Center (LUMC), and Leiden Institute for Brain and Cognition (LIBC) supported this research. We are grateful to Jos Pacilly (LUCL language laboratories) for support involving signal processing, stimulus recording and technical setup for the experiment. We also wish to thank the participants of this study for their participation.

Chapter **5**

Cue-weighting in the perception of music and prosody with cochlear implant simulations

Abstract

Cochlear implant (CI) users have difficulty perceiving music and prosody. Musical training has been found to transfer to language perception. However, it is not known whether auditory cues can be separately trained and transferred after implantation. Two groups of normally hearing (NH) listeners were trained in perceiving either pitch or temporal cues in music under simulated CI conditions (vocoding). They were subsequently tested on another music test (Familiar Melody Identification, FMI) and two prosody tests (Emotion Discrimination, ED; Focus Discrimination, FD), each in conditions with only pitch cues, only temporal cues or both cues available. We hypothesized cue-specific training-related reliance, and possibly cross-cue and cross-domain (music to language) training transfer. Tendencies towards training-related cue reliance and individual participant-level cross-cue or cross-domain correlations for pitch and cross-cue plus cross-domain correlations for temporal cues were revealed. There were no correlations between scores and musical background or listening habits. Participants relied on temporal cues for FMI, mostly on pitch cues for ED and approximately equally on

pitch and temporal cues for FD. Vocoding makes listeners weight temporal cues more heavily. The results show a potential for post-implantation musical training in enhancing both music and prosody perception for different cues.

5.1 Introduction

For users of cochlear implants (CI), the perception of music and speech prosody poses considerable challenges. These perceptual domains represent central aspects of enjoyment and communication in life. Being able to enjoy music has been found to correlate with quality of life for people with CIs, depending on the quality of the sound provided (Lassaletta et al., 2007). The quality of the sound being lower than that for normal hearing, performance on a number of specific tasks has been found to be compromised for CI recipients. Among other issues, they have difficulty, to a greater or lesser extent, with the identification of melodic contours (Galvin, Fu, & Shannon, 2009), the distinction of timbres or instruments (Galvin et al., 2009; Gfeller, Witt, Woodworth, Mehr, & Knutson, 2002), the recognition of familiar melodies (Gfeller, Turner, et al., 2002; Kong, Cruz, Jones, & Zeng, 2004) and emotions (Hopyan, Manno III, Papsin, & Gordon, 2016; Shirvani, Jafari, Sheibanizadeh, Motasaddi Zarandy, & Jalaie, 2014) in music, and they have a higher threshold for distinguishing melodic intervals (Luo, Masterson, & Wu, 2014). Looi, Gfeller, and Driscoll (2012) concluded in a review that CI users have a lower appraisal of music than people with normal hearing (NH), and avoid listening to music more than they did before implantation.

Prosody refers to the variation in the way a specific string of consonants and vowels (segments) that make up an utterance can be pronounced (Lehiste, 1976). This variation occurs primarily in the dimensions of frequency (e.g., intonation), intensity (stress), and duration (pauses, phrasing by timing). The functions of prosody can be classified into linguistic and emotional functions. Linguistic prosody signals aspects of the meaning of an utterance, such as the grouping of words, and the way specific words relate to the context, such as by marking new information. Emotional prosody signals the emotional or attitudinal state of the speaker. In contrast with the processing of the segments of speech, CI users have trouble perceiving prosody. Meister et al. (2007) showed that implanted

participants scored lower than controls with normal hearing on the recognition of six types of linguistic prosody. The disadvantage was largest for intonational word and sentence accent and sentence type (question or statement), and was smallest for minimal word pairs differing in duration (or duration and spectrum) of a phoneme and for phrasing by timing. These results suggest that perception based on a timing cue is less problematic than that based on frequency cues. In a study by Luo, Fu, and Galvin (2007), CI recipients and NH controls decided whether semantically neutral sentences were pronounced with an angry, a happy, a sad, an anxious or a neutral emotion. Whereas the controls scored around 90% correct, the CI recipients' performance was around 40% correct. Taken together, these studies could entail that CI recipients potentially miss out on aspects of the meaning of the utterances and the emotion of the speakers. This might be one of the causes underlying an atypical socio-emotional development in the case of children with CIs (Wiefferink, Rieffe, Ketelaar, De Raeve, & Frijns, 2013).

The difficulties with the perception of music and prosody that CI users experience most likely stem from the limited transmission of pitch provided by the device. CIs typically transmit the temporal dynamic envelope of a limited number of spectral bands, modulating a train of electric pulses with a fixed rate, to tonotopically corresponding locations in the cochlea. This procedure removes the signal's fine-structure. The mechanisms of pitch perception that this is theoretically compatible with, allow pitch perception only to a very restricted degree, for a number of reasons. First of all, for pitch by cochlear location, the number of effective bands appears to be limited to around eight, due to spectral overlap (e.g., Friesen, Shannon, Baskent, & Wang, 2001). Second, pitch by stimulation rate works only up to 300 to 500 Hz (Carlyon, Deeks, & McKay, 2010). Finally, pitch can be derived from the temporal envelope, but this is limited by the envelope detector's cut-off frequency and the stimulation rate (Busby, Tong, & Clark, 1993; Xin & Fu, 2004). In practice, these mechanisms together allow a Just Noticeable Difference of

approximately half an octave with much variation depending on the task and the individual, which is considerably more than the one semitone or less reported for NH listeners (Kang et al., 2009; O'Halpin, 2009; Wang, Zhou, & Xu, 2011).

As a result of the poor pitch perceptual abilities, CI recipients attend differently to the available cues in music and prosody than NH listeners do. CI users in one of the experiments by Kong et al. (2004) had greater difficulty recognizing familiar melodies when both rhythmic and tonal cues (in one condition) or when only tonal cues (in another condition) were available than NH controls did, but the difference between the groups was much larger in the latter than in the former condition. Children with CIs recognized familiar songs based on rhythm as accurately as NH peers but performed more poorly than the latter when having to rely on tone (Bartov & Most, 2014). In a set of experiments by O'Halpin (2009) children with and without CIs decided whether utterances were compounds (with stress on the first element, e.g., *bluebottle*) or phrases (with stress on the second element, e.g. *blue bottle*) and identified which word in a phrase carried a focal accent. The author compared scores on those tasks to the participants' difference limens for F0, intensity and duration of nonsense syllables, which were synthetically incrementally manipulated. She concluded that the implanted children pay least attention to F0 cues, more to amplitude cues and most to duration cues. Marx et al. (2014) studied cue weighting in question/statement discrimination with either monotonous F0 or with neutralized amplitude and duration by NH and CI listeners with (CI-combined) or without (CI-only) an additional hearing aid. For CI-only users, scores were affected by removal of amplitude/temporal cues but not by removal of F0 cues, whereas for the other groups it was the other way around. This suggests that F0 cues were not available for CI users. The above studies together seem to indicate that compared to NH listeners, implanted listeners rely more on temporal and intensity cues and less on spectral cues.

Performance on music and prosody perception tasks has been found to be enhanced by musical training, where musical training either refers to theoretical or practical music lessons that an individual had some time before taking part in a study (long-term), or to relatively short task-relevant training designed as part of the study (short-term). In NH people, benefits related to being a musician that have been reported include more fine-grained temporal processing, smaller difference limens for pitch, more efficient segregation of speech from noise, improved recognition of lexical tones and timbres as well as enhanced reading skills and working memory (for reviews, see Moreno & Bidelman, 2014; Patel, 2014). For CI users, long- or short-term musical training has been shown to facilitate pitch discrimination (Chen et al., 2010; Vandali, Sly, Cowan, & van Hoesel, 2015), melodic contour identification (Fu, Galvin, Wang, & Wu, 2015; Galvin, Eskridge, Oba, & Fu, 2012), and prosodic processing such as that of stress, compounds versus phrasal prosody (which could effectively be signaled by stress), F0, and contrastive focus (Patel, 2014; Torppa et al., 2014; Torppa, Faulkner, Vainio, & Järvikivi, 2010) (for a review on musical training, see Looi et al., 2012). It is still an open question at this point whether the benefit of musical training is merely correlational or also causal (Moreno & Bidelman, 2014). Nevertheless, Limb and Roy (2014) concluded that musical training might prove the best way to improve music listening for CI users. More recently, Fuller, Galvin, Maat, Free, and Baskent (2014) studied the positive influence of musicianship on auditory processing, which they called the ‘musician effect’, under the degraded spectral condition of CI hearing. With simulated CI hearing, they showed that musicians had an advantage over non-musicians in emotion perception for speech and even stronger for melodic contour identification, but not as much for word identification. They interpreted these results as suggesting that the more the task requires pitch perception, the larger the musician effect is. Apparently, they argued, the effect operates on a relatively specific, lower level (i.e., not on a more general cognitive level).

Two conclusions about CI perception can be drawn from this overview. First of all, performance on music-related tasks is positively influenced by short- or long-term musical training. Second, the effect can transfer to non-musical, speech-related tasks. The transfer of short-term musical training, however, has only just begun to be studied (Patel, 2014; Yucel, Sennaroglu, & Belgin, 2009). Moreno and Bidelman (2014) made a distinction between near and far transfer, where near transfer refers to transfer between closely related psychophysical features such as cues, and where far transfer denotes transfer between different cognitive domains such as language versus music. In the present study, we aimed to test both types of transfer in a single setup. Given the existence of the musician effect, we asked ourselves in the present study if this effect also works for separate cues. That is, with a hearing situation like that of CI users, is it possible to train listeners to improve their perception of one specific cue, without enhancing the competence on another cue? If this is the case, the range of the training effect is highly specific (i.e., restricted to that very cue); if not, the effect operates on a more general cognitive or auditory level. In order to find out more about the level on which the effect operates, if at all, we also tested the effect on a non-musical domain, viz. the perception of prosody. There were thus two orthogonal psychophysical or cognitive levels on which transfer of musical cue training could take place: within or beyond the same cue (tone or temporal) and within or beyond the same domain (music or language), corresponding to near and far transfer, respectively.

5.2 Methods

In order to test the effect of music training on music and prosody tasks by CI users, we conducted tests with NH listeners using vocoder simulations. Participants were divided into two groups, one which followed temporal (rhythmic) training and another which followed pitch (melodic) training. All participants completed seven tests, three

for training (called the Trainings) and four for post-training testing (the Tests). The Tests were identical for everybody, but the Trainings differed per group. The Trainings were three variants of melody identification and the Tests comprised a Familiar Melody Identification (FMI) test, another musical task in which participants reported where they felt an ambiguous melody started (the Ambiguous Melody (AM) test), and two prosody discrimination tasks, an Emotion Discrimination (ED) test and a Focus Discrimination (FD) test. The goal of the AMT was to assess whether participants attended more to melody or to rhythm when listening to melodies, and which would enable us to rule out a potential confound of attention (instead of competence). The FMI test and the prosody tests contained conditions in which either temporal or pitch cues, or both cues simultaneously, were present. With this design, two groups were trained either in musical rhythm or musical melody perception, and were subsequently tested on identical music and prosody tasks in which their pitch vs. temporal cue weighting was assessed. Trainings and Tests were performed with vocoded stimuli. The prosodic Tests also included a condition with non-vocoded stimuli.

5.2.1 Participants

Fifty-two higher-education students (47 women, 5 men) with normal hearing participated as volunteers or for credits. They had a mean age of 20 years and 5 months (henceforth, '20;5') (SD: 3;7). Candidates were excluded if they had hearing problems, if they were not native speakers of Dutch or if they were professional musicians. They performed a tone audiometry test at the octaves from 0,125 to 8 kHz (Audio Console 3.3.2, Inmedico A/S, Lystrup, Denmark) and were rejected if they had thresholds elevated more than 40 dB above normal at any of the frequencies. One candidate was excluded on this basis. All participants signed an informed consent form and all but three per group completed a brief questionnaire about their education level and musical background, adapted from the Salk/McGill music inventory (Levitin et al., 2004). Participants were randomly assigned to one of

Table 1. Frequencies and means (plus standard deviations) of demographic variables for the Temporal and the Pitch groups. In each group, three participants did not fill in the music background questionnaire, such that the responses to the questions a-h are based on 23 respondents per group. For questions d, e, and h, values of *0* were imputed for participants for whom the questions were not applicable. For questions f and g, the participants were not included if the questions were not applicable. Included are results of χ^2-tests (for the frequencies) and independent samples *t*-tests (for the means) for the outcome variables. No group differences were significant according to these tests.

	Group				
Personal or demographic variable	**Temporal**	**Pitch**			
	count	*count*	χ^2	*df*	*p*
Male/female	2/24	3/23			1.00[1]
Right-/left-handed	22/4	20/6	.50	1	.48
(a) Do you play an instrument or sing? Yes/no	5/18	7/16	.45	1	.50
(b) Did you receive practical training in playing/singing? Yes/no	14/9	14/9	.00	1	1.00
(c) Did you receive theoretical training in music? Yes/no	9/14	13/10	1.39	1	.24
	Mean (SD)	*Mean (SD)*	*t*	*df*	*p*
(d) How many hours do you play/sing per week?[2]	1.8 (4.9)	1.0 (2.4)	.69	44	.50
(e) For how many years have you played/sung?[2]	2.3 (5.0)	2.6 (4.8)	-.19	44	.85
(f) At what age did you start playing/singing?[3]	10.0 (5.6)	11.1 (3.8)	-.43	10	.68
(g) How many years ago did you last receive the training?[3]	3.7 (3.2)	5.1 (3.1)	-1.2	27	.24
(h) How many hours per week do you listen to music?[2]	14.6 (11.1)	14.8 (13.7)	-.047	44	.96

[1]Fischer's exact; [2]'0' as an answer allowed; [3]only if applicable (therefore, *df* was reduced compared to other variables)

two groups: a group receiving temporal cue training (Temporal group) and a group receiving pitch cue training (Pitch group) (both $N = 26$). Table 1 shows personal, demographic and musical background characteristics of the two groups, as well as results of χ^2-tests and *t*-tests of differences in frequencies and means, respectively. Groups did not differ statistically on any of these variables. In absolute terms, the

Pitch group had a slight advantage in number of people playing an instrument or singing and in having received theoretical music training. On the other hand, the Temporal group performed music for more hours per week and (if applicable) had last received training more recently. The study was approved by the ethical committee of the Faculty of Humanities of Leiden University.

5.2.2 Stimuli

Trainings. Stimuli for the Trainings were ten five-note, 4/4 measure melodic piano contours with approximate ranges of one octave around A4 (440 Hz), composed by the authors for the current purpose. Notes were 500 ms long (intensity decay of around 0,030 dB/ms as simulated by the software) and had no rests in between. There was variation in the interval size and direction of the melodies, in order to ensure that there was a range of melodies with more and less salient pitch changes. These ten contours served as templates to create variants for both sets of three Trainings per group. All stimuli for the first two sets were created as wav files with MuseScore[1] (Schweer, 2012); for the third set, the melodies from MuseScore were further processed with *Praat version 5* (Boersma & Weenink, 2014). Music scores and schemas of stimuli of all Trainings are displayed in Figure 1. In the first set, the notes were kept as quarter notes for the Pitch group, but were created as ten rhythmic variants for the Temporal group. The rhythmic variants covered a range of more and less salient patterns. This combined procedure yielded one hundred shapes (ten rhythmic variants for each of ten melodies). From that pool, the Pitch group was to discriminate different melodies with equal (but varying between trials) rhythms, whereas the Temporal group was to discriminate different rhythms with equal (but varying between trials) melodies. In this way, the same stimuli were used for both groups, but they were trained for different cues while ignoring another cue. A similar procedure was followed for the second and

[1] https://musescore.org/

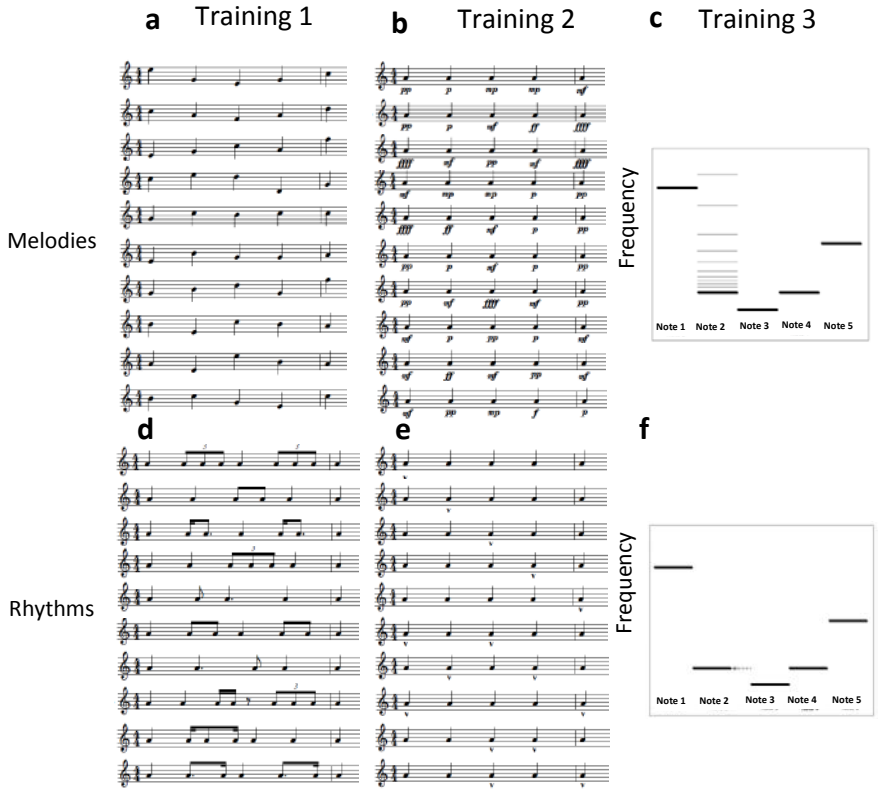

Figure 1. Scores (Trainings 1 and 2) and diagrams (Training 3) of the templates of musical stimuli composed and created for the Trainings. See the text for an explanation of the Trainings. (a) The 10 melodic contours for Pitch Training 1. The ones shown all have the same rhythm, although in the experiment varying rhythms were used. (b) The 10 dynamic contours for Pitch Training 2. The experiment used the rhythm shown but with the varying melodies from panel a. Loudness is symbolized with the increasing scale *pp-p-mp-mf-f-ff-ffff*, spanning an approximate range of 24 dB. (c) Schematic display of the first melody of panel a (as an example), showing the 10 possible incremental variants of pitch of one of its notes, used for Pitch Training 3. The fat lines represent the original; the thin lines represent the variants. In the experiment, all variants on all notes of six of the melodies were used. (d) The 10 rhythmic patterns for Temporal Training 1. The ones shown are on a single note, whereas in the experiment, varying melodies were used. (e) 10 accent patterns for Temporal Training 2. Notes marked by the 'v' sign are accented, having at the attack a loudness of between approximately 2 and 13 dB more than surrounding peaks, depending on the position and pitch of the notes involved. (f) Schematic display of the first melody of panel a (as an example), showing the 10 possible incremental variants of duration of one of its notes, used for Temporal Training 3. The fat lines represent the original; the thin lines represent the variants. In the experiment, all variants on all notes of six of the melodies were used.

third Training. For the second Training, ten patterns of increasing (crescendo) and decreasing (decrescendo) loudness were generated, for the Pitch set (Figure 1, Panel b), as well as ten patterns of one or two single-note accents per melody, for the Temporal set (Figure 1, Panel e). The (de)crescendo patterns were believed to represent more of a melodic aspect of the contour than the accents because they extended over the entire melody, whereas the accents establish a beat, which is more of a temporal feature. The third Training was a modified melody task (Swanson, Dawson, & Mcdermott, 2009). Variants were created in *Praat* using the Pitch Synchronous Overlay and Add (PSOLA) technique (Moulines & Verhelst, 1995). For each of the contour's five note positions, the deviant note was higher in pitch by 3, 5, 7, 10, 15, 20, 30, 40, 60, or 80% (for the Pitch group, Figure 1, Panel c), or longer by 5, 6, 7, 8, 10, 15, 20, 30, 40, or 50% (for the Temporal group, Figure 1, Panel f). We ensured by means of visual and auditory inspection that no signal distortions were introduced by the processing. Note that some of the temporal increments, and consequently the total range, are smaller than the pitch increments and range because with CI hearing and hearing through vocoders the temporal resolution is higher than the frequency resolution. Since the aim of this study is not to test temporal vs. frequency resolution but to test the reliance on those cues, the temporal dimension was not to have a (too large) perceptual advantage. Per Training, the stimulus set was divided into easy and difficult contrasts, where a contrast refers to the difference between the stimuli that are to be distinguished from each other within a trial. Easy are those for which the difference between the stimuli is relatively large, and difficult are those for which the difference is relatively small. For Trainings 1 and 2, this was based on differences in shape and intervals. For Training 3, the five largest increments formed an easy contrast with the original melody, and the five smallest increments formed a difficult contrast. The purpose of this distinction was to have participants switch to difficult contrasts in case they reached a ceiling with the easy contrasts. In each of the six

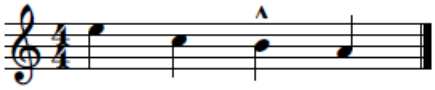

Figure 2. Example of a contour for the Ambiguous Melody test. The accented note, marked by '^', is, essentially, not the lowest or highest of the four.

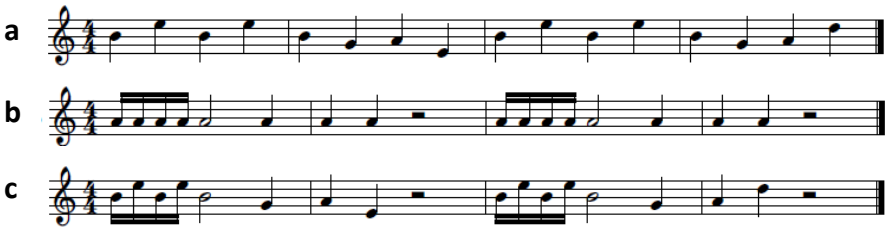

Figure 3. Example of variants for the Familiar Melody Identification test. The example shown is for the melody 'Morricone – The good, the bad and the ugly'. (a) With intact melody, but neutralized rhythm, (b) with intact rhythm, but neutralized melody, (c) with intact melody and rhythm.

Trainings, a subset of 60 of the 100 created shapes was used in the experiment.

Musical tests. Stimuli of the two musical tests, the Familiar Melody Identification Test (FMI; the theoretically central test of this study) and the Ambiguous Melody (AM) test, were created in a way similar to that of the Trainings. For the AM test, four-note melodic contours were created, in which one note was loudness-accented (changing the overall amplitude but not the spectral slope) but whereby that accented note was never the highest or lowest not of the four. Of each contour, a chain of sixteen repetitions was formed as a single file. Participants were asked to indicate on which note they felt the contour started. In general, either the accented, the highest or the lowest note was most likely to be perceived as the first note of each contour. An example of a contour is shown in Figure 2. Visual and

auditory checks made sure that no boundaries between repetitions could be perceived. The FMI test consisted of (excerpts of) ten well-known Dutch and international melodies which in a pilot study were found to be familiar for all participants. They were: 'Beethoven – 5th symphony' (slowed down), 'Bach – menuet', 'Mozart – Eine Kleine Nachtmusik', 'Morricone – The good, the bad and the ugly', 'Jingle bells', 'Happy birthday', 'Nokia ringtone', 'Hoedje van papier', and 'Sinterklaas kapoentje'. The melodies were of different duration and number of notes. Three variants of each tune were created, one maintaining only the pitch, one maintaining only the rhythm and one maintaining both the pitch and the rhythm. This was done by changing all individual notes into quarter notes, by changing all pitches to a single pitch (A4), and by not changing anything, respectively. An example of these variants is shown in Figure 3.

Prosodic tests. For the two prosodic tests, the Emotion Discrimination (ED) test and the Focus Discrimination (FD) test, sentences with durations between 1.5 and 2 seconds were recorded as natural stimuli in a sound-treated booth by a professional linguist (CL) with a sampling frequency of 44,100 kHz and a sampling depth of 32 bit. For the ED test, the sentences were twelve article-color-noun phrases (e.g., *een rode stoel*, 'a red chair'), each in three variants: (1) with no particular emotion (neutral), (2) with a happy-sounding emotion and (3) with a sad-sounding emotion. For the FD test, the sentences were of the form article-color-noun-*en een* (e.g., *een gele bloem en een*, 'a yellow flower and a'). The purpose of the words *en een* was to avoid phrase-final prosody on the preceding noun and to implicitly evoke a continuation containing a contrasting object or color supporting the interpretation of focus. Mirroring the FD test's stimuli, the sentences were recorded in three variants: (1) with equal focus on the adjective and the noun, (2) half of them once with narrow focus on the color and (3) the other half once with narrow focus on the noun. For the stimuli of both tests, in order to prevent ceiling-level performance in discrimination due to global sentence-level rhythmic or durational differences between variants, we asked the speaker to

keep the general speaking rate more or less constant across the variants. Following recording, for all stimuli of both tests, we spliced the relevant aspects of the prosody from the emotional or focused utterance onto the neutral variant of the same phrase on an phone by phone basis, again using the PSOLA algorithm incorporated in *Praat*. We thus created three resynthesized variants, respectively importing from the non-neutral phrases (1) the pitch contour (Pitch condition), (2) the phone durations (Temporal condition), and (3) both the pitch contour and the phone durations (Total condition).

Vocoding. As the final step in stimulus processing, we simulated cochlear implant hearing by applying an 8-channel sinewave vocoder modelled on Continuous Interleaved Sampling (CIS), using the *AngelSim*^{TM} software (Fu, 2013). In the procedure, the signal is band-passed between 200 to 7,000 Hz with 24 dB/octave filter slopes, with cut-off frequencies based on Fuller et al. (2014). Of each band the amplitude envelope is detected with a cut-off frequency of 240 Hz (24 dB/octave). A sinewave instead of a noise vocoder was chosen because it leaves the spectral information of the signal more intact, without which the tasks might have become infeasible (Fuller et al., 2014). It has to be noted, however, that noise vocoders might be more realistic simulations of CI hearing.

5.2.3 Procedure

Participants performed all components of the experiment in a single session, which lasted around two hours including breaks. A session had the following setup for all participants. They first completed either the Pitch or the Temporal Trainings 1, 2 and 3 (each 15 minutes), followed by the AM test (10 minutes), the FMI test (20 minutes) and, counterbalanced per group, the ED an FD tests (each 12 minutes). All these components, except the AM test, were run with *E-Prime 2.0* (Psychology Software Tools, Pittsburgh, PA, USA; Schneider, Eschman, & Zuccolotto, 2012) in a sound-treated booth using headphones (Beyerdynamic DT770 PRO), at a distance of 70 cm from the screen. The music tests were conducted with vocoded

stimuli and the prosody tests both with non-vocoded and vocoded stimuli. The vocoded conditions were the focus of the study since we wanted to mimic the possible effect of training on hearing in CI users; the non-vocoded condition in the prosody tests was included for comparison with analyses not reported here. In all components, accuracy and reaction time data were registered unless stated otherwise.

Trainings. The procedures of all Trainings were identical. Participants passed through a short practice phase familiarizing them with the task and vocoded stimuli. The task objective was to indicate by button-press which of three melodic contours heard was different from the other two. Trials had the following structure: a fixation cross (on screen for 1,000 ms), consecutive playing of three contours (their respective durations), feedback (only for practicing; visible for 1,500 ms after the response), inter-stimulus interval (500 ms). The time to respond was 4,000 ms measured from the onset of the third contour. The subsequent experimental phase consisted of two blocks of 30 trials, with a break in between. These were either twice the same easy block, if participants scored less than 90% correct in the first block, or alternatively, one easy followed by one difficult block, if they scored at least 90% correct in the first block. Participants received feedback about the accuracy after each block as well as the written remark that they should attain at least 85% correct. The order of stimuli was randomized for each participant and the position of the target contour (first, second or third) was counterbalanced.

Musical tests. In the AM test, participants indicated for each of the eight contour chains on which of the four notes they felt that a repetition started. They did this by tapping on the desk in sync with the pattern that they experienced. They were told to ignore the beginning of the file as the chain started at a random position, and were asked to wait for six or seven repetitions before deciding. The experimenter manually realized fade-in with a volume button to further obscure the start of the chain. The experimenter scored the note position (1, 2, 3, or 4) that the participant synchronized with. If it

was not clear, e.g. if the participant failed to tap at a regular pace, he/she repeated the trial. The FMI test started with a familiarization phase where all melodies were played both vocoded and non-vocoded, with the tune's name printed on the screen. Participants had the option of replaying them as often as they wanted to, and were explicitly encouraged to do so until they felt they knew them very well. Following this, there was a short practice phase to learn the task. The task involved identifying the melody that was played by choosing from three options shown on the screen (3AFC). The structure of a trial was as follows: fixation cross (on screen 500 ms), playing of the target melody (duration depending on the melody), inter-stimulus interval (500 ms). The time to respond was 11,000 ms, taking into account the longest of the melodies (8 s), but the trial jumped to the next as soon as a response was registered. The three response options were shown on the screen from the onset of the melody, from left to right (on one line of text). The target position was randomized. The experimental phase was divided into three blocks, one with only pitch as a cue (Pitch condition), one with only note durations as a cue (Temporal condition), and one with both F0 and duration as cues (Total condition). Each block consisted of thirty trials where each of the ten melodies served as a target three times, with varying competitors. Blocks alternated with breaks and their order was counterbalanced between participants.

Prosody tests. The prosody tests were 2AFC tasks starting with a practice phase including both vocoded and non-vocoded stimuli. Participants heard a sentence which carried happy or sad prosody (in the ED test) or where the color or the noun (FD test) was focused, and pressed a corresponding button based on options shown on the screen to the left and right. These options were 'sad' and 'happy' (in Dutch; screen position counterbalanced), and the color and the noun (screen position not counterbalanced, to avoid a conflict with the linear position in the sentence) for the two tests, respectively. A picture of the object mentioned in the sentence was also shown to support understanding of the sentence. The trials were made up of a fixation

cross (1,250 ms), the stimulus sound plus time to response (4,000 ms) and an inter-stimulus interval (200 ms). The experimental part consisted of three (ED test) or two (FD test) blocks with pauses in between. The order of conditions (Pitch, Temporal, Total) was counterbalanced across participants and the order of the stimuli was randomized. Vocoded stimuli preceded non-vocoded stimuli to avoid habituation to relatively normal stimuli before hearing the less intelligible stimuli. The FD test included a phonetic cue condition without prosodic resynthesis both for the non-vocoded and vocoded stimuli. The total number of experimental stimuli in the ED test was 12 sentences × 2 emotions × 3 phonetic cues × 2 vocoding conditions = 144 items, and in the FD test 6 sentences × 2 focus positions × 4 phonetic cues × 2 vocoding conditions = 96 items.

5.2.4 Statistics

Statistical analyses were carried out using *SPSS version 21* (IBM Corp, Armonk, NY). Demographic and musical background differences were tested with independent samples *t*-tests and Pearson's χ^2 tests, depending on the type of variable. Separate Repeated Measures (RM) Analyses Of Variance were run for each Training and Test except the AM test, with, where relevant, Group as a between-subjects variable and Vocoding and Cue as within-subjects variables. The AM test results, defined in number of times that each of the four note positions was selected per participant, were subjected to Pearson's χ^2 tests. Post-hoc tests were Bonferroni-corrected.

