The processing of Dutch prosody with cochlear implants and vocoder simulations

Velde, D.J. van de

Chapter **3**

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

This chapter is based on:

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

Abstract

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

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

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

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

3.1 Introduction

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

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

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

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

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

Channel interaction refers to the overlap of spectral regions of adjacent electrodes as a result of spread of excitation, which is known to occur in CIs. Interaction between channels reduces the spectral detail of a signal (‘spectral smearing’) and, as a consequence, it compromises speech recognition performance (Crew et al., 2012; Henry and Turner, 2003). Previous studies have tested the effects of spectral smearing by varying the slope of the noise filters. Shannon et al. (1998) tested the effect of spectral smearing on speech recognition by comparing 3, 6, and 18 dB/octave filters with a standard condition with almost no channel overlap. They found that the performance on the recognition of sentences, vowels, and consonants was still very high with the steepest slope, but that the performance with the 3 and 6 dB/octave filters, although above chance, was significantly below that of the standard condition. Fu and Nogaki (2005) showed that speech recognition is better with a 24 dB/octave filter than with a 6 dB/octave filter. In a study by Litvak et al. (2007), the number of channels (15) was kept constant and four different filter slopes were used (5, 10, 20, and 40 dB/octave). Recognition scores for synthetic vowels and consonants decreased with shallower filter slopes. Comparing the results to data from Saoji et al. (2005), they concluded that the slopes in the region from 4 to 30 dB/octave matched the CI performance well (i.e., there was no significant difference). Still, the CI patients performed slightly worse than the NH subjects, which was hypothesized to be due to the NH listeners benefiting from dynamic and temporal cues more than CI users do. More recently, the effects of channel interaction were investigated in a non-speech pitch task. Crew et al. (2012) used sinewave vocoder simulations with 16 band-pass filters and 3 filter slopes (24, 12, and 6 dB/octave) to test musical pitch perception. They replicated the findings of speech perception studies with steeper slopes producing more channel interaction and poorer performance in melodic contour identification. The authors suggested that the results were comparable to those of CI users in Zhu et al. (2011) and to those in Luo et al. (2007) who used sine-wave vocoder simulations as well. It should be noted here, however, that, as explained above, noise-band vocoder simulations might be more representative of a CI on pitch-related tasks due to the limited spectral information provided.

Although studies have been devoted separately to intonation perception and to channel interaction with vocoders, there is a paucity of research on the combination of the two. Therefore, the present study is concerned with the effect of channel interaction on intonation perception. In linguistics, intonation is analyzed as a series of pitch accents, i.e., connected F0 targets lending prominence to some of the syllables in an utterance (Ladd, 1996). Although different acoustical parameters (cues) covary in the production of accents, pitch movements have been claimed to be by far the most important perceptual cue to the presence of an accent. This prominence of pitch movements for accents is in contrast with the perception of stress, for which durational and intensity cues are relatively important (van Heuven and Sluijter, 1996). This difference makes intonation more suitable than stress for the examination of linguistic pitch pattern perception. Moreover, we chose a type of pitch accent in Dutch, the variants of which are believed to be distinguished from each other by pitch movements only: accents with the pragmatic meanings of news, surprise, and predictability (Rietveld and van Heuven, 2009). The drawback of multi-cue phenomena would be that researchers can be less certain that stimuli in which only one of those cues is manipulated are processed as in natural language perception, since in natural language processing the cues would not be isolated. Pure pitch intonation is not uncommon in languages as, for instance, the distinction between questions and declarative sentences can also fall in this category (Rietveld and van Heuven, 2016). Because this type of intonation is expected to be especially difficult to perceive for CI users, the problem of intonation perception by CI users is a real issue. Because the perception of speech melody requires more spectral detail than the perception of the segmental (vowels and consonants) layer of speech (Smith et al., 2002), we used relatively steep filter slopes.

3.2 Methods

The identification of speech intonation contours was examined using a 15-channel noise-band vocoder with a 40 and a 20 dB/octave noise filter. The simulation algorithm used for the present study was the same as the one used in Litvak et al. (2007). Listeners performed an intonation identification task listening to vocoded and unprocessed speech. The stimuli were three basic Dutch melodic shapes, which are thought to be conveyed solely by F0. The intensity and the duration of the stimuli were kept constant by using the same recorded phrase as a basis for superposition of all three intonational variants, whereas the filter slope was systematically varied in order to manipulate the amount of spectral detail available to the participants. Our first hypothesis was that NH listeners would have more difficulty in discriminating melodies in the vocoded than in the unprocessed condition. Our second hypothesis was that a steeper slope (40 dB/octave) should induce better recognition than a shallower slope owing to the smaller amount of channel interaction (Litvak et al., 2007; Shannon et al., 1995; Souza et al., 2011). If participants would be unable to perform the task with these settings, this would imply that more extreme filter slopes, possibly a higher number of electrodes and/or additional (non-pitch related) cues were required for the identification of intonation.

