Speech perception with F0mod, a cochlear implant pitch coding strategy

Cite as: Francart T, Osses A, Wouters J. Speech perception with F0mod, a cochlear implant pitch coding strategy. Int J Audiol. 2015 Feb 20:1-9

Running title: Speech perception with F0mod

Key words: cochlear implants, pitch perception, sound coding, speech recognition, noise

Tom Francart ¹

ExpORL, Dept. Neurosciences, K.U.Leuven

Alejandro Osses

ExpORL, Dept. Neurosciences, K.U.Leuven

Jan Wouters ²

ExpORL, Dept. Neurosciences, K.U.Leuven

Tom Francart and Alejandro Osses declare equal contributions to this study and should be considered joint first authors

Corresponding author:

Tom Francart

ExpORL, Dept. Neurosciences, KU Leuven,

O & N 2,

Herestraat 49 bus 721,

B-3000 Leuven, Belgium

tel: +32 (0)16 37 98 40

fax: +32 (0)16 33 04 86

tom.francart@med.kuleuven.be

¹tom.francart@med.kuleuven.be

²jan.wouters@med.kuleuven.be

List of acronyms

ACE Advanced Combination Encoder
ACE OP ACE with own processor
CI Cochlear implant
CVC consonant vowel consonant
F0 Fundamental frequency
LIST Leuven intelligibility sentence test
PB Paul Bragshaw sentence database
SNR Signal to noise ratio
SRT Speech recognition threshold

gErr gross-error rate
uvErr unvoiced-error rate
vErr voiced-error rate

Abstract

Objective The fundamental frequency modulation (F0mod) sound processing strategy was developed to improve pitch perception with cochlear implants. In previous work it has been shown to improve performance in a number of pitch-related tasks such as pitch ranking, familiar melody identification and Mandarin Chinese tone identification. The objective of the current study was to compare speech perception with F0mod and the standard clinical Advanced Combination Encoder (ACE) strategy.

Design F0mod was implemented on a real-time system. Speech recognition in quiet and noise was measured for 7 cochlear-implant listeners, comparing F0mod with ACE, using three different Dutch-language speech materials. Additionally the F0 estimator used was evaluated physically, and pitch ranking performance was compared between F0mod and ACE.

Results Immediately after switch-on of the F0mod strategy, speech recognition in quiet and noise were similar for ACE and F0mod, for 4 out of 7 listeners. The remaining 3 listeners were subjected to a short training protocol with F0mod, after which their performance was reassessed, and a significant improvement was found.

Conclusions As F0mod improves pitch perception, for the 7 subjects tested did not interfere with speech recognition in quiet and noise, and has a low computational complexity, it seems promising for implementation in a clinical sound processor.

1 Introduction

Compared to normal-hearing (NH) listeners, users of a cochlear implant (CI) have poor pitch perception. To perceive pitch, which is related to the fundamental frequency (F0) of a periodic sound, NH listeners can make use of three different acoustic cues: temporal fine structure, periodicity and spectral information. Temporal fine structure provides the most salient pitch, and periodicity a rather weak one. With CI stimulation, temporal fine structure cannot be perceived and periodicity cues are not always appropriately coded. Spectral cues on the one hand have a poor resolution and on the other hand are influenced by speech coding such as the filter bank used and therefore may not clearly reflect F0.

In the Advanced Combination Encoder (ACE) strategy (Vandali et al., 2000), with is the default strategy of sound processors manufactured by Cochlear, the acoustic signal is sent through a filter bank and in each band the envelope is extracted, discarding all temporal fine structure. Of these envelopes the N largest ones are selected and used to modulate fixed-rate pulse trains. To what extent envelope modulations due to F0 are coded depends on stimulus properties such as F0, spectral content and level in combination with strategy parameters such as filter bank and automatic gain control.

