The processing of Dutch prosody with cochlear implants and vocoder simulations

Velde, D.J. van de

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/

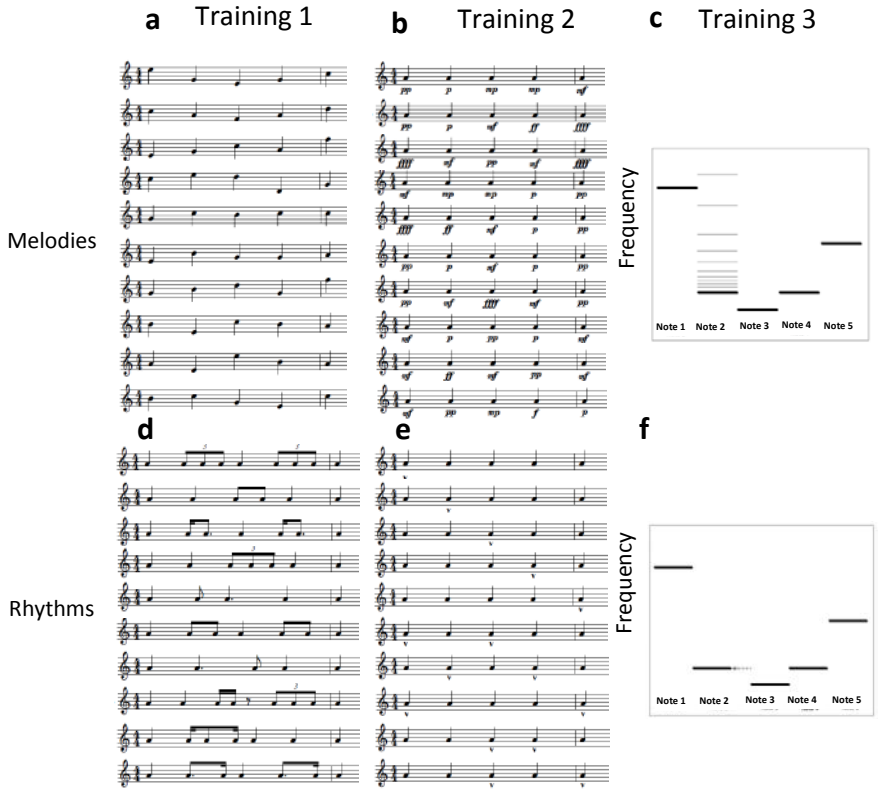

Figure 1. Scores (Trainings 1 and 2) and diagrams (Training 3) of the templates of musical stimuli composed and created for the Trainings. See the text for an explanation of the Trainings. (a) The 10 melodic contours for Pitch Training 1. The ones shown all have the same rhythm, although in the experiment 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

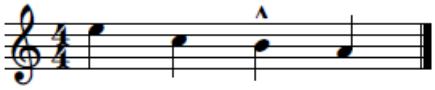

Figure 2. Example of a contour for the Ambiguous Melody test. The accented note, marked by '^', is, essentially, not the lowest or highest of the four.

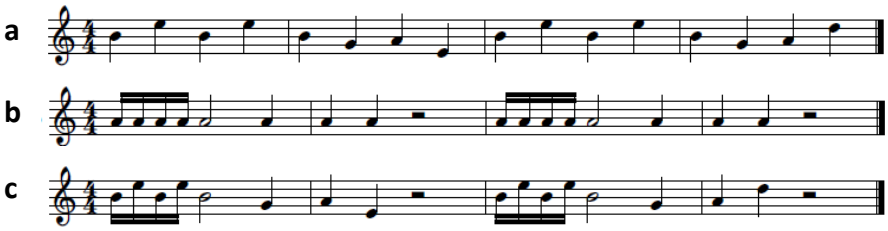

Figure 3. Example of variants for the Familiar Melody Identification test. The example shown is for the melody 'Morricone – The good, the bad and the ugly'. (a) With intact melody, but neutralized rhythm, (b) with intact rhythm, but neutralized melody, (c) with intact melody and rhythm.