3.2.1 Participants

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

3.2.2 Stimuli

Seven different short phrases were recorded by a male native Dutch speaker (DV) using a Sennheiser MKH416T condenser type microphone and Adobe Audition 1.5 (Adobe Systems, San Jose, CA, USA): *een verbanddoos* (a first aid kit), *een cadeaubon* (a gift certificate), *morgenavond* (tomorrow evening), *over een uur al* (in an hour already), *naar de Veluwe* (to the Veluwe), *naar Leeuwarden* (to Leeuwarden), and *een agenda* (an agenda). This last phrase was only used as a practice stimulus. The phrases were selected because they were semantically and pragmatically logical utterances that could be produced with the three intonation types envisioned, consisted mainly of voiced segments and had a similar stress pattern (viz. main stress on the penultimate syllable, or the antepenultimate syllable if the final syllable was the reduction vowel 'schwa'). The recording sampling rate was 44,100 Hz, and the sampling resolution was 16 bit. The phrases were originally produced by the speaker with a rise–fall intonation to express the information as 'news'. The tokens were stylized using *Praat* software, *Version 5.3* (Boersma & Weenink, 2012) and then re-synthesized to obtain the three different F0 contours. For this manipulation, a rising or falling F0 contour was superimposed on the natural declination of the utterance. The contours were created as pitch accents, so they had the approximate duration of a syllable. The duration of the accent and of the whole phrase was the same for all three contour types per phrase. The phrase durations varied between 740 and 1000 ms. All manipulations had a nine semitone range between the high and low declination. In the rising contour, the range was twelve semitones at the end of the utterance. Since the upper line does not decline towards the end of the utterance, the distance between the lower declination line and the upper line became larger than nine semitones, yet the rise itself still covered only nine semitones. The dynamic range was scaled to 0.99 and the durations of all tokens of each phrase were equalized. This process yielded three versions of each of the seven phrases varying only in the shape of the pitch contour: a falling, a rising, and a rising–falling contour. The resulting set of re-synthesized stimuli comprised 18 stimuli. Importantly, the use of naturally produced re-synthesized stimuli, in comparison with previous studies where synthetic phonemes were used (Litvak et al., 2007; Shannon et al., 1995; Souza et al., 2011), ensured that the stimuli were relatively realistic.

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

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

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

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

3.2.3 Procedure

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

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

In the experimental phase, the entire stimulus set was presented twice and in pairs. Each trial contained one pair of stimuli. This yielded 54 pairs (1 pair for every 1 of 6 phrases × 3 contours × 3 processing conditions) × 2 repetitions = 108 trials. The total set was divided into three blocks of 36 pairs each. In each block, there was a different target pattern to identify. The order of the six blocks was counterbalanced across listeners (four participants for each block). In each trial, two stimuli of the same vocoder condition but a different pitch contour were presented sequentially. The order of the presentation of the two stimuli (target and non-target) was counterbalanced over contour pair within blocks. Conditions within each block were randomized. Each stimulus had a duration of 1000 ms, and the silent interval between trials, during which responses were collected, was 4000 ms. Stimuli were presented at loudness levels of normal speech listening (around 65 dB SPL). The listener had to indicate which of the two stimuli expressed the target pattern in a two-way alternative forced choice task by pressing number 1 for the first or number 2 for the second sound, on the numerical part of the keyboard, with two different fingers. The response accuracy and reaction time were collected. Reaction times were measured from the onset of the second of the pair of phrases played instead of at the end because decision making could in principle start during (not after) the second phrase. Although the durations of the phrases varied, reaction times were not corrected for this because all stimuli were equal between processing conditions.