While there have been many attempts to give CI users access to temporal fine structure cues, success has been very limited so far (reviewed by Francart & McDermott, 2013). There have also been attempts to improve spectral resolution by changing parameters of the filter bank (Geurts & Wouters, 2004; Kasturi & Loizou, 2007; Laneau et al., 2004), or using current steering (Donaldson et al., 2011; Luo et al., 2010; Wu & Luo, 2013), with varying degrees of success. Finally, there are several strategies that aim to improve periodicity pitch perception by increasing modulation depth due to F0 and modulating electrodes synchronously (Green et al., 2004, 2005; Laneau et al., 2006; Milczynski et al., 2012, 2009; Vandali et al., 2005; Vandali & van Hoesel, 2011, 2012), also with varying degrees of success in terms of effect on pitch perception and speech recognition.

Pitch strategies that aim to improve periodicity pitch are generally based on the notion that for good periodicity pitch perception with CIs, there need to be sufficiently deep modulations (Geurts & Wouters, 2001; McKay et al., 1995) and the modulations need to be synchronised across channels to some extent (Geurts & Wouters, 2001). Most strategies work by first removing existing modulations above a certain cut-off frequency (e.g., 30 Hz), as mainly very low modulation frequencies are important for speech perception (Shannon et al., 1995), and then reintroducing modulations based on a modulation waveform constructed based on an estimate of F0. The following modulator shapes have been used: the enhanced broadband envelope (MEM, Vandali et al., 2005), rectangular gating (F0Sync, Vandali et al., 2005), a sine wave (Laneau et al., 2006; Milczynski et al., 2012, 2009), exponential decay (Vandali & van Hoesel, 2011, 2012) and a sawtooth (Green et al., 2004, 2005). These artificially introduced modulations are always deep and in most strategies synchronised across all channels.

While such strategies have been shown to improve pitch perception, it should be noted that they do not yield performance comparable to NH listeners. This is due to a lack of temporal fine structure information, lack of resolution of spectral cues, but also the rate limitation found for temporal pitch: most CI users cannot distinguish pitches above around 300 Hz purely based on rate information (e.g., Shannon, 1992).

The Sawsharp strategy (Green et al., 2004, 2005) improved identification of intonation and of pitch movement in diphthongs. It had a detrimental effect on vowel recognition and formant frequency discrimination. The authors hypothesised that this might improve with perceptual adaptation to the strategy.

Vandali et al. (2005) compared the clinical standard Advanced Combinator Encoder (ACE)

strategy to four different experimental strategies: Peak Derived Timing (PDT), Modulation Depth Enhancement (MDE), F0 Synchronized ACE (F0Sync), and Multi-channel Envelope Modulation (MEM). Compared to ACE, performance on a pitch ranking task was improved for MEM, MDE and F0Sync. No difference in speech perception was found between ACE, MDE, and MEM for English sentence tests in multi-talker babble for a signal to noise ratio (SNR) fixed at 10 dB. All strategies were implemented on a real-time platform and did not use the clean speech signal.

The eTone strategy (Vandali & van Hoesel, 2011, 2012) mixes the original signal in every band with a version modulated with an exponential decay shape, with a mixing ratio based on an estimate of harmonicity. Compared to ACE, it was found to improve pitch ranking for stimuli three semitones apart. No effect was found on recognition of English sentences in multi-talker babble, with a subject-dependent fixed SNR, ranging from 4 to 12 dB.