5.3 Results

Trainings. Responses with a latency of less than 500 ms were considered unreliably fast and were not analyzed (5.6% of data). Further analyses were run on the remaining data. Mean accuracy scores and 95% confidence intervals are shown in Figure 4. Continuous lines with triangles indicate the results of the Pitch group

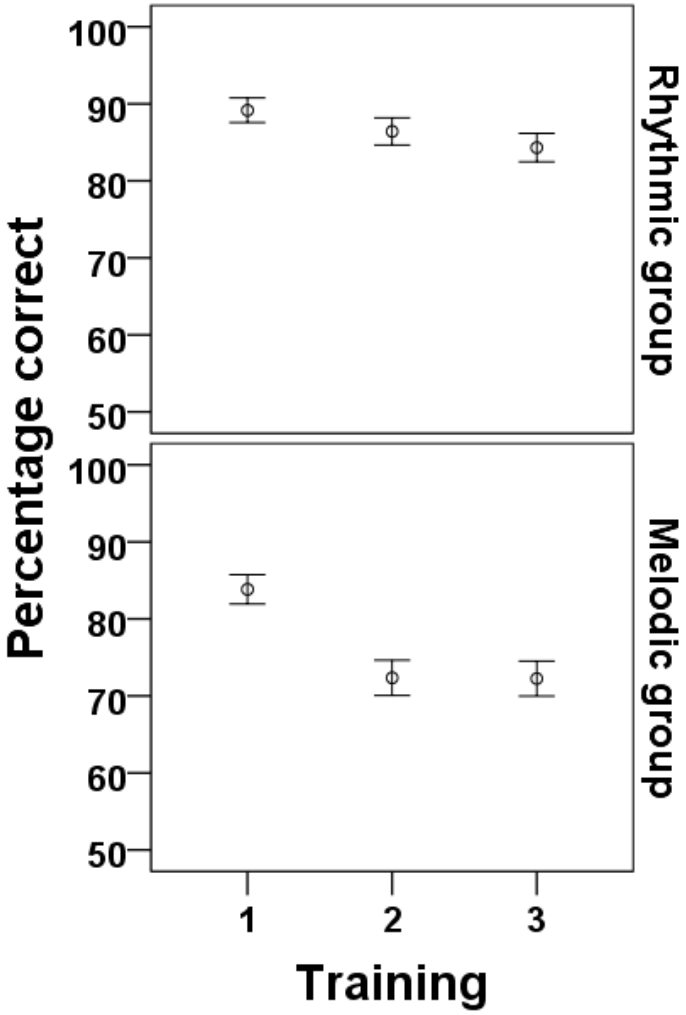

Figure 4. Mean accuracy results of the six Trainings, three for each Group, in percentage correct. Chance level is at 33.3%. Error bars represent 95% confidence intervals.

and dashed lines with circles those of the Temporal group; Errors bars represent 95% confidence intervals (this holds for all figures). The

results of the Trainings will not be analyzed thoroughly, since their results are not intended to answer research questions by themselves but serve only as a possible source of an effect on the Tests. What is relevant is that scores on all Trainings were well between chance and ceiling level, indicating that they were neither too easy nor too difficult. Performance dropped from the first to the last Training in both groups, which ensures that participants remained challenged throughout the training. The overall difficulty for Temporal Trainings (88%, 86%, and 84%, respectively) was higher than that for Pitch Trainings (84%, 73%, 72%).

Musical Tests. In the AM test, we counted the number of responses per possible note position judged as starting notes, per participant. As an example, of the eight contours, a participant might have judged two of them to start on (what was composed as) the first note, one on the second, four on the third, and one on the fourth. We compared the difference in frequency distribution between accent-position (rhythmically marked) responses and, its complement, non-accent (non-rhythmically marked) position responses. This difference was not significant by Pearson's χ^2 ($\chi^2(1) = 1.63$, $p = .20$). The difference in distribution across all four positions, however, was significant ($\chi^2(3) = 12.45$, $p = .006$). The Pitch group more often indicated the two positions straddling the accented one than the Temporal group, whereas the Temporal group indicated more often the accented position and the one two positions away from it. These results suggest that the two groups listened to the contours in different ways, but did not pay attention to rhythmic accents to a different degree. The results do not reveal, however, in what way the listening strategies did differ.

In the FMI test, the data of one participant in each group were unavailable because they used the wrong response buttons. Null responses were not analyzed (1.1% of data of analyzable participants). We ran Repeated Measures (RM) ANOVAs on the remaining data with Cue as a within-subjects factor and Group as a between-subjects

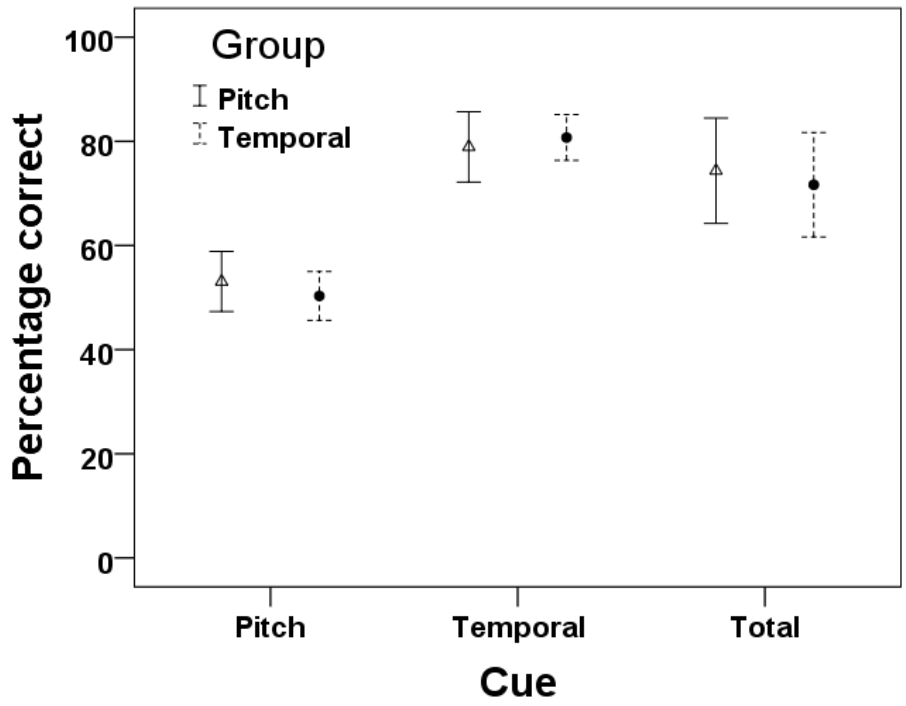

Figure 5. Mean accuracy (percentage correct) and 95% confidence intervals (errors bars) of the Familiar Melody Identification test, split by Cue and by Group. In the Pitch condition, only tone height information was available for identifying melodies. In the Temporal condition, only note duration was available. In the Total condition, both cues were available (i.e., melody and timing were unchanged). Chance level is at 33.3%.

factor. Figure 5 and Table 2 summarize the results in terms of mean accuracies, standard deviations and confidence intervals. Figure 5 shows the scores per Group (line types) and per cue (abscissa). Degrees of freedom were Greenhouse-Geisser corrected to compensate for possible violation of the assumption of sphericity. The effect of Cue was significant ($F(1.44, 69.10) = 39.48$, $p < .001$), but not the effect of Group ($F(1,48) = 0.11$, $p = .74$) nor the interaction between Cue and Group ($F(1.44, 69.10) = .30$, $p = .74$). Bonferroni-corrected post-hoc tests revealed that the effect of Cue was significant for the comparisons Pitch vs. Temporal and Pitch vs. Total (both $p < .001$), but not for Temporal vs. Total ($p = .13$). The results show that

Table 2. Means (and standard deviations) of accuracy (percentage correct) results of the Familiar Melody Identification test. Shown are the values per Group, per Cue, as well as their subtotals and totals. Note that 'Total' refers to the Total condition.

Group	Cue			
	Pitch Mean % (SD)	**Temporal Mean % (SD)**	**Total Mean % (SD)**	**Overall Mean % (SD)**
Pitch	53.1 (14.3)	79.0 (16.4)	74.4 (25.0)	**68.7 (22.0)**
Temporal	50.3 (11.4)	80.7 (10.7)	71.7 (24.3)	**67.6 (20.9)**
Overall	**51.7 (12.9)**	**79.8 (13.7)**	**73.0 (24.5)**	**68.1 (21.4)**

melodies were easier to identify when only temporal information was present (79.8%) than when only pitch information was present (51.7%). Although participants were able to identify melodies solely based on pitch information, as testified by above-chance performance in that condition, the addition of pitch to temporal information (the Total condition) did not aid identification, as the performance in the Total condition (73.0%) was not significantly different from that in the Temporal condition (79.8%).

The indicates that the cost of vocoding is more severe for pitch than for temporal information. When participants recognize the presence of temporal information, that is what they base their responses on, without attending to pitch. The lack of a Group effect indicates that the Trainings were not sufficient to induce a Group differentiation in terms of cue-specific perception competences. Importantly, a trend is nevertheless visible in the expected direction, with the Temporal Group performing worse than the Pitch Group in the Pitch condition, but with the Pitch Group performing worse in the Temporal condition. The Temporal group also performed worse, however, in the Total condition, where we could, in fact, have expected the trends to cancel each other out.

It must be noted that the three options that participants chose from in each trial did not differ only in pitch or temporal (rhythm) information or both, but also in the absolute length in seconds or number of notes, creating a confound in cue availability. However, this confound is not different between manipulated cues since the same stimuli were used in all conditions. Nevertheless, the effect of note duration or number of notes could vary between cues. We therefore investigated the effect of the smallest difference in duration (MDD) and smallest difference in number of notes (MDN) found in the three pairs among the three options per trial on accuracy. In other words, if participants used these latent cues, they would have at least had to detect the smallest difference of two of the response options. We conducted item RM ANOVAs across groups with either MDN or MNN as covariates. In both cases, the pattern of results was identical to that in the original analysis in terms of significance values. The effects of Cue with MDN ($F(2,16) = 5.64$, $p = .035$) and with MNN ($F(2,16) = 4.72$, $p = .05$) as a covariate were still significant, although to a lesser degree. Bonferroni-corrected post-hoc tests showed that with presence of MDN and MNN the comparison between Pitch and Temporal (both $p < .001$) and Pitch and Total (both $p = .001$) were significant but not between Temporal and Total (MDN: $p = .14$; MNN: $p = .21$), as without the confound. We conclude from this discussion that although participants did rely to some extent on differences in total duration and numbers of notes between melodies, that did not significantly change the pattern of effects. Possible training effects were also investigated by computing one-tailed Spearman's *rho* correlations between, on the hand, the per-participant mean percentage correct for all Trainings or the difference in score between the first block of the first Training and the second block of the third Training, and on the other hand, the mean accuracies on the FMI test for the three Cues, for combined and separate Groups. Spearman's *rho* was used because at least one of the variables was not normally distributed according to the Kolmogorov-Smirnov test. The only significant correlations were between Trainings mean and the

Test's Pitch condition for the Temporal Group (ρ = .66, p < .001) and for the combined Groups (ρ = .24, p = .049). In the remaining cases, the lowest p-level in any of the Group by Cue cells was 0.063 and the highest coefficient was 0.315. The correlations with the Trainings mean for the Temporal group, which is most probably also responsible for the combined groups correlation, could, however, reflect either a training effect or an effect inherent to the stimulus type (temporal) because a comparable correlation was not found for the Pitch group.

Prosody tests. In the ED test, 1.2% of the data were not analyzed because they had a null response or a response time faster than 500 ms. An RM ANOVA was conducted with Group (Pitch group, Temporal group) as a between-subjects factor and Vocoding (Vocoded, Non-vocoded) and Cue (Pitch, Temporal, Both) as within-subjects factors. Results are summarized in Figure 6 (error bar graph of accuracy means split by Cue, Group, and Vocoding), Table 3 (accuracy means and standard deviations of cells, subtotals and totals) and Table 4 (RM ANOVA results of main effects, interactions and post-hoc tests). The results show that Vocoding introduces a 17-point drop in overall accuracy (83% for Non-vocoded vs. 67% for Vocoded) in the discrimination of emotions, but this effect is different for the Pitch (97% vs. 65%), Temporal (54% vs. 61%), and the Total (99% vs. 75%) conditions. Thus, for Non-vocoded stimuli, performance was better (near ceiling) in the Pitch than in the Temporal condition (near chance), and as good in the Total as in the Pitch condition. For vocoded stimuli, on the other hand, performance in the Pitch and Total conditions dropped, but more so in the former than in the latter, whereas the Temporal condition improved somewhat. These results together indicate that emotion discrimination is based on the manipulated Pitch (F0) and not on the manipulated Temporal features, and that Vocoding affects only Pitch. Therefore, cue weighting is shifted when stimuli are vocoded, i.e., for non-vocoded stimuli, discrimination is entirely based on Pitch, whereas for vocoded stimuli, reliance shifts more towards Temporal features. Importantly, the near-ceiling scores in the Non-vocoded condition confirm that the emotions

were perceived as intended by the speaker and that the task was feasible. Groups did not perform significantly differently. However, the Temporal group tended towards higher accuracies, and more so in the Vocoded than in the Non-vocoded condition. This is in line with a

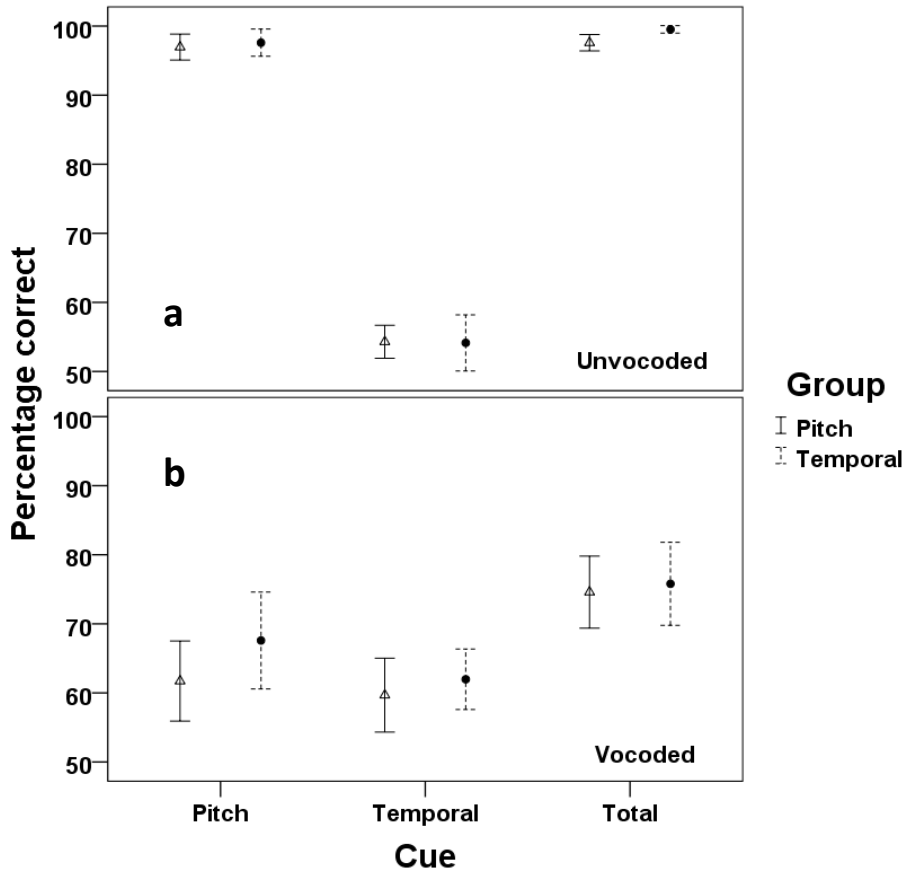

Figure 6. Mean accuracy (percentage correct) and 95% confidence intervals (errors bars) of the Emotion Discrimination test, split by Cue, Group and Vocoding conditions. In the Pitch condition, only tone height (intonation) information was available for identifying melodies. In the Temporal condition, only segment duration information was available. In the Total condition, both cues were available. Chance level is at 50%. (a) Results for the Non-vocoded condition, in which the prosody of the stimuli was resynthesized but where the stimuli were not vocoded. (b) Results for the Vocoded condition, in which the prosody of the stimuli was resynthesized and subsequently sinewave vocoded (see the section Methods for details).

Table 3. Means (and standard deviations) of accuracy (percentage correct) results of the Emotion Discrimination test. Shown are the values per Vocoding Condition, per Group, per Cue, as well as their subtotals and overall values.

		Cue			
		Pitch	**Temporal**	**Total**	**Overall**
Processing	**Group**	**Mean % (SD)**	**Mean % (SD)**	**Mean % (SD)**	**Mean % (SD)**
Unvocoded	**Pitch**	96.96 (4.65)	54.3 (5.9)	97.59 (2.93)	**82.95 (20.9)**
	Temporal	97.60 (4.88)	54.14 (10.06)	99.52 (1.36)	**83.75 (22.04)**
	Both	**97.28 (4.73)**	**54.22 (8.16)**	**98.55 (2.46)**	**83.35 (21.41)**
Vocoded	**Pitch**	61.73 (14.35)	59.67 (13.26)	74.59 (12.93)	**65.33 (14.91)**
	Temporal	67.59 (17.35)	61.97 (10.83)	75.79 (14.87)	**68.45 (15.5)**
	Both	**64.66 (16.04)**	**60.82 (12.04)**	**75.19 (13.81)**	**66.89 (15.24)**
Overall	**Pitch**	79.34 (20.69)	56.99 (10.52)	86.09 (14.87)	**74.14 (20.14)**
	Temporal	82.59 (19.72)	58.06 (11.07)	87.66 (15.9)	**76.1 (20.49)**
	Both	**80.97 (20.18)**	**57.52 (10.76)**	**86.87 (15.34)**	**75.12 (20.3)**

more pronounced reliance on temporal features in the former than in the latter condition. One-tailed Spearman's *rho* computations between per-participant Training means and improvement, on the one hand, and mean ED test scores, on the other, showed that the lowest *p*-level in any of the (combined and separate) Group-by-Cue cells was 0.10 and the highest absolute coefficient was 0.178. We therefore conclude that there was no effect of Training on ED at the individual participant level.

The FD data (0.2% excluded) were analyzed by the same RM ANOVA design as used for the ED test. Results are summarized in

Table 4. RM ANOVA results of the effects of Group, Vocoding, Cue, their interactions, and, if applicable, the pairwise comparisons of percentage correct scores, in the ED test. Post-hoc tests were Bonferroni-corrected. They are not shown for non-significant main effects. Significant results are in bold. The subject df was always 50 and was therefore not specified.

Factor, interaction or comparison	*F*	*Group df*	*p*
Group	1.87	1	0.18
Vocoding	159.93	1	< **.001**[1]
Cue	237.13	2	< **.001**[1]
Pitch vs. Temporal (overall)	193.60	1	< **.001**[2]
Pitch vs Temporal (Unvocoded)	1182.11	1	< **.001**[2]
Pitch vs Temporal (Vocoded)	1.64	1	.62
Pitch vs. Total	42.41	1	< **.001**[2]
Pitch vs Total (Unvocoded)	3.41	1	.21
Pitch vs *Total (Vocoded)	37.58	1	< **.001**[2]
Temporal vs Total	353.61	1	< **.001**[2]
Temporal vs Total (Unvocoded)	1415.24	1	< **.001**[2]
Temporal vs Total (Vocoded)	27.20	1	< **.001**[2]
Group × Vocoding	.79	1	.39
Group × Cue	.32	2	.73
Vocoding × Cue	117,47	2	< **.001**[1]
Pitch vs Temporal	158.82	1	< **.001**[3]
Pitch vs *Total	23.97	1	< **.001**[3]
Temporal vs *Total	109.38	1	< **.001**[3]
Group × Vocoding × Cue	.62	2	.54

[1]Significant at the $p = .05$ level
[2]Significant at the $p = .008$ level. The *p*-threshold was Bonferroni-corrected by 6 and rounded to .005 in order to correct for multiple comparisons.
[3]Significant at the $p = .015$ level. The *p*-threshold was Bonferroni-corrected by 3 and rounded to .015 in order to correct for multiple comparisons.

Figure 7 (as Figure 6), Table 5 (as Table 3) and Table 6 (as Table 4). Vocoding introduces a 3-point drop in overall accuracy (68% for Non-vocoded vs. 65% for Vocoded) in the discrimination of focus, which

effect is stronger for Pitch (72% vs. 62%) than for Total (77% vs. 73%), but in the reverse direction for Temporal (54% vs. 60%). Thus, pitch was most affected and Total remained approximately equal, whereas Temporal was enhanced. The effect was, however, only marginally significant. This is mainly because discrimination was difficult even in the Non-vocoded condition, so that vocoding could not compromise it much further. Performance was significantly different between Pitch and Temporal in the Vocoded but not in the Non-vocoded condition, whereas it was the other way around for Pitch vs. Total. The results suggest that, as in the ED test, temporal information was least useful in the Non-vocoded condition, but was more relied on in the Vocoded condition. Pitch was, however, less informative than in the ED test, but it could be almost entirely compensated for by Temporal information when vocoded. It has to be noted that, as shown by overall scores, the FD test was more difficult than the ED test. Scores on extra conditions (not shown here) with vocoded versus human (i.e. neither resynthesized nor vocoded) stimuli, however, added after a pilot for that purpose, revealed that the focus positions, were identified by the listeners as intended by the speakers – as in the ED test although somewhat lower. The Pitch group had a mean accuracy of 92% for vocoded and 96% correct for non-vocoded stimuli, and the Temporal group had an accuracy 92% and 94% correct, respectively. As in the ED test results, there was no significant effect of or interaction with Group, but there was a trend of an advantage for the Temporal group for the Non-vocoded Pitch and Total conditions. As in the FMI and the ED test, one-tailed participant-level Spearman's *rho* correlations were run between Trainings mean and improvement scores vs. vocoded Cue mean scores (for combined and separated Groups). None of the twelve correlations were significant (p = 0.054 or higher), except for the correlation between Temporal Group's Trainings means and the Test's Total condition (ρ = .46, p = .010), between Pitch Group's Trainings means and the Test's Pitch (ρ = .64, p < .001) and Total (ρ = .50, p = .005) conditions, as well as between combined Groups' Trainings mean and

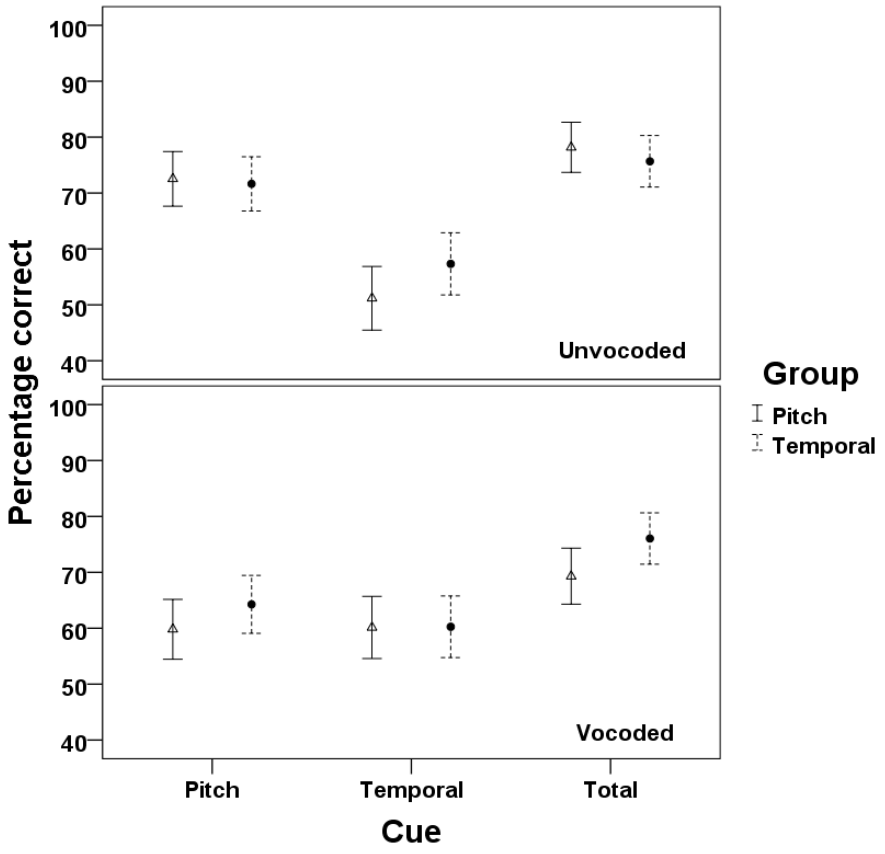

Figure 7. Mean accuracy (percentage correct) and 95% confidence intervals (errors bars) of the Focus Discrimination test, split by Cue, Group and Vocoding conditions. The description is the same as for Figure 6. (a) Results for the Unvocoded condition. (b) Results for the vocoded condition. The description is the same as for Figure 6.

the Test's Total condition (ρ = 0.28, p = .021), suggesting a relationship between basic musical perception and Focus perception, but not necessarily specific to the level of the trained cue.

As a further exploration of effects of musical training on scores in the Tests, analyses were conducted with the cohort split according to, or with Pearson's r correlations based on, personal characteristics reported in the musical background questionnaire (completed by 46

Table 5. Means (and standard deviations) of accuracy results of the Emotion Discrimination test, in percentage correct. See Table 3 for the description.

		Cue			
		Pitch	**Temporal**	**Total**	**Overall**
Processing	**Group**	**Mean % (SD)**	**Mean % (SD)**	**Mean % (SD)**	**Mean % (SD)**
	Pitch	71,60 (19,01)	51,08 (18,94)	77,84 (15,03)	**66,84 (20,96)**
Unvocoded	**Temporal**	71,50 (15,10)	57,35 (12,66)	75,69 (16,06)	**68,18 (16,50)**
	Overall	**71,55 (17,00)**	**54,22 (16,26)**	**76,76 (15,44)**	**67,51 (18,82)**
	Pitch	59,02 (13,53)	60,31 (12,24)	69,45 (14,64)	**62,93 (14,12)**
Vocoded	**Temporal**	64,43 (13,11)	60,23 (16,13)	76,15 (16,37)	**66,94 (16,53)**
	Totals	**61,73 (13,47)**	**60,27 (14,17)**	**72,80 (15,74)**	**64,93 (15,45)**
	Pitch	65,31 (17,53)	55,70 (16,46)	73,65 (15,29)	**64,89 (17,92)**
Overall	**Temporal**	67,97 (14,45)	58,79 (14,43)	75,92 (16,06)	**67,56 (16,47)**
	Overall	**66,64 (16,04)**	**57,24 (15,48)**	**74,78 (15,64)**	**66,22 (17,24)**

participants) which would not create very unequal subgroup sizes. In different analyses, the combined group of participants was divided according to the question if they had received formal practical instrument playing or singing lessons (Yes: $N = 27$, No: $N = 19$) and if they had received theoretical music lessons (Yes: $N = 22$, No: $N = 24$). Correlations were run based on the number of hours of playing/singing per week, number of years having played/sung, and the number of hours per week of listening to music. No significant effects on or interactions with (Pitch/Temporal) Group were found, nor any except very low correlations for any of the tests. Finally, we ran Spearman's *rho* correlations to compare individual scores between

the three Tests. For none of the Group-by-Cue cells (six per test) were correlations significant (maximally ρ = .259, p = .11), except for the correlation between the FMI and FD tests for the Temporal Group in the Temporal condition (ρ = .34, p = .050), between the FMI and ED tests for Temporal Group in the Total condition (ρ = .49, p = .007), between the ED and FD tests for the Pitch Group in the Temporal condition (ρ = .46, p = .010), between the FMI and FD tests for combined Groups in the Temporal condition (ρ = .28, p = .026), and between the FMI and ED tests for the combined Groups in the Total condition (ρ = .30, p = .019). Thus, we see both cross-cue and cross-domain correlations.

Summarizing the results, all Training components were scored on well between chance and ceiling level, but the Temporal Training was easier overall than the Pitch Training. As suggested by the results of the AM test, the Trainings made the groups listen differently to melodies, implying that the Trainings differentiated the Groups. Familiar Melody Identification was performed primarily based on Temporal information and Groups did not differ significantly in this, but did show a trend in the expected cue-specific direction. At the individual participant level, mainly the Temporal Group scores in the Pitch condition increased when the score in the combined Trainings also increased. Emotion discrimination was based on pitch information, which was highly informative, but this was partly compensated for by elevated reliance on temporal information when stimuli were vocoded. For focus discrimination pitch was less informative, but for vocoded stimuli there was more compensation by temporal information such that weighting of pitch and temporal information was balanced. At the individual participant level, there was no advantage of Training performance on ED, but for FD there were advantages for both Training programs for either the corresponding (Pitch Training and FD Pitch) or the non-corresponding (Temporal Training and FD Total) cue. There was also correlation between some tests (FMI, ED, FD) for some of the Group-by-Cue cells, but not bounded by domain (music or language) or cue. No

effects of differences in biographical musical background on any of the Tests were found.

Table 6. RM ANOVA results of the effects of Group, Vocoding, Cue, their interactions, and, if applicable, the pairwise comparisons on percentage correct scores, in the FD test. See Table 4 for further details of the description.

Factor, interaction or comparison	*F*	*Group df*	*p*
Group	1.01	1	0.32
Vocoding	3.14	1	.083
Cue	39.04	2	< **.001**[1]
Pitch vs. Temporal	16.51	1	**.001**[2]
Pitch vs. Temporal (Unvocoded)	32.55	1	< **.001**[2]
Pitch vs. Temporal (Vocoded)	.28	1	1.00
Pitch vs. Total	24.96	1	< **.001**[2]
Pitch vs. Total (Unvocoded)	8.75	1	.014
Pitch vs. Total (Vocoded)	20.59	1	< **.001**[2]
Temporal*Total	80.18	1	< **.001**[2]
Temporal vs. Total (Unvocoded)	61.21	1	< **.001**[2]
Temporal vs. Total (Vocoded)	24.92	1	< **.001**[2]
Group* Vocoding	.84	1	.36
Group*Cue	.021	2	.98
Vocoding*Cue	11.58	2	< **.001**[1]
Pitch*Temporal	20.92	1	< **.001**[3]
Pitch*Total	4.58	1	0.037
Temporal*Total	7.24	1	**0.010**[3]
Group*Vocoding*Cue	2.87	2	.062

[1]Significant at the $p = .05$ level
[2]Significant at the $p = .008$ level. The p-threshold was Bonferroni-corrected by 6 and rounded to .005 in order to correct for multiple comparisons.
[3]Significant at the $p = .015$ level. The p-threshold was Bonferroni-corrected by 3 and rounded to .015 in order to correct for multiple comparisons.

5.4 Discussion

The aim of this study was to explore the role of musical training in the weighting of pitch and temporal cues on music and linguistic (prosodic) perception under conditions (sine wave vocoding)

mimicking those experienced by cochlear implant users. By orthogonally assessing performance with the separate availability of pitch and temporal cues both in musical and linguistic perception, the level(s) at which possible training transfer can take place can be narrowed down. These two levels were referred to as near (the cue level) and far (the domain level, i.e., music vs. language) transfer by Moreno and Bidelman (2014). The most important findings of the current study were that there was some evidence for a positive relationship between short-term cue-specific vocoded (but not long-term) music training and vocoded music and prosody perception, and that emotional and linguistic prosody were perceived with different cue-weightings.

5.4.1 Effect of short-term training

No significant effect of short-term musical training (i.e., training completed as part of the study) was observed on the group level. This is in contrast with earlier findings. CI users in a study by Galvin, Fu, and Nogaki (2007) were trained for half an hour (or three hours, for one participant) per day on Melodic Contour Identification (MCI) for a period ranging between one week and two months, and tested pre- and post-training on MCI and FMI. Improvement was observed with as little as one week of training. In another study, NH participants completed one of three vocoder simulation training programs of fifteen twelve-minute lessons divided over five weeks, differing in the nature of feedback, in which they learned to discriminate instruments (Driscoll, Oleson, Jiang, & Gfeller, 2009). Participants showed better post- than pre-training performance, and the improvement was more pronounced if the training involved more explicit feedback. In a study by Loebach, Pisoni, and Svirsky (2009), two groups of NH participants were trained by transcribing 100 sentences under vocoded (experimental group) or unprocessed (control group) conditions, respectively, and tested before and after training on the same task with 20 (different) sentences. Post-testing also included speaker gender and identity discrimination and environmental sound identification. All

training and testing together took around one hour to complete. Performance on the transcription test significantly increased after training and more so for the experimental than for the control group. Speaker gender and identity perception scores did not differ between groups, but the experimental group outperformed the control group on experimental sound identification. One of the very few studies concerning cross-domain transfer of short-term musical training (Patel, 2014) involved preliminary data of two non-musician CI users who practiced for ten hours spread over one month playing five-note melodies. Before and after training, they were tested on sentence in noise recognition, MCI, and a linguistic prosody test, for which they were asked to discriminate between instances of the word *popcorn* resynthesized with either question or statement intonation. One participant improved in sentence recognition but not in prosody discrimination, and the other participant showed some improvement in prosody discrimination but none in sentence recognition. Despite the inconsistency between participants, the results confirmed the possibility of cross-domain transfer. In another study, however (Yucel et al., 2009), musically trained (2-year study-related keyboard practice) children showed no speech development advantage over non-trained controls in speech processing except for an interactive game, which could also be explained by general developmental factors. Together, the above studies show that short-term musical training under vocoded conditions can improve performance on musical and probably linguistic tasks. What is more, linguistic training of less than an hour can benefit non-linguistic perception, showing very fast cross-domain transfer.