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

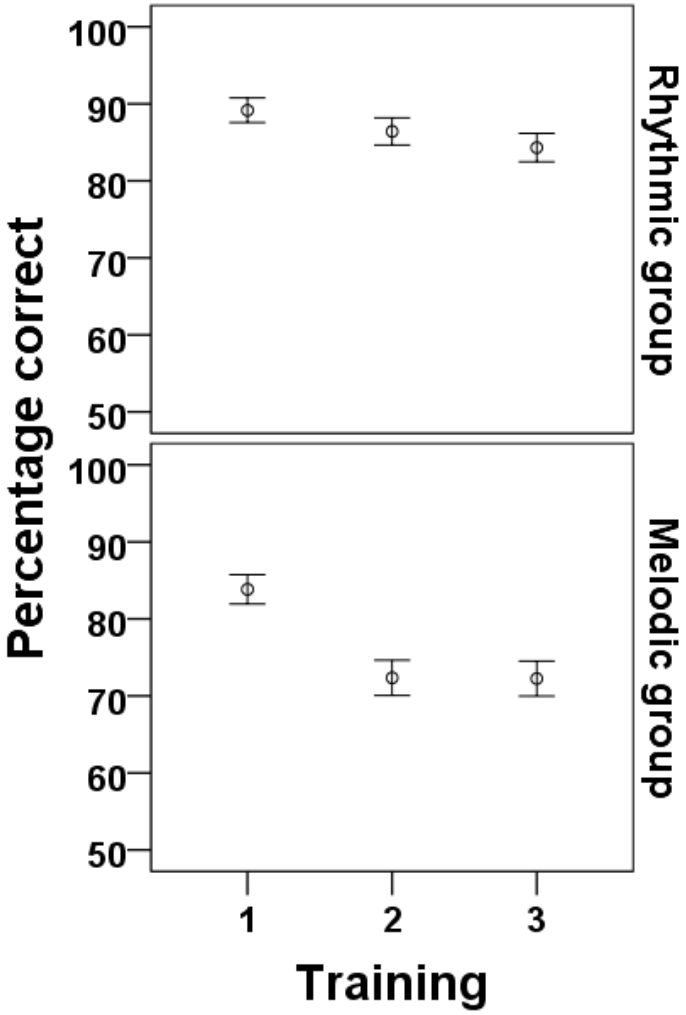

Figure 4. Mean accuracy results of the six Trainings, three for each Group, in percentage correct. Chance level is at 33.3%. Error bars represent 95% confidence intervals.

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

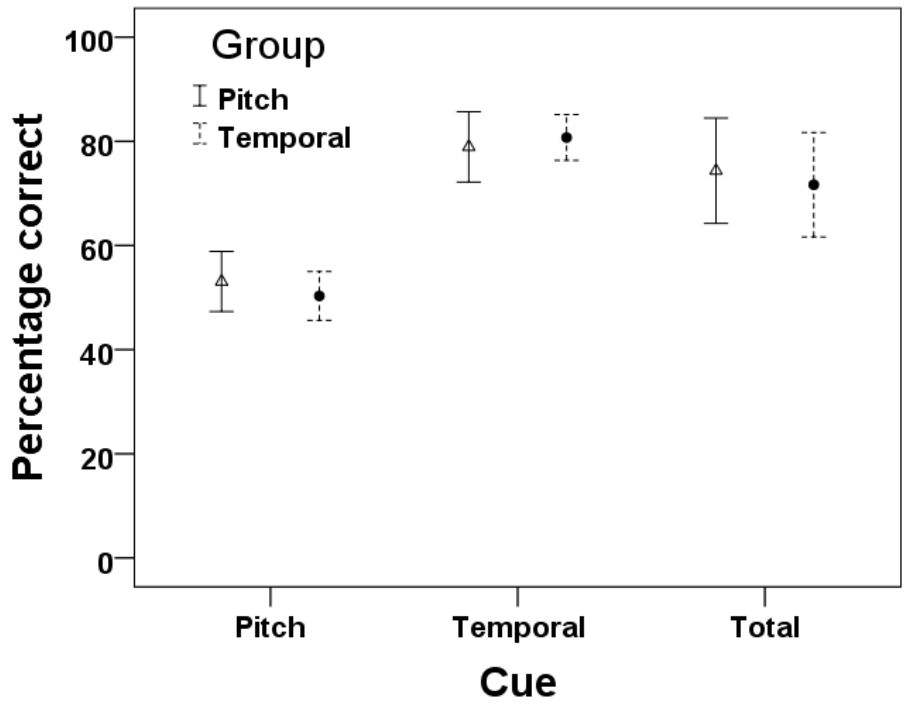

Figure 5. Mean accuracy (percentage correct) and 95% confidence intervals (errors bars) of the Familiar Melody Identification test, split by Cue and by Group. In the Pitch condition, only tone height information was available for identifying melodies. In the Temporal condition, only note duration was 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