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

3.3 Results

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

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

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

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

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

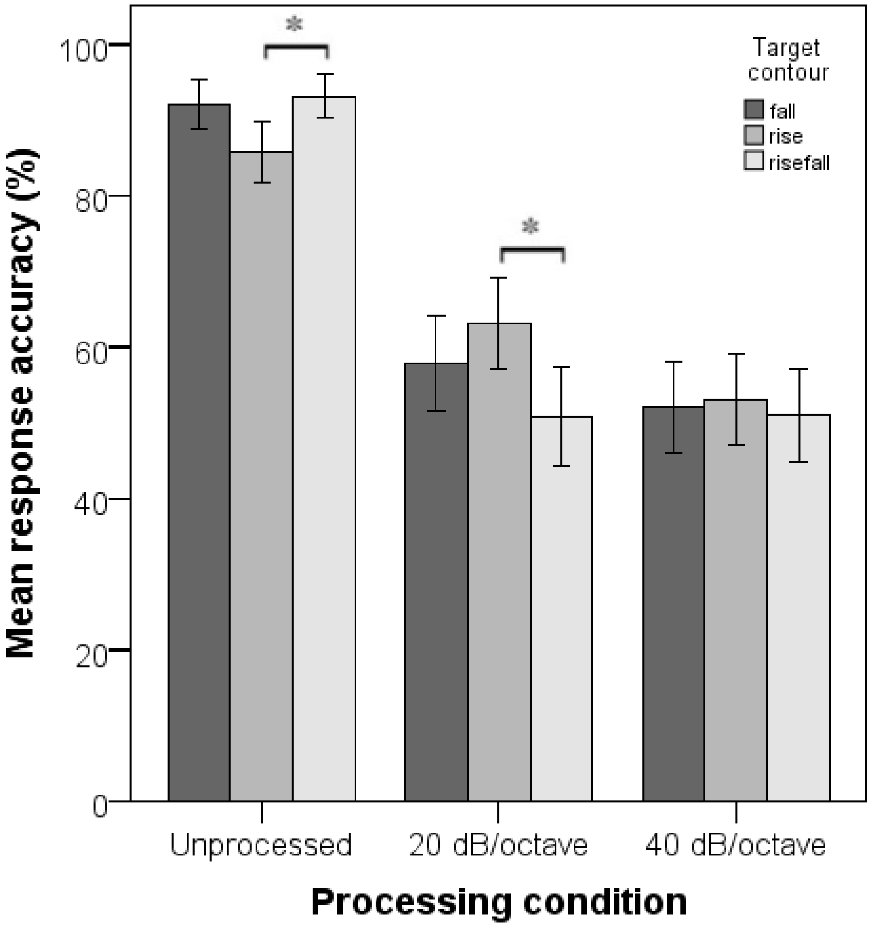

Figure 1. Response accuracy (%, percentage correct) for three intonation contours as a function of processing condition. Error bars display 95% confidence intervals. The overall effect of processing condition between the unprocessed and the 20 dB/octave condition and between the unprocessed and the 40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