The current report did not clearly confirm the cross-domain transfer as a short-term training effect found in the literature. This discrepancy could be due to a number of factors. First of all, our training session was, with around 45 minutes, very brief. Previous musical training was at least several hours divided over multiple days. The training in Loebach et al. (2009) was very short (less than an hour) but it was linguistic instead of musical. It might be the case that

vocoded musical training requires more time than non-musical training for transfer to different tasks and/or different domains to take place. Second, as a novelty, our training was cue-specific, aimed to improve perception of one aspect of vocoded listening. It could be the case that in vocoded settings, cues cannot be trained in isolation, i.e., without improving vocoded intra- or cross-domain perception in general. The findings by Fuller et al. (2014) that musicians have a greater advantage the more the task requires pitch perception, supports the hypothesis of cue-specific abilities, although to our knowledge rhythmic and pitch training have to date not been systematically compared. We did not include pre-training testing because we hypothesized an interaction between groups and cues, and we cannot determine, therefore, if the training had a cue-specific intra-domain transfer effect. Third, a training does not work if it is too easy or too difficult. The test scores were rather evenly distributed across the entire range with no specific concentration of scores towards either chance or ceiling levels. Therefore we feel safe to say that bottom or ceiling effects cannot explain the absence of cross-domain transfer in our results. Fourth, it is possible that an effect would have been obtained if we had applied more feedback, since Driscoll et al. (2009) found a stronger effect for trainings featuring more explicit feedback. It has to be determined in future work adopting more elaborate training programs which of these explanations is most likely.

Despite the lack of a cue-specific training effect on FMI and prosody tests, the results of the AM test, although intended only as a control test, suggest that the two groups listened in different ways. There was a significant difference in the general distribution of the number of times they perceived the melodies to start on each of the four note positions, but not in terms of the rhythmic versus non-rhythmic positions. This suggests that the different way of listening is not necessarily a matter of just rhythmic versus non-rhythmic attention but it could for instance reflect training-induced enhanced versus repressed attention to pitch or, alternatively, to positions surrounding the accented note. Given that the groups attended

differently to stimuli, this suggests that they did differ in their listening strategy but that cue-specific training resulted in null-effects because they did not differentiate groups to a sufficient degree. It is likely that the hypothesized effects in the FMI and prosody tests were real but required larger power. This is supported by tendencies of group differences and interactions between groups and cue conditions in those tests. In the FMI test, there was a tendency towards enhanced performance in the Pitch cue condition for the Pitch group, but in the Temporal cue condition for the Temporal group. If this reflects a genuine effect, cue-specific training is feasible and is expected to generate larger effects when it is more elaborate. It has to be noted that there was also a tendency towards a lower performance in the Total cue condition for the Temporal group. This is not expected if both cues can be equally relied upon. Apparently, pitch is the more salient or reliable cue and even if participants are not trained on that cue, they rely on it thus failing to benefit from the trained cue (Temporal). In the ED and FD tests, different tendencies were shown. The Temporal group had an advantage over the Pitch group in most conditions, especially in conditions in which Pitch was present (Pitch or Both). This suggests that vocoded prosody perception benefits more from temporal than from pitch training. A possible account for this is to assume that what is important in pitch prosody is fine temporal structure and segmental alignment of the intonation contour, whereas for musical melody perception, there is no temporal variation in the Pitch condition such that there would be no benefit of enhanced temporal processing abilities. Importantly, in the prosody tests, group differences were smaller or different in the Non-vocoded than in the Vocoded condition. This observation suggests that the Vocoded tendencies were not due to inherent group differences in stimulus processing that were already present before training, but were a result induced by the training.

5.4.3 Effect of musicianship

We found no effect of long-term training in the form of playing an instrument or singing (i.e., musicianship), having received theoretical music lessons or a correlation with the number of hours of music listening per week. This result is dissonant with a previous study on the musician effect for stimuli vocoded in a very comparable manner to ours (Fuller et al., 2014), where musicians versus non-musicians were tested on three tasks which the authors interpreted as demanding increasing reliance on pitch information: repetition of words and sentences heard with varying signal-to-noise ratios, identification of emotions with or without normalized amplitude and duration, and Melodic Contour Identification (MCI). Musicians performed as well as non-musicians on speech repetition, slightly better on emotion recognition and much better on MCI. Musicians thus had a greater advantage the more pitch reliance was required, suggesting that the musician effect functions on the relatively low level of the auditory system instead of on a higher cognitive level. Our contrasting result of a lack of a musician effect could be due to a number of reasons. First of all, we did not select for musicianship with stringent criteria, whereas Fuller et al. selected participants who had started musical training before the age of seven and had received it for at least ten years including the last three years regularly. A strict selection of musicians vs. non-musicians might have brought task result differences to light in our study. A second explanation for the discrepancy is the nature of the stimuli of the emotion perception test (the only test that can be compared because it was present in both studies). The stimuli of the emotion test in Fuller et al.'s study comprised four emotions pronounced by four actors, which with all cues available in the non-vocoded conditions were recognized at an average of around 90% correct, whereas we used two emotions from one speaker, which could be discriminated at (near-) ceiling level with pitch alone. Because our task was apparently easier and could also be based on discrimination strategies, this may have obscured any possible sensitivity difference between musicians and non-musicians.

In the emotion recognition task in (Fuller et al., 2014), performance was significantly compromised when amplitude and duration information were removed, for musicians and non-musicians alike. The negative effect of removing (intensity and) temporal information is in line with our finding that the ED test scores in the Total condition were higher than those in the Pitch condition, further strengthening the conclusion that pitch is the most important cue for emotion perception but that temporal information is additive. The removal of temporal information was done differently in the two studies. Fuller et al. removed temporal information by normalizing only the total duration of sentences by linear time compression/expansion, possibly introducing word-internal conflicts between segment durations, whereas we copied individual segment-by-segment durations from an emotional variant onto a neutral variant of the same phrase. Fuller et al. quite probably failed to remove all temporal information, thereby partly obscuring the pitch advantage. Further, the results by Fuller et al. (2014) seem to indicate that long-term musical training does not change cue reliance (for the cues tested) because there was no interaction between musicianship and cue availability. Although this would account for our lack of a significant training effect, it does not preclude that more elaborate training could reveal a cue-specific or cue-general benefit of training one cue versus the other, as suggested by the tendencies found.

5.4.3 Correlations on the level of the individual participant

Although no significant effect of training was found, there were significant correlations between Training and Test performances on the level of individual participants. These correlations do not echo the training effect but do reveal the level at which the discrimination competence functions. For the Pitch group, if a participant had a higher performance in the Training, this was also the case in the FD test, so the competence generalized across domains (music and language). With a weaker correlation, this was also true for the Temporal group between FMI and FD tests. Within domains, cross-

cue generalization occurred for the Temporal group in music (Training and FMI) and, although weaker, for the pitch group in language (ED and FD). Interestingly, though, these relationships were not accompanied by within-cue correlations, meaning that participants shifted instead of broadened their attention. Finally, cross-cue cross-domain correlations existed for the Temporal group between Training and FD and between FMI and ED. These correlations occurred in the Total conditions. Because they were not accompanied by (high) cue-specific correlations, we assume that participants used both cues in the Total condition, therefore counting them as cue-general correlations. We conclude, first of all, that competence can generalize both across cues and across domains, and, second, that temporal perception acuity seems to be more generalizable (across cues and domains separately and concurrently), whereas pitch perception acuity is only generalizable across domains and only to a lesser degree across cues. In a review, Moreno and Bidelman (2014) concluded that musical training can transfer to other skills in various ways, both different auditory skills within and outside music (near vs. far transfer), as well as different perceptual levels, from low-level (other auditory processing) to high-level (outside auditory processing, assuming generalization to a more general cognitive level). In their terminology and assuming that correlations can be equated with transfer, our findings would correspond to (although not equate with) high-level transfer (on the 'processing level' dimension) for cross-cue generalization and far transfer (on the 'transfer' dimension) for cross-domain generalization.

A small number of studies have addressed the question whether perception abilities of certain cues underlie both music and prosody. Wang et al. (2011) observed a strong correlation between CI users' performance on a pitch discrimination task with varying intervals in a melody and a lexical tone identification task, suggesting pitch perception acuity as an underlying ability for the two domains. In a study by See, Driscoll, Gfeller, Kliethermes, and Oleson (2013) on pediatric CI recipients, pitch ranking abilities predicted

performance in direction discrimination of intonational and musical contours. Tao et al. (2014), on the other hand, found no correlation between lexical tone recognition and MCI performance. However, scores on the MCI test were very low, possibly preventing sufficient variation to base correlations on. Recently, Kalathottukaren, Purdy, and Ballard (2015) assessed prosody perception tests from the Profiling Elements of Prosody in Speech Communication (PEPS-C;Peppe & McCann, 2003) vocal affect recognition from the Diagnosis of Nonverbal Accuracy 2 (DANVA 2; Baum & Nowicki, 1998) and the Montreal Battery of Evaluation of Amusia (MBEA; Peretz, Champod, & Hyde, 2003) in twelve CI users. No correlations were revealed between language and music tests. However, the authors attributed this to low power and suggested that pitch perception abilities were at the base of problems with prosody and music perception. The above studies show that focus has been on the frequency (pitch) dimension, but that the temporal dimension has been relatively neglected. Nevertheless, they at least suggest that pitch perception is an important factor linking prosody and music perception in the same listeners. Studies devoted to psychophysical correlates of either domain separately or linking music to segmental speech have shown that temporal perception performance is also a predictor (Chatterjee & Peng, 2008; Luo, Fu, Wei, & Cao, 2008; O'Halpin, 2009). Another study, however, found only pitch but not temporal perception to predict either music, language or the correlation between the two domains (Won, Drennan, Kang, & Rubinstein, 2010). Given the cue-general and domain-general correlations that we found, the present study adds to this literature by supporting views claiming that both pitch and temporal perception abilities underlie both music and prosody perception under vocoded conditions.

5.4.4 Relevance for cochlear implant users

Speech and music together constitute two of the most important types of auditory signals in many people’s lifes. Cochlear implant users

achieve high levels of speech understanding but have much difficulty enjoying music, which is to a large extent due to compromised pitch perception (Looi et al., 2012). Given the findings that musicians and short-term trained people experience an advantage in perception of pitch, music and language in normal and degraded auditory circumstances, post-surgery music training is likely to benefit cochlear implant users' music and speech enjoyment and use, as was concluded in several reviews (Limb & Roy, 2014; Looi et al., 2012; Patel, 2014). Caution is warranted, though, in the generalization of results of simulations to actual CI hearing. CI recipients have a different hearing background, have much more experience with CI input and perceive auditory input altogether in a different way than NH listeners in an experiment. Although the training program in this study was presumably not elaborate enough to have sufficient power to show clear training effects, the results suggests that cross-cue and cross-domain relationships exist. That is, listeners who rely on one cue within a domain can also rely on that cue in the other domain, and alternatively, they can rely on the other cue in the same domain or in the other domain. More particularly, pitch cue reliance is limited to either within-cue cross-domain transfer or cross-cue within-domain transfer, whereas temporal cue reliance can also function cross-cue cross-domain. Training CI users by means of musical exercises therefore has the potential to not only benefit musical experience but also prosody perception. Practising both pitch and temporal cues is likely to have the broadest effect. Further, this research shows that with vocoded hearing, familiar melody recognition is most successful with temporal cues, as pitch cues have been severely affected by vocoding. Emotional prosody discrimination, on the other hand, relies more on pitch and less on temporal cues, the latter of which compensate for the loss of the former by vocoding. Focus prosody discrimination, finally, relies less on pitch and more on temporal cues than emotional prosody. This implies that CI users weight cues differently (more or less reliance on temporal cues) than NH listeners and the weighting varies per type of signal.

Conclusions

This study investigated the possible transfer effect of musical training of pitch versus temporal cues on the same (pitch to pitch or temporal to temporal) and the other (pitch to temporal or vice versa) cue, as well as within the same domain (music) and another domain (prosody). This research used a compact training program, but the tendencies reflecting the hypothesized interaction between training group and cue availability, as well as a difference in listening strategy shown by the Ambiguous Melody test are promising in the sense that a more extended training is likely to have a larger effect. It must be noted that we did not include a pre-training baseline test because we hypothesized an interaction between training group and performance with selective availability of the respective cues, but inclusion of such a test would yield valuable extra information about possible cue-general improvement differences between groups. The primary findings were the following.

1) Musical cue-specific pitch and temporal cue training with vocoded stimuli as short as 45 minutes showed tendencies towards corresponding cue reliance in familiar melody recognition and towards an advantage for temporal training for prosody perception. More elaborate training has the potential to show larger effects.
2) There was no relationship between years of practical or theoretical training or weekly hours of music listening and performance on familiar melody recognition, emotional or linguistic prosody perception
3) Listeners relied almost entirely on temporal cues for familiar melody recognition, more on pitch than on temporal cues for emotion discrimination and approximately to an equal degree on the two cues for focus discrimination. Vocoding maked reliance shift more towards temporal cues.

4) There were within-cue cross-domain (i.e., far transfer between music and prosody) and within-domain cross-cue (i.e., high-level transfer between pitch and temporal cues) correlations for pitch perception, and cross-cue cross-domain correlations for temporal cue perception.

Chapter **6**

Prosody perception and production by children with cochlear implants

Abstract

Cochlear implant (CI) users have been reported to have difficulty perceiving and producing prosody. In this study the perception and production of emotional and linguistic (focus) prosody were compared in children with CIs and normally hearing (NH) peers.

Thirteen CI and Thirteen hearing-age (HA) matched NH children (HAs between 3;8 and 9;5) performed, as baseline tests, non-verbal emotion understanding tests (for general emotional development), a non-word-word repetition test (for general linguistic development) and stimulus identification and naming tests (for basic task understanding). Main tests were verbal emotion (happy, sad) discrimination, verbal focus position (color or noun) discrimination in simple color + noun sentences, acted emotion production and focus production (elicitation of corrective focus). Accuracy scores were compared across groups and correlations between tests were computed. Emotion and focus productions were evaluated by a group of 10 adult Dutch listeners with normal hearing.

The focus perception test could not be analyzed. Scores for the two groups were comparable for all tests, except a lower score for the CI group in the Non-word repetition test. On the individual participant level, emotional prosody perception and production scores were weakly and moderately significantly correlated for CI children but uncorrelated for NH children. In both groups, emotion production, but not emotion perception, was weakly predicted by hearing age. Non-verbal emotion (but not linguistic) prosody understanding performance, predicted CI children's emotion perception and production scores, but not the controls'.

Given the comparable overall scores, CI children catch-up with their peers no later than towards the end of primary school. Increasing time in sound facilitates vocal emotional expression, which possibly requires independently maturing emotion perception skills. CI and NH children apply the same cue-weighting strategies for emotion perception, relying almost exclusively on F0 information.

6.1 Introduction

Children with cochlear implants (CI) experience delays or deviations in their oral (productive and perceptual) linguistic and socio-emotional development relative to normally (NH) hearing peers. This is, first of all, because the onset of their oral language acquisition process is delayed until the moment of implantation (usually at least at one year of age). Second, due to the fact that the quality of the linguistic input that can be received after implantation is degraded compared to what NH peers can perceive, a full appreciation of phonetic nuances important for linguistic and paralinguistic information is hindered. For instance, CI users have been found to have problems with identifying vowels (Dorman & Loizou, 1998; Garrapa, 2014; Valimaa, Maatta, Lopponen & Sorri, 2002; Välimaa, Sorri, Laitakari, Sivonen & Muhli, 2011; however, see Iverson, Smith & Evans, 2006), distinguishing questions from statements (Meister, Landwehr, Pyschny, Walger & von Wedel, 2009; Peng, Lu & Chatterjee, 2009; Straatman, Rietveld, Beijen, Mylanus & Mens, 2010), understanding speech in noise (Gfeller, Turner, Oleson, Zhang, Gantz, Froman & Olszewski, 2007; Neuman, 2014), identifying emotions in speech (Geers, Davidson, Uchanski & Nicholas, 2013; Luo, Fu & Galvin, 2007) and discriminating speaker gender and identity (Fu, Chinchilla, Nogaki & Galvin, 2005; Fuller, Gaudrain, Clarke, Galvin, Fu, Free & Baskent, 2014; however, see Meister et al., 2009). Problems with the production of speech have also been observed, including voice quality (Ubrig, Goffi-Gomez, Weber, Menezes, Nemr, Tsuji & Tsuji, 2011), articulation (Van Lierde, Vinck, Baudonck, De Vel & Dhooge, 2005), lexical tone production (Han, Zhou, Li, Chen, Zhao & Xu, 2007), emotion imitation (Nakata, Trehub & Kanda, 2012; Wang, Trehub, Volkova & van Lieshout, 2013), intelligibility (Chin, Tsai & Gao, 2003), and the quality, content and efficiency of retold stories (Boons, De Raeve, Langereis, Peeraer, Wouters & van Wieringen, 2013). However, vocal characteristics within the norm have also been reported (Souza, Bevilacqua, Brasolotto & Coelho, 2012).

According to one series of studies testing 181 implanted children, speech perception and production performance have been shown to explain 42% of overall total language scores and as much as 63% when split for overall spoken language scores and (Geers, Nicholas & Sedey, 2003), showing the importance of speech perception and production for children's linguistic development. Furthermore, problems in those areas have been associated with delays in socio-emotional development. Wiefferink, Rieffe, Ketelaar, De Raeve and Frijns (2013) tested Dutch CI and NH two-and-a-halve- to five-year-old children on facial and situational emotion understanding and general expressive and receptive language development. For the recipients, performance on all tests was poorer than for the control group and showed positive correlations between language and emotion tests that require verbal processing. These results showed that CI children experience delays in verbal as well as non-verbal emotion understanding and that linguistic development can predict aspects of emotional development. Mancini, Giallini, Prosperini, D'Alessandro H, Guerzoni, Murri, Cuda, Ruoppolo, De Vincentiis and Nicastri (2016), however, found that 79% of their cohort of 72 CI children, aged 4 to 11 years, showed normal emotion understanding skills. The differences with Wiefferink et al.'s results were attributed to discrepancies between the participant groups: Mancini et al.'s cohort had a wider age range and a larger percentage of children with an exclusively oral language use. It might be the case that CI children catch up for their delay in emotional development when they are at school age. Nevertheless, similarly to Wiefferink et al. (2013), Mancini et al. (2016) also reported a link between emotional and linguistic development of CI children.

One area of speech that has been relatively little studied in the research on the linguistic development of CI children and adults with CIs is prosody. Prosody is defined as the speech information which cannot be reduced to the individual segments (consonants and vowels) or their juxtaposition (Rietveld & Van Heuven, 2016). It is an essential component of speech because it conveys both message-

related (meaning) and speaker-related (emotion and attitude) information. These types are referred to as linguistic and emotional prosody, respectively. For a number of possible reasons, linguistic and emotional prosody may develop differently in a language learner. First of all, their neurolinguistic processing is most likely partly lateralized, with emotional prosody being associated mostly with the right hemisphere and linguistic prosody with both hemispheres (Witteman, van Ijzendoorn, van de Velde, van Heuven & Schiller, 2011); second, they are phonetically different (linguistic information is discrete whereas emotional information is gradient); and third, the production of linguistic prosody plausibly requires knowledge of linguistic rules whereas that of emotional prosody, being more intuitive, might not, and might thus depend less on perception.

Whereas comprehension of sentences by pediatric and adult CI users has been found to be relatively intact (e.g., Helms, Müller, Schön, Moser, Arnold, Janssen, Ramsden, Von Ilberg, Kiefer & Pfennigdorf, 1997), several aspects of the perception and production of linguistic and emotional prosody have proven more problematic. As for the perception of linguistic prosody, Meister, Tepeli, Wagner, Hess, Walger, von Wedel and Lang-Roth (2007) reported poorer performance for adult CI users than for NH controls on the identification of word and sentence accent position and sentence type (question vs. statement), but not on discrimination of durational minimal pairs of words, and sentential phrasing with any available cue (e.g., *Die Oma schaukelt das Mädchen nicht.* vs. *Die Oma schaukelt. Das Mädchen nicht.*, lit. 'Grandma swings the girl not' vs. 'Grandma swings. The girl not.'). Children with CIs were outperformed by peers with hearing aids (HA) in the discrimination of questions vs. statements and lexical stress position on bisyllables, but groups performed equally on the identification of words' syllable number and sentence stress (narrow focus) position (Most & Peled, 2007). O'Halpin (2009) found lower performance for school-going children than for NH peers for phrasal discrimination (*blue bottle* vs. *bluebottle*) and identification of two-way (*It's a BLUE book* vs. *It's a*

blue BOOK, where capitals demark accent) and three-way sentence accent position (*The BOY is painting a boat* vs. *The boy is PAINTING a boat* vs. *The boy is painting a BOAT*). Combined, these studies suggest that CI users have difficulty perceiving some but not all aspects of linguistic prosody, with a notable disadvantage for the identification of the position of accents on syllables and words (for evidence for similar difficulties by NH adults, see Schiller, 2006).

As for emotional prosody perception, Volkova, Trehub, Schellenberg, Papsin and Gordon (2013) found that five- to seven-year old implanted children discriminated happy and sad utterances with a score above chance but less accurately than NH peers. Children with CIs aged between seven and thirteen years in Hopyan-Misakyan, Gordon, Dennis and Papsin (2009) performed worse than NH peers when identifying the emotion (happy, angry, sad, fearful) of emotionally pronounced variants of semantically neutral sentences but the two groups performed equally on affective facial recognition, showing that difficulties with vocal emotion recognition could not be explained by more general delays in emotion understanding. In a study by Luo et al. (2007), adult recipients' scores were poorer than those of a NH control group when identifying the emotion (happy, angry, sad, fearful or neutral) of sentences. These studies show that CI recipients of various ages have difficulty identifying emotions in speech.

The main phonetic dimensions by which prosodic information is conveyed – dynamic, temporal and intonational (F0, fundamental frequency) variation – have been investigated to explain the mechanism behind CI users' prosody perception capabilities. Meister, Landwehr, Pyschny, Wagner and Walger (2011) measured difference limens (DL) and incrementally manipulated the F0, intensity and duration of accented syllables. They found that CI users had difficulty when F0 and intensity cues were made available but not when duration was made available, indicating that duration was more reliable for them than the other cues. These results were consistent with the findings that DLs for duration were comparable between

groups (51 ms for CI vs. 40 ms for NH) but worse for the recipients for F0 (5.8 vs. 1.5 semitones) and for intensity (3.9 dB vs. 1.8 dB). The CI children in O'Halpin (2009) showed larger DLs than the control group in detection of F0 manipulated *baba* bisyllables but less so for intensity and duration. The variation in their performance was however large, with some participants showing smaller DLs than the smallest of the control group for intensity and duration. DLs per cue correlated with performance on the perception of phrasal accents reviewed above, which suggests that the children apply their successful psychophysical capabilities for prosodic perception. Taken together, it can be concluded from this research that CI users have problems discriminating variation in the intonational domain, but less in the dynamic and probably even less in the temporal domain and that this has repercussions for the type of prosodic information that they adequately receive.

A small number of studies have addressed the issue of prosody production by CI users. Lyxell, Wass, Sahlen, Samuelsson, Asker-Arnason, Ibertsson, Maki-Torkko, Larsby and Hallgren (2009) observed poorer performance for school-going CI children than for NH peers on the perception and production of word and phrase level prosody, but did not fully specify the task and phonetic analysis of the recorded data. Japanese children with CIs aged 5 to 13 years produced less appropriate imitations of disappointed and surprised utterances than a NH control group and their performance pattern was correlated to their impaired identification of emotions (i.e., happy, sad or angry) in semantically neutral sentences (Nakata et al., 2012). A below-normal performance but no correlation was found for six- to ten-year-old recipients between the Beginner's Intelligibility Test, a sentence imitation test for CI users (Osberger, 1994), and the Prosodic Utterance Production test, an imitation test for sentences with happy, sad, interrogative and declarative moods (Bergeson & Chin, 2008). Phonetic differences between CI relative to NH children's productions were found such as inadequate speech rate (longer utterances, longer pauses and schwas, more breath groups), inappropriate stress

production and vocal resonance quality, a smaller F0 range and a shallower F0 declination, i.e., the natural downward F0 slope over an utterance (Clark, 2007; Lenden & Flipsen, 2007). Relative to NH peers, declarations and question produced by implanted children and young adults were less accurately identified as such (74% vs. 97%) and rated as less appropriate (3.1 vs. 4.5 on a scale from 1 to 5) by NH raters (Peng, Tomblin & Turner, 2008). In her study on school-going recipients, O'Halpin (2009) reported no correlation between most of the perception scores and production appropriateness of narrow focus position. The CI children in Holt (2013) produced phrasal emphasis (focus) sometimes with different accent types in terms of the autosegmental framework (Gussenhoven, 2004; Pierrehumbert, 1980) and with different syllabic alignments and temporal phrasing. In as far as they were able to produce the accents correctly, however, they did this without being able to discriminate between the accent types according to perception experiments, suggesting that accurate perception is not a prerequisite for reasonable production. In conclusion, as for perception, the production of both linguistic and emotional prosody by CI users of different ages deviates from the NH norm in several aspects. There is, however, mixed evidence regarding the question if good perception skills are required for good production skills.

The current research aimed at filling in this gap by testing the perception and production of linguistic and emotional prosody in the same group of implanted children and compare them to a control group of NH peers. The processing of linguistic and emotional prosody by implanted children has never been clearly contrasted. This line of research needs to be undertaken because the perceptional capabilities of CI children may have different repercussions for both the perception and production of the two types of prosody. Whereas the perception of both types may be affected by the degraded input (be it in a different manner or to a different degree), the production of emotional prosody is expected to be less affected than that of linguistic prosody due to its relatively intuitive, less rule-based nature.

In order to control for a number of known possible confounds, information about general linguistic level, emotion understanding and the family's socio-economic status was also gathered. We tested the following predictions.

> (A1) Prosody perception and production scores within participants are correlated. Such an effect would suggest that reasonable production skills require reasonable perception skills for a comparable task. (A2) That effect is larger within than across the prosody type (linguistic vs. emotional) and (A3) larger for linguistic than for emotional prosody because emotional production, due its supposedly relatively intuitive and less rule-based nature, is expected to be less dependent on perception skills.
>
> (B1) Scores per prosody type (linguistic or emotional) are influenced by their respective general scores for linguistic and emotional capacities, (B2) but this effect is larger for linguistic than for emotional prosody.
>
> (C1) Assuming a possible effect of more general maturation on linguistic, including prosodic, skills (hypothesis B), CI activation age negatively correlates with prosody processing capacities, but (C2) this effect is larger for linguistic than for emotional prosody.
>
> (D1) For the perception of prosody, CI participants rely more heavily on temporal cues as opposed to F0 cues than NH participants do. For NH participants, this reliance would be more equal between cues or the other way around. (D2) We expect that this effect is stronger for linguistic than for emotional prosody.

In summary, we investigated if scores on perception and production of prosody were related to each other per participant and if this relationship differed between linguistic and emotional prosody. We also studied to what extent these scores were related to more general

linguistic and emotional capacities, and if CI users used different cues for prosody perception and production than the NH control group.

6.2 Methods

All children were tested on emotion perception, focus perception, emotion production and focus production, the order of which was randomized across participants. This block of four main tests was preceded by a familiarization phase, in which participants were acquainted with the names of the stimuli (colors and objects). Additionally, there were four baseline tests with the purpose of assessing the levels of possibly confounding competences: non-verbal emotion understanding, stimulus identification and naming, and non-word repetition, the first three of which took place before the main tests and the last of which after them, if the child's concentration allowed. The non-verbal emotional understanding comprised two tests from a battery designed to assess social-emotional development in normally hearing and children with special (linguistic) developmental or language backgrounds such as those with cochlear implants (Wiefferink, de Vries & Ketelaar, 2015). This test was included to ensure that all participants had a basic understanding of emotions, tested without the requirement of good verbal expression. All other tests were developed by the authors for the current research. The stimulus identification and naming tests were used as a baseline assessment of the capability to understand and name the stimuli to be used in the main tests. The non-word repetition test was included as a proxy for general linguistic capacities, which might or might not correlate with scores on tests gauging prosody processing capacities. The parents or caretakers were asked to complete a questionnaire about their socio-economic status (SES) and the child's linguistic and medical background. The study was approved by the Leiden University Medical Center's (LUMC) medical ethical committee (NL46040.058.13).

It should be noted at this point that, due to a technical error, no data for the focus perception test had been collected. The description of the methodology will therefore focus on the other tests.

6.2.1 Participants

Thirteen implanted children and thirteen children with normal hearing (NH) participated in this study. They were matched on gender (in both groups eleven boys) and hearing age, defined as the time since the onset of stable hearing, which is implant activation date for recipients and the date of birth for controls. The CI group's mean hearing age was 6;10 (years;months) (ranging between 3;8 and 9;5 and with an SD of 1;9) and the NH group's mean hearing age was 6;9 (range: 4;5-9;4; SD: 1;6). The CI group's mean chronological age was 9;1 (range: 6;1-12;3; SD: 2;0) and that of the NH group was by definition identical to its hearing age. Chronological age is defined as the time since birth. We used the following inclusion criteria for participants (both CI and NH unless not applicable): at least three years gross of CI experience, unilateral implantation, no reported medical problems related to the CI, Dutch as the only first language, no attested psychosocial and (only NH) audiological or speech problems. NH children were not subjected to audiological testing since their hearing was supposed to be better than that of the CI children to begin with. Participant characteristics are shown in Table 1.

6.2.2 Stimuli

Speech stimuli for all tests were recorded as natural utterances in an anechoic booth with a sampling rate of 44.100 Hz and a sampling depth of 16 bit and were pronounced by a child language acquisition expert (CL). She was asked to pronounce stimuli at a regular pace and with specific prosody such that, where applicable, emotions and focused words would be clear for young children.

In the emotion perception test, all trials were based on six object names and six color names in Dutch: 'auto' (car), 'bal' (ball), 'ballon' (balloon), 'bloem' (flower), 'schoen' (shoe), 'stoel' (chair),

Table 2. Demographic and implant characteristics of CI recipients. Hearing age refers to the time since implantation. 'AB' is the Advanced Bionics HiRes 90k HiFocus 1j implant; 'Nucleus' is the Nucleus Freedom Contour Advance implant. Abbreviations: x;y – years;months; mos.: months.

Subject number (gender)	Chro-nolog-ical age	Estimated age at hearing loss onset	Estimated duration of deafness (months)	Age at first CI active-tion	Hea-ring age (mos.)	Etiology	Im-plan-ted ear(s)	Current implant type	Current speech proces-ssor
1 (M)	10;1	0;0	11	0;11	109	congenital, hereditary	Both	AB	Neptune
2 (M)	8;0	0;0	15	1;3	80	unknown (sudden)	Both	AB	Neptune
3 (M)	11;10	unknown	unknown	8;1	44	unknown	Right	AB	Neptune
4 (F)	8;2	0;0	13	1;1	84	congenital	Right	AB	Neptune
5 (M)	12;3	unknown	unknown	4;10	88	unknown	Left	AB	Neptune
6 (M)	10;7	0;3	9	1;2	113	unknown (sudden)	Both	AB	Neptune
7 (M)	10;8	unknown	unknown	5;1	67	unknown	Left	Nucleus	Cochlear CP810
8 (F)	6;6	0;0	21	1;9	57	Chudley Mc-Cullough	Left	AB	Neptune
9 (M)	8;1	0;0	14	1;2	83	congenital	Both	AB	Neptune
10 (M)	10;10	0;0	21	1;9	109	congenital	Both	AB	Neptune
11 (M)	6;1	0;0	11	0;11	61	congenital	Both	AB	Neptune
12 (M)	8;1	0;0	14	1;2	83	congenital, hereditary	Both	AB	Neptune
13 (M)	7;2	0;0	12	1;0	73	congenital	Both	AB	Neptune

'blauw' (blue), 'geel' (yellow), 'groen' (geen), and 'rood' (red). These words were chosen on the basis of a number of criteria: (1) they consisted mainly of voiced segments such that the intonation pattern would be least interrupted; (2) they were supposedly not semantically biased towards any emotion; (3) they had no inherent color bias, to avoid anomalies such as green bananas and blue trees; (4) nouns had common neuter, so they had the same article and adjectival declination; and (5) the nouns were known by at least 86% of children aged 2;3 years as tested by a questionnaire with 961 (pairs of) parents and listed in the Lexilijst (Schlichting & Lutje Spelberg, 2002). According to that questionnaire, the colors were known by between 47% and 63% of children of that age. However, they were the four most frequent colors known by young children, our participants had a higher hearing age than 2;3 years and they were familiarized with the

stimuli before the test phase. Words ending in voiceless segments were dispreferred because they interrupt the intonation contour but in our choice of stimuli priority was given to the criteria of familiarity and natural color-neutralness. Therefore, some voiceless segments are present in the list. Auditory stimuli had normalized amplitudes by scaling to peak (0.99). All stimuli were prerecorded because we wanted to prevent inter-token variation in the stimuli. They were presented in auditory-only modality to prevent clues from lip-reading, for which the experimental group might have an advantage.

