The processing of Dutch prosody with cochlear implants and vocoder simulations

Velde, D.J. van de

Summary of research chapters

This thesis investigated the processing of prosody by users of cochlear implants (CIs). Prosody is the speech information that cannot be reduced to information predictable from individual segments and sequences of segments. It notably varies in fundamental frequency (F0), intensity and durations of parts of an utterance and, among other types of information, functions to convey aspects of information structure (such as the marking of new information, or focus), phrasing of sentences, sentence type (question or statement), as well as about the emotion or attitude with which the speaker has pronounced an utterance. It is both important in speech comprehension and (some aspects of it) notoriously difficult for CI users to perceive, making it an important object of research in this population. Three types of participants were subjected to experiments, namely children with CIs, normally hearing (NH) children without CIs and NH adults listening to simulations (vocoders) of CI hearing (and to non-vocoded stimuli, as a control condition). This topic was approached from five different angles: (1) linguistic vs. emotional prosody, (2) perception and production of prosody, (3) prosody and music, (4) cue weighting, and (5) the prosody processing capacities in children. These five angles were divided over five studies, presented in five respective research chapters. Each of those are summarized below.

Chapter 2 studied the differences, if any, between basic prosodic F0 and duration measures in spontaneous speech of early and

late implanted children with NH peers, at three intervals of hearing age (18, 24 and 30 months after implantation or birth, respectively). The hypotheses were (1) that deviations in CI children's prosodic F0 measures would be relatively large and that those in duration measures would be smallest, reflecting the relative difficulties of these acoustic dimensions in their perception; (2) that late implanted children would show stronger deviations than early implanted children; and (3) that deviations would diminish with increasing hearing age. The first two hypotheses were not supported by the results, as no systematic differences in deviations were observed between prosodic measures nor between clinical groups. However, the results suggested that CI children showed more deviance on parameters that require control of the pronunciation of prosody relatively to those which could be considered as automatic by-products of speech. This could be a reflection of perception difficulties. The third hypothesis was supported by the results because performance on most parameters became less deviant for later test moments.

In Chapter 3, a study is reported where the perception of intonation contours was tested in NH adults listening to vocoded stimuli. Stimuli were naturally recorded short Dutch phrases (e.g., *een agenda*, 'an agenda') between which the only difference was the F0 contour. The F0 contours were stylized versions of variants of phrases expressing surprise, news or disappointment. Subsequently, stimuli with vocoded with 20 dB/octave and 40 dB/octave filter slopes. In three conditions (the two filter slope conditions as well as an unprocessed condition), participants were asked to indicate which type they thought was expressed. Performance in the vocoded conditions was inferior (at chance level) to that in the unprocessed condition (around 90% correct), but there was no difference between the two filter slope conditions. These results showed that this type of vocoding compromised the perception pure F0 prosodic contrasts, but that, most probably, above-chance level performance and differences in performance between filter slope conditions would only be shown for even steeper filter slopes.

The study described in Chapter 4 is an extension of that in Chapter 3. Instead of only two filter slope conditions (20 and 40 dB/octave), five slopes were tested (5, 20, 80, 120, and 160 dB/octave). Stimuli were composed of short phrases of the template 'ARTICLE ADJECTIVE NOUN' (e.g., *een blauwe bal*, 'a blue ball'), produced in five variants, viz. with two emotions (sad and happy), with two focus positions (on the adjective and on the noun), and a neutral variant (as much as possible a neutral emotion and equal focus on the adjective and the noun). These were recorded as natural stimuli, and subsequently either the F0 contour, the segment durations, or both, were used to replace those of the neutral variant with, yielding three new half-natural variants for each of two tests (the emotion test and the focus test). Thus, per test the only information available for the discrimination of emotions (in one test) or focus positions (in another test) was the replaced cue. Stimuli were finally vocoded using a 15-channel noise vocoder. In six conditions comprising five filter slopes and a control condition with no vocoding, participants were asked to decide which emotion, or, in a separate test, which focus position was heard. Without vocoding, performance was near ceiling, showing that the emotions and focus positions were successfully conveyed by the speaker. With vocoding, performance ranged from near-chance level for the shallowest slope (5 dB/octave) to high performance at 120 dB/octave, although in general performance for the focus test was lower than for the emotion test. At 160 dB/octave, scores were comparable to those at 80 dB/octave, lower than at 120 dB/octave. For emotion perception, the pattern of scores in the condition including both F0 and duration cues was closest to that including only F0 cues, whereas for focus perception it was closest, albeit less close, to the condition including only duration cues. Together, these results show that steepening the filter slope has positive effects for prosody perception until values as extreme as 120 dB/octave, but that this effect is stronger for emotion than for focus perception because (with the current stimuli) for the former F0 cues are more informative than for the latter. The filter slope of 120

