The processing of Dutch prosody with cochlear implants and vocoder simulations

Velde, D.J. van de

Chapter **4**

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

This chapter is based on:

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

Abstract

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

4.1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2 Methods

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

4.2.1 Participants

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

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

4.2.2 Stimuli

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

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

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

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

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

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

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

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

A. Emotion test

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

B. Focus test

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

C. Durations of adjectives and nouns in the focus test

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

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

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

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

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

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

4.2.3 Procedure

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

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

4.2.4 Statistics

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

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

4.3 Results

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

4.3.1 Neutral stimuli

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

4.3.2 Non-vocoded stimuli

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

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

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

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

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

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

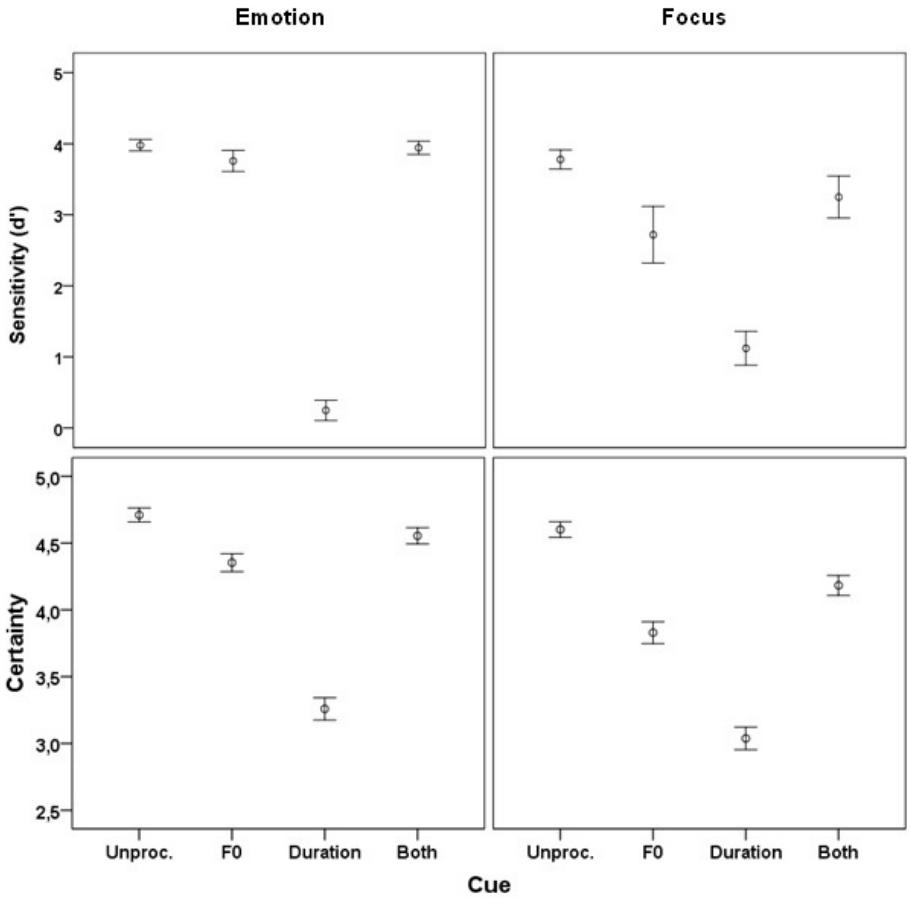

Figure 1. *d'* scores (top panels) and Certainty (bottom panels) scores per Cue (abscissa) and per Test (columns) for the non-vocoded stimuli. Error bars represent 95% confidence intervals. Unproc (Unprocessed) refers to non-resynthesized stimuli. In the F0 condition, F0 information was available for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

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

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

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

4.3.3 Vocoded stimuli

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

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

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

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

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

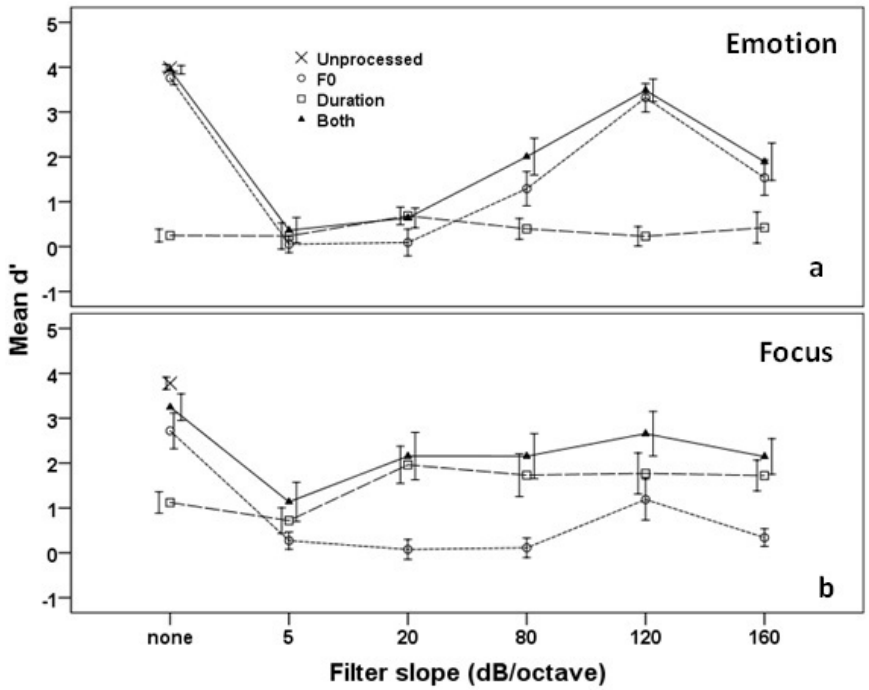

Figure 2. *d'* scores per Filter slope (abscissa) and Cue (line types), for each Emotion discrimination (a) and Focus discrimination (b) tests in the vocoded conditions. Included are the results for the unprocessed condition (crosses) which is only relevant for the 'none' filter slope (non-vocoded condition),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

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

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

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

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

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

4.4 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4.3 Implications for CI users

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

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

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

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

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

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

4.4.4 Limitations

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

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

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

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

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

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

Conclusions

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

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

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

Acknowledgements

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