In the Emotion perception test, all 24 combinations of the six objects and four colors were produced in a happy and a sad variant. The phrases followed the template 'een' [color] [N], where 'een' is the singular indefinite article. They were between 1.38 and 1.93 seconds long, with an average duration of 1.72 seconds for happy and 1.62 for sad phrases. It has been reported elsewhere (van de Velde, Schiller, van Heuven, van Ginkel, Briaire, Beers & Frijns, forthcoming) that the emotions, taken into account possible response biases, could be discriminated at near-ceiling level in the unprocessed condition by NH listeners, ensuring that the intended emotions and focus positions were successfully conveyed.

Sentences were all manipulated into three extra variants by cross-splicing aspects of the prosody from the non-neutral stimuli to the same neutral equivalents (the Cue condition): (1) only the F0 contour (F0 condition); (2) only the durations of the allophones (Duration condition); and (3) both the F0 contour and the allophone durations (Both condition). This was done in order to control the cues available to the participants. Because unique neutral variants (i.e., one single variant for the two emotions) constituted the bases of the stimuli, judgements by participants could only be based on F0, allophone durations, or both, respectively. Except for these cues, the two emotions were identical, since the underlying segmental material was identical for both emotion variants of a given phrase.

In all relevant tests, response options were represented with additional images. Pictures recurring in different tests were those

depicting the auditory noun and color stimuli. They were based on the database of the Max Planck Institute in Nijmegen and were controlled for the number of pixels, name agreement, picture familiarity and age of acquisition for five- to six-year-old children (Cycowicz, Friedman, Rothstein & Snodgrass, 1997). These original line drawings were filled with basic colors using Microsoft Paint in order to be able to contrast colored objects with each other. All children were familiarized with the visual stimuli before testing by showing all color and object pictures as well as their combinations one by one and in groups, and inviting them to name them, the researcher correcting and asking to repeat whenever necessary. Pictures were controlled for the total number of pixels per picture.

In baseline test 1, Non-verbal emotion understanding, the stimuli and procedure in this test were developed by Wiefferink et al. (2015), to which we refer for details about stimuli. In the baseline tests 2 and 3, Stimulus identification and naming, the stimuli consisted of the auditory and visual materials that were also used in the four main tests, i.e., (subsets of) the 24 color/object combinations. The auditory stimuli were always the identical tokens of the same phrase and the visual materials the exact same pictures. In the emotion perception and production tests, there were, additionally, simple line drawings of a happy and a sad face.

In baseline test 4, the Non-word repetition test, stimuli consisted of nonsense words in a carrier phrase presented as the supposed words for phantasy toys of which colored photos accompanied the auditory stimuli. These photos were taken from a database developed for non-word repetition tests, designed to avoid associations with known objects or with emotions, particularly by children (Horst & Hout, 2015). The nonsense words were four stimuli of each word length from one to five syllables. They were based on De Bree, Rispens and Gerrits (2007), but adapted for children with a linguistic age of 3;0 years. The criteria for the phonological composition of the nonsense words, based on Dollaghan and Campbell (1998), were as follows: (1) they began and ended with

consonants (Cs); (2) they contained no consonant clusters; (3) to ensure that non-word repetition would not be affected by a participant's vocabulary knowledge, non-words were constructed such that none of their individual syllables (CV or CVC) corresponded to a Dutch word; (4) they only contained phonemes that even atypically developing children with a chronological age of 2;8 years have acquired according to Beers (1995), and excluding the 'late eight' (i.e.consonants that are acquired late; Shriberg & Kwiatkowski, 1994), except for /s/ (which would have left too few possibilities to work with); (5) they contained only tense vowels, as these are perceptually more salient and less likely to being reduced to schwa than lax vowels; (6) to limit syllabic positional predictability, consonants, except /s/, occupied only positions in which they occurred less than 32% of their occurrences (Van Oostendorp, personal communication); (7) for independent recall of all consonants, they appeared only once in a word. Practice stimuli were different from the experimental stimuli. The carrier phrase of all non-words was the exact same token of 'Kijk! Een [word], een [word]. Kan jij dat zeggen?' ('Look! A [word], a [word], can you say that?'). The target words were spliced into the indicated slots. The complete lists of non-words can be found in Appendix C.

6.2.3 Procedure

Testing took place in the children's homes, at the Leiden University phonetics laboratory or at the Leiden University Medical Center, depending on the parents' preference. Testing was divided over multiple sessions if time and concentration limits requested so. Combined visits had a duration of between one and two and a half hours. Testing started with the Non-verbal emotion understanding test and was followed by color/object identification and naming to familiarize the children with the stimuli and the paradigms at hand. Subsequently, we administered the four main tests, emotion and focus perception and production, in a counterbalanced order across participants. Finally, depending on time and motivation of the

children, non-word repetition was tested. All tests except the non-verbal emotion understanding and stimulus identification and naming were preceded by practice stimuli that could be repeated if deemed necessary by the experimenter. All but the Non-verbal emotion understanding test were performed on a touchscreen computer. If the child pointed without touching, the experimenter selected the intended option for the child. There was no time limit for trials in any of the tests. The experimenter globally supervised the procedure throughout by explaining the tests and continuing to a next trial whenever this was not automatic. In all computer tests, the experimental part was preceded by a practice phase of between two and four trials, repeated maximally once when the experimenter thought the child did not understand the task well enough. In the practice phase, responses prompted feedback in the form of a happy or a doubtful smiley, all in greyscale to prevent biases towards any experimental color.

All tests except the Non-verbal emotion understanding test were run on a Lenovo 15 inch touchscreen laptop with the keyboard flipped backwards so children could easily reach the screen. Stimuli were played through a single Behringer MS16 speaker placed centrally over the screen. The distance from the speaker to the tip of the child's nose was set at 61.5 cm at zero degrees azimuth at the start of testing. Hardware settings were adapted for every participant to calibrate the sound level at 65 dBSPL at the ear using a Trotec BS 06 sound meter. This portable meter was calibrated to a high-quality A-weighted sound level meter on the basis of a one-minute steady stretch of noise with the same spectrum as that a large portion of the combined stimuli (thus from the same speaker) of the experiments. Note that the usage of headphones was not an option as they would interfere with children's implants. Presentation of auditory stimuli was mediated by a Roland UA 55 external sound card. In the prosody production and Non-word repetition tests, speech was recorded using a Sennheiser PC 131 microphone as input to a Cakewalk UA-1G USB audio interface. All computer tests were run with *E-Prime 2.0 Professional* (Psychology Software Tools, Pittsburgh, PA, USA;

Schneider, Eschman, & Zuccolotto, 2012) and *Powerpoint 2010* on a *Windows 8.1* operating system.

Baseline test 1, Non-verbal emotion understanding. This test consisted of the subtests Face discrimination, Face identification and Expression. The first involved sorting four series of eight line drawings into one of two categories: cars or faces, faces with or without glasses, faces with a negative (angry, sad) or positive (happy) emotion, and sad or angry faces, respectively. In the first and third series only, the first two trials were done by the experimenter as an example. In the second subtest, divided over two pages, there were two instances of line drawings of faces for each of the emotions happy (twice on one page), sad, angry fearful (twice on the other page). The child was asked to indicate consecutively which face showed each of these emotions, and, for each emotion, if another face showed that as well. In these two subtests, numbers of correct responses were recorded. In the third subtest, the child was presented with eight line drawings of emotion evoking situations (two of each of the emotions happy, sad, angry and scared) and was asked to tell how the protagonist, always shown from behind the head to avoid cues from the facial expression, felt, to match one of four emotional faces to it and to tell why the protagonist felt that way. In case he or she did not respond, each question was repeated once. The verbal and drawn emotion chosen were recorded as well as the verbatim response.

Baseline tests 2 and 3, Stimulus identification and naming. In the first of these two tests, stimulus identification, the child consecutively identified each of all of the 24 auditory object/color combinations by selecting a picture on screen. The target position was counterbalanced, as were the position and type of the distractors (only different color, only different object, both different). There was no time limit. Performance was calculated as percentage correct. Also, to prevent unnecessary proliferation of the number of trials, only six of the possible fifteen object contrasts were used, namely car-flower, ball-shoe, balloon-chair, flower-ball, shoe-car, chair-balloon (the first one being the target). These pairs were both conceptually and (in

Dutch) phonologically well distinctive. All objects in this shortlist functioned exactly once as a target and once as an object distractor. To make the task easy and to circumvent red-green color blindness, only two color contrasts were used, namely blue-red and green-yellow (twelve times each). In the second test, stimulus naming, subsequently, the same stimuli as in the identification test appeared as pictures on screen and the child was asked to name them as a color/object noun phrase (e.g., *Een rode bal*, 'A red ball') using the vocabulary from the identification test and trained for in the familiarization test. Responses were recorded as audio files and scored as accurate or inaccurate (wrong, unclear or no response), neglecting the presence or choice of a determiner.

Baseline test 4, Non-word repetition. This test consisted of twenty trials in series of four for each of the lengths from one to five syllables (four times five), consecutively. Children were asked to repeat the word they heard once. Responses were recorded to be scored later. Pictures and auditory stimuli for a trial were presented simultaneously. The picture remained visible until the next trial started.

Main test 1, Emotion perception. In this test, participants heard a phrase pronounced in either a happy or a sad manner. They were asked to indicate which emotion was conveyed by touching or pointing at the corresponding picture of an emotional face on the screen. There were three counterbalanced blocks of 24 randomized trials separated by breaks, differing in Cue and each preceded by two warm-up trials. A trial consisted of a fixation animation (1,250 ms), the stimulus presentation (indefinite time) and an inter-stimulus interval (ITI, 200 ms). During stimulus presentation, the two response options were shown on the screen to the left and right, as well as a depiction of the pronounced phrase (e.g., a blue ball). The response option positions were swapped halfway through the test for counterbalancing, which was indicated by an animation of the faces moving to their new position.

Main test 2, Emotion production. In this test, children were asked to act emotions using the words and emotion depicted. For instance, if they saw a picture of a red chair and a happy face, they were required to say 'red chair' in a happy way. Variants with different articles and plurals were accepted. There was no time limit for a trial. There were eight trials, namely two objects to be named with each of the emotions 'happy', 'sad', 'angry' and 'scared'. There were no warm-up trials.

Main test 4, Focus production. The children verbally responded to prerecorded questions eliciting focus prosody. The questions of the form 'Is this a [color] [N]?' either matched (half of the stimuli) a picture they produced or contrasted in the color or in the noun (both a quarter of the stimuli). There were 24 stimuli on a single block, preceded by two warm-up trials. Trials were similar in setup to those of the emotion and focus perception tests.

6.2.4 Data analysis

Group comparisons (CI vs. NH) were, when single values per participant were compared, performed with non-parametric tests because of the small sample sizes. A significance level of $p = 0.05$ was adopted. Analyses were performed using *SPSS version 23.0* (IBM Corp, Armonk, NY). Effect sizes are reported for two-way comparisons, as less fine-grained comparisons were not the endpoints of interest.

Baseline test 1, Non-verbal emotion understanding. In the Face discrimination and the Face identification tasks, the groups' mean numbers of correct responses were computed and compared for all trials pooled together, using the Mann-Whitney *U* test for independent samples. In the Face discrimination task, this was done for all test components pooled as well as for each component separately, i.e., by addition of numbers of correct responses for both response options of an object or face pair (cars vs. flowers, faces with glasses vs. hats, faces with positive vs. negative expressions, and faces with sad vs. angry expressions). In the Expression task, mean response accuracy

was compared between groups, separately for the verbal and the pointing responses. For both these response types, a distinction was made between strict and tolerant evaluation policies. In the strict policy, each trial was assigned one of four expected (prototypical) emotions (happy, angry, sad, scared) and a response counted as accurate if and only if that exact emotion was chosen. In tolerant policy, only a distinction between positive (happy) and negative (angry, sad, scared) emotions was made. Positive or negative vocabulary other than the expected emotion labels were tolerated as well. For both these policies, analyses were performed.

Baseline tests 2 and 3, Stimulus identification and naming. These data were analyzed by computing percentages correct. For the Stimulus identification test, this involved the percentage of accurately identified phrases by selecting the picture on the screen corresponding to the phrase heard. For the subsequent Naming test, this involved overtly naming the picture shown on screen using the vocabulary encountered in the Identification test. Responses were recorded to allow evaluation of naming accuracy (by the first author).

Baseline test 4, Non-word repetition. All responses were transcribed using broad IPA (International Phonetic Alphabet) transcription by the first author as well as, for a reliability check, 130 items (25%; equally drawn from all participants and as equally as possible from all items) by a trained Dutch phonologist unaware of the target pronunciations. Based on guidelines by Dollaghan and Campbell (1998), they were scored on a phoneme by phoneme basis, every omission or, contrary to Dollaghan and Campbell (1998), addition of a phoneme and substitution by another phoneme counting as an error. In case of omitted or added syllables, utterance were aligned with the target in such a way as to minimize the number of errors. Subsequently, the numbers of phonemes repeated correctly was divided by the total number of target phonemes per word yielding a Percentage of Phonemes Correct (PPC) per stimulus length in number of words (ranging between one and five) (Dollaghan & Campbell, 1998). These measures were compared between groups (CI and NH).

Main tests 1, Emotion perception test. Because in the Emotion test, only two response options were available, following Signal Detection Theory, scores were transformed into hitrates, with one value per subject per Phonetic Parameter (Stanislaw & Todorov, 1999). In this way, possible response biases were accounted for. Following Macmillan and Kaplan (1985), perfect scores for a subject in a cell, which are not computable, were replaced by 100%/2*N*, where *N* is the number of items in the cell (24). Results are presented as *d'* scores. Data were subsequently subjected to a Repeated Measures ANOVA, with Phonetic Parameter as the within-subjects variable and Group as the between-subjects variable.

Main tests 2 and 3, Emotion production and Focus production. Participants' verbal responses in the Emotion and Focus production tests were evaluated by a single panel of 10 Dutch adults with a mean age of 27.3 years who did not present a hearing loss of over 40 dBHL at any of the octave frequencies between 0.125 and 8 kHz, as audiometrically assessed (Audio Console 3.3.2, Inmedico A/S, Lystrup, Denmark). In the Emotion test, listeners judged by button-press which of four emotions (happy, angry, sad, scared) was conveyed independent of the contents of the utterance. In the Focus test, they judged which of three focus positions (color, object or both) was accented. Another condition of the Focus production, in which the question posed to the children corresponded in both color and object to the image displayed, was not further analyzed. In this test, listeners were explained the procedure of the production task and asked to imagine to which question the speaker's utterances were a response to, so that they would judge the phrasal accents as corrective focus realizations (which is how they were intended by the speakers). In both evaluation tests, the order of response options was counterbalanced between two different versions. The order of the two tests per listener was also counterbalanced.

For every trial, for each participant, ten correct or incorrect responses were considered, according to the evaluations by the panel of ten adult listeners. A child's production counted as correct when the

emotion it was prompted to produce in the task corresponded to the emotion perceived by an adult listener, and counted as incorrect otherwise. This yielded 1,910 data points in the Emotion production test and 2,780 data points in the Focus production test. Percentages correct were calculated over this entire dataset and compared between Groups and Emotions. No *d'* scores were calculated, as is common for alternative forced choice (AFC) tasks with more than two options (Macmillan & Creelman, 2004).

6.3 Results

Parent questionnaire. Parents of NH children reported Dutch to be their own first language as well as the mother tongue and first language of their child, used at home, at school and with friends. One child had been treated for hearing problems and one other child had received speech therapy. No NH children had been treated by a neurologist or for social problems and none had problems with their sight. The average SES, computed as the sum of the questionnaire ranks of the two parents' highest level finished education and their income category, of this group was 19.4, ranging between 17 and 21 and with a SD of 1.6. Parents of CI children also reported Dutch as their first language. Their child's first linguistic input was reported as Sign Language of the Netherlands (SNL) received from parents who learned it as a second language or from Dutch Sign Language (DSL) teachers. Three parents indicated that the acquisition of Dutch was simultaneous with that of DSL and that two parents had not reported DSL's acquisition onset age. All parents of CI children indicated that communication with their child before implantation was more frequent (answers of three parents missing) and (except for two parents) easier using sign language and all of them reported (except one missing answer) that after implantation spoken language communication was more frequent and easier, showing that implantation had successfully given access to spoken language. One

CI recipient had been treated by a neurologist but no children had been treated for social problems. One CI recipient had problems with his/her sight. The average SES of this group's parents was 18.0, ranging between 12 and 22 and with a SD of 3.4.

Baseline test 1, Non−verbal emotion understanding. In the Face discrimination task, mean numbers of correct responses were not different between groups for all object or face pairs together (U = 1230.5, z = −1.17, p = .24, r = −.23) or any of the pairs separately according to Mann-Whitney U tests (cars vs. flowers: U = 84.5, z = 0; p = 1, r = 0; faces with glasses vs. hats: U = 71.5, z = −1.44, p = .51, r = −.28; negative vs. positive faces: U – 77.0, z = −.56, p = .72, r = −.11; angry vs. sad faces: U = 74.5, z = −.56, p = .61, r = −.11; exact significance). In the Face identification task, no effect of group on the number of correct responses was found either (U = 53.0, z = −1.8, p = .11, r = −.35; exact significance). In the Expression task, no effect of group on mean accuracy scores was found for strict (U = 4724.5, z = −1.0, p = .32, r = −.20) and tolerant (U = 4892.5, z = −1.8, p = .074, r = −.35) verbal responses, nor for the strict (U = 5267.5, z = −.26, p = .79, r = −0.051) and tolerant (U = 5253.0, z = −1.4, p = .16, r = −.27) pointed responses. These results suggest that, to the degree tested, the two groups have comparable levels of non-verbal emotion understanding.

Baseline tests 2 and 3, Stimulus identification and naming. In the Identification test, the CI group scored 98.7% correct and the NH group a 100%. In the Naming test, CI group's accuracy was 100% and the NH group's accuracy 99.4%. There were no missing cases. These results show that both groups were sufficiently able to perform the kind of tasks that the main part of the study consisted of, namely identification and verbal responding. Moreover, the results show that subjects knew the words corresponding to the pictures used.

Baseline test 4, Non-word repetition. In the Non-word repetition test, 3 out of 520 productions (0.006%) were missing. The second rater's transcription of 20% of data corresponded for 93.8% to those by the first rater with disagreement occurring almost exclusively

at the phonetic level of individual phonemes such as voicing, showing the first rater's transcription to be reliable. Of the remaining data, Table 2 and Figure 1 summarize the results, showing mean percentages of phonemes correct (i.e., correctly repeated) per group and per item length, in number of syllables. The two groups show a parallel downward pattern with increasing item length, but the CI recipients consistently show a lower score by around 5%. The relatively low percentages for the one- and two-syllable words is due to the relatively large percentage of mispronounced (or misheard) final nasal consonants in those words. The overall score was statistically significantly different between the two groups according to a *t*-test with equal variances not assumed ($t(1,515) = -3.2$, $p = .001$, $r = .69$). The NH group was therefore somewhat more accurate at repeating non-words than the CI group.

Main test 1, Emotion perception. Table 3 and Figure 2 show *d'* scores in the Emotion perception test, split by Phonetic parameter (Intonation, Temporal or Both) and subject group. Repeated measures ANOVA on the *d'* scores revealed a main effect of Phonetic parameter ($F(2,22) = 49.79$, $p < .001$), but no effect of Group ($F(1,23)$

Table 2. Mean percentage phonemes correct and standard deviations (in parentheses) correct per syllable length (in number of syllables) and per participant group (CI or NH) in the Non-word repetition test.

Group	Mean Percentage Phonemes Correct (SD)					
	Item length (number of syllables)					
	1	2	3	4	5	**Total**
CI	88.5 (17.3)	87.3 (13.7)	92.3 (12.2)	85.0 (18.4)	66.8 (23.5)	**84.0 (19.5)**
NH	92.2 (14.3)	91.5 (11.4)	97.8 (7.7)	89.3 (12.1)	74.3 (21.2)	**89.0 (16.1)**
Total	**90.3 (15.9)**	**89.4 (12.8)**	**95.1 (10.5)**	**87.1 (15.7)**	**70.5 (22.6)**	**86.5 (18)**

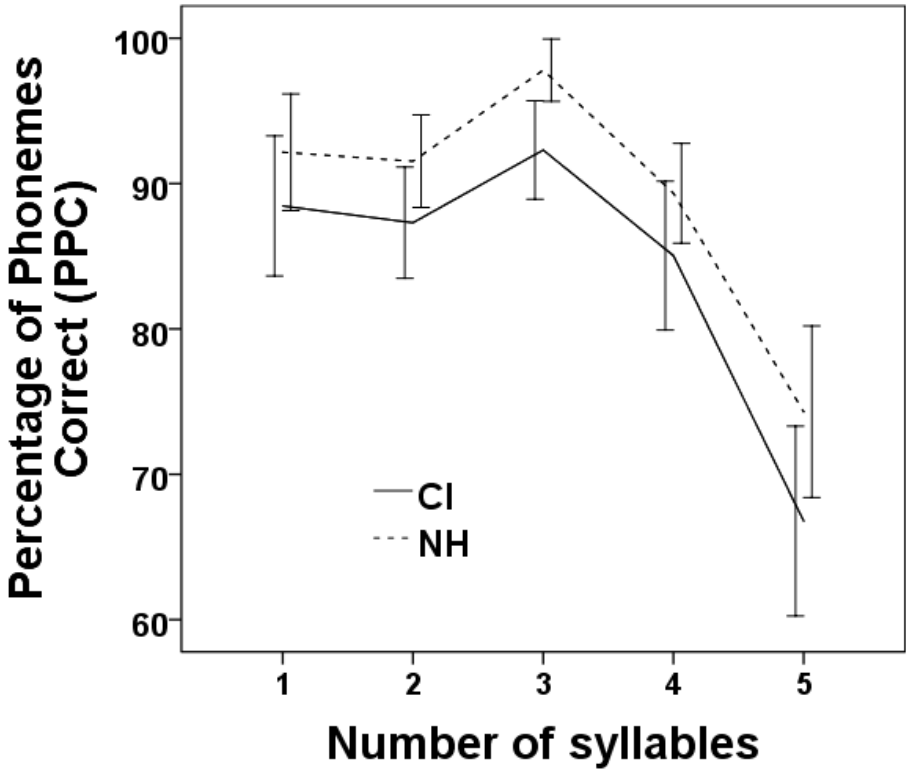

Figure 1. Percentage of Phonemes Correct per number of syllables in the Non-word repetition test. Percentages correct represent percentages of correctly repeated phonemes per non-word. Additions, omissions and substitutions of phonemes counted as errors.

= .18, p = .68, r = .39), nor an interaction between Phonetic parameter and Group (F(2,22) = .29, p = .97). Post-hoc analyses revealed that of the three Phonetic parameters, scores on the Temporal condition differed highly significantly from both Intonation (t(24) = 7.61, p < .001, r = .84) and Both (t(25) = −10.70, p < .001, r = .91), but the Intonation and Both conditions were not significantly different from each other (t(24) = −1.79, p = .086, r = .34) given a Bonferroni-corrected p-criterion of .05/3. These results suggest that CI and NH groups were equally capable of discriminating the two emotions and that they do that applying the same cue weighting strategy.

Table 3. Mean *d'* scores split by Phonetic parameter and by participant group (CI or NH) in the Emotion perception test. Participants judged if prerecorded utterances were pronounced with a happy or sad emotion. Phonetic parameters indicate which type of phonetic information was available in the stimulus.

Group	*d'*			
	Phonetic parameter			
	Intonation	Temporal	Both	Total
CI	2,40 (1,26)	0,52 (0,65)	2,80 (1,05)	**1,91 (1,41)**
NH	2,32 (1,19)	0,26 (0,55)	2,64 (1,2)	**1,72 (1,47)**
Total	**2,36 (1,2)**	**0,39 (0,61)**	**2,72 (1,11)**	**1,82 (1,43)**

Main test 2, Emotion production. In the Emotion production test, of all trials, 3.8% were missing (missing response or technical error). Table 4 and Figure 3 show mean percentages correct of the four emotions in both participant groups (CI and NH). The overall accuracy of the CI group (62.3%) was somewhat higher than that of the NH group (57.8%) but the group difference varied across emotions. According to two- and four-way ANOVAs, respectively, there was a very small but significant effect of Group ($F(1,1902) = 7.06$, $p = .008$, $r = .061$) and one of Emotion ($F(3,1902) = 45.43$, $p < .001$), as well as a significant interaction between Group and Emotion ($F(3,1902) = 7.82$, $p < .001$). Bonferroni-corrected post-hoc tests showed that all levels of Emotions differed highly significantly ($p < .001$), except angry and sad ($p = 1$). Separate Group comparisons for each emotion showed that the CI group scored higher than the NH group on scared ($F(1,438) = 10.06$, $p = .002$, $r = .15$) and angry ($F(1,478) = 14.01$, $p < .001$, $r = .17$) responses, that the NH scored better on happy responses ($F(1,298) = 5.11$, $p = .024$, $r = .13$), but that there was no difference for sad responses ($F(1,488) = .017$, $p = .90$, $r = .019$). These results indicate that the two groups have different specialties when it comes to the production of emotions, but that in

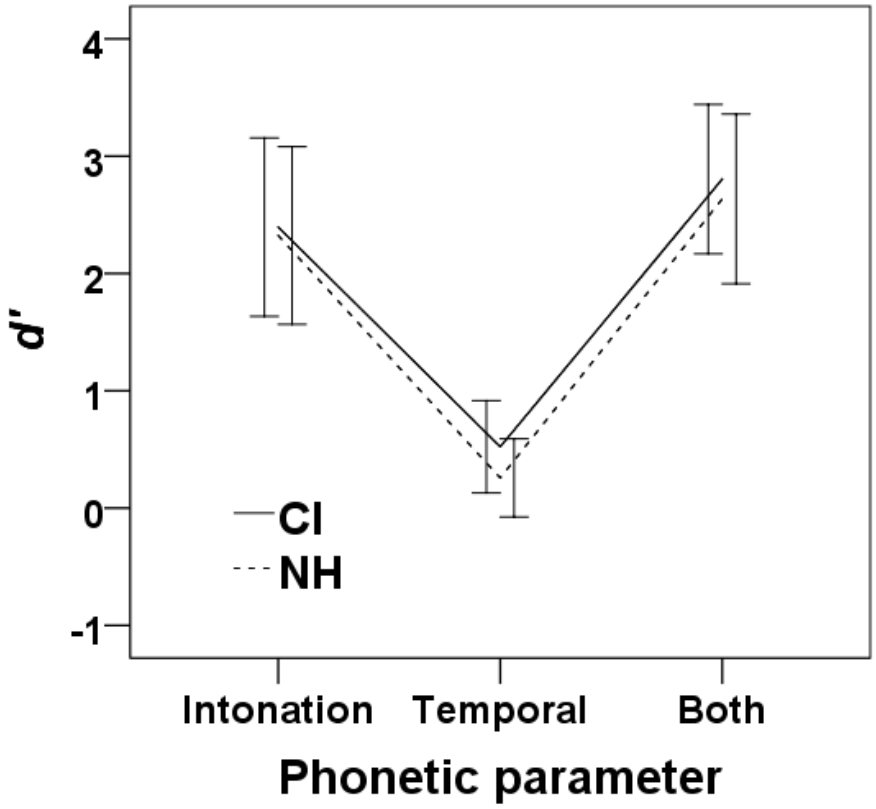

Figure 2. Mean *d'* scores split by Phonetic parameter and by participant group (CI or NH) in the Emotion perception test. Participants judged if prerecorded utterances were pronounced with a happy or sad emotion. Phonetic parameters indicate which type of phonetic information was available in the stimulus.

general the groups are almost equally good at distinguishing them. *Main test 3, Focus production.* In the Focus production test, of all trials, 10.9% were missing (missing response or technical error). Table 5 and Figure 4 show mean percentages correct of the three focus positions tested evaluated by a panel of listeners in both participant groups (CI and NH). The mean percentage correct for the CI group was 58.1% and for the NH group 60.4%. A main effect of Focus was 58.1% and for the NH group 60.4%. A main effect of Focus position was found ($F(2, 2774) = 57.00$, $p < .001$, $r = .14$), but not of Group

Table 4. Mean percentages correct and standard deviations (in parentheses) per focus position and per participant group (CI or NH) of focus position conveyed in dummy phrases in the Focus production test.

Emotion	Accuracy mean (SD)		
	CI	NH	**Total**
happy	70.4 (46.7)	79.2 (40.7)	**74.6 (43.6)**
angry	70.6 (46.1)	53.2 (50.0)	**62.1 (48.6)**
sad	62.3 (49.6)	61.7 (48.7)	**62.0 (48.6)**
scared	46.4 (50.0)	31.6 (46.6)	**40 (49.0)**
Total	**62.3 (48.5)**	**57.8 (49.4)**	**60.3 (48.9)**

($F(1,2774) = 1.94$, $p = .026$) nor an interaction between Focus position and Group ($F(2,2774) = .94$, $p = .39$). These results indicate that the two groups were equally effective at distinguishing the focus positions in their output and that they most likely produced them with similar strategies, given that they were similarly judged by the panel of listeners.

Correlations among tests and between age and test scores. Two-tailed correlations between six scores of Non-verbal emotion understanding test and the scores of the Non-word repetition, Emotion perception, Emotion production, and Focus production tests were tested per Group. The six scores of the Non-verbal emotion understanding test were (1) total scores (in numbers of correct responses) for the Face discrimination task (i.e., averaged scores over all four test components) and (2) the Face identification task (total number of items correct over all trials) as well as (3 through 6) percentage correct scores for verbal and pointed responses according to strict and tolerant policies. These subscores of components within Non-verbal emotion understanding were not tested for correlations

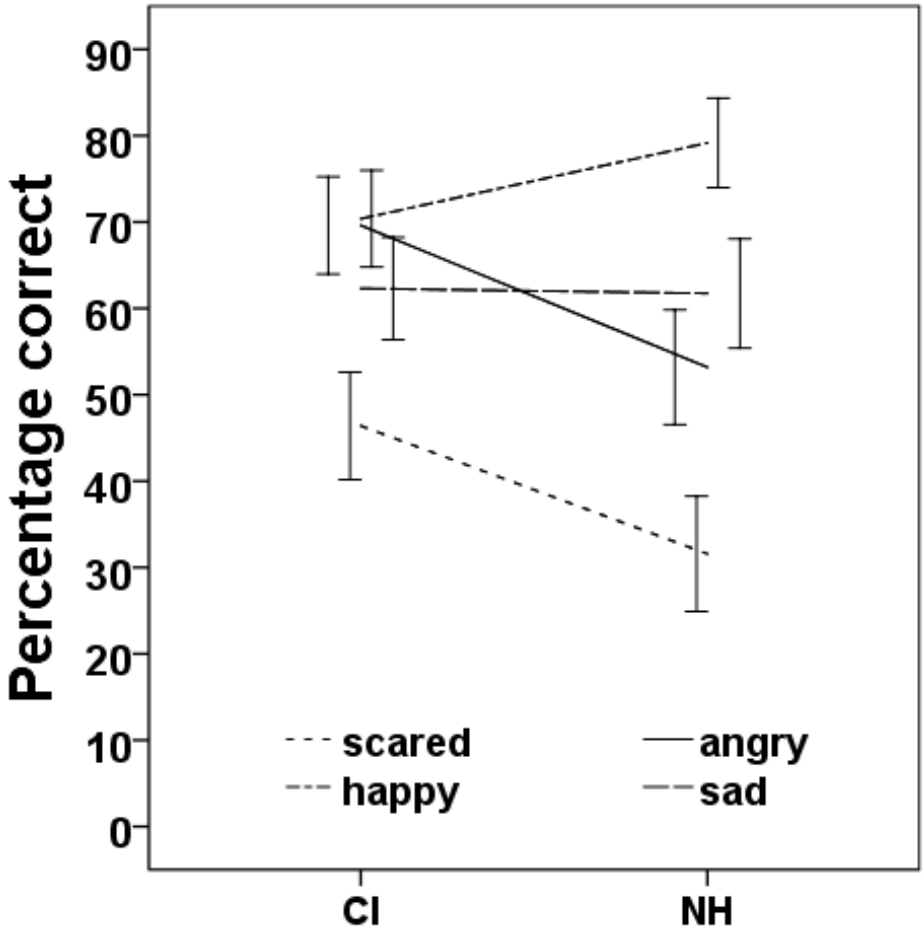

Figure 3. Mean percentages correct per emotion and per participant group (CI or NH) of emotions conveyed in dummy phrases in the Emotion production test. Percentages correct were computed by averaging judgements of emotions perceived by a panel of ten naïve adult Dutch listeners with normal hearing.

among each other nor with the Non-word repetition test, but only for correlations with the main tests. Based on Kolmogorov-Smirnov tests and visual inspection of Q-Q plots, we assumed that the distributions per group for Non-word repetition, Emotion perception, Emotion production, and Focus production were largely normal and that distribution for the other scores were not normal. It should be noted

Table 5. Mean percentages correct and standard deviations (in parentheses) per focus position and per participant group (CI or NH) of focus position conveyed in dummy phrases in the Focus production test.