The F0mod strategy uses a sinusoidal modulator, synchronous across all channels, based on an estimate of F0 (Geurts & Wouters, 2001; Laneau et al., 2006). Laneau et al. (2006) found that compared to ACE, F0mod improved F0 discrimination of musical notes for different instruments, melody recognition of familiar Flemish songs (with all rhythm cues removed), and estimates of musical pitch intervals. No effect was found for F0 discrimination of single-formant stimuli. In a more tightly controlled paradigm, Milczynski et al. (2009) found that compared to ACE, F0mod improved pitch ranking, melodic contour identification, and familiar melody identification. In a follow-up study, Milczynski et al. (2012) compared ACE and F0mod for speech recognition of Mandarin Chinese, which is a tonal language, in which pitch determines the lexical meaning of certain phonemes. An off-line implementation of F0mod was used, which made use of the clean speech signal to estimate F0. Significantly better lexical tone perception was found with the F0mod strategy than with ACE for the male voice. No significant difference in recognition of Mandarin Chinese sentences in noise was found between F0mod and ACE. This can mean that more training with F0mod was required than provided in this acute experiment, that there was no effect on speech recognition in noise, or that effects of improved tone identification and decreased overall speech recognition counteracted each other.

In the current study, we implemented the F0mod strategy on a real-time system and tested speech perception in quiet and noise for a large range of Dutch-language speech materials. Dutch is a non-tonal language, so no improvement due to better access to tonal information can be expected. Therefore our hypothesis was that there was no significant difference in speech perception between ACE and F0mod. We also conducted a brief pitch ranking test, to reproduce the results obtained by Milczynski et al. (2009). In the current real-time implementation the clean speech signal was not available to the strategy, yielding more realistic circumstances.

2 General Methods

2.1 Strategy

The F0mod strategy is described in detail by Milczynski et al. (2009). Briefly, for unvoiced segments of the signal, it is identical to ACE, and for voiced segments the envelopes in all channels are modulated synchronously based on an estimate of F0. An autocorrelation-based F0-extractor is used, which for each time frame yields an estimate of F0 and of the harmonicity. The latter is used to decide whether the signal was voiced and therefore whether the modulations should be introduced. When modulations are introduced, existing modulations are reduced by low-pass filtering the envelope with a cut-off frequency of 60 Hz.

Figure 1: Schematic representation of the test setup

2.2 Apparatus

In previous studies (Laneau et al., 2006; Milczynski et al., 2012, 2009) an off-line MATLAB implementation of ACE was used. For the current study, we implemented F0mod on a real-time system based on a model of ACE provided by Cochlear Ltd. This means that the ACE processing was identical to the processing in clinical devices. A schematic overview of the test setup is shown in figure 1.

A Simulink/xPC system was used, consisting of a host PC, running MATLAB and Simulink, which compiled the Simulink model into real-time code, which was sent to and executed by a dedicated real-time target machine (Speedgoat GMBH, Liebfeld, Switzerland). Signal processing parameters could be changed in real-time from the Simulink environment. The hardware was set up according to figure 1: the microphone signal from a CP810 sound processor housing was captured by the audio interface of the target machine. It was then processed according to the ACE or F0mod strategy as implemented in the Simulink model. Then the signal was sent to a custom Stingen box provided by Cochlear, which generated the RF signal on a coil to let the subject's implant provide the required electrical stimulation. As a reference, the same tests were conducted with the subject's own processor.

Acoustic stimuli were presented using the APEX 3 program developed in our lab (Francart

et al., 2008), which controlled an RME Multiface II sound card, connected to a Genelec 8020A loudspeaker. The stimulation level was calibrated using a Bruel & Kjaer 2250 sound level meter and 4189 microphone. All tests were conducted in a sound-proof room.

2.3 Subjects

Seven subjects were recruited from the clinical population of the University Hospital Leuven (UZLeuven). They were volunteers and signed an informed consent form. This study was approved by the local medical ethical committee. Various subject information is listed in table I. The clinical processors were of the Freedom or CP810 type. S16 and S17 had some limited residual hearing in the non-implanted ear. During testing their hearing aid was switched off and for the pitch ranking test the non-implanted ear was plugged. The subjects participated in 2 to 4 test sessions of 90 to 120 minutes, including breaks.