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

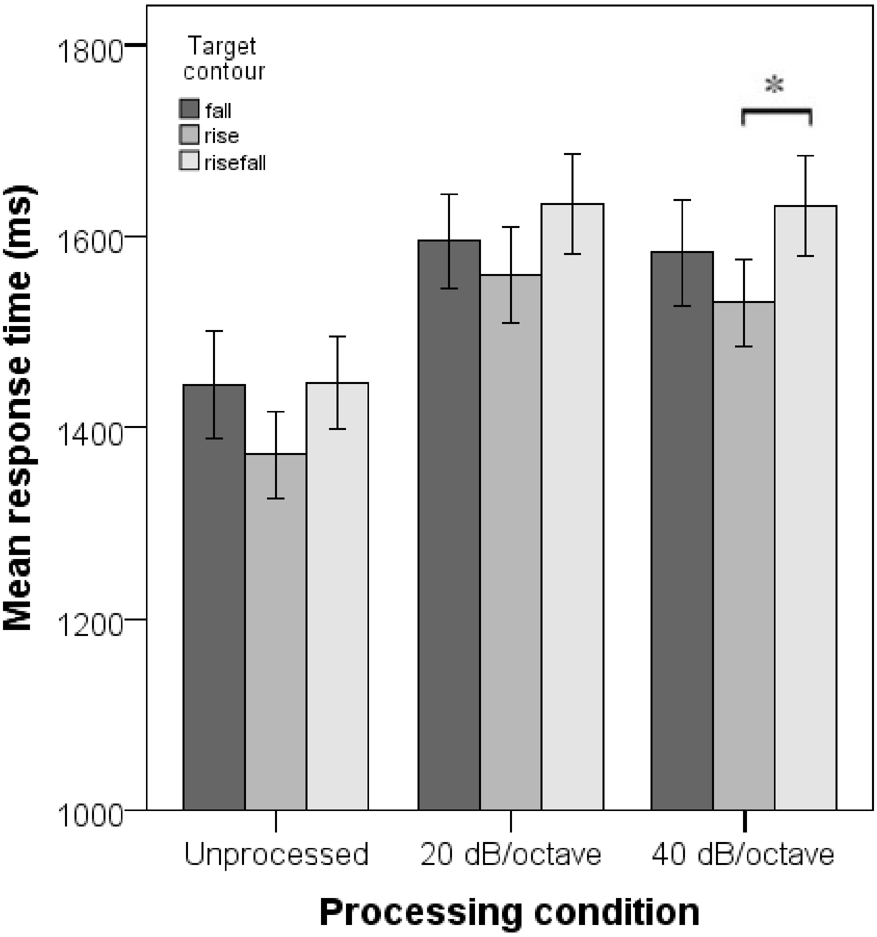

Figure 2. Reaction times (ms) for three intonation contours as a function of processing condition. Error bars display 95% confidence intervals. The overall effect of processing condition between the unprocessed and the 20 dB/octave condition and between the unprocessed and the 40 dB/octave condition,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

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

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

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

* significant (*p* < .05)

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

A test of the three-way interaction between the phrase, the contour, and the processing condition only revealed a significant effect of the phrase in the unprocessed condition, for response accuracy (*p* < 0.001) where the rise–fall was better than the rise, but not for reaction time. Pairwise comparisons showed that the performance for the phrase *een cadeaubon* was significantly poorer than the performance for all the other phrases. A repeated-measure ANOVA revealed that in the unprocessed condition, the performance on the phrase *een cadeaubon* was significantly lower in the rise (48.9%) than in the fall (81.8%) and the rise–fall (82.2%) (*F*(2,134) = 4.60, *p* = 0.001). When *een cadeaubon* was omitted from the analysis, the interaction between the phrase and the contour was no longer significant. In addition, there was no longer a significant interaction between slope condition and target contour. Apparently, the effect of the phrase in the unprocessed condition is the same as the effect of the contour. Therefore, the interaction between the contour and the condition can well be interpreted as an interaction between the phrase and the condition. No interaction between the phrase and the contour was found in the 20 dB/octave filter slope.

3.4 Discussion

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

It is known that NH participants listening to vocoder simulations can take advantage of dynamic and temporal cues for intonation identification. However, since intensity and duration were kept constant in the present study, NH listeners had to exclusively rely on F0 information. When this information becomes obscure, intonation identification turns out to be a(n) (unsurmountable) challenge for vocoder listeners. In contrast, the hypothesis that a steeper noise filter slope (40 dB/octave) would produce better intonation identification than a shallow slope (20 dB/octave) as a result of less channel overlap was not confirmed. This indicates that there were no cues available in the signal that could be effectively used by the listeners. Presuming that the vocoder processing did not eliminate all differences between the contours, any such differences were not sufficient for contour discrimination. Thus, the current type of processing eliminated any effective cues, be it spectral or temporal.

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

The performance of NH listeners in intonation identification with the two filter slopes is not (directly) in agreement with the results of Litvak et al. (2007) who found that subjects could identify synthetic vowels and consonants with slopes as sharp as (20 and 40 dB/octave) or shallower (5 and 10 dB/ octave) than ours. Performance increased when slopes were steeper. These results suggest that identification of purely F0-related intonation was more difficult than segment recognition with the use of noise-band vocoders and with the current settings. This is not entirely surprising, given that listeners can use multiple cues for vowel and consonant identification; not only dynamic or temporal cues, but also additional cues in the frequency domain, e.g. the burst spectra of stop consonants or the noise spectra of fricative consonants (Dorman et al., 1997). This is confirmed in Shannon et al. (1995), who found that considerable reduction of spectral cues still allows for a surprisingly high level of phoneme recognition. Possible reliance on non-spectral cues might explain the effect of sharpening the noise filters from 20 to 40 dB/octave found by Litvak et al. (2007) compared to the present study. The present results were also not in line with those of Crew et al. (2012), who tested musical pitch contour discrimination using a tone (sine-wave) vocoder simulation. The stimuli were relatively similar to those of the current study, as they were controlled for duration and amplitude and a comparable number of band-pass filters were used (16). They found that increasing the filter slopes from 6 through 12 to 24 dB/octave had a positive effect on the performance. This would suggest that a difference could also be found between slopes of 20 and 40 dB/octave, as in our study, since those slopes should enhance spectral differentiation even more. However, since that was not borne out, the discrepancy between the results of Crew et al. (2012) and the present study could be due either to the different vocoder type employed (sine wave vs. noise vocoder) or to the stimuli used (musical vs. linguistic stimuli) or both.