Emotion	Accuracy mean (SD)		
	CI	NH	**Total**
adjective	70 (45.9)	72.1 (44.9)	**71.0 (45.4)**
noun	59.6 (49.1)	59.3 (49.2)	**59.4 (49.1)**
both	44.2 (49.7)	50.1 (50)	**47.9 (49.9)**
Total	**58.1 (49.4)**	**60.4 (48.9)**	**59.2 (49.2)**

that for the CI group the distribution of scores on the Non-word repetition test was marginally significant (p = .063).

Only the following correlations were significant. For the CI group, scores of the Face discrimination were marginally significantly and weakly correlated with Emotion perception (r = .387, p = .046), those of the Face identification task were moderately correlated with Emotion production (r .52, p = .015), strictly judged verbal responses on the Expression task were moderately or weakly correlated with Emotion perception (r = .453, p = .025) and, marginally significantly, Focus production (r = .308, p = .089), respectively, and strictly judged pointed responses were weakly correlated with Emotion production (r = .398, p = .039). In the NH group, strictly judged pointed responses were weakly to moderately and marginally significantly correlated to Focus production (r = .423, p = .054). The correlation between Emotion perception and Emotion production was marginally significant for the CI group (r = .523, p = .067) whereas it was not for the NH group (r = −1.44, p = .656), and the correlation between Non-word repetition and Emotion production was marginally significant for NH group (r = .543, p = .068) whereas it was not for the CI group (r = .017, p = .96).

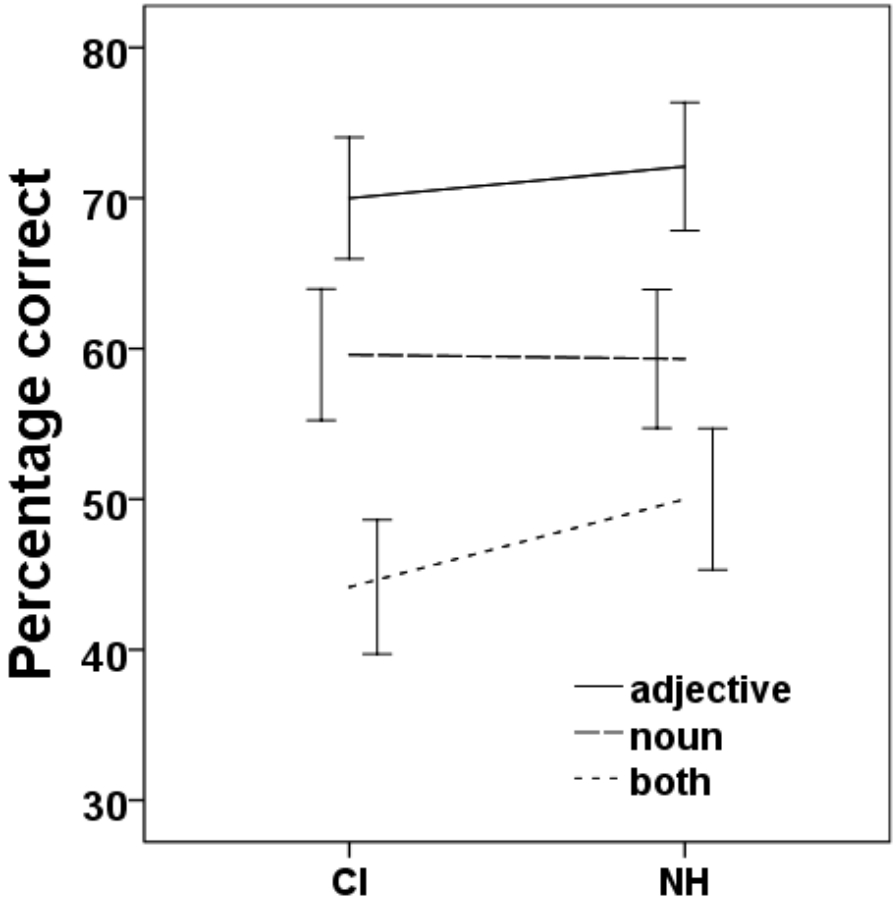

Figure 4. Mean percentages correct per focus position and per participant group (CI or NH) of focus positions conveyed in dummy phrases in the Focus production test. Percentages correct were computed by averaging judgements of emotions perceived by a panel of ten naïve adult Dutch listeners with normal hearing.

Finally, correlations were run between main test scores on the one hand, and activation age, hearing and chronological age, on the other hand. In the CI group, the only significant correlation was between hearing age and Emotion production (r = .028, p = .542). In the NH group (where chronological age is by definition equivalent to

hearing age), there was a marginally significant correlation between age and Emotion production (r = .470, p = .061).

These results show that in general, the capacities tested in the different main tasks seem unrelated to each other but that there is a trend towards emotion production being predicted by emotion perception skills for CI (but not NH) children and by Non-word repetition skills for the NH (but not the CI) group. Moreover, partly as a trend, scores on the Emotion perception and production and Focus production are to a limited degree predicted by some non-verbal emotion understanding scores, although more so for the CI than for the NH group. Age measures were to some extent correlated with Emotion production but not with other main test scores.

6.4 Discussion

The aim of this study was to test the capabilities and cue-weighting strategies of a small group of children with cochlear implants and a control group of normally hearing children, matched for hearing age, on both the perception and production of (emotional and linguistic) prosody, controlling for the level of non-verbal emotional understanding and general linguistic capacities, and to test for correlations between scores of main tests and between baseline and main tests. To our knowledge, this study was the first to test perception and production of emotional (and linguistic) prosody in the same cohort of pediatric CI recipients. Moreover, effectiveness of the non-imitative production of emotions was never tested in this population. Although, contrary to our hypotheses, the two groups performed generally in a similar way, some differences were observed that coincided with our expectations.

Our first set of hypotheses was (A1) that prosody perception and production scores within participants were correlated (A2) that that effect was larger within than across prosody type (linguistic vs. emotional) and (A3) that that effect was larger for linguistic than for

emotional prosody. Hypothesis A1 was confirmed to a limited degree. In the CI group, but not the NH group, Emotion perception performance moderately and marginally significantly predicted Emotion production performance. Other correlations, however, were either very weak and/or not significant. This result is in support of hypothesis A2, since the only correlation of any significance involves within-prosody type (emotional linguistic) and between-prosody correlations were not found. As Focus perception, however, could not be analyzed, it remains unknown if this holds for linguistic prosody as well. For the same reason, Hypothesis A3 cannot be confirmed nor rejected.

The trend for a link between emotion perception and production in the CI group supports results by Nakata et al. (2012) who found a correlation for five- to thirteen-year-old recipients between imitative emotion perception and production scores. The trend, if reflecting an actual effect, provides some support for the view that in this population better prosody perception skills allow better prosody production skills (e.g., Nakata et al., 2012), at least within the domain of emotion. This would entail that the production of emotions cannot develop and function entirely independently from their perception, whereby the independence stance would stem from the idea that the way to distinguish vocal emotions in production is not (sociolinguistically) acquired but innate (Scherer, Banse & Wallbott, 2001). Instead, thus, our results argue in favor of the opposite hypothesis, stating that vocal expression of emotions is at least partly learned. This is also consistent with the fact that in the present data for the NH children, no such trend was observed, as, naturally, they have received normal input since birth and their variation in skills of emotion perception and production distinction might be due to other factors, such as personality factors, instead of perceptual acuity. It has to be kept in mind, however, that due to the small sample size, personality factors, for instance, might have played a role in the experimental group as well, for instance representing a wider variation than for the NH group.

As our second set of hypotheses, we expected (B1) that scores per prosody type (linguistic or emotional) were influenced by their respective general scores for linguistic and emotional capacities, (B2) but that this effect was larger for linguistic than for emotional prosody. As for prediction by the linguistic baseline test (Non-word repetition), a marginal correlation between Non-word repetition and Emotion production was found in the NH group, but no other correlations. As for the emotional baseline test (Non-verbal emotion understanding), in the CI group, Emotion perception and Emotion production each correlated with scores on two of the Non-verbal emotion understanding subtests, namely Face discrimination and strictly judged verbal responses for the Expression task, on the one hand, and Face identification and strictly judged pointed responses in the Expression task, on the other hand, respectively. In the NH group, there was only a marginal correlation between Focus production and strictly judged pointed responses in the Expression task. Therefore, Hypotheses B1 (concerning emotion tasks) and B2 are partly confirmed for the CI group, but not for the NH group.

These results complement and in part contradict those by Wiefferink et al. (2013), from whose battery of non-verbal emotion understanding tests those in the present study were adopted. Whereas Wiefferink et al. (2013) reported a delay in emotional development in CI children, in the present study, no difference was found. This lack of a difference is, however, consonant with the normal emotional development found by Mancini et al. (2016). As in the study by Mancini and colleagues, whose implanted participants were between 4 and 11 years old, it is highly probable that this is due to the fact that the current study tested older children (between 6 and 12 years old). Although these older children did not reach ceiling level performance on all subtests, the tests might be less sensitive to possibly more nuanced differences in emotional capacities at these ages. Nevertheless, importantly, the similarity in performance of the two groups on non-verbal emotional understanding tests suggests that CI

children's emotional capacities (partly) are at a level comparable with that of their peers near – at the latest – the end of primary school.

A difference with both the studies by Wiefferink et al. (2013) and Mancini et al. (2016) is that for CI children no correlation between performance on verbal emotion tasks and general language level was found. A connection between those faculties was explained by Flom and Bahrick (2007), cited in Mancini et al. (2016), by assuming that language helps naming emotions and linking them to external referents (objects and events) and that the temporal synchrony between vocally (i.e., prosodically) and non-vocally expressed emotions (e.g., faces) and external referents is required for leaning to distinguish emotions. Given this hypothesis, the present lack of a correlation between general linguistic and verbal emotion tasks might be accounted for by assuming that the linguistic experience of the CI children, whose linguistic level was less advanced than that of the control group but as a small effect, was sufficient for talking about emotions, linking them to external referents and learning the synchrony with facial expressions. It has to be noted that CI and NH participants were matched for hearing age (not chronological age) and the experimental group therefore had more chance to gain experience than the control group with learning to distinguish and express – verbal and non-verbal – emotions.

Our third set of hypotheses was (C1) that CI activation age would negatively correlate with prosody capacities, but (C2) that this effect would be larger for linguistic than for emotional prosody. In the CI group, CI activation age nor chronological age was found to predict outcomes, but hearing age did moderately correlate with Emotion production. In the NH group, age was marginally significantly and weakly to moderately correlated with Emotion production. The hypotheses are therefore not confirmed because any effect observed is related to emotional and not linguistic prosody processing. Increasing experience with the implant (hearing age) did improve emotion perception, but hearing age negatively correlated with activation age,

obscuring conclusions about the separate effect of either factor. Samples were too small to perform partial correlations.

These results suggest that increasing time in sound and possibly an earlier onset of stable hearing (in this study defined as birth for the NH group and age at activation for the CI group) help improve emotion production performance. The fact that emotion perception was not found to be predicted by this factor would suggest that emotion perception capacities mature independently of hearing experience, whereas emotion production capacities do develop as a function of it. The finding that emotion production is at most indirectly dependent on other capacities resonates with results from the study by Bergeson and Chin (2008), whose six- to ten-year-old implanted children's emotion prosody imitation performance showed no correlations with intelligence scores and general linguistic level. This at first glance seems inconsistent with our other result that emotion production skills are correlated with emotion perception skills; however, the two lines of results could be reconciled by the assumption that emotion perception capacities are a necessary but not a sufficient requirement for emotion production capacities, whereby one of the other possible requirements are sufficient emotional capacities (as also observed in the present study).

Our final set of hypotheses was (D1) that for the perception of prosody, CI participants would rely more heavily on temporal cues as opposed to F0 cues than NH participants do, but (D2) that this effect would be stronger for linguistic than for emotional prosody. In as far as this study tested these hypotheses, they were not confirmed. The recipients weighted their cues in the same way as the children from the control group, namely by relying almost entirely on F0 cues, disconfirming Hypothesis D1. Hypothesis D2 was not tested because the Focus perception test did not yield analyzable results. That hypothesis therefore remains to be investigated in future research.

The similarity in emotion perception performance and cue weighting strategy between CI and NH children is in marked contrast with earlier research, where children of different ages showed poorer

emotion perception performance (Geers et al., 2013; Luo et al., 2007; Nakata et al., 2012) and a heavier reliance on temporal vs. F0 information by CI users (Meister et al., 2011; O'Halpin, 2009). It might be the case that the happy and sad stimuli used happened to have relatively pronounced differences in intonation contour and/or register, allowing even CI users, who have poor F0 resolution, to reach ceiling level when only F0 information was present, possibly also diverting their attention from temporal information when only temporal information was present (in the Duration condition). More difficult tasks, with less exaggerated renderings of emotions and/or with more emotions, might bring to light more subtle differences in cue weighting strategies between these groups. Nevertheless, it is remarkable that in this study the F0 information was sufficient for CI children to distinguish emotions at a level equal to that of their NH peers.

Shortcomings and suggestions for future research

A number of shortcomings of this study have to be taken into account that warrant prudence in interpreting the data. First of all, the sample size, with two groups of thirteen participants, was small as a result of the limited availability of implanted children passing the inclusion criteria, compromising generalizability to other pediatric recipients.

Second, possibly, the cohort has self-selected for the better-performing children because parents who feared their children might perform sub-optimally might for that reason not have responded to the invitation. This may have contributed to the fact that the implanted children performed within the norm on several (sub)tests. The fact that effects and tendencies have been found in the present sample suggests that stronger effects could be found in larger studies.

For the above reasons, research using larger sample sizes with implanted children with a broad range of linguistic and maturational developmental levels (i.e., chronological and hearing ages) is likely to

extend and strengthen the present results. Further, a study in which early and late implanted children, compared to three groups matched for hearing age and chronological age to both experimental groups would further complement the current study by separately testing the roles of duration of CI experience and general maturational development for emotional and linguistic language skills.

Third, a limited set of stimuli was used in all tests so that the results of the tests would allow a comparison. Moreover, the stimuli's variants with different cue availability (intonation, temporal information or both) were highly controlled in order to test the role of the respective types of phonetic information irrespective of the contents of the stimuli. The sole speaker recruited to record the stimuli, however, will have idiosyncratic prosodic characteristics (Kraayeveld, 1997); the production by another speaker or of other stimuli might have brought about different weightings of temporal and intonation information. It can therefore not be excluded that the cue reliance mechanism found in this study is specific to the stimuli used.

The baseline test for non-verbal emotion understanding used in this study might not have been sensitive nuances in emotional development that could influence performance on linguistic prosody tests. More challenging tests, such as involving higher emotional skills such as beliefs and moral values, in combination with the tests of the perception and production of irony, surprise and deception and acoustic measurements of elicited utterances, might capture fine-grained differences in emotional verbal development in this population.

Conclusions

In this research, we tested the perception and production of emotional and linguistic prosody by six- to twelve-year-old children with cochlear implants and normally hearing, hearing age matched children. It has to be noted that linguistic prosody perception (focus

perception) could not be analyzed. The following conclusions resulted from the study.

1) Emotional prosody perception and production scores were weakly and moderately significantly correlated for CI children but uncorrelated for NH children, suggesting that higher perception skills allow higher production skills and that emotion production is partly learned (as opposed to innate).
2) For CI children but not NH children, emotion perception and production scores were predicted by non-verbal emotional understanding performance. No such correlation was found for linguistic prosody. For NH children, only marginal contra-modal (from emotion to focus and vice versa) correlations were found. Our data showed no overall performance level difference between groups, suggesting that these children either never experienced deviations for the tested capacities or if they had had any delay, that has become irrelevant by the age at testing.
3) Hearing age (itself correlated with activation age for the CI children) weakly predicted emotion production, but not emotion perception, performance in both groups, suggesting that increasing time in sound has a favorable effect on vocal emotional expression, possibly requiring independently maturing emotion perception skills.
4) For emotion perception, CI and NH children adopt the same cue-weighting strategies, relying almost entirely on F0 information as opposed to temporal information, and perform at the same level of accuracy.

Acknowledgments

We are grateful to Jos Pacilly, engineer at the Phonetics Laboratory at Leiden University, for his technical support at many stages of this research. We are also grateful to Walter Verlaan, engineer at Leiden University Medical Center, for his assistance in developing the experimental setup. We also thank all participants for their willingness to cooperate.

Chapter 7

Conclusions

The aim of this thesis was to increase insight into the mechanism by which users of cochlear implants (CIs) perceive and produce prosody and to investigate how prosody is perceived with vocoder simulations. This was investigated in five separate studies using Dutch children with CIs and, as controls, normally hearing (NH) adults and children by testing their capability to distinguish and to produce utterances with different emotions (emotional prosody) and focus positions (linguistic prosody). The research aim was approached from five research perspectives with corresponding hypotheses: (1) differences between linguistic and emotional linguistics; (2) the distinction and relationship between the perception and production of prosody; (3) the relationship between prosody and music perception; (4) the cue weighting mechanism employed by CI users in perceiving prosody; and (5) the prosody processing capacities by children with CIs.

One study involved the analysis of basic prosodic parameters of spontaneous utterances by children with CIs (Chapter 2). Two studies (Chapters 3 and 4) tested the influence of cue availability (duration and F0 cues) and the slope of the synthesis filter in vocoder simulations of CIs on the discriminability of emotions and focus

positions by NH adults. Chapter 5 additionally tested if the weighting of these cues would be affected by a short training with vocoded materials and if the training effect, if present, would transfer to other cues and/or outside of the domain of language (viz., music). The final study (Chapter 6) investigated differences in cue weighting in perception and effectiveness in the production of emotional and linguistic prosody by five- to eleven-year-old children with CIs with their hearing-age matched peers, controlling for general level of emotional and linguistic capacities. Below the hypotheses related to the research themes will be revisited in light of the results of the different studies.

7.1 Perspective 1. Linguistic and emotional prosody

We hypothesized that emotional prosody would be recognized (Hypothesis 1a) and realized (Hypothesis 1b) using different cues than linguistic prosody, that emotional prosody perception would be less correlated to music processing than linguistic prosody (Hypothesis 1c) and that emotional linguistic prosody perception and production would be less correlated with each other than linguistic prosody perception and production would (Hypothesis 1d). These hypotheses were addressed in Chapters 4, 5 and 6.

In a pair of experiments (Chapter 4) testing the effect of a wide range of synthesis filter slopes as well as, orthogonally, the availability of duration vs. F0 cues, on the discrimination of happy vs. sad phrases (emotional prosody) and phrases with sentential focus on either the adjective or on the noun (linguistic focus), using vocoder simulations of cochlear implants it was shown that listeners relied more on the F0 cues than on the duration cues in emotional prosody and more on the duration cues than on the F0 cues in linguistic prosody. Another study (Chapter 5), using vocoder simulations with NH participants (not the same individuals as in Chapter 4) to test the effect of cue-specific training on cue-weighting in prosody and music

perception, found a comparable cue-weighting strategy. A study testing children with and without cochlear implants (Chapter 6) found the same cue-weighting for emotional prosody perception for both groups; however, testing of linguistic prosody did not succeed and therefore did not allow conclusions about the listening mechanism. This cue weighting strategy found in several of the studies most likely reflected that in the emotional stimuli F0 cues were more important relative to duration cues than they were in the focused stimuli, while at the same time the vocoder algorithm was more detrimental to F0 cues than to duration cues, thereby compromising the discrimination of emotional stimuli more than that of focused stimuli. In Chapter 6, it was found that children with and without CIs adopted the same cue-weighting strategies. This evidence together supports Hypothesis 1a. It has to be noted, however, that the same stimuli were used in all studies and therefore this conclusion cannot be generalized to other stimuli without caution.

Hypothesis 1c received some support from the study described in Chapter 5. Performance in short-term training in discriminating unfamiliar melodic contours based on melodic, in one participant group, or rhythmic, in another participant group, properties (weakly) correlated with scores in linguistic (focus position) but not emotional prosodic perception. Correlations were also observed between scores on familiar melody recognition and focus perception and emotion. Thus, correlations between music perception performance with linguistic prosody performance were more consistently reported than those with emotional prosody performance. This could have to do with the correspondence between the musical stimuli and the linguistic stimuli related to the expression of focus position that in both types of stimuli most of the variations (except for *crescendi* and *diminuendi* in one of the training sets for the melodic training group) were of a grammatical nature. That is, accents, durational differences, and note heights (in music), on the one hand, and sentential accents (in speech), on the other hand, were bound to a specific position in the stimulus. Emotional prosody, by contrast, was not of a grammatical

nature but pertained to extra-linguistic characteristics of a sentence. This type of expression would be more related to global pitch register, pace or intensity variations in music, but that type of 'musical prosody' was not used in the stimuli.

The realization of linguistic and emotional prosody (Hypothesis 1b) and its relationship to perception (Hypothesis 1d) were addressed in a set of studies described in Chapter 6. Emotional prosody perception and production were correlated for CI children but uncorrelated for NH children, supporting Hypothesis 1d for the clinical group, but not for the control group. Linguistic and emotional prosody contrasts were conveyed with equal success by both groups, as assessed by a panel of ten naïve NH Dutch adults, suggesting that the children did not have more difficulty with producing one type prosody over the other. This is in dissonance with Hypothesis 1b, although the results might reflect a ceiling effect, in that the groups' scores, in case they had been more different when they were younger, might have had the time to converge due to the participants' relatively advanced age and that for younger children a difference between a clinical and a control group might have been observed.

7.2 Perspective 2. Perception and production

We hypothesized that both perception (Hypothesis 2a) and production (Hypothesis 2b) would be deviant in CI users because they might develop as an integrated system, which would surface as a within-participant correlation between perception and production scores (Hypothesis 2c).

In vocoder simulations of CIs (Chapters 3 and 4), the perception of prosody was shown to be affected by the vocoding of the stimuli. Relative to conditions without vocoding, performance was compromised when participants were asked to discriminate between stimuli that differed only in synthesized intonation contour (Chapter 3) or that differed in emotion or focus position (Chapter 4). Moreover,

the experiments in Chapter 4 showed that under vocoder conditions emotion perception relied relatively heavily on F0 cues in comparison with duration cues and that this F0 reliance was less pronounced for focus perception. Under non-vocoded conditions, this relative reliance on F0 and duration cues was comparable for emotion perception, but reversed for focus perception, showing that signal degrading that mimics CI hearing, apart from compromising performance, can induce a change in listening strategy. This supports Hypothesis 2a for this group of participants. However, children with and without CIs performed with comparable accuracy and listening strategy (cue weighting) (Chapter 6). This pattern of results suggests that children with CIs have learned to adopt the same listening strategy as NH peers, whereas vocoder simulations elicit a different strategy in NH listeners than they would adopt when listening to non-vocoded stimuli.

In production of prosody, in two different studies (Chapters 2 and 6), no differences except tendencies in the speech of CI children relative to that of NH peers were observed. Basic prosodic measures in late implanted (after two years of age; mean chronological age: 6;8) and early implanted (before two years of age; mean chronological age: 2;10) CI children did not significantly deviate, although they did improve with increasing implant experience (Chapter 2). In the same line of evidence, the production of emotions and focus positions was equally successful between six- to twelve-year-old CI and hearing-age matched NH children (Chapter 6). Therefore, no evidence for Hypothesis 2b was found.

In the study on the production of basic prosodic measures, the expected stronger deviation for F0 than for duration measures (the first of which is more problematic for CI users than the second) was not found. One possible interpretation of the findings, however, was that measures requiring a relatively high degree of articulatory or laryngeal control – such as articulation rate, ratio between voiced and voiceless parts of the utterance, and mean F0 of the utterance – showed a tendency towards being more deviant than parameters that

could be considered as a by-product of speaking – such as F0 declination and the F0 variability. In another study (Chapter 6), prosody perception and production performance were found to be correlated in CI children, whereas this correlation was not found for their NH peers. These results lend some support for Hypothesis 2c. Although prosody production scores by CI children were not in general found to be lower than those of NH children, as would be expected based on their degraded input, they first of all did show tendencies towards differences in parameters that might require intact input as opposed to parameters that are automatic by-products of speaking, and second, their relationship between production and perception scores was stronger than for NH children.

7.3 Perspective 3. Prosody and music

We predicted that NH listeners could be cue-specifically trained with musical materials to recognize musical melodies based on either melody or rhythm cues (Hypothesis 3). This training effect might transfer to a cue weighting strategy in which participants rely on the non-trained cue in melody perception (cross-cue transfer), on the trained cues in prosody perception (cross-domain transfer) or on prosody perception for both cues (cross-cue plus cross-domain transfer). This last issue was called the Transfer Issue, since there was no hypothesis into one of the directions of the effect.

Hypothesis 3 was not clearly confirmed. No significant cue-specific effect of musical training on prosody perception was found, but only a tendency of a temporal training effect on temporal prosody perception. Most likely the lack of effects is due to the brevity of the training (45 minutes); however, the tendencies do suggest that more elaborate training could have a more robust transfer effect. Regarding the Transfer Issue, within-domain cross-cue, cross-domain within-cue as well as cross-domain cross-cue correlations on the level of individual participants' performances were found. These might reflect

individual sensitivity variations to a training effect, although, because no pre-training baseline tests were performed, it cannot be excluded that they reflect more general sensitivity variations (such as for temporal cues, F0 cues, musical stimuli or prosodic stimuli) that surfaced in different experiments of the study.

7.4 Perspective 4. Cue weighting

We hypothesized that in the perception of prosody CI users would rely relatively heavily on temporal cues as opposed to F0 cues, as compared to their NH peers (Hypothesis 4a). According to Hypothesis 4b, this cue weighting would be reflected in speakers' speech output in that F0 related basic prosodic measures of CI users would deviate more from speech of NH peers than temporal prosodic measures. Further, it was predicted that reduced channel interaction, realized by manipulating the steepness of channel filter slopes in vocoder simulations, would improve F0 perception, but not temporal perception (Hypothesis 4c).

Hypothesis 4a was supported for linguistic but not for emotional prosody. In a pair of experiments using vocoder simulations (Chapter 4), cue-weighting was balanced towards a relatively heavy reliance on duration as opposed to F0 cues when compared to the control condition with non-vocoded stimuli, where this weighting was reversed. However, emotional prosody perception, F0 cues were dominant both in the vocoded and in the unvocoded conditions. The supposed relative reliance on temporal (duration) cues was not reflected in basic prosodic measures of CI children's speech output; i.e., F0 parameters were not more deviant than temporal parameters (Chapter 2). Therefore, no support for Hypothesis 4b was found. Reducing channel interaction in vocoder simulations from 5 dB/octave to 160 dB/octave improved emotional and linguistic prosody perception, but only up to 120 dB/octave (performance with 160 dB/octave slopes was lower than with 120 dB/octave slopes).

Increasing the filter slope steepness had more effect on the reliance on F0 than on duration cues in emotion perception, whereby most likely duration cues were little informative for emotion discrimination with the given stimuli to begin with. In focus discrimination (linguistic prosody), however, changing the slopes only improved reliance on temporal cues when steepened from 5 dB/octave to 20 dB/octave and only improved reliance on F0 cues when steepened from 80 dB/octave to 120 dB/octave (from 120 dB/octave to 160 dB performance using that cue reduced again). This pattern of results therefore lends partial support to Hypothesis 4c, since it is confirmed for emotional prosody (with the stimuli used in the relevant experiments), but the effect depends on the filter slope value for linguistic prosody perception.

7.5 Perspective 5. The prosody processing capacities of children

We conjectured that CI children's language acquisition would be delayed relative to that of NH peers by as much as the time until implantation (Hypothesis 5a), but that this delay would be longer for prosody perception than for prosody production (Hypothesis 5b) and longer for linguistic prosody than for emotional prosody (Hypothesis 5c), and finally that CI children would (partially) catch up with increasing implant experience (Hypothesis 5d).

Basic prosodic measures did not significantly deviate from those of hearing-age matched NH peers (Chapter 2), nor did they differ between early and late implanted children. There were, however, tendencies towards deviant capacities, whereby the CI recipients shower lower scores than the control group on some measures but higher scores on other measures. Performance did, however, increase with increasing implant experience. Presuming that the tendencies reflected an actual effect, they might suggest that in prosody production some parameters develop from the onset of stable hearing while others mature from birth. Emotional and linguistic prosody perception were found not to deviate in school-aged children

relative to hearing-age matched NH children (Chapter 6), suggesting either that CI input was sufficient for normal performance or, if they had had a delay, they caught up with their peers. Together, these results do not provide evidence for Hypotheses 5a, 5b and 5c, but they do tentatively support Hypothesis 5d.

7.6 Vocoders and cochlear implants

In some of the chapters in this thesis, vocoded stimuli were used as simulations of cochlear implant percepts. This was done for two major reasons. First of all, vocoders allow the manipulation of signal processing parameters that cannot be varied and therefore neither be tested in actual CI users since some of their settings are fixed. They could, however, be adapted for future implant designs. Second, the usage of vocoders allows for the recruitment of a more easily accessible and audiologically more uniform participant sample.

At the same time, however, as discussed in various chapters, it needs to be pointed out that vocoder simulations provide only an approximation of actual CI hearing. This is for a number of reasons. First, the frequency and spectral resolution of CI hearing roughly correspond to that achieve by a maximum of around eight channels (Friesen, Shannon, Baskent & Wang, 2001) in vocoders and filter slopes of around 5 dB/octave (Litvak, Spahr, Saoji & Fridman, 2007). CI users base their discrimination of these signal dimensions on temporal information, whereas NH listeners can combine F0, spectral, and intensity cues. Second, CI users' amplitude range corresponds to as little as a third of the of NH listeners (Bingabr, Espinoza-Varas & Loizou, 2008). Moreover, very steep filter slopes may activate only a very focused region of neurons, reducing amplitude. Third, the electrode-neuron interface is irregular in that dead regions on the hearing nerve disrupt neuron activation. Fourth, there exists much variation in both the audiological background, device hardware and software and psychophysical and cognitive performance of CI users.

Finally, CI users benefit from their experience with their device and learn to exploit subtle cues that NH listeners ignore when first confronted with vocoded signals.

These limitations beg the question how relevant vocoder simulations are for performance with CIs. In this dissertation, two types of vocoders were used, a 15-channel noise vocoder (Chapters 3 and 4) and an 8-channel sinewave vocoder (Chapter 5). Taking into account the psychophysical differences between vocoder simulations and CI hearing mentioned above, the performances reported in the respective chapters might be optimistic relative to the expected performance by CI users. However, they might still be relatively realistic when considering that CI users' device experience may compensate for their degraded input by more efficiently exploiting the fewer cues that they can rely on. Finally, the relevance of the simulations could be that they most accurately approximate the performance by excellent CI users and the performance with possible future improvements of CIs, such as with increased effective numbers of electrodes and increased effective filter slopes.

