The processing of Dutch prosody with cochlear implants and vocoder simulations

Velde, D.J. van de

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.