The present data show that NH listeners were not able to use the additional spectral detail provided in the 40 dB/octave condition effectively in intonation identification. Indeed, unsolicited comments by the majority of the subjects suggested that most of the vocoded stimuli given for pattern identification sounded identical. It is possible that more extreme filter slopes might have revealed a clearer effect. A slope of 20 dB/octave, used here as the relatively narrow filter, is still steep compared to (parts of) the settings adopted in some of the previous studies and also on the steep side compared to settings of actual CIs (Fu and Nogaki, 2005; Litvak et al., 2007; Shannon et al., 1998). It should be noted that a majority of subjects (75%) showed better performance on the 20 dB/octave than on the 40 dB/octave condition. Moreover, the overall performance in the 20 dB/octave condition was significantly above chance, whereas the performance on the 40 dB/ octave condition was not. However, because the difference between the significant result and the non-significant result (i.e., the comparison between slope conditions) was not itself significant, these results are not interpreted as reflecting a real performance difference. The finding that there was an interaction between the performance on the target contour in the 20 dB/octave and the unprocessed condition, can be accounted for by assuming that contour identification could have been driven by general processing restrictions: the double pitch change that is involved in the rise–fall contour (one change for the rise and another for the fall) would be easier to recognize than a single rise or fall because there are two points for identification instead of one. However, two types of result speak against this hypothesis. First, the mean reaction time, although just a trend, for rise–fall was slower (1447 ms) than for the rise (1372 ms), suggesting that the rise–fall was more difficult to process (see below). And second, this contour effect in the unprocessed condition was much larger for the phrase *een cadeaubon* than for the other phrases. The effect disappeared when *een cadeaubon* was removed from the analysis, so either the double pitch change explanation does not hold or it explains only the effect found for *een cadeaubon*. The 20 dB/octave condition showed the opposite effect. The better performance for the rise than the rise–fall in this condition suggests that the double pitch change explanation cannot account for the data or that a different mechanism is responsible for processing in the 20 dB/octave condition than in the unprocessed condition. A possible explanation is in terms of processing time. Due to the poor spectral definition, the listener needed additional processing time. This time could have been available in the rise but not in the rise–fall, since the rise–fall accent was a relatively short part of the phrase (160–350 ms depending on the phrase), whereas the rise allowed the total time of the phrase (700–850 ms) for analysis. The trend in the reaction times in the 20 dB/octave condition were in line with this account, because their mean was longer for the rise–fall (1634 ms) than for the rise (1560 ms). Since the rise did not score uniformly better than the other contours, the bigger tone range at the end of the rise (as compared to the other contours) was not an important cue to identification. However, since participants could be using different strategies for different processing conditions, it could be the case that the tone range was crucial in the 20 dB/octave condition but not in the unprocessed condition. Yet, it is still unlikely that it plays a role because the larger tone range is only virtual, in that it is only larger if the declination of the phrase is extrapolated.

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

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

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

Conclusions

The present results confirmed that the limited F0 resolution provided by noise-band vocoder simulations reduces the ability to identify intonation patterns by means of pitch alone. On average, no significant difference was found between a 20 dB/octave filter and a 40 dB/octave filter. These slopes were relatively steep compared to conditions from previous research that were found to suffice for discrimination of segmental speech information. This study therefore suggests that even relatively extreme filter slopes do not provide sufficient spectral resolution for the identification of intonation that is conveyed purely by F0 movements. No direct or indirect cues to F0, such as the temporal envelope of higher harmonics that could be in the signal were effective. There are CI devices that use comparable settings, apart from probably shallower slopes. This has implications for explanations of the perception of intonation by the relevant users. Our message in essence is that if intonation perception by CI users succeeds, it is plausible that an extremely shallow slope is a benefit for pure intonation perception and/or that the users exploit not just pitch but additional bottom-up cues (such as phrase level temporal or dynamic cues) and/or top-down (contextual) cues. An alternative for increasing the spectral resolution worth exploring is to increase the number of electrodes and use accordingly steep filter slopes. We further suggest that future research address a broader range of slope conditions and compare them to conditions in which alternative or additional cues are present. Finally, it is recommended that tests also be carried out with actual CI users in order to examine which vocoder setting comes closest to their performance.

Acknowledgements

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