7.7 Directions for future research

This thesis clears the ground for several lines of research in the area of language processing by (pediatric) users of cochlear implants. First of all, when prosody processing is studied, the distinction between emotional and linguistic prosody should be taken into account. This thesis suggests that the two types are processed differently, i.e., with different cue weighting strategies. In vocoder simulations, linguistic (focus) prosody discrimination relies relatively heavily on temporal (duration) cues, whereas emotional prosody discrimination seems to rely relatively heavily on F0 cues. The fact, however, that this strategy was not found to differ in actual CI users (children) compared with NH peers, warrants extensions of research in at least two different directions. First of all, different stimuli than the ones used in this

thesis have to be tested, i.e., using more languages, more speakers (for recording stimuli), more stimuli, and more prosody types, such as different emotions and different linguistic functions (e.g., stress and phrasing). Second, more language user groups have to be tested, such as children with a wider variety of chronological and implantation ages (or times in sound), as well as adults, in order to develop a more fine-grained model of language development in the population of CI users, the role of prosody and the interplay of demographic factors involved in that development.

The tendencies towards effects of short cue-specific musical training with vocoders on prosody perception and the cross-domain and cross-cue correlations between music and prosody perception and between temporal and F0 cue reliance suggest that longer training might have a stronger effect. Studies using more extensive cue-specific musical training are therefore warranted. In order to distinguish between within-participant correlations between subtests and true training effects future studies should incorporate a pre-training baseline assessment of performance on musical and prosody tests as well as cue-weighting strategies. Such an effect would pave the way for rehabilitation strategies aimed at improving prosody processing by users of CIs.

As a follow-up on both the study investigating basic prosodic measures of spontaneous speech and the study investigating the accuracy of acted emotions and sentences with specific focus positions by children with CIs, future studies should measure possible deviances in the prosodic parameters of productions in the latter type of study. Whereas we did not find significant differences between basic prosodic measures in spontaneous speech of CI recipients as compared to NH peers, these differences might be present when children are prompted to produce emotional utterances or answer a specific question. That is, the accuracy of their productions, as assessed by an independent panel of NH listeners, might show relatively much variation in parameters used to express those linguistic and paralinguistic attributes. This variation might correlate

with the effectiveness of the attributes conveyed. Such a correlation might reflect a search for the most effective production strategy. If, moreover, CI children's productions are equally as effective as those of NH children but they highlight different prosodic parameters, this would reveal a compensation strategy on the part of the speaker, the listener or both.

Finally, the results on the effect of varying the filter slopes of vocoders on the discriminability of emotions and focus positions when only duration and/or F0 cues were available, could be an incentive to explore the effect of a wide range of filter slopes with different vocoding algorithms on performance in different listening tasks, such as speech understanding and music appreciation. One question would be if the pattern of results whereby the 120 dB/octave condition shows better performance than both steeper and less steep slopes, would be replicated when other tasks and other vocoder algorithms would be used. Another question is what the cause underlying this pattern is and what the information source, if not temporal or spectral hearing, is by means of which listeners can discriminate prosodic minimal pairs. A final question would be whether this theoretical target value can ever be obtained in the processing by CIs and whether their users could perform like the NH listeners using vocoders.

Bibliography

Aantal implantaties in Nederland. (2016). Retrieved from http://www.opciweb.nl/ci-centra/aantal-implantaties-in-nederland/

Amir, O., & Grinfeld, D. (2011). Articulation rate in childhood and adolescence: Hebrew speakers. *Language and Speech, 54*, 225-240.

Anderson, E. S., Oxenham, A. J., Nelson, P. B., & Nelson, D. A. (2012). Assessing the role of spectral and intensity cues in spectral ripple detection and discrimination in cochlear-implant users. *Journal of the Acoustical Society of America, 132*, 3925-3934. Doi: 10.1121/1.4763999

Anderson, I., Weichbold, V., D'Haese, P. S. C., Szuchnik, J., Quevedo, M. S., Martin, J., . . . Phillips, L. (2004). Cochlear implantation in children under the age of two - what do the outcomes show us? *International Journal of Pediatric Otorhinolaryngology, 68*, 425-431. Doi: 10.1016/j.ijporl.2003.11.013

American Speech-Language-Hearing Association (1993). Definitions of communication disorders and variations [Relevant Paper]. Retrieved from www.asha.org/policy

AuBuchon, A. M., Pisoni, D. B., & Kronenberger, W. G. (2015). Verbal processing speed and executive functioning in long-term cochlear implant users. *Journal of Speech, Language, and*

Hearing Research, 58, 151-162. Doi: 10.1044/2014_jslhr-h-13-0259

Baayen R.H., Milin P. 2010. Analyzing reaction times. *International Journal of Psychological Research, 3*: 12–28.

Ball, C., & Ison, K. T. (1984). Speech production with electrocochlear stimulation. *British Journal of Audiology, 18*, 251.

Baskent, D. (2006). Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels. *Journal of the Acoustical Society of America, 120*, 2908-2925. Doi: 10.1121/1.2354017

Baudonck, N., D'Haeseleer, E., Dhooge, I., & van Lierde, K. (2011). Objective vocal quality in children using cochlear implants: a multiparameter approach. *Journal of voice: official journal of the Voice Foundation, 25*, 683-691.

Baudonck, N., van Lierde, K., Dhooge, I., & Corthals, P. (2011). A comparison of vowel productions in prelingually deaf children using cochlear implants, severe hearing-impaired children using conventional hearing aids and normally-hearing children. *Folia Phoniatrica Et Logopaedica, 63*, 154-160.

Baudonck, N., van Lierde, K., D'Haeseleer, E., & Dhooge, I. (2015). Nasalance and nasality in children with cochlear implants and children with hearing aids. *International Journal of Pediatric Otorhinolaryngology, 79*, 541-545. Doi: 10.1016/j.ijporl.2015.01.025

Baum, K. M., & Nowicki, S. (1998). Perception of emotion: Measuring decoding accuracy of adult prosodic cues varying in intensity. *Journal of Nonverbal Behavior, 22*, 89-107. Doi: 10.1023/A:1022954014365

Beadle, E. A. R., McKinley, D. J., Nikolopoulos, T. P., Brough, J., O'Donoghue, G. M., & Archbold, S. M. (2005). Long-term functional outcomes and academic-occupational status in implanted children after 10 to 14 years of cochlear implant use. *Otology & Neurotology, 26*, 1152-1160. Doi: 10.1097/01.mao.0000180483.16619.8f

Beers, M. (1995). *The phonology of normally developing and language-impaired children* (Doctoral dissertation). University of Amsterdam, Amsterdam.

Bergeson, T., & Chin, S. (2008). *Prosodic utterance production*. Unpublished manuscript, Indiana University School of Medicine.

Bilger, R. (1977). Evaluation of subjects presently fitted with implanted auditory protheses. *Annals of Otology Rhinology and Laryngology, 86*, 1-176.

Bingabr, M., Espinoza-Varas, B., & Loizou, P. C. (2008). Simulating the effect of spread of excitation in cochlear implants. *Hearing Research, 241*, 73-79. Doi: 10.1016/j.heares.2008.04.012

Blamey, P., Artieres, F., Başkent, D., Bergeron, F., Beynon, A., Burke, E., . . . Gallégo, S. (2013). Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: an update with 2251 patients. *Audiology and Neurotology, 18*, 36-47.

Blamey, P., Sarant, J. Z., Paatsch, L. E., Barry, J. G., Bow, C. P., Wales, R. J., . . . Tooher, R. (2001). Relationships among speech perception, production, language, hearing loss, and age in children with impaired hearing. *Journal of Speech, Language, and Hearing Research, 44*, 264-285.

Blume, S. S. (1999). Histories of cochlear implantation.. *Social Science and Medicine, 49*, 1257-1268. Doi: 10.1016/s0277-9536(99)00164-1

Boersma, P., & Weenink, D. (2014). *Praat*: doing phonetics by computer [Computer program]. Retrieved from www.praat.org

Boons, T., Brokx, J. P., Dhooge, I., Frijns, J. H., Peeraer, L., Vermeulen, A., . . . van Wieringen, A. (2012). Predictors of spoken language development following pediatric cochlear implantation. *Ear and Hearing, 33*, 617-639.

Boons, T., De Raeve, L., Langereis, M., Peeraer, L., Wouters, J., & van Wieringen, A. (2013). Narrative spoken language skills in severely hearing impaired school-aged children with cochlear

implants. *Research in Developmental Disabilities, 34*, 3833-3846. Doi: S0891-4222(13)00334-X

Burkholder, R. A., & Pisoni, D. B. (2003). Speech timing and working memory in profoundly deaf children after cochlear implantation. Journal of Experimental Child Psychology, 85, 63-88. Doi: 10.1016/S0022-0965(03)00033-X

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America, 103*, 3153-3161.

Busby, P. A., Tong, Y. C., & Clark, G. M. (1993). The perception of temporal modulations by cochlear implant patients. *Journal of the Acoustical Society of America, 94*, 124-131. Doi: 10.1121/1.408212

Campisi, P., Low, A., Papsin, B., Mount, R., Cohen-Kerem, R., & Harrison, R. (2005). Acoustic analysis of the voice in pediatric cochlear implant recipients: a longitudinal study. *Laryngoscope, 115*, 1046-1050. Doi: 10.1097/01.MLG.0000163343.10549.4C

Carlyon, R. P., Deeks, J. M., & McKay, C. M. (2010). The upper limit of temporal pitch for cochlear-implant listeners: stimulus duration, conditioner pulses, and the number of electrodes stimulated. *Journal of the Acoustical Society of America, 127*, 1469-1478. Doi: 10.1121/1.3291981

Chatterjee, M., & Peng, S. C. (2008). Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition. *Hearing Research, 235*, 143-156. Doi: S0378-5955(07)00264-X

Chen, J. K.-C., Chuang, A. Y. C., McMahon, C., Hsieh, J.-C., Tung, T.-H., & Li, L. P.-H. (2010). Music training improves pitch perception in prelingually deafened children with cochlear implants. *Pediatrics 125*, 793-800.Doi: 10.1542/peds.2008-3620

Christiansen, J. B., & Leigh, I. W. (2004). Children with cochlear implants: Changing parent and deaf community perspectives.

Archives of Otolaryngology–Head & Neck Surgery, 130, 673-677.

Coelho, A. C., Brasolotto, A. G., & Bevilacqua, M. C. (2012). Systematic analysis of the benefits of cochlear implants on voice production. *Jornal da Sociedade Brasileira de Fonoaudiologia, 24*, 395-402. Doi: S2179-64912012000400018

Colletti, L., Mandalà, M., Zoccante, L., Shannon, R. V., & Colletti, V. (2011). Infants versus older children fitted with cochlear implants: performance over 10 years. *International Journal of Pediatric Otorhinolaryngology, 75*, 504-509.

Connor, C. M., Craig, H. K., Raudenbush, S. W., Heavner, K., & Zwolan, T. A. (2006). The age at which young deaf children receive cochlear implants and their vocabulary and speech-production growth: Is there an added value for early implantation? *Ear and Hearing*, *27*, 628-644.

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*, 618-626. Doi: 10.1097/Aud.0b013e318174e787

Crew, J. D., Galvin, J. J., & Fu, Q. J. (2012). Channel interaction limits melodic pitch perception in simulated cochlear implants. *The Journal of the Acoustical Society of America, 132*, EL429-EL435. Doi: 10.1121/1.4758770

Crosson, J., & Geers, A. (2001). Analysis of narrative ability in children with cochlear implants. *Ear and Hearing, 22*, 381-394.

Cysneiros, H. R. S., Leal, M. d. C., Lucena, J. A., & Muniz, L. F. (2016). Relationship between auditory perception and vocal production in cochlear implantees: a systematic review. *Codas. 28*. Doi: 10.1590/2317-1782/20162015165

de Hoog, B. E., Langereis, M. C., van Weerdenburg, M., Keuning, J., Knoors, H., & Verhoeven, L. (2016). Auditory and verbal memory predictors of spoken language skills in children with

cochlear implants. *Research in Developmental Disabilities, 57*, 112-124. Doi: 10.1016/j.ridd.2016.06.019

Deafness and hearing loss. (2015, March). Retrieved from http://www.who.int/mediacentre/factsheets/fs300/en/

Deliyski, D. D. (1993). *Acoustic model and evaluation of pathological voice production*. Paper presented at the Eurospeech Conference.

Dellwo, V., Fourcin, A., & Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. In J. Trouvain, W. J. Barry., *Proceedings of the International Congress of Phonetic Sciences (ICPhS) XVI, August 6–10, Saarbrücken* (pp. 1129–1132).

Djourno, A., & Eyries, C. (1957). [Auditory prosthesis by means of a distant electrical stimulation of the sensory nerve with the use of an indwelt coiling]. *Presse Medicale, 65*, 1417.

Dorman, M. F., & Loizou, P. C. (1997). Speech intelligibility as a function of the number of channels of stimulation for normally-hearing listeners and patients with cochlear implants. *American Journal of Otology, 18*, S113-S114.

Dorman, M. F., & Loizou, P. C. (1998). The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normally-hearing subjects using simulations of processors with two to nine channels. *Ear and Hearing, 19*, 162–166.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America, 102*, 2403–2411.

Dornan, D., Hickson, L., Murdoch, B., & Houston, T. (2009). Longitudinal study of speech perception, speech, and language for children with hearing loss in an auditory-verbal therapy program. *Volta Review, 109*, 61-85.

Drennan, W. R., Won, J. H., Nie, K., Jameyson, E., & Rubinstein, J. T. (2010). Sensitivity of psychophysical measures to signal processor modifications in cochlear implant users. *Hearing Research, 262*, 1–8.

Driscoll, V. D., Oleson, J., Jiang, D., & Gfeller, K. (2009). Effects of training on recognition of musical instruments presented through cochlear implant simulations. *Journal of the American Academy of Audiology, 20*, 71-82.

Edwards, L. C. (2007). Children with cochlear implants and complex needs: a review of outcome research and psychological practice. *Journal of deaf studies and deaf education, 12*, 258-268.

Eisen, M. D. (2009). The history of cochlear implants. In J. Niparko (Ed.), *Cochlear implants: Principles & Practices.* Lippincott Williams & Wilkins.

Evans, M. K., & Deliyski, D. D. (2007). Acoustic voice analysis of prelingually deaf adults before and after cochlear implantation. *Journal of Voice, 21*(6), 669-682. Doi: S0892-1997(06)00089-0

Fagan, M. K., & Pisoni, D. B. (2010). Hearing experience and receptive vocabulary development in deaf children with cochlear implants. *The Journal of Deaf Studies and Deaf Education, 15*, 149-161. Doi: 10.1093/deafed/enq001

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: implications for cochlear implants. *Journal of the Acoustical Society of America, 108*, 1877–1887.

Fikkert, P. (1994). *On the acquisition of prosodic structure* (HIL dissertations 6). The Hague: Holland Academic Graphics.

Finke, M., Buchner, A., Ruigendijk, E., Meyer, M., & Sandmann, P. (2016). On the relationship between auditory cognition and speech intelligibility in cochlear implant users: An ERP study. *Neuropsychologia, 87*, 169-181. Doi: 10.1016/j.neuropsychologia.2016.05.019

Firszt, J. B., Koch, D. B., Downing, M., & Litvak, L. (2007). Current steering creates additional pitch percepts in adult cochlear implant recipients. *Otology & Neurotology, 28*, 629-636.

Fizmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). New Jersey: John Wiley & Sons, Inc.

Flipsen, P. (2002). Longitudinal changes in articulation rate and phonetic phrase length in children with speech delay. *Journal of Speech Language and Hearing Research, 45*, 100-110. Doi: 10.1044/1092-4388(2002/008)

Fourcin, A., Abberton, E., Richardson, K., & Shaw, T. (2011). Aspects of voice measurement with young users of cochlear implants. *Seminars in Hearing, 32*, 42-52.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. USA: Sage Publications, Inc.

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. *Journal of the Acoustical Society of America, 110*, 1150-1163.

Fu, Q.-J. (2013). AngelSim: Cochlear implant and hearing loss simulator [Computer program], Version 1.08.01. Retrieved from http://www.tigerspeech.com/angelsim/

Fu, Q. J., Galvin, J. J., Wang, X. S., & Wu, J. L. (2015). Benefits of music training in mandarin-speaking pediatric cochlear implant users. *Journal of Speech Language and Hearing Research, 58*, 163-169. Doi: 10.1044/2014_Jslhr-H-14-0127

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, 19-27. Doi: 10.1007/s10162-004-5024-3

Fu, Q. J., Nogaki, G., & Galvin, J. J., 3rd. (2005). Auditory training with spectrally shifted speech: implications for cochlear

implant patient auditory rehabilitation. *Journal of the Association for Research in Otolaryngology, 6*, 180–189.

Fu, Q. J., & Shannon, R. V. (2002). Frequency mapping in cochlear implants. *Ear and Hearing, 23*, 339-348. Doi: 10.1097/01.Aud.0000027432.18827.07

Fuller, C. D., Galvin, J. J., 3rd, Maat, B., Free, R. H., & Baskent, D. (2014). The musician effect: does it persist under degraded pitch conditions of cochlear implant simulations? *Frontiers in Neuroscience, 8*, 179. Doi: 10.3389/fnins.2014.00179

Galvin, J. J., Fu, Q. J., & Nogaki, G. (2007). Melodic contour identification by cochlear implant listeners. *Ear and Hearing, 28*, 302-319. Doi: 10.1097/01.aud.0000261689.35445.20

Galvin, J. J., Fu, Q. J., & Shannon, R. V. (2009). Melodic Contour Identification and Music Perception by Cochlear Implant Users. Annals of the New York Academy of Sciences, 1169, 518-533. Doi: 10.1111/j.1749-6632.2009.04551.x

Galvin, J. J., Eskridge, E., Oba, S., & Fu, Q.-J. (2012). Melodic contour identification training in cochlear implant users with and without a competing instrument. *Seminars in Hearing, 33*, 399-409.

Geers, A. (2003). Predictors of reading skill development in children with early cochlear implantation. *Ear and Hearing, 24*, 59S-68S.

Geers, A., Brenner, C., & Davidson, L. (2003). Factors associated with development of speech perception skills in children implanted by age five. *Ear and Hearing, 24*, 24s-35s. Doi: 10.1097/01.Aud.0000051687.99218.0f

Geers, A., Davidson, L. S., Uchanski, R. M., & Nicholas, J.G. (2013). Interdependence of linguistic and indexical speech perception skills in school-age children with early cochlear implantation. *Ear and Hearing, 34*, 562–574. Doi:10.1097/Aud.0b013e31828d2bd6.

Geers, A., & Moog, J. (1994). Effectiveness of cochlear implants and tactile aids for deaf-children - the sensory aids study at central institute for the deaf - foreword. *Volta Review, 96*, R5-R6.

Geers, A., Nicholas, J. G., & Sedey, A. L. (2003). Language skills of children with early cochlear implantation. *Ear and Hearing, 24*, 46s-58s. Doi: 10.1097/01.Aud.0000051689.57380.1b

Geers, A., Nicholas, J., Tobey, E., & Davidson, L. (2016). Persistent language delay versus late language emergence in children with early cochlear implantation. *Journal of Speech, Language, and Hearing Research, 59*, 155-170.

Geers, A., Tobey, E., Moog, J., & Brenner, C. (2008). Long-term outcomes of cochlear implantation in the preschool years: From elementary grades to high school. *International Journal of Audiology, 47*, S21-S30. Doi: 10.1080/14992020802339167

Gfeller, K., Turner, C., Mehr, M., Woodworth, G., Fearn, R., Knutson, J. F., . . . Stordahl, J. (2002). Recognition of familiar melodies by adult cochlear implant recipients and normally-hearing adults. *Cochlear Implants International, 3*, 29-53. Doi: 10.1002/cii.50

Gfeller, K., Witt, S., Woodworth, G., Mehr, M. A., & Knutson, J. (2002). Effects of frequency, instrumental family, and cochlear implant type on timbre recognition and appraisal. *Annals of Otology, Rhinology and Laryngology, 111*, 349-356.

Giezen, M. R., Escudero, P., & Baker, A. (2010). Use of acoustic cues by children with cochlear implants. *Journal of Speech Language and Hearing Research, 53*, 1440-1457. Doi: 10.1044/1092-4388(2010/09-0252)

Gilbers, S., Fuller, C., Gilbers, D., Broersma, M., Goudbeek, M., Free, R., & Baskent, D. (2015). Normally-hearing listeners' and cochlear implant users' perception of pitch cues in emotional speech. *Iperception, 6*, 1-19. Doi: 10.1177/0301006615599139

Glick, H., & Sharma, A. (2016). Cross-modal plasticity in developmental and age-related hearing loss: clinical

implications. *Hearing Research, 343*, 191-201. Doi: 10.1016/j.heares.2016.08.012

Goedemans, R., van der Hulst, H., & Visch, E. (Eds) (1996). Stress patterns of the world. Part I: Background. Leiden: Holland Institute of Generative Linguistics.

Goffman, L., Ertmer, D. J., & Erdle, C. (2002). Changes in speech production in a child with a cochlear implant: Acoustic and kinematic evidence. *Journal of Speech Language and Hearing Research, 45*, 891-901. Doi: 10.1044/1092-4388(2002/072)

Goldman-Eisler, F. (1968). *Psycholinguistics. Experiments in spontaneous speech*. London: Academic Press.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffen.

Goy, H., Fernandes, D. N., Pichora-Fuller, M. K., & van Lieshout, P. (2013). Normative voice data for younger and older adults. *Journal of Voice, 27*, 545-555. Doi: 10.1016/j.jvoice.2013.03.002

Gravel, J. S., & Tocci, L. L. (1998). Setting the stage for universal newborn hearing screening. In: L. G. Spivak (Ed.), *Universal Newborn Hearing Screening*. New York, NY, USA: Thieme Medical Publishers, Inc.

Graven, S. N., & Browne, J. V. (2008). Auditory development in the fetus and infant. *Newborn and infant nursing reviews, 8*, 187-193.

Green, T., Faulkner, A., & Rosen, S. (2004). Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. *Journal of the Acoustical Society of America, 116*, 2298–2310.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders, 39*, 350-365. Doi: S0021-9924(06)00058-X

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying

syllable production. *Brain and Language, 96,* 280-301. Doi: S0093-934X(05)00115-X

Hammer, A. (2010). *The acquisition of verbal morphology in cochlear-implanted and specific language impaired children* (Doctoral dissertation). LOT, Utrecht.

Harrison, R. V., Gordon, K. A., & Mount, R. J. (2005). Is there a critical period for cochlear implantation in congenitally deaf children? Analyses of hearing and speech perception performance after implantation. *Developmental Psychobiology, 46*, 252-261. Doi: 10.1002/Dev.20052

Hassan, S. M., Malki, K. H., Mesallam, T. A., Farahat, M., Bukhari, M., & Murry, T. (2011a). The effect of cochlear implantation and post-operative rehabilitation on acoustic voice analysis in post-lingual hearing impaired adults. *European Archives of Oto-Rhino-Laryngology, 268*, 1437-1442. Doi: 10.1007/s00405-011-1501-6

Hassan, S. M., Malki, K. H., Mesallam, T. A., Farahat, M., Bukhari, M., & Murry, T. (2011b). The effect of cochlear implantation on nasalance of speech in postlingually hearing-impaired adults. *Journal of Voice, 26,* 669.e17–669.e22. Doi: S0892-1997(11)00120-2

Hayes, H., Geers, A. E., Treiman, R., & Moog, J. S. (2009). Receptive vocabulary development in deaf children with cochlear implants: achievement in an intensive auditory-oral educational setting. *Ear and Hearing, 30,* 128-135. Doi: 10.1097/AUD.0b013e3181926524

Henry, B. A., & Turner, C. W. (2003). The resolution of complex spectral patterns by cochlear implant and normally-hearing listeners. *Journal of the Acoustical Society of America, 113*, 2861–2873.

Higgins, M. B., McCleary, E. A., & Schulte, L. (2001). Articulatory changes with short-term deactivation of the cochlear implants of two prelingually deafened children. *Ear and Hearing, 22*, 29-45.

Hillenbrand, J. (1987). A methodological study of perturbation and additive noise in synthetically generated voice signals. *Journal of Speech and Hearing Research, 30*, 448-461.

Hocevar-Boltezar, I., Radsel, Z., Vatovec, J., Geczy, B., Cernelc, S., Gros, A., . . . Zargi, M. (2006). Change of phonation control after cochlear implantation. *Otology & Neurotology, 27*, 499-503.

Hocevar-Boltezar, I., Vatovec, J., Gros, A., & Zargi, M. (2005). The influence of cochlear implantation on some voice parameters. *International Journal of Pediatric Otorhinolaryngology, 69*, 1635-1640. Doi: 10.1016/j.ijporl.2005.03.045

Holden, L. K., Finley, C. C., Firszt, J. B., Holden, T. A., Brenner, C., Potts, L. G., . . . Heydebrand, G. (2013). Factors affecting open-set word recognition in adults with cochlear implants. *Ear and Hearing, 34*, 342-360.

Holler, T., Campisi, P., Allegro, J., Chadha, N. K., Harrison, R. V., Papsin, B., & Gordon, K. (2010). Abnormal voicing in children using cochlear implants. *Archives of Otolaryngology – Head and Neck Surgery, 136*, 17-21. Doi: 10.1001/archoto.2009.194

Holt, R. F., & Svirsky, M. A. (2008). An exploratory look at pediatric cochlear implantation: Is earliest always best? *Ear and Hearing, 29*, 492-511. Doi: 10.1097/Aud.0b013e31816c409f

Hopyan, T., Manno, 3rd, F. A., Papsin, B. C., & Gordon, K. A. (2016). Sad and happy emotion discrimination in music by children with cochlear implants. *Child Neuropsychology, 22*: 366-380, 1-15. Doi: 10.1080/09297049.2014.992400

Horga, D., & Liker, M. (2006). Voice and pronunciation of cochlear implant speakers. *Clinical Linguistics & Phonetics, 20*, 211-217. Doi: XX61748T86823365

Hsu, H. W., Fang, T. J., Lee, L. A., Tsou, Y. T., Chen, S. H., & Wu, C. M. (2013). Multidimensional evaluation of vocal quality in children with cochlear implants: a cross-sectional, case-

controlled study. *Clinical Otolaryngology, 39*, 32-38. Doi: 10.1111/coa.12213

IBM Corp, Released 2014. SPSS Statistics for Windows, Version 23.0. Armonk, NY, USA: IBM Corp.

Johnson, C., & Goswami, U. (2010). Phonological awareness, vocabulary, and reading in deaf children with cochlear implants. *Journal of Speech, Language, and Hearing Research, 53*, 237-261.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548-567.

Jongkees, L. (1978). Doven weer horen. *Nederlands Tijdschrift Voor Geneeskunde, 122*, 1621.

Kadi-Hanifi, K., & Howell, P. (1992). Syntactic analysis of the spontaneous speech of normally fluent and stuttering children. *Journal of Fluency Disorders, 17*, 151-170. Doi: 10.1016/0094-730x(92)90008-E

Kalathottukaren, R. T., Purdy, S. C., & Ballard, E. (2015). Prosody perception and musical pitch discrimination in adults using cochlear implants. *International Journal of Audiology, 54*, 444-452, 444-452. Doi: 10.3109/14992027.2014.997314

Kane, M. O. L., Schopmeyer, B., Mellon, N. K., Wang, N.-Y., & Niparko, J. K. (2004). Prelinguistic communication and subsequent language acquisition in children with cochlear implants. *Archives of Otolaryngology–Head & Neck Surgery, 130*, 619-623.

Kang, R., Nimmons, G. L., Drennan, W., Longnion, J., Ruffin, C., Nie, K. B., . . . Rubinstein, J. (2009). Development and validation of the university of washington clinical assessment of music perception test. *Ear and Hearing, 30*, 411-418.

Kent, R. D. (1976). Anatomical and neuromuscular maturation of speech mechanism - evidence from acoustic studies. *Journal of Speech and Hearing Research, 19*, 421-447.

Kishon-Rabin, L., Taitelbaum, R., Tobin, Y., & Hildesheimer, M. (1999). The effect of partially restored hearing on speech production of postlingually deafened adults with multichannel cochlear implants. *Journal of the Acoustical Society of America, 106*, 2843-2857.

Knoors, H. (2008). Cochleaire implantatie bij dove kinderen: effecten op de ontwikkeling en mogelijke gevolgen voor pedagogisch beleid. In T. van der Lem & G. Spaai (Eds.), *Effecten van cochleaire implantatie bij kinderen: een breed perspectief*. Deventer: van Tricht.

Kong, Y. Y., Cruz, R., Jones, J. A., & Zeng, F. G. (2004). Music perception with temporal cues in acoustic and electric hearing. *Ear and Hearing, 25*, 173-185. Doi: 10.1097/01.Aud.0000120365.97792.2f

Kraayeveld, H. (1997). *Idiosyncrasy in prosody* (Doctoral dissertation). University of Nijmegen.

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication, 34*, 391-405.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America, 100*, 2425-2438.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*, F13-F21.

Kwon, B. J., & van den Honert, C. (2006). Effect of electrode configuration on psychophysical forward masking in cochlear implant listeners. *The Journal of the Acoustical Society of America, 119*, 2994-3002.

Ladd, B. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Lamoré, P. (2016). Doofheid - Algemene informatie. Retrieved from http://www.audiologieboek.nl/htm/hfd7/7-3-1.htm#7314

Lane, H., Matthies, M., Perkell, J., Vick, J., & Zandipour, M. (2001). The effects of changes in hearing status in cochlear implant users on the acoustic vowel space and CV coarticulation. *Journal of Speech, Language, and Hearing Research, 44*, 552-563.

Lane, H., Perkell, J., Wozniak, J., Manzella, J., Guiod, P., Matthies, M., . . . Vick, J. (1998). The effect of changes in hearing status on speech sound level and speech breathing: a study conducted with cochlear implant users and NF-2 patients. *Journal of the Acoustical Society of America, 104*, 3059-3069.

Laneau, J., Moonen, M., & Wouters, J. (2006). Factors affecting the use of noise-band vocoders as acoustic models for-pitch perception in cochlear implants. *Journal of the Acoustical Society of America, 119*, 491–506.

Laneau, J., & Wouters, J. (2004). Multichannel place pitch sensitivity in cochlear implant recipients. *Journal of the Association for Research in Otolaryngology, 5*, 285-294. Doi: 10.1007/s10162-004-4049-y

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

Lang, H. G. (2002). Higher education for deaf students: Research priorities in the new millennium. *Journal of deaf studies and deaf education, 7*, 267-280.

Lassaletta, L., Castro, A., Bastarrica, M., Pérez-Mora, R., Madero, R., De Sarriá, J., & Gavilán, J. (2007). Does music perception have an impact on quality of life following cochlear implantation? *Acta Oto-Laryngologica, 127*, 682-686.

Lawrence, M. (1964). Direct stimulation of auditory nerve fibers. *Archives of otolaryngology, 80*, 367-368.

Lazard, D. S., Vincent, C., Venail, F., van De Heyning, P., Truy, E., Sterkers, O., … Blamey, P. J. (2012). Pre-, per-and postoperative factors affecting performance of postlinguistically deaf adults using cochlear implants: a new

conceptual model over time. *PLoS ONE, 7*, e48739. Doi:10.1371/journal.pone.0048739.

Leder, S. B., Spitzer, J. B., Kirchner, J. C., Flevaris-Phillips, C., Milner, P., & Richardson, F. (1987). Speaking rate of adventitiously deaf male cochlear implant candidates. *Journal of the Acoustical Society of America, 82*, 843-846.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, I. (1976). Suprasegmental features of speech. In N. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 225-239). London: Academic Press.

Leigh, J., Dettman, S., Dowell, R., & Briggs, R. (2013). Communication development in children who receive a cochlear implant by 12 months of age. *Otology & Neurotology, 34*, 443-450.

Lenden, J. M., & Flipsen, P. (2007). Prosody and voice characteristics of children with cochlear implants. *Journal of Communication Disorders, 40*, 66-81. Doi: 10.1016/j.jcomdis.2006.04.004

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Levelt, C. (1994). *On the acquisition of place* (HIL dissertations 8). Dordrecht: ICG Printing.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41-104.

Levitin, D. J., Cole, K., Chiles, M., Lai, Z., Lincoln, A., & Bellugi, U. (2004). Characterizing the musical phenotype in individuals with Williams syndrome. *Child Neuropsychology, 10*, 223-247.

Lieberman, P. (1986). The acquisition of intonation by infants: physiology and neural control. In C. H. Ltd. (Ed.), *Intonation in discourse* (pp. 239-257). London: Johns-Lewis, E. C.

Limb, C. J., & Roy, A. T. (2014). Technological, biological, and acoustical constraints to music perception in cochlear implant users. *Hearing Research, 308*, 13-26. Doi: 10.1016/j.heares.2013.04.009

Litvak, L. M., Spahr, A. J., Saoji, A. A., & Fridman, G. Y. (2007). Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners. *Journal of the Acoustical Society of America, 122*, 982-991. Doi: 10.1121/1.2749413

Liu, H. M., Kuhl, P. K., & Tsao, F. M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science, 6*, F1-F10.