2.4 Fitting

All sound processor fitting parameters were obtained from the subjects' clinical maps, and were identical for the ACE and F0mod conditions. The following parameters were the same for all subjects: a stimulation rate of 900 pps, 8 maxima selected, monopolar (MP1+2) stimulation, a phase width of 25 μ s, and an inter-phase gap of 8 μ s. Threshold and comfort levels were subject dependent. Note that Milczynski et al. (2009) used a stimulation rate of 1800 pps, contrary to the 900 pps used by default clinically. For S12 one electrode was disabled, and for the other subjects all 22 electrodes were enabled. In the xPC conditions all front-end processing (auto-sensitivity, ADRO, noise suppression, etc.) was disabled. With the subject's own processor the front-end processing was not modified.

When first using the experimental processor the ACE strategy was switched on and the volume (i.e., reduction of comfort level by a fraction of the dynamic range, starting at 50%) was gradually increased until the subject reported the same loudness as with their own processor for running speech. During all experiments the same volume setting was used for ACE and F0mod.

3 F0 estimator evaluation

F0mod uses an autocorrelation-based F0-estimator. In this section the results of an evaluation of the F0 estimator using instrumental measures are described.

3.1 Methods

We used two databases of recorded speech to evaluate the F0-estimator: the database recorded by Bagshaw et al. (1993) (PB) and the LIST sentence material (van Wieringen & Wouters, 2008). The database recorded by Bagshaw et al. (1993) is widely used to evaluate F0 estimators. It contains a set of 50 sentences uttered by a male and a female English speaker, with corresponding F0 contours derived from laryngograph recordings. F0 estimates outside the range of 75 to 400 Hz were considered as unvoiced, corresponding to the evaluation of the eTone F0 estimator (Vandali & van Hoesel, 2011). The estimated and reference F0 contours were optimally aligned. The sentences were either used in quiet or in steady state white noise. While a speech spectrum weighted noise is more perceptually relevant, we used a white noise here to be able to compare our results to those reported in the literature. The use of speech weighted noise reduced the global performance of the F0 estimation by 0.7% to 2.8%.

We also used the LIST sentence material (van Wieringen & Wouters, 2008), as it was also used in the perceptual evaluation. For this material no laryngograph recordings are available, so we used the autocorrelation pitch extractor available in Praat (Boersma & Weenink, 2001)³, applied to the clean speech signal. This pitch extractor was chosen because (1) it has been used in other evaluations of pitch determination algorithms (de Cheveigne & Kawahara, 2002) and; (2) it is a robust implementation of an autocorrelation F0 extractor and includes sophisticated heuristics (Boersma, 1993). However, when interpreting the results, one should keep in mind that the Praat estimates may differ from the true F0. F0 estimates outside the range of 75 to 300 Hz were considered as unvoiced, corresponding to the default settings of the F0mod F0-estimator (Milczynski et al., 2012, 2009). To obtain error measures relevant to the perceptual results, no alignment was done between reference and estimated F0 contours. The sentences were either used in quiet or steady state noise with the spectrum corresponding to the long term average spectrum of the speech material, the latter in contrast to the PB sentences where white noise was used.

To compare between F0mod and the reference F0 contours, four different error measures were used: voiced-to-unvoiced error rate, unvoiced-to-voiced error rate, gross error rate (Rabiner et al., 1976) and error rate (de Cheveigne & Kawahara, 2002; Vandali & van Hoesel, 2011). These error measures were applied to frame-by-frame F0 estimates with a frame length of 22.5 ms and a hop size of 9.2 ms.

The voiced-error rate (vErr, %) is defined as the number of voiced-to-unvoiced errors (i.e., reference yields a voiced decision and F0mod an unvoiced decision) normalised to the total number of the voiced segments. The effect of a vErr on a signal processed by F0mod would be that a voiced segment is not modulated (i.e., the standard ACE processing is active). Similarly, the unvoiced-error rate (uvErr, %) is defined as the number of unvoiced-to-voiced errors normalised to the total number of the unvoiced segments. The effect of an uvErr on a signal processed by F0mod would be that an unvoiced segment is erroneously modulated. As this type of error could have a negative effect on speech perception, it was minimised by the choice of parameters for the F0 estimator.