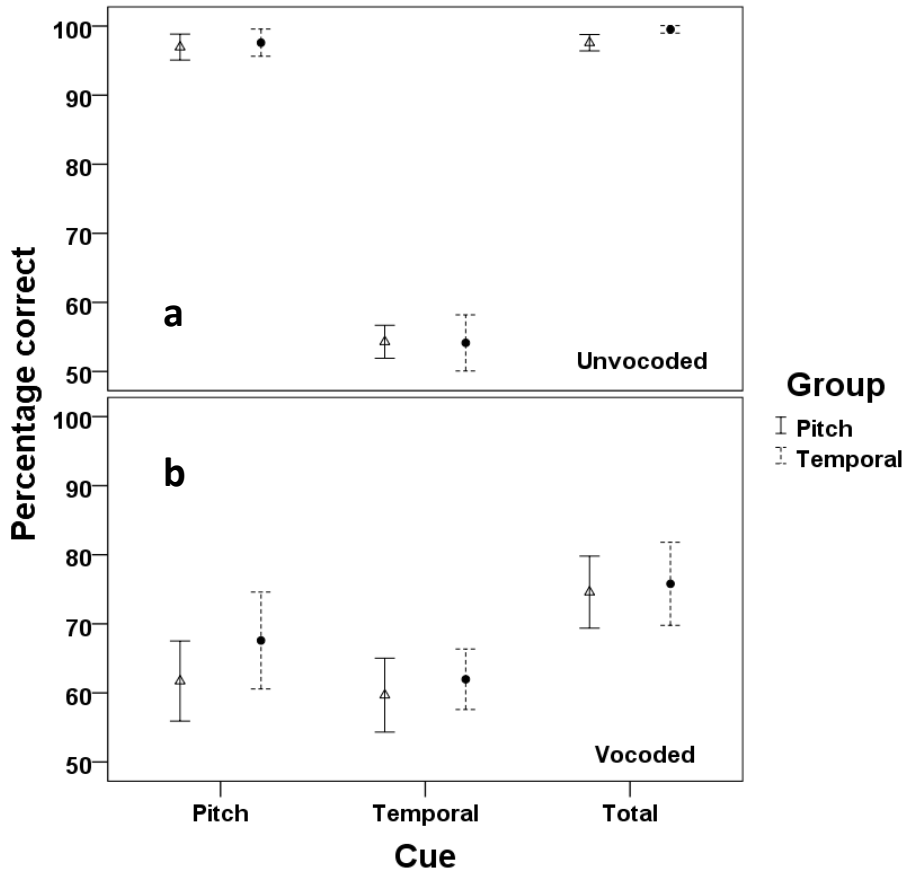

Figure 6. Mean accuracy (percentage correct) and 95% confidence intervals (errors bars) of the Emotion Discrimination test, split by Cue, Group and Vocoding conditions. In the Pitch condition, only tone height (intonation) information was available for identifying melodies. In the Temporal condition, 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

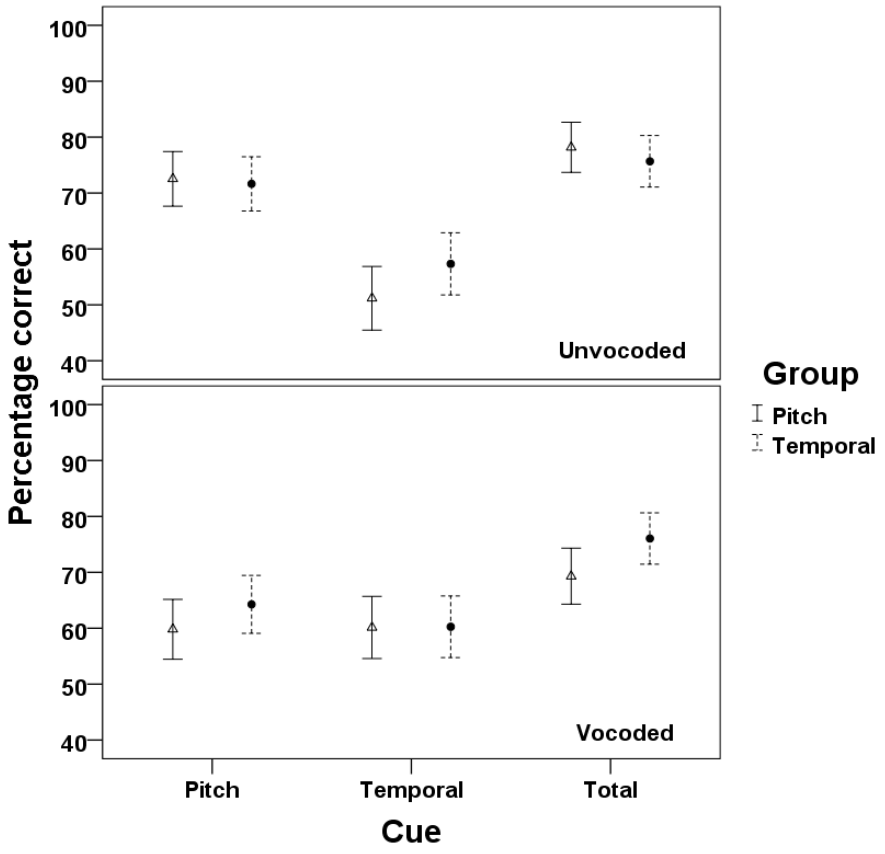

Figure 7. Mean accuracy (percentage correct) and 95% confidence intervals (errors bars) of the Focus Discrimination test, split by Cue, Group and Vocoding conditions. The description is the same as for Figure 6. (a) Results for the Unvocoded condition. (b) Results for the vocoded condition. The description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