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*, 662-674.

Loizou, P. C. (2006). Speech processing in vocoder-centric cochlear implants. *Advances in Oto-Rhino-Laryngology, 64*, 109-143.

Looi, V., Gfeller, K., & Driscoll, V. (2012). Music Appreciation and Training for Cochlear Implant Recipients: A Review. *Seminars in Hearing, 33*, 307-334. Doi: 10.1055/s-0032-1329222

Luo, X., Fu, Q. J., & Galvin, J. J., 3rd. (2007). Vocal emotion recognition by normally-hearing listeners and cochlear implant users. *Trends in amplification, 11*, 301-315. Doi: 11/4/301

Luo, X., Fu, Q. J., Wei, C. G., & Cao, K. L. (2008). Speech recognition and temporal amplitude modulation processing by mandarin-speaking cochlear implant users. *Ear and Hearing, 29*, 957-970. Doi: 10.1097/AUD.0b013e3181888f61

Luo, X., Masterson, M. E., & Wu, C. C. (2014). Melodic interval perception by normally-hearing listeners and cochlear implant users. *Journal of the Acoustical Society of America, 136*, 1831-1844. Doi: 10.1121/1.4894738

Lyxell, B., Wass, M., Sahlen, B., Samuelsson, C., Asker-Arnason, L., Ibertsson, T., . . . Hallgren, M. (2009). Cognitive development, reading and prosodic skills in children with cochlear implants. *Scandinavian Journal of Psychology, 50*, 463-474. Doi: 10.1111/j.1467-9450.2009.00754.x

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98*, 185-199.

Marschark, M., Lang, H. G., & Albertini, J. A. (2002). *Educating deaf students: From research to practice.* New York, NY, USA: Oxford University Press.

Marschark, M., Rhoten, C., & Fabich, M. (2007). Effects of cochlear implants on children's reading and academic achievement. *Journal of Deaf Studies and Deaf Education, 12*, 269-282. Doi: 10.1093/deafed/enm013

Marx, M., James, C., Foxton, J., Capber, A., Fraysse, B., Barone, P., & Deguine, O. (2014). Speech prosody perception in cochlear implant users with and without residual hearing. *Ear and Hearing, 36*, 239-248. Doi: 10.1097/AUD.0000000000000105

Massida, Z., Belin, P., James, C., Rouger, J., Fraysse, B., Barone, P., & Deguine, O. (2011). Voice discrimination in cochlear-implanted deaf subjects. *Hearing Research, 275*, 120-129.

McConkey Robbins, A., Green, J. E., & Waltzman, S. B. (2004). Bilingual oral language proficiency in children with cochlear implants. *Archives of Otolaryngology -- Head and Neck Surgery, 130*, 644-647. Doi: 10.1001/archotol.130.5.644

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*, 143-178.

Meister, H. (2011). Processing prosodic cues with cochlear implants. *Sprache-Stimme-Gehör, 35*, E99-E104. Doi: 10.1055/s-0031-1284405

Meister, H., Fursen, K., Streicher, B., Lang-Roth, R., & Walger, M. (2016). The use of voice cues for speaker gender recognition in cochlear implant recipients. *Journal of Speech, Language, and*

Hearing Research, 59, 546-556. Doi: 10.1044/2015_jslhr-h-15-0128

Meister, H., Tepeli, D., Wagner, P., Hess, W., Walger, M., von Wedel, H., & Lang-Roth, R. (2007). Experimente zur Perzeption prosodischer Merkmale mit Kochleaimplantaten [Experiments on prosody perception with cochlear implants]. *HNO, 55*, 264-270. Doi: 10.1007/s00106-006-1452-1

Ménard, L., Polak, M., Denny, M., Burton, E., Lane, H., Matthies, M. L., . . . Vick, J. (2007). Interactions of speaking condition and auditory feedback on vowel production in postlingually deaf adults with cochlear implants. *Journal of the Acoustical Society of America, 121*, 3790-3801. Doi: 10.1121/1.2710963

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied Multivariate Research*. Thousand Oaks, CA, USA: Sage Publications.

Mitchell, R. E., & Karchmer, M. A. (2004). Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies, 4*, 138-163.

Monini, S., Banci, G., Barbara, M., Argiro, M. T., & Filipo, R. (1997). Clarion cochlear implant: short-term effects on voice parameters. *American Journal of Otology, 18*, 719-725.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant behavior and development, 16*, 495-500.

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced in utero affects vowel perception after birth: a two-country study. *Acta Pediatrica, 102*, 156-160.

Moon, I. S., Park, S., Kim, H.-N., Lee, W.-S., Kim, S. H., Kim, J.-H., & Choi, J. Y. (2014). Is there a deafness duration limit for cochlear implants in post-lingual deaf adults? *Acta Oto-Laryngologica, 134*, 173-180.

Moore, B. C. (2003). Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. *Otology & Neurotology, 24*, 243-254.

Moreno, S., & Bidelman, G. M. (2014). Examining neural plasticity and cognitive benefit through the unique lens of musical training. *Hearing Research, 308*, 84-97. Doi: 10.1016/j.heares.2013.09.012

Most, T., & Michaelis, H. (2012). Auditory, visual, and auditory-visual perceptions of emotions by young children with hearing loss versus children with normal hearing. *Journal of Speech Language and Hearing Research, 55*, 1148-1162. Doi: 10.1044/1092-4388(2011/11-0060)

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous wave-form processing techniques for text-to- speech synthesis using diphones. *Speech Communication, 9*, 453–467.

Moulines, E., & Verhelst, W. (1995). Time-domain and frequency-domain techniques for prosodic modification of speech. In: W. B. Kleijn & K. K. Paliwal (Eds.), *Speech coding and synthesis* (pp. 519-555). New York, NY, USA; Elsevier Science Inc.

Nakata, T., Trehub, S. E., & Kanda, Y. (2012). Effect of cochlear implants on children's perception and production of speech prosody. *Journal of the Acoustical Society of America, 131*, 1307-1314. Doi: 10.1121/1.3672697

Netten, A. P., Dekker, F. W., Rieffe, C., Soede, W., Briaire, J. J., & Frijns, J. H. (2017). Missing data in the field of otorhinolaryngology and head & neck surgery: need for improvement. *Ear and Hearing, 38*, 1-6.

Neumeyer, V., Harrington, J., & Draxler, C. (2010). An acoustic analysis of the vowel space in young and old cochlear-implant speakers. *Clinical Linguist & Phonetics, 24*, 734-741.

Nguyen, L. H., Allegro, J., Low, A., Papsin, B., & Campisi, P. (2008). Effect of cochlear implantation on nasality in children. *Ear, Nose, and Throat Journal, 87*, 138, 140-133.

Nikolopoulos, T. P., Dyar, D., Archbold, S., & O'Donoghue, G. M. (2004). Development of spoken language grammar following cochlear implantation in prelingually deaf children. *Archives of Otolaryngology–Head & Neck Surgery, 130*, 629-633.

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. *Journal of the Acoustical Society of America, 95*, 1085-1099. Doi:10.1121/1.408469

Niparko, J. K., Lingua, C., & Carpenter, R. M. (2009). Assessment of candidacy for cochlear implantation. In J. Niparko (Ed.), *Cochlear implants: Principles & Practices*. Philadelphia, PA, USA: Lippincott Williams & Wilkins.

Niparko, J. K., Tobey, E. A., Thal, D. J., Eisenberg, L. S., Wang, N. Y., Quittner, A. L., & Fink, N. E. (2010). Spoken language development in children following cochlear implantation. *Journal of the American Medical Association, 303*, 1498-1506. Doi: 303/15/1498

Nittrouer, S., Caldwell-Tarr, A., & Lowenstein, J. H. (2013). Working memory in children with cochlear implants: Problems are in storage, not processing. *International Journal of Pediatric Otorhinolaryngology, 77*, 1886-1898.

O'Halpin, R. (2009). *The perception and production of stress and intonation by children with cochlear implants* (Doctoral dissertations). London: UCL. Retrieved from http://eprints.ucl.ac.uk/20406/1/20406.pdf

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development, 59*, 441-449.

Osberger, M. J. (1994). Speech intelligibility of children with cochlear implants. *Volta Review, 96*, 169-180.

Osberger, M. J., & McGarr, N. S. (1982). Speech production characteristics of the hearing impaired. In N. Lass (Ed.), *Speech and language: advances in basic research and practice* (pp. 227-288). New York, NY, USA: Academic Press.

Oster, A.-M. (1987). Some effects of cochlear implantation on speech production. *Speech Transmission Laboratory – Quarterly Progress and Status Report, 28*, 81-89.

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research, 308*, 98-108.

Peng, S. C. (2005). *Perception and production of speech intonation in pediatric cochlear implant recipients and children with normal hearing* (Doctoral dissertation). Iowa City, IA, USA: University of Iowa.

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 Neuro-Otology, 14*, 327–337. Doi:10.1159/000212112.

Peng, S. C., Tomblin, J. B., Cheung, H., Lin, Y. S., & Wang, L. S. (2004). Perception and production of mandarin tones in prelingually deaf children with cochlear implants. *Ear and Hearing, 25*, 251-264. Doi: 00003446-200406000-00006

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*, 336-351. Doi: 10.1097/AUD.0b013e318168d94d

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*, 345-354.

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders - The Montreal battery of evaluation of amusia. *Neurosciences and Music, 999*, 58-75. Doi: 10.1196/annals.1284.006

Perkell, J., Lane, H., Denny, M., Matthies, M. L., Tiede, M., Zandipour, M., . . . Burton, E. (2007). Time course of speech changes in response to unanticipated short-term changes in

hearing state. *Journal of the Acoustical Society of America, 121*, 2296-2311. Doi: 10.1121/1.2642349

Perkell, J., Lane, H., Svirsky, M., & Webster, J. (1992). Speech of cochlear implant patients: a longitudinal study of vowel production. *Journal of the Acoustical Society of America, 91*, 2961-2978.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication, 22*, 227-250.

Perrin, E., Berger-Vachon, C., Topouzkhanian, A., Truy, E., & Morgon, A. (1999). Evaluation of cochlear implanted children's voices. *International Journal of Pediatric Otorhinolaryngology, 47*, 181-186.

Peterson, N. R., Pisoni, D. B., & Miyamoto, R. T. (2010). Cochlear implants and spoken language processing abilities: Review and assessment of the literature. *Restorative neurology and neuroscience, 28*, 237-250.

Pfingst, B. E., Zwolan, T. A., & Holloway, L. A. (1997). Effects of stimulus configuration on psychophysical operating levels and on speech recognition with cochlear implants. *Hearing Research, 112*, 247-260. Doi: 10.1016/S0378-5955(97)00122-6

Pfingst, B. E., Franck, K. H., Xu, L., Bauer, E. M., & Zwolan, T. A. (2001). Effects of electrode configuration and place of stimulation on speech perception with cochlear prostheses. *Journal of the Association for Research in Otolaryngology, 2*, 87-103.

Pisoni, D. B. (2000). Cognitive factors and cochlear implants: Some thoughts on perception, learning, and memory in speech perception. *Ear and Hearing, 21*, 70-78. Doi; 10.1097/00003446-200002000-00010

Pisoni, D. B., Kronenberger, W. G., Roman, A. S., & Geers, A. E. (2011). Measures of digit span and verbal rehearsal speed in

deaf children after more than 10 years of cochlear implantation. Ear *and Hearing, 32*, 60S-74S. Doi: 10.1097/AUD.0b013e3181ffd58e

Poissant, S. F., Peters, K. A., & Robb, M. P. (2006). Acoustic and perceptual appraisal of speech production in pediatric cochlear implant users. *International Journal of Pediatric Otorhinolaryngology, 70*, 1195-1203. Doi: 10.1016/j.ijporl.2005.12.008

Punch, R., & Hyde, M. (2011). Social participation of children and adolescents with cochlear implants: a qualitative analysis of parent, teacher, and child interviews. *Journal of Deaf Studies and Deaf Education, 16*, 474-493. Doi: 10.1093/deafed/enr001

Purcell, D. W., & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, *119*, 2288-2297.

Qin, M. K., & Oxenham, A..J. (2005). Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. *Ear and Hearing*, *26*, 451–460.

Richardson, L. M., Busby, P. A., Blamey, P. J., Dowell, R. C., & Clark, G. M. (1993). The effects of auditory feedback from the nucleus cochlear implant on the vowel formant frequencies produced by children and adults. *Ear and Hearing*, *14*, 339-349.

Rietveld, A. C. M., & van Heuven, V. J. (2016). *Algemene fonetiek* (4th ed.). Bussum: Coutinho.

Robinson, K. (1998). Implications of developmental plasticity for the language acquisition of deaf children with cochlear implants. *International Journal of Pediatric Otorhinolaryngology, 46*, 71-80.

Saoji, A., Litvak, L., Emadi, G., Spahr, T., & Greenslade, K. (2005). *Spectral modulation transfer functions in cochlear implant listeners*. Poster presented at the Conference on Implantable Auditory Prostheses, Pacific Grove, CA, USA.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion, 15*, 123-148. Doi: 10.1007/Bf00995674

Schorr, E. A., Roth, F. P., & Fox, N. A. (2009). Quality of life for children with cochlear implants: perceived benefits and problems and the perception of single words and emotional sounds. *Journal of Speech, Language, and Hearing Research, 52*, 141-152.

Schweer, W. (2012). MuseScore [Computer program], Version 1.2. Retrieved from http://musescore.org/

See, R. L., Driscoll, V. D., Gfeller, K., Kliethermes, S., & Oleson, J. (2013). Speech intonation and melodic contour recognition in children with cochlear implants and with normal hearing. *Otology & Neurotology, 34*, 490-498. Doi: 10.1097/MAO.0b013e318287c985

Seifert, E., Oswald, M., Bruns, U., Vischer, M., Kompis, M., & Haeusler, R. (2002). Changes of voice and articulation in children with cochlear implants. *International Journal of Pediatric Otorhinolaryngology, 66*, 115-123.

Shannon, R. V. (2002). The relative importance of amplitude, temporal, and spectral cues for cochlear implant processor design. *American Journal of Audiology, 11*, 124-127.

Shannon, R. V., Fu, Q.-J., & Galvin, 3rd, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Oto-Laryngologica, 124*, 50-54.

Shannon, R. V., Zeng, F. G., Kamath, V.,Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303–304.

Shannon, R. V., Zeng, F. G., & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America, 104*, 2467–2476. Doi: 10.1121/1.423774

Sharma, A., Dorman, M. F., & Kral, A. (2005). The influence of a sensitive period on central auditory development in children with unilateral and bilateral cochlear implants. *Hearing Research, 203*, 134-143.

Sharma, A., Tobey, E., Dorman, M., Bharadwaj, S., Martin, K., Gilley, P., & Kunkel, F. (2004). Central auditory maturation and babbling development in infants with cochlear implants. *Archives of Otolaryngology – Head and Neck Surgery, 130*, 511-516. Doi: 10.1001/archotol.130.5.511

Shin, M. S., Song, J. J., Han, K. H., Lee, H. J., Do, R. M., Kim, B. J., & Oh, S. H. (2015). The effect of psychosocial factors on outcomes of cochlear implantation. *Acta Otolaryngologica, 135*, 572-577. Doi: 10.3109/00016489.2015.1006336

Shirvani, S., Jafari, Z., Sheibanizadeh, A., Motasaddi Zarandy, M., & Jalaie, S. (2014). Emotional perception of music in children with unilateral cochlear implants. *Iranian Journal of Otorhinolaryngology, 26*, 225-233.

Simmons, F. B. (1966). Electrical stimulation of the auditory nerve in man. *Archives of Otolaryngology, 84*, 2-54.

Smith, A. B., Roberts, J., Smith, S. L., Locke, J. L., & Bennett, J. (2006). Reduced speaking rate as an early predictor of reading disability. *American Journal of Speech-Language Pathology, 15*, 289-297. Doi: 10.1044/1058-0360(2006/027)

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*, 87–90.

Snel-Bongers, J., Netten, A. P., Boermans, P. P., Briaire, J. J., & Frijns, J. H. (submitted). Evidence-based inclusion criteria for cochlear implantation in postlingual deafened patients. *Ear and Hearing*.

Snow, D., & Ertmer, D. (2009). The development of intonation in young children with cochlear implants: A preliminary study of the influence of age at implantation and length of implant

experience. *Clinical Linguistics & Phonetics, 23*, 665-679. Doi: 10.1080/02699200903026555

Snow, D., & Ertmer, D. (2012). Children's development of intonation during the first year of cochlear implant experience. *Clinical Linguistics & Phonetics, 26*, 51-70. Doi: 10.3109/02699206.2011.588371

Soderstrom, M., Seidl, A., Nelson, D. G. K., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49*, 249-267. Doi: 10.1016/S0749-596x(03)00024-X

Souza, P., Arehart, K., Miller, C. W., & Muralimanohar, R. K. (2011). Effects of age on F0 discrimination and intonation perception in simulated electric and electroacoustic hearing. *Ear and Hearing, 32*, 75–83. Doi:10.1097/AUD.0b013e3181eccfe9.

Souza, P., & Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *Journal of the Acoustical Society of America, 126*, 792-805.

Spencer, L. J., Gantz, B. J., & Knutson, J. F. (2004). Outcomes and achievement of students who grew up with access to cochlear implants. *The Laryngoscope, 114*, 1576-1581.

Stacey, P. C., Fortnum, H. M., Barton, G. R., & Summerfield, A. Q. (2006). Hearing-impaired children in the United Kingdom, I: Auditory performance, communication skills, educational achievements, quality of life, and cochlear implantation. *Ear and Hearing, 27*, 161-186. Doi: 10.1097/01.aud.0000202353.37567.b4

Stafford, R. C., Stafford, J. W., Wells, J. D., Loizou, P. C., & Keller, M. D. (2014). Vocoder simulations of highly focused cochlear stimulation with limited dynamic range and discriminable steps. *Ear and Hearing, 35*, 262-270.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137-149.

Stephens, D., & Kerr, P. (2000). Discapacidad auditiva: Una actualization [Auditory Disablements: An Update]. *Audiology, 39,* 322-332.

Stickney, G. S., Loizou, P. C., Mishra, L. N., Assmann, P. F., Shannon, R. V., & Opie, J. M. (2006). Effects of electrode design and configuration on channel interactions. *Hearing Research, 211,* 33-45.

Stone, M. A., Fuellgrabe, C., & Moore, B. C. J. (2008). Benefit of high-rate envelope cues in vocoder processing: effect of number of channels and spectral region. *Journal of the Acoustical Society of America, 124,* 2272–2282.

Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Demonet, J. F., Deguine, O., & Barone, P. (2015). Increased audiovisual integration in cochlear-implanted deaf patients: independent components analysis of longitudinal positron emission tomography data. *European Journal of Neuroscience, 41,* 677-685. Doi: 10.1111/ejn.12827

Strik, H. (1994). *Physiological control and behavior of the voice source in the production of prosody* (Doctoral dissertation). University of Nijmegen.

Summerfield, A., & Marshall, D. (1995). *Cochlear implantation in the UK 1990–1994. Report by the mrc institute of hearing research on the evaluation of the national cochlear implant programme.* London: HMSO Books.

Svirsky, M. A., Jones, D., Osberger, M. J., & Miyamoto, R. T. (1998). The effect of auditory feedback on the control of oral-nasal balance by pediatric cochlear implant users. *Ear and Hearing, 19,* 385-393.

Svirsky, M. A., Lane, H., Perkell, J. S., & Wozniak, J. (1992). Effects of short-term auditory deprivation on speech production in adult cochlear implant users. *Journal of the Acoustical Society of America, 92,* 1284-1300.

Svirsky, M. A., Stallings, L. M., Lento, C. L., Ying, E., & Leonard, L. B. (2002). Grammatical morphologic development in pediatric

cochlear implant users may be affected by the perceptual prominence of the relevant markers. *Annals of Otology Rhinology and Laryngology, 111*, 109-112.

Swanson, B., Dawson, P., & Mcdermott, H. (2009). Investigating cochlear implant place-pitch perception with the Modified Melodies test. *Cochlear Implants International, 10*, 100-104.

Szyfter, W., Pruszewicz, A., Woznica, B., Swidzinski, P., Szymiec, E., & Karlik, M. (1996). The acoustic analysis of voice in patients with multi-channel cochlear implant. *Revue de Laryngologie Otologie Rhinologie, 117*, 225-227.

Tang, Q., Benítez, R., & Zeng, F.-G. (2011). Spatial channel interactions in cochlear implants. *Journal of neural engineering, 8*, 046029.

Tao, D., Deng, R., Jiang, Y., Galvin, J. J., 3rd, Fu, Q. J., & Chen, B. (2014). Melodic pitch perception and lexical tone perception in mandarin-speaking cochlear implant users. *Ear and Hearing, 36*, 102-110. Doi: 10.1097/AUD.0000000000000086

‘t Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge [England], New York: Cambridge University Press.

Theunissen, S. C. P. M. (2013). *Psychopathology in hearing-impaired children* (Doctoral dissertation). Department of Otorhinolaryngology, Faculty of Medicine, Leiden University Medical Center (LUMC), Leiden University.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7*, 53-71.

Thoutenhoofd, E. (2006). Cochlear implanted pupils in Scottish schools: 4-year school attainment data (2000–2004). *Journal of Deaf Studies and Deaf Education, 11*, 171-188.

Tobey, E. A., Angelette, S., Murchison, C., Nicosia, J., Sprague, S., Staller, S. J., . . . Beiter, A. L. (1991). Speech production performance in children with multichannel cochlear implants. *American Journal of Otology, 12*, 165-173.

Torppa, R., Faulkner, A., Huotilainen, M., Jarvikivi, J., Lipsanen, J., Laasonen, M., & Vainio, M. (2014). The perception of prosody and associated auditory cues in early-implanted children: The role of auditory working memory and musical activities. *International Journal of Audiology, 53*, 182-191. Doi: 10.3109/14992027.2013.872302

Torppa, R., Faulkner, A., Vainio, M., & Järvikivi, J. (2010). Acquisition of focus by normal hearing and Cochlear implanted children. In: *Proceedings of the 5th international Conference on Speech Prosody*. Chicago, Il, USA.

Tye-Murray, N., Spencer, L., Bedia, E. G., & Woodworth, G. (1996). Differences in children's sound production when speaking with a cochlear implant turned on and turned off. *Journal of Speech and Hearing Research, 39*, 604-610.

Tye-Murray, N., Spencher, L., & Woodworth, G. G. (1995). Acquisition of speech by children who have prolonged cochlear implant experience. *Journal of Speech and Hearing Research, 38*, 327-337.

Ubrig, M. T., Goffi-Gomez, M. V., Weber, R., Menezes, M. H., Nemr, N. K., Tsuji, D. H., & Tsuji, R. K. (2011). Voice Analysis of Postlingually Deaf Adults Pre- and Postcochlear Implantation. *Journal of Voice, 25*, 692-299. Doi: S0892-1997(10)00123-2

Uchanski, R. M., & Geers, A. E. (2003). Acoustic characteristics of the speech of young cochlear implant users: a comparison with normally-hearing age-mates. *Ear and Hearing, 24*, 90S-105S. Doi: 10.1097/01.AUD.0000051744.24290.C1

Vaccari, C., & Marschark, M. (1997). Communication between Parents and Deaf Children: Implications for Social-emotional Development. *Journal of Child Psychology and Psychiatry, 38*, 793-801.

Valero Garcia, J., Rovira, J. M., & Sanvicens, L. G. (2010). The influence of the auditory prosthesis type on deaf children's voice quality. *International Journal of Pediatric*

Otorhinolaryngology, 74, 843-848. Doi: S0165-5876(10)00202-8

Vandali, A., Sly, D., Cowan, R., & van Hoesel, R. (2015). Training of cochlear implant users to improve pitch perception in the presence of competing place cues. *Ear and Hearing, 36*, E1-E13.

van de Velde, D. J., Dritsakis, G., Frijns, J. H., van Heuven, V. J., & Schiller, N. O. (2015). The effect of spectral smearing on the identification of pure F0 intonation contours in vocoder simulations of cochlear implants. *Cochlear Implants International, 16*, 77-87. Doi: 10.1179/1754762814Y.0000000086

van Dijkhuizen, J. N., Beers, M., Boermans, P. P., Briaire, J. J., & Frijns, J. H. (2011). Speech intelligibility as a predictor of cochlear implant outcome in prelingually deafened adults. *Ear and Hearing, 32*, 445-458. Doi: 10.1097/AUD.0b013e31820510b7

van Dijkhuizen, J. N., Boermans, P.-P. B., Briaire, J. J., & Frijns, J. H. (2016). Intelligibility of the patient's speech predicts the likelihood of cochlear implant success in prelingually Deaf Adults. *Ear and Hearing, 37*, e302-e310.

van Heuven, V. J. J. P., & Sluijter, A. M. C. (1996). Notes on the phonetics of word prosody. In R. Goedemans, H. van der Hulst, & E. Visch (Eds.), *Stress patterns of the world, part 1: Background* (pp. 233-269). The Hague: Holland Academic Graphics.

van Lierde, K. M., Vinck, B. M., Baudonck, N., De Vel, E., & Dhooge, I. (2005). Comparison of the overall intelligibility, articulation, resonance, and voice characteristics between children using cochlear implants and those using bilateral hearing aids: A pilot study. I*nternational Journal of Audiology, 44*, 452-465. Doi: 10.1080/14992020500189146

Velten, E. (1968). A laboratory task for induction of mood states. *Behavior Research and Therapy, 6*, 473-482.

Vogel, I., & Raimy, E. (2002). The acquisition of compound vs. phrasal stress: the role of prosodic constituents. *Journal of Child Language, 29*, 225-250. Doi: 10.1017/S0305000902005020

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during earl childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America, 117*, 338-350. Doi: 10.1121/1.1835958

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech Language and Hearing Research, 50*, 1510-1545. Doi: 10.1044/1092-4388(2007/104)

Waldstein, R. S. (1990). Effects of postlingual deafness on speech production: implications for the role of auditory feedback. *Journal of the Acoustical Society of America, 88*, 2099-2114.

Wang, W., Zhou, N., & Xu, L. (2011). Musical pitch and lexical tone perception with cochlear implants. *International Journal of Audiology, 50*, 270-278. Doi: 10.3109/14992027.2010.542490

Waters, T. (1986). Speech therapy with cochlear implant wearers. *British Journal of Audiology, 20*, 35-43. Doi: 10.3109/03005368609078996

Wells, B., Peppé, S., & Goulandris, N. (2004). Intonation development from five to thirteen. *Journal of Child Language, 31*, 749-778.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development, 7*, 49-63.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Psychology, 66*, 173-196.

Whitmal, N. A., Poissant, S. F., Freyman, R. L., & Helfer, K. S. (2007). Speech intelligibility in cochlear implant simulations: effects of carrier type, interfering noise, and subject

experience. *Journal of the Acoustical Society of America, 122*, 2376–2388. Doi:10.1121/1.2773993.

Wiefferink, C. H., Rieffe, C., Ketelaar, L., De Raeve, L., & Frijns, J. H. M. (2013). Emotion Understanding in Deaf Children with a Cochlear Implant. *Journal of deaf studies and deaf education, 18*, 175-186. Doi: 10.1093/deafed/ens042

Wilson, B., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., & Rabinowitz, W. M. (1991). Better speech recognition with cochlear implants. *Nature, 352*, 236-238. Doi: 10.1038/352236a0

Wilson, B. (2006). Speech processing strategies. In H. R. Cooper & L. C. Craddock (Eds.), *Cochlear Implants: A Practical Guide* (2nd ed., pp. 21–69). Hoboken, NJ, USA: John Wiley & Sons.

Wilson, B. S., & Dorman, M. F. (2007). The surprising performance of present-day cochlear implants. *IEEE Transactions on Biomedical Engineering, 54*, 969-972. Doi: 10.1109/tbme.2007.893505

Wilson, B., & Dorman, M. F. (2008). Cochlear implants: a remarkable past and a brilliant future. *Hearing Research, 242*, 3-21. Doi: S0378-5955(08)00125-1

Wilson, B., & Dorman, M. (2009). The design of cochlear implants. In J. Niparko (Ed.), *Cochlear implants: Principles & Practices*. Philadelphia, PA, USA: Lippincott Williams & Wilkins.

Wilson, B. (2014). Getting a decent (but sparse) signal to the brain for users of cochlear implants. *Hearing Research, 322*, 24-38. Doi: 10.1016/j.heares.2014.11.009

Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology 4*, 1-13.

Witteman, J., van IJzendoorn, M. H., van de Velde, D., van Heuven, V. J. J. P., & Schiller, N. O. (2011). The nature of hemispheric specialization for linguistic and emotional prosodic perception:

A meta-analysis of the lesion literature. *Neuropsychologia, 49*, 3722-3738. Doi: 10.1016/j.neuropsychologia.2011.09.028

Won, J. H., Drennan, W. R., Kang, R. S., & Rubinstein, J. T. (2010). Psychoacoustic abilities associated with music perception in cochlear implant users. *Ear and hearing, 31*, 796-805. Doi: 10.1097/Aud.0b013e3181e8b7bd

Won, J. H., Nie, K., Drennan, W. R., & Rubinstein, J. T. (2012). Maximizing the spectral and temporal benefits of two clinically used sound processing strategies for cochlear implants. *Trends in Amplification, 16*, 201–210.

Xin, L., & Fu, Q. J. (2004). Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants. *Journal of the Acoustical Society of America, 116*, 3659-3667. Doi: 10.1121/1.1783352

Xu, L., Thompson, C. S., & Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *Journal of the Acoustical Society of America, 117*, 3255–3267.

Yucel, E., Sennaroglu, G., & Belgin, E. (2009). The family oriented musical training for children with cochlear implants: speech and musical perception results of two year follow-up. *International Journal of Pediatric Otorhinolaryngology, 73*, 1043-1052.

Zeng, F. G. (2002). Temporal pitch in electric hearing. *Hearing Research, 174*, 101-106. Doi: 10.1016/S0378-5955(02)00644-5

Zhou, N., Huang, J., Chen, X., & Xu, L. (2013). Relationship Between Tone Perception and Production in Prelingually Deafened Children With Cochlear Implants. *Otology & Neurotology 34*, 499–506.

Zhu, M., Chen, B., Galvin, J. J., 3rd, Fu Q.-J. (2011). Influence of pitch, timbre and timing cues on melodic contour identification with a competing masker (L). *Journal of the Acoustical Society of America, 130*, 3562–3565. Doi: 10.1121/1.3658474

Zhu, Z., Tang, Q., Zeng, F.-G., Guan, T., & Ye, D. (2012). Cochlear-implant spatial selectivity with monopolar, bipolar and tripolar stimulation. *Hearing Research, 283*, 45-58.

Zwolan, T. A., Kileny, P. R., Ashbaugh, C., & Telian, S. A. (1996). Patient performance with the Cochlear Corporation “20+ 2” implant: bipolar versus monopolar activation. *Otology & Neurotology, 17*, 717-723.

Summary of research chapters

This thesis investigated the processing of prosody by users of cochlear implants (CIs). Prosody is the speech information that cannot be reduced to information predictable from individual segments and sequences of segments. It notably varies in fundamental frequency (F0), intensity and durations of parts of an utterance and, among other types of information, functions to convey aspects of information structure (such as the marking of new information, or focus), phrasing of sentences, sentence type (question or statement), as well as about the emotion or attitude with which the speaker has pronounced an utterance. It is both important in speech comprehension and (some aspects of it) notoriously difficult for CI users to perceive, making it an important object of research in this population. Three types of participants were subjected to experiments, namely children with CIs, normally hearing (NH) children without CIs and NH adults listening to simulations (vocoders) of CI hearing (and to non-vocoded stimuli, as a control condition). This topic was approached from five different angles: (1) linguistic vs. emotional prosody, (2) perception and production of prosody, (3) prosody and music, (4) cue weighting, and (5) the prosody processing capacities in children. These five angles were divided over five studies, presented in five respective research chapters. Each of those are summarized below.

Chapter 2 studied the differences, if any, between basic prosodic F0 and duration measures in spontaneous speech of early and

late implanted children with NH peers, at three intervals of hearing age (18, 24 and 30 months after implantation or birth, respectively). The hypotheses were (1) that deviations in CI children's prosodic F0 measures would be relatively large and that those in duration measures would be smallest, reflecting the relative difficulties of these acoustic dimensions in their perception; (2) that late implanted children would show stronger deviations than early implanted children; and (3) that deviations would diminish with increasing hearing age. The first two hypotheses were not supported by the results, as no systematic differences in deviations were observed between prosodic measures nor between clinical groups. However, the results suggested that CI children showed more deviance on parameters that require control of the pronunciation of prosody relatively to those which could be considered as automatic by-products of speech. This could be a reflection of perception difficulties. The third hypothesis was supported by the results because performance on most parameters became less deviant for later test moments.