A gross error is defined as a difference in F0 estimate of 20% or more between the reference and F0mod, when both the reference and F0mod yield a voiced decision. Correspondingly, the gross-error rate (gErr, %) is defined as the number of gross errors normalised to the total number of the voiced segments. Finally, the error rate (Err, %) is defined as the number of gross and voiced-to-unvoiced errors normalised to the number of voiced segments of the database. This definition was also used by de Cheveigne & Kawahara (2002) and Vandali & van Hoesel (2011).

These four measures were calculated for both speech databases in quiet and in noise with SNRs ranging from -5 to 20 dB.

3.2 Results and discussion

In Figure 2 and 3 the different error measures are shown for the two speech databases. While the magnitude of errors differs, the tendencies are the same for all speech materials. vErr, gErr and error rate increase with decreasing SNR, and uvErr decreases with increasing SNR. Errors were generally larger for LIST than for PB. This could be due to (1) errors in the reference, as no laryngograph data was available, (2) the misalignment between reference and test F0 contours and (3) the use of speech spectrum weighted instead of white noise. However, the error due to misalignment was small. When optimally aligning the test and reference contours, uvErr was most affected and improved by 5% in quiet, and by less than 2% for SNRs below 10 dB.

³The frequency contours were obtained by using the following Praat command: To pitch (ac)... 0.01 75 15 no 0.01 0.45 0.01 0.35 0.14 400

To compare our results using the PB database (average between male and female speakers) with those obtained by Van Immerseel & Martens (1992) with the SIFT algorithm, we renormalised the error measures to total speech time. The results for the conditions in quiet and at an SNR of 0 dB were 3.25% and 12.59% for vErr rate, 5.05% and 4.04% for uvErr rate and 2.00% and 6.69% for gErr rate, respectively, resulting in a global error of 10.3% in quiet and 23.32% for the condition in noise. This is comparable to the results obtained for the SIFT algorithm, where for the same conditions a global error of about 9% and 27% was reported. Furthermore, the uvErr showed a similar tendency with decreasing values for decreasing SNR.

Our tests with the PB database were designed such that the our results are directly comparable to those obtained with eTone (Vandali & van Hoesel, 2011). In quiet, the error rate for their “standard criterion” for male and female PB speakers was 2.38% and 3.98% respectively compared to the 11.09% and 11.78% for F0mod. At an SNR of 0 dB the reported percentages were 7% and 10% for eTone, while for F0mod they are 46.11% and 32.95% respectively. The use of a harmonic sieve in eTone might be more noise robust than finding the maximal autocorrelation as is done for F0mod.

For both speech materials in quiet the error rate falls within the range reported by de Cheveigne & Kawahara (2002), who evaluated eleven pitch estimation algorithms. The average across the male and female speakers using the PB database (abbreviated as DB2) for a total of eleven algorithms was 8.4% with minimum and maximum errors of 1.4% and 19.0%. It should be noted that error rate is a performance measure intended to evaluate accuracy of an F0 extractor more than to evaluate a voiced/unvoiced decision. Therefore, the error rate is probably not the most informative error measure in the current context.

The F0 estimator implemented in F0mod has a good performance at positive SNRs and moderate performance at low SNRs, especially compared to eTone, but it is much less computationally expensive and given the general poor periodicity pitch discrimination ability of CI listeners, this may not make a difference in practice. To our knowledge, there are no studies evaluating the perceptual tolerance to F0 error measures and its influence on speech perception.

4 Experiment 1

4.1 Methods

4.1.1 Speech recognition

Speech recognition was measured for three processing strategies: ACE with the subject’s own processor (ACE OP), ACE implemented on the real-time xPC system with pre-processing disabled (ACE xPC), and F0mod implemented on xPC, with three different speech materials. Speech was always presented at a level of 60 dB A, and when required noise was set to the appropriate level to obtain the desired SNR.