dB/octave could be used a theoretical target value for future speech processing algorithms in CIs.

Chapter 5 reports a study where NH adults received a brief 45-minute training in perceiving either temporal (one group) or melodic contrasts (another group) in vocoded musical stimuli. The goal of this study was to test if this cue-specific training would induce greater reliance on that cue as opposed to the other (non-trained) cue in prosody perception and/or musical melody recognition. A questionnaire filled in before the training showed that the groups did not differ in musical background. After training, participants performed the focus and emotion test described in Chapter 4, a familiar melody recognition test with duration cues, F0 cues or both available, as well as a test assessing if they had a rhythm or melody listening bias when segmenting four-note sequences with ambiguous starting points (the highest note or the loudest note). No significant cross-domain (music to prosody) or cross-cue (duration cues to F0 or melodic cues, or vice versa) training effects were found, although there was a tendency towards a within-cue training effect on familiar melody recognition and, for temporal training, on prosody perception. However, groups did show a segmentation bias in the ambiguous melody test corresponding with the cue they were trained in. Moreover, individual participant-level cross-cue and cross-domain correlations were found. Together, these results suggest that longer cue-specific trainings would have the potential to show positive within- and cross-domain effects improving perception of melodies and prosody.

In Chapter 6, four out of five perspectives of the thesis come together. Six-to-twelve-year-old children with CIs and NH hearing-age matched children were tested on cue usage in four tests on a computer covering perception and production of linguistic and emotional prosody sharing highly comparable stimuli. Besides the core quartet of tests (perception and production of both linguistic and emotional prosody), their general non-verbal emotional and linguistic capacities were tested by means of affective phrases and emotion-

inducing situations, and by means of non-word repetition, respectively. Performance on these tests did not differ significantly between groups showing similar baseline capacities. Before the core tests, children were familiarized with the procedure and the stimuli by means of simple naming and identification tasks; both groups scored near or at ceiling level. In the core tests, the linguistic and emotional prosody perception tests were similar to those described in Chapter 4, including the exact stimuli and the cue availability. In the linguistic prosody (focus) production test, children responded to a question of the form *Is dit een blauwe bal?* ('Is this a blue ball?') where either the adjective, the noun or both contrasted with a picture on a screen. In the emotion production test, children were asked to describe an object picture (e.g., a red chair) and say it in a sad or happy manner depending on the face accompanying the object picture. The emotions and focus positions of the productions were judged by an independent panel of ten Dutch adults. The results showed no difference in cue weighting strategy between groups, nor in the effectiveness of the emotion and focus position productions (this holds for emotion mainly, as the focus perception results could not be analyzed). However, weak correlations between emotional prosody perception and production as well as between emotional prosody perception and production, on the one hand, and non-verbal emotional understanding performance, on the other hand, were found in CI but not, or to a lesser degree, in NH children. Finally, hearing age weakly predicted emotion production but not perception in both groups. Together, these results suggest that CI children at this age, despite being compromised and delayed by a hearing disadvantage, have caught up with their peers when it comes to prosody perception and production.