In Chapter 3, a study is reported where the perception of intonation contours was tested in NH adults listening to vocoded stimuli. Stimuli were naturally recorded short Dutch phrases (e.g., *een agenda*, 'an agenda') between which the only difference was the F0 contour. The F0 contours were stylized versions of variants of phrases expressing surprise, news or disappointment. Subsequently, stimuli with vocoded with 20 dB/octave and 40 dB/octave filter slopes. In three conditions (the two filter slope conditions as well as an unprocessed condition), participants were asked to indicate which type they thought was expressed. Performance in the vocoded conditions was inferior (at chance level) to that in the unprocessed condition (around 90% correct), but there was no difference between the two filter slope conditions. These results showed that this type of vocoding compromised the perception pure F0 prosodic contrasts, but that, most probably, above-chance level performance and differences in performance between filter slope conditions would only be shown for even steeper filter slopes.

The study described in Chapter 4 is an extension of that in Chapter 3. Instead of only two filter slope conditions (20 and 40 dB/octave), five slopes were tested (5, 20, 80, 120, and 160 dB/octave). Stimuli were composed of short phrases of the template 'ARTICLE ADJECTIVE NOUN' (e.g., *een blauwe bal*, 'a blue ball'), produced in five variants, viz. with two emotions (sad and happy), with two focus positions (on the adjective and on the noun), and a neutral variant (as much as possible a neutral emotion and equal focus on the adjective and the noun). These were recorded as natural stimuli, and subsequently either the F0 contour, the segment durations, or both, were used to replace those of the neutral variant with, yielding three new half-natural variants for each of two tests (the emotion test and the focus test). Thus, per test the only information available for the discrimination of emotions (in one test) or focus positions (in another test) was the replaced cue. Stimuli were finally vocoded using a 15-channel noise vocoder. In six conditions comprising five filter slopes and a control condition with no vocoding, participants were asked to decide which emotion, or, in a separate test, which focus position was heard. Without vocoding, performance was near ceiling, showing that the emotions and focus positions were successfully conveyed by the speaker. With vocoding, performance ranged from near-chance level for the shallowest slope (5 dB/octave) to high performance at 120 dB/octave, although in general performance for the focus test was lower than for the emotion test. At 160 dB/octave, scores were comparable to those at 80 dB/octave, lower than at 120 dB/octave. For emotion perception, the pattern of scores in the condition including both F0 and duration cues was closest to that including only F0 cues, whereas for focus perception it was closest, albeit less close, to the condition including only duration cues. Together, these results show that steepening the filter slope has positive effects for prosody perception until values as extreme as 120 dB/octave, but that this effect is stronger for emotion than for focus perception because (with the current stimuli) for the former F0 cues are more informative than for the latter. The filter slope of 120

dB/octave could be used a theoretical target value for future speech processing algorithms in CIs.

Chapter 5 reports a study where NH adults received a brief 45-minute training in perceiving either temporal (one group) or melodic contrasts (another group) in vocoded musical stimuli. The goal of this study was to test if this cue-specific training would induce greater reliance on that cue as opposed to the other (non-trained) cue in prosody perception and/or musical melody recognition. A questionnaire filled in before the training showed that the groups did not differ in musical background. After training, participants performed the focus and emotion test described in Chapter 4, a familiar melody recognition test with duration cues, F0 cues or both available, as well as a test assessing if they had a rhythm or melody listening bias when segmenting four-note sequences with ambiguous starting points (the highest note or the loudest note). No significant cross-domain (music to prosody) or cross-cue (duration cues to F0 or melodic cues, or vice versa) training effects were found, although there was a tendency towards a within-cue training effect on familiar melody recognition and, for temporal training, on prosody perception. However, groups did show a segmentation bias in the ambiguous melody test corresponding with the cue they were trained in. Moreover, individual participant-level cross-cue and cross-domain correlations were found. Together, these results suggest that longer cue-specific trainings would have the potential to show positive within- and cross-domain effects improving perception of melodies and prosody.

In Chapter 6, four out of five perspectives of the thesis come together. Six-to-twelve-year-old children with CIs and NH hearing-age matched children were tested on cue usage in four tests on a computer covering perception and production of linguistic and emotional prosody sharing highly comparable stimuli. Besides the core quartet of tests (perception and production of both linguistic and emotional prosody), their general non-verbal emotional and linguistic capacities were tested by means of affective phrases and emotion-

inducing situations, and by means of non-word repetition, respectively. Performance on these tests did not differ significantly between groups showing similar baseline capacities. Before the core tests, children were familiarized with the procedure and the stimuli by means of simple naming and identification tasks; both groups scored near or at ceiling level. In the core tests, the linguistic and emotional prosody perception tests were similar to those described in Chapter 4, including the exact stimuli and the cue availability. In the linguistic prosody (focus) production test, children responded to a question of the form *Is dit een blauwe bal?* ('Is this a blue ball?') where either the adjective, the noun or both contrasted with a picture on a screen. In the emotion production test, children were asked to describe an object picture (e.g., a red chair) and say it in a sad or happy manner depending on the face accompanying the object picture. The emotions and focus positions of the productions were judged by an independent panel of ten Dutch adults. The results showed no difference in cue weighting strategy between groups, nor in the effectiveness of the emotion and focus position productions (this holds for emotion mainly, as the focus perception results could not be analyzed). However, weak correlations between emotional prosody perception and production as well as between emotional prosody perception and production, on the one hand, and non-verbal emotional understanding performance, on the other hand, were found in CI but not, or to a lesser degree, in NH children. Finally, hearing age weakly predicted emotion production but not perception in both groups. Together, these results suggest that CI children at this age, despite being compromised and delayed by a hearing disadvantage, have caught up with their peers when it comes to prosody perception and production.

Samenvatting in het Nederlands

In dit proefschrift is de verwerking van prosodie door gebruikers van cochleair implantaten (CI) onderzocht. Prosodie is spraakinformatie die niet gereduceerd kan worden tot individuele segmenten en opeenvolgingen van segmenten. De voornaamste dimensies waarin het varieert zijn de fundamentele frequentie (F0), intensiteit en duur van delen van uitingen. Het geeft onder meer informatie over informatiestructuur (zoals de het markeren van nieuwe informatie, ook wel focus genoemd), de manier waarop zinnen in de uitspraak in woordgroepen worden onderverdeeld (frasering), zinstype (vraag of mededeling), alsook de emotie en attitude waarmee de spreker een uiting realiseert. Prosodie is tegelijkertijd belangrijk voor spraakbegrip en (wat betreft bepaalde aspecten ervan) vormt een berucht struikelblok voor CI-gebruikers. Daarmee is het een belangrijk onderwerp voor onderzoek naar die populatie. In de onderzoeken voor dit proefschrift zijn drie soorten deelnemers onderzocht, namelijk kinderen met CI's, normaalhorende (NH) kinderen zonder CI's en NH volwassenen die luisteren naar simulaties (vocoders) van het horen met een CI (alsook, als controleconditie, naar niet-gevocoderde stimuli). Het onderzoeksonderwerp van dit proefschrift is benaderd vanuit vijf perspectieven: (1) taalkundige versus emotionele prosodie, (2) de perceptie en de productie van prosodie, (3) prosodie en muziek, (4) de weging van cues bij de waarneming van prosodie en (5) de ontwikkeling van prosodie bij kinderen. Deze vijf perspectieven zijn

verdeeld over vijf onderzoeken en worden behandeld in vijf onderzoekshoofdstukken. Elk van die hoofdstukken wordt hieronder samengevat.

In Hoofdstuk 2 zijn de eventuele verschillen tussen basale prosodische maten op het gebied van F0- en duurvariaties, in de spontane spraak van vroeg- en geïmplanteerde kinderen vergeleken met die van NH kinderen, van wie de hoorleeftijd correspondeerde met die van de klinische groep. Metingen hebben plaatsgevonden op drie tijdsmomenten – 18, 24 en 30 maanden – na implantatie (voor de CI-groep), dan wel na geboorte (voor de controlegroep). De hypotheses waren (1) dat afwijkingen in de prosodische F0-maten bij CI-kinderen groter zouden zijn dan die in duurmaten, (2) dat de spraak van laatgeïmplanteerde sterkere afwijkingen zou vertonen dan die van vroeggeïmplanteerde kinderen en (3) dat afwijkingen zouden afnemen als functie van de hoorleeftijd. De eerste twee hypotheses werden niet ondersteund door de resultaten, omdat geen systematische verschillen in afwijkingen waren geobserveerd tussen prosodische maten noch tussen de twee klinische groepen. De resultaten veronderstelden echter dat CI-kinderen meer afwijkingen vertoonden als het ging om parameters die een grotere mate van beheersing vragen van de uitspraak van prosodie ten opzichte van maten die kunnen worden beschouwd als automatische neveneffecten van spraak. Dit zou een reflectie kunnen zijn van waarnemingsproblemen. De derde hypothese werd ondersteund door de resultaten, omdat de prestaties van de meeste maten minder afwijkend werden op latere testmomenten.

In Hoofdstuk 3 wordt verslag gedaan van een onderzoek waarbij de perceptie van gevocoderde intonatiecontouren door NH volwassenen is getest. De stimuli bestonden uit varianten als natuurlijke uitingen opgenomen korte Nederlandse frases, zoals *een agenda*, die alleen door de F0-contour van elkaar te onderscheiden waren. De F0-contouren waren gestileerde versies van varianten de frases waarin respectievelijk verrassing, nieuws en teleurstelling werd uitgedrukt. Vervolgens waren de stimuli gevocoderd met filterhellingen van 20 dB/octaaf en 40 dB/octaaf. Participanten werd

in drie verschillende (de twee filterhellingcondities en een conditie zonder vocoding) gevraagd om aan te geven welk van die types was uitgedrukt. De prestaties in de gevocoderde condities waren lager (op kansniveau) dan die in de niet-gevocoderde conditie (rond de 90% correct), maar er was geen verschil tussen de twee filterhellingcondities onderling. Deze resultaten tonen aan dat dit type vocoding de waarneming van prosodische contrasten op basis van alleen de F0, onmogelijk maakte maar dat voor scores boven kansniveau hoogstwaarschijnlijk scherpere filterhellingen nodig zouden zijn.

Het onderzoek dat in Hoofdstuk 4 is beschreven is een uitbreiding op het onderzoek uit Hoofdstuk 4. In plaats van slechts twee filterhellingen te testen (20 en 40 dB/octaaf) zijn nu vijf hellingen getest (5, 20, 80, 120, en 160 dB/octaaf). De stimuli bestonden uit korte frases van de vorm 'LIDWOORD-BIJVOEGLIJK NAAMWOORD-ZELFSTANDIG NAAMWOORD' (zoals *een blauwe bal*), opgenomen in vijf varianten, te weten twee emoties (verdrietig en blij), twee focusposities (op het bijvoeglijk naamwoord en op het zelfstandig naamwoord) en een neutrale variant (zo veel mogelijk zonder specifieke emotie en met gelijkwaardige focus op het bijvoeglijk naamwoord en het zelfstandig naamwoord). Deze stimuli waren opgenomen als natuurlijke uitingen en vervolgens was ofwel de F0-contour, waren de segmentduren of waren beide types informatie van de neutrale variant vervangen door die van de niet-neutrale varianten. Zo waren voor gebruik van elk deelexperiment (een emotietest en een focustest) drie nieuwe half-synthetische varianten gecreëerd. In elk van beide testen was de enige beschikbare informatie om de emoties (in de ene test) dan wel de focusposities (in de andere test) van elkaar te onderscheiden de vervangen cue. Als laatste stap waren de stimuli gevocoderd met een 15-kanaals-ruisvocoder. Deelnemers werd gevraagd om voor elk van zes condities bestaande uit vijf filterhellingen en een conditie zonder vocoding aan te geven welke emotie dan wel, in een aparte test, welke focuspositie ze dachten dat was uitgedrukt. De prestaties in de conditie zonder

vocoding waren tegen het plafondniveau aan, wat aantoont dat de spreker de emoties en focusposities succesvol had uitgedrukt. In de vocodercondities varieerden de scores van nabij het kansniveau in het geval van het minst scherpe filter (5 dB/octaaf) tot hoge scores in het geval van filters van 120 dB/octaaf; scores waren in het algemeen in de focustest echter lager dan in de emotietest. De scores bij 160 dB/octaaf waren vergelijkbaar met die bij 80 dB/octaaf, beide lager dan bij 120 dB/octaaf. In de emotietest was het scorepatroon in de conditie met beide cues beschikbaar het gelijkst aan die met alleen F0-informatie beschikbaar, terwijl die in de focustest het gelijkst was aan die met alleen duurinformatie beschikbaar. Samengenomen tonen deze resultaten aan dat verscherping van de filters een positief effect heeft op prosodiewaarneming tot de extreme waarde van 120 dB/octaaf, maar dat dit effect (gegeven de stimuli die in dit experiment waren gebruikt) sterker is voor focuspositiewaarneming dan voor emotiewaarneming, omdat de F0-cues voor emoties informatiever zijn dan voor focuspositie. De filterhelling van 120 dB/octaaf kan als theoretische doelwaarde worden gebruikt voor spraakverwerkingsalgoritmes voor CI's in de toekomst.

Hoofdstuk 5 bevat het verslag van een onderzoek waarbij NH volwassenen een korte training van 45 minuten krijgen om ofwel temporele (één groep) ofwel melodische (een tweede groep) contrasten beter leerden waarnemen in gevocoderde muzikale stimuli. Het doel van dit onderzoek was om te testen of zulke cue-specifieke training een groter vertrouwen op de getrainde cue dan op de niet-getrainde cue zou teweegbrengen. De antwoorden op een vooraf ingevulde vragenlijsten lieten zien dat de muzikale achtergronden van de groepen niet van elkaar verschilden. Na de training voerden de deelnemers de testen naar de waarneming van emotie en focustest uit die in Hoofdstuk 4 zijn beschreven, alsook een herkenningstesten van bekende muzikale melodieën met ofwel duur- ofwel F0- ofwel beide cues beschikbaar, en als derde een test waarin werd onderzocht of ze een voorkeur hadden om sequenties van vier noten met ambigue startposities (dat wil zeggen, op de luidste of op de hoogste noot) op

basis van ritme of op basis van melodie te segmenteren. In dit onderzoek zijn geen trainingseffecten van het ene domein naar het andere (van muziek naar prosodie) noch van de ene cue naar de andere (van duurcues naar F0- of melodische cues of andersom) gevonden, hoewel er wel een tendens van een trainingseffect voor de getrainde cue bij de bekende-melodieëntest is geobserveerd. Ook is een segmentatievoorkeur in de ambigue-melodieëntest geconstateerd die overeenkwam met de getrainde cue. Tot slot zijn correlaties op het gebied van de individuele deelnemers gevonden voor prestaties bij beschikbaarheid van ongelijke cues en van ongelijk domein. Deze resultaten laten bij elkaar zien dat langere cue-specifieke training een positief effect zou kunnen hebben op effecten binnen en tussen cues en domeinen waarbij de perceptie van muzikale melodieën en prosodie wordt verbeterd.

In Hoofdstuk 6 komen vier van de vijf perspectieven van het proefschrift samen. Van zes tot twaalf jaar oude kinderen met CI's en in hoorleeftijd overeenkomende NH kinderen werd het gebruik van cues onderzocht in vier verschillende computertesten op het gebied de perceptie en productie van taalkundige en emotionele prosodie, alle gebruikmakend van in hoge mate overeenkomstige stimuli. Behalve het kwartet aan kerntesten (de perceptie en productie van zowel taalkundige als emotionele prosodie), is hun algemene nonverbale emotionele en taalkundige ontwikkelingsniveau getest door middel van respectievelijk zinnen en situaties die bepaalde emoties uitdrukken en nonwoordherhaling. De prestaties verschilden bij deze testen niet significant tussen de deelnemersgroepen, wat aangeeft dat ze vergelijkbare basisontwikkelingen hadden. Voorafgaand aan de kerntesten werden kinderen bekendgemaakt met de procedure en de stimuli door middel van simpele benoemings- en identificatietaken; beide groepen scoorden op of nabij het plafondniveau. Van de kerntesten waren die van de perceptie van emotie en focuspositie gelijk aan die uit Hoofdstuk 4, inclusief de exacte stimuli en beschikbaarheid van verschillende cues. In de focusproductietest werd deelnemers gevraagd antwoord te geven op een vraag van de vorm *Is*

dit een blauwe bal?, waarbij ofwel het bijvoeglijk naamwoord, ofwel het zelfstandig naamwoord of beide contrasteerde met een afbeelding op het scherm. In de emotieproductietest werd kinderen gevraagd om een afbeelding te beschrijven (bijvoorbeeld van een rode stoel) en dat op een verdrietige of blije manier te zeggen, afhankelijk van de emotie van een gezichtje dat erbij op het scherm stond. De geproduceerde emoties en focusposities werden naderhand als zodanig beoordeeld door een onafhankelijk panel van tien Nederlandse volwassenen. De resultaten lieten geen verschil in cue-wegingstrategie tussen de twee groepen zien, noch in de effectiviteit van de emoties en focusposities in hun producties (dit geldt met name voor emotionele prosodie, omdat de resultaten van de focusperceptietest niet konden worden geanalyseerd). Er zijn echter zwakke correlaties gevonden tussen emotionele-prosodieperceptie en -productie alsook tussen, aan de ene kant, emotionele-prosodieperceptie en -productie en, aan de andere kant, nonverbale-emotiebegrip in de CI-groep maar niet, of in mindere mate, in de NH groep. Tot slot bleek hoorleeftijd in beide groepen de scores in de emotieproductietest maar niet die in de emotieperceptietest in beperkte mate te voorspellen. Samengenomen veronderstellen deze resultaten dat CI-kinderen op deze leeftijd, ondanks de problemen en uitstel die ze kunnen ervaren door hun nadelige gehoorsituatie, de achterstand op het gebied van de perceptie en productie van prosodie op hun leeftijdsgenoten hebben ingehaald.

Appendix A

The participant questionnaire assessing language and musical training background, partly after the Salk/McGill Music Inventory (SAMMI; Levitin et al., 2004), used, used in the study reported in Chapter 4. The writing space available for providing was more ample in the original layout.

Vragenlijst muziekachtergrond	
1. Zingt u of bespeelt u een muziekinstrument of heeft u dat gedaan?	Ja / Nee (onderstreep)
1.1 Zo ja, hoeveel uur per week doet of deed u dat?	……..… uur
1.2 Zo ja, hoeveel jaar heeft u gezongen/een instrument bespeeld?	……..... jaar
2. Luistert u naar muziek?	Ja / Nee (onderstreep)
2.1 Zo ja, hoeveel uur per week luistert u ongeveer naar muziek?	…… uur
5. Wat is de hoogst genoten opleiding die u doet of heeft gedaan?	Vakgebied…….. Niveau:……....

6. Houdt u zich bezig met verwerking van geluidsmateriaal (bijvoorbeeld als ingenieur, geluidstechnicus, onderzoeker, spraakwetenschapper, audioloog, logopedist, componist)?	Ja / Nee (onderstreep) Zo ja, namelijk als: ……

Appendix B

The participant questionnaire assessing language and musical training background, partly after the Salk/McGill Music Inventory (SAMMI; Levitin et al., 2004), used in the study reported in Chapter 5. The writing space available for providing was more ample in the original layout.

Vragenlijst muziekachtergrond	
1. Zingt u of bespeelt u een muziekinstrument of heeft u dat gedaan?	Ja / Nee (onderstreep)
1.1 Zo ja, hoeveel uur per week doet of deed u dat?	…… uur
1.2 Zo ja, hoeveel jaar heeft u gezongen/een instrument bespeeld?	……..… jaar
1.3 Zo ja, hoeveel jaar zingt u al / bespeelt u al een instrument?	……..…… jaar
2. Heeft u formele training in zingen of het bespelen van een instrument gehad?	Ja / Nee (onderstreep)
2.1 Hoe lang geleden heeft u voor het laatst les gehad in zingen/het bespelen van een instrument (afgerond tot 1 jaar nauwkeurig)?	….……jaar geleden
3. Heeft u formele training in muziektheorie gehad (zoals lessen in het lezen, analyseren of	Ja / Nee (onderstreep)

componeren van muziek)?

4. Luistert u naar muziek? Ja / Nee (onderstreep)

4.1 Zo ja, hoeveel uur per week luistert u ongeveer naar muziek? …… uur

5. Wat is uw moedertaal of wat zijn uw moedertalen (taal/talen waarmee u van jongs af aan bent opgegroeid)? ……..……..
……...

6. Beheerst u andere talen? Ja / Nee (onderstreep)

6.1. Zo ja, welke en op welk niveau op een schaal tussen *1 (nauwelijks) -------10 (moedertaalniveau)*?

	Taal	Niveau van spreken/schrijven	Niveau van verstaan/lezen
1			
2			
3			
4			
5			

7. Heeft u langer dan drie maanden achtereen in het buitenland doorgebracht? Ja / Nee (onder-streep)

7.1. Zo ja, in welk(e) land(en) was dat en hoe lang?

	Land	Duur van verblijf
1		
2		
3		

8. Bent u ooit behandeld voor problemen met uw gehoor of uw spraak?	Ja / Nee (onder-streep)
8.1. Zo ja, kunt u het probleem kort beschrijven, inclusief of het om het linker- or het rechteroor ging (indien van toepassing)?	…… …… ……
9. Hebt u momenteel problemen met uw gehoor?	Ja / Nee (onde-rstreep)
9.1. Zo ja, kunt u het probleem kort beschrijven, inclusief of het om het linker- or het rechteroor ging (indien van toepassing)?	…… …… ……

Appendix C

Non-word repetition stimuli used in Chapter 6

The stimuli used in the non-word repetition task that was part of the study reported in Chapter 6. No experimental items functioned as practice items. There were no practice items consisting of one, two or four syllables.

1 syllable	**2 syllables**	**3 syllables**	**4 syllables**	**5 syllables**
Practice	**Practice**	**Practice**	**Practice**	**Practice**
-	-	Nietoofoes	-	Piefoesaanooteem
Experi-mental	**Experimental**	**Experimental**	**Experimental**	**Experimental**
Noos	Noetiem	Sootienoem	Naafiesooteem	Taanoosoefiepeem
Fiem	Saapoom	Taapienoes	Siepootaafoem	Saanoepootiefeem
Taas	Tienees	Piefoenoom	Foepietoonees	Faanietoosoepeem
Poem	Pietoes	Faapooties	Poofienaatees	Niepoofaanoetees

Appendix D

The parent questionnaire assessing language and demographic background and aspects of medical history of one or both parents/caretakers and his/her/their participating child(ren), used in the study reported in Chapter 6.

Vragenlijst voor ouders/verzorgers van kinderen met normaal gehoor of met CI

Opmerkingen vooraf voor de ouder(s)/verzorger(s):

a) Het kind dat de spreek- en luistertesten doet wordt hieronder ‘de deelnemer’ genoemd.
b) Probeert u het zo volledig mogelijk in te vullen. Als u een vraag heeft kunt u die aan de proefleider stellen. Doet u dit alstublieft als de testen met de deelnemer

1. Welke taal/talen spreken de ouders/verzorgers van de deelnemer? En hoe goed beheersen ze die taal/talen? U kunt zowel gesproken als gebarentalen invullen.

Eerste ouder/verzorger

Taal:
……..……..……...……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:
……..……..……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:
……..……..……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:
……..……..……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

---

Indien van toepassing: Tweede ouder/verzorger

Taal:
……..……..……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:
……..……..……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:
……..……..……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:
……..……..……

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

2. Welke taal of talen heeft de deelnemer geleerd? Vanaf welke leeftijd en op welke manier? En hoe goed vindt u dat hij/zij elk die talen beheerst? U kunt gesproken talen maar ook gebarentaal invullen. Bij manier van leren kunt u denken aan: thuis, op school, speciale taalles, van televisie of computer, enz.

Taal:

……

Geleerd vanaf (leeftijd): ……..

Hoe geleerd?:

……..

……..……

Beheersing:

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:

……

Geleerd vanaf (leeftijd): ……..

Hoe geleerd?:

……..

……..……

Beheersing:

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:

……

Geleerd vanaf (leeftijd): ……..

Hoe geleerd?:

……..

……..……

Beheersing:

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

Taal:

……

Geleerd vanaf (leeftijd): ……..

Hoe geleerd?:

……..

……..……

Beheersing:

☐ Zeer beperkt ☐ Beperkt ☐ Matig ☐ Functioneel ☐ Goed ☐ Zeer goed ☐Vloeiend

3. Welke taal of talen spreekt de deelnemer met de ouder(s)/verzorger(s)? Hoeveel procent van de tijd ongeveer?

Taal:

……..……

Percentage:…….

Taal:

……..……

Percentage:…….

Taal:

……..……

Percentage:…….

Taal:

……

Percentage:…….

4. Indien van toepassing: Welke taal of talen spreekt de deelnemer op school? Hoeveel procent van de tijd ongeveer?

Taal:
……
Percentage:…….

Taal:
……
Percentage:…….

Taal:
……
Percentage:…….

Taal:
……
Percentage:…….

5. Welke taal of talen spreekt de deelnemer met (sommige) vriendjes of vriendinnetjes? Hoeveel procent van de tijd ongeveer?

Taal:
……
Percentage:…….

Taal:
……
Percentage:…….

Taal:
……
Percentage:…….

Taal:
……
Percentage:…….

6. Is de deelnemer ooit langer dan drie maanden achter elkaar in het buitenland geweest? Zo ja, welke taal heeft hij/zij daar gesproken?

☐ Nee

☐ Ja, namelijk:

Land:
……...……
Taal:…….…….

Land:
……...……
Taal:…….…….

Land:
……...……
Taal:…….…….

Land:
……...……
Taal:…….…….

De volgende vragen gaan over emoties bij de deelnemertijdens de <u>afgelopen 2 maanden</u>. U kunt aankruisen wat u het meest van toepassing vindt. Wilt u alstublieft bij elke vraag één antwoord aankruisen?

7. In welk van de talen uit <u>vraag 2</u> drukt de deelnemer emoties meestal uit?

……
……
……
……

8. Kunt u goed inschatten welke emotie de deelnemer voelt?

☐ (Bijna) nooit ☐ Zelden ☐ Soms ☐ Vaak ☐ (Bijna) altijd

9. Kan de deelnemer de emoties van anderen goed inschatten?

☐ (Bijna) nooit ☐ Zelden ☐ Soms ☐ Vaak ☐ (Bijna) altijd

10. Hoe vaak toont de deelnemer emoties?

☐ (Bijna) nooit ☐ Zelden ☐ Soms ☐ Vaak ☐ (Bijna) altijd

De volgende twee vragen zijn alleen voor de ouder(s)/verzorger(s) van deelnemers met een cochleair implantaat. Alle andere ouder(s)/verzorger(s) kunnen meteen naar vraag 13 (blz. 7).

DEEL VOOR OUDER(S)/VERZORGER(S) VAN DEELNEMERS MET EEN COCHLEAIR IMPLANTAAT

11. Gebruikte de deelnemer vóór implantatie gebarentaal én gesproken taal? En hoe is dat sinds de implantatie? Hoeveel procent ongeveer?

Vóór implantatie:

☐ Nee, maar één taal

☐ Ja, namelijk:

Percentage gebarentaal:……..

Percentage gesproken taal:…….

Na implantatie:

☐ Nee, maar één taal

☐ Ja, namelijk:

Percentage gebarentaal:…….

Percentage gesproken taal:…….

12. In welke taal, gebarentaal of gesproken taal, verloopt de communicatie gemakkelijker, vóór en na implantatie?

Vóór implantatie:……..……

Na implantatie:……..……...

EINDE VAN HET DEEL VOOR OUDER(S)/VERZORGER(S) VAN DEELNEMERS MET EEN COCHLEAIR IMPLANTAAT

De volgende twee vragen zijn alleen voor de ouder(s)/verzorger(s) van normaalhorende deelnemers. Alle andere ouder(s)/verzorger(s) kunnen meteen naar vraag 15 (blz. 8).

DEEL VOOR OUDER(S)/VERZORGER(S) VAN NORMAALHORENDE DEELNEMERS

13. Is de deelnemer ooit behandeld voor problemen met horen? Aan welk oor of welke oren?

☐ Nee

☐ Ja, het linkeroor

☐ Ja, het rechteroor

☐ Ja, beide oren

☐ Ja, ik weet niet welk oor/welke oren

14. Is de deelnemer ooit behandeld door een spraaktherapeut?

☐ Nee

☐ Ja

EINDE VAN HET DEEL VOOR OUDER(S)/VERZORGER(S) VAN NORMAALHORENDE DEELNEMERS

Vragen voor alle ouders/verzorgers

15. Is de deelnemer ooit behandeld door een neuroloog (hersenarts)?

☐ Nee

☐ Ja

16. Is de deelnemer ooit behandeld voor een sociale stoornis?

☐ Nee

☐ Ja

17. Heeft de deelnemer problemen met zien?

☐ Nee

☐ Ja, maar met een bril of contactlenzen ziet hij/zij goed

☐ Ja en hij/zij gebruikt geen bril of contactlenzen

Wat is het probleem met zien?
……..

☐ Ja, zelfs met een bril of contactlenzen zijn er problemen

Wat is het probleem met zien?
……..

18. Wat is de hoogste opleiding die de ouder(s)/verzorger(s) van de deelnemer gevolgd heeft/hebben?

Eerste ouder/verzorger	Indien van toepassing: Tweede ouder/verzorger
☐ Lagere school/basisschool	☐ Lagere school/basisschool
☐ MAVO/VMBO	☐ MAVO/VMBO
☐ HAVO	☐ HAVO
☐ VWO/Gymnasium	☐ VWO/Gymnasium
☐ Lager Beroepsonderwijs	☐ Lager Beroepsonderwijs
☐ Middelbaar beroepsonderwijs	☐ Middelbaar beroepsonderwijs
☐ Hoger beroepsonderwijs	☐ Hoger beroepsonderwijs
☐ Universitair	☐ Universitair
☐ Weet ik niet of wil ik niet zeggen	☐ Weet ik niet of wil ik niet zeggen

19. Wat is huidige jaarinkomen van het huishouden van de ouder(s)/verzorger(s)?

☐ Lager dan €15.000	☐ Tussen €45.000 en €60.000
☐ Tussen €15.000 en €30.000	☐ Hoger dan €60.000
☐ Tussen €30.000 en €45.000	☐ Weet ik niet of wil ik niet zeggen

20. Heeft u opmerkingen over één of meerdere van de volgende punten? Dan kunt u die hier invullen.

- De taalachtergrond van de deelnemer
- Het taalgebruik van de deelnemer?
- Uw eigen taalachtergrond
- Uw eigen taalgebruik ten opzichte van de deelnemer?
- De algemene ontwikkeling van de deelnemer?
- … Overige dingen die u belangrijk vindt in verband met het onderzoek.

……..……..……

……..……..……

……..……..……

……..……..……

……..……..……

Dit is het einde van de vragenlijst. Hartelijk dank voor het invullen.

Curriculum vitae

Daan van de Velde (1984) was born in Amsterdam in 1984 and grew up in Warmenhuizen, North Holland. From 1996 to 2002, he attended the Willem Blaeu College in Alkmaar, specializing in science, biology and classical languages. In 2002 and 2003 he studied Astronomy at Leiden University, but continued by starting Linguistics and French Languages and Cultures at the same university in 2003 and 2004, respectively. In 2007 and 2009, he obtained his bachelor's and research master's (Structure and Variation in the Languages of the World) degrees in Linguistics and in 2009 and 2016 his bachelor's and master's degrees in French Languages and Cultures, respectively. He additionally completed a minor in Music. For his research master's programme in Linguistics he followed courses of the research master Cognitive Neuroscience at Nijmegen University. His thesis for this master's programme was on the phonological representation of adult Dutch cochlear implant users. Since 2012, he has taught classes in the areas of phonetics, general linguistics, psycholinguistics, language acquisition, statistics and academic skills at the department of Linguistics at Leiden University and at the department of French at Utrecht University. In 2010 he started his PhD research on the processing of prosody by Dutch cochlear implant users at Leiden University Centre for Linguistics. This thesis is the result of that research.