Word recognition was tested using the Flemish Lilliput consonant-vowel-consonant (CVC) words. They are uttered by a female speaker and presented in lists of 10 words. Afterwards a percentage phonemes correct score is calculated. Word recognition was tested in quiet, and in noise at a fixed SNR of 10 dB, using steady state noise with spectrum matched to the spectrum of the speech material. Two lists of 10 words in quiet and in noise were administered to each subject in each sound processing condition.

Closed set sentence recognition in noise was tested using the Flemish Matrix test (e.g., Hochmuth et al., 2012; Wagener & Brand, 2005). Steady state noise with spectrum matched to the long term average spectrum of the speech was used. The tests consist of sentences constructed from a

Figure 2: Error measures: vErr, uvErr, gErr for the evaluated speech materials. Error bars indicate the standard deviation of the average across sentences.

Figure 3: Error rate (Err) measure for complete set of PB and LIST sentences.

matrix of ten names, verbs, numerals, nouns and adjectives. Any combination of these elements is grammatically correct and the subject is instructed to repeat the sentence he or she heard, bearing in mind the matrix of possibilities. While there is a small procedural learning effect to such Matrix tests, the advantage is that thereafter the test can be repeated infinitely, in contrast to open set materials where the same sentence can only be used once for each subject. We conducted the Matrix test in quiet and at an SNR of 10 dB. Two lists of 10 sentences were conducted in each of these two conditions. For the Matrix test, condition ACE OP was not conducted to save testing time.

Open set sentence recognition in noise was tested using the LIST-f sentences (van Wieringen & Wouters, 2008). Steady state noise with spectrum matched to the long term average spectrum of the speech was used. The sentences are uttered by a female speaker and are very clearly articulated at a low speech rate. This means that most CI users can understand nearly 100% correct in quiet and speech reception thresholds (SRTs) in noise can be quite low. The average SRT for NH listeners is -7.8 dB. We determined the SRT (i.e., the estimated SNR at which the score is 50% correct) using an adaptive procedure, based on sentence scores. A sentence was considered correct if all keywords in a sentence were correctly identified (van Wieringen & Wouters, 2008). The speech level was held constant at 60 dB A. Starting at -4 dB, the SNR of the first sentence of each list was increased in steps of 2 dB until the sentence was identified correctly. Subsequently, the SNR was varied in steps of 2 dB in a one-up, one-down procedure (Plomp & Mimpen, 1979). The SRT was calculated as the average of the last 6 SNRs, including the SNR calculated based on the last response. If the standard deviation of the last 6 SNRs was larger than 2.5 dB, the measurement was excluded and the test was repeated. This happened twice across all experiments. At least two lists of 10 sentences were conducted for each subject and condition. No practice tests were conducted.

The order of testing (speech material) was randomly determined for each subject, but in blocks of 2-3 tests per processing strategy, to avoid switching strategies too often. Condition ACE OP was conducted first, followed by randomly either ACE xPC or F0Mod.

4.1.2 Pitch ranking

To validate whether the same results could be obtained with the current real-time implementation as previously obtained by Milczynski et al. (2009), we repeated their pitch ranking experiment, in which they found a significant improvement in pitch ranking with F0mod compared to ACE for reference tones of 131Hz and 165Hz, and no significant differences for 104Hz, 205Hz and 262Hz. The stimuli and procedures, including loudness balancing, were exactly the same, except that they adjusted stimulus level by varying the volume (maximum current level on each electrode in percentage of dynamic range), while we varied the overall level of the acoustical stimulus.

For details about stimuli and procedures we refer to Milczynski et al. (2009). In summary, the stimuli were 500-ms long harmonic complexes with F0s ranging from 104 to 330 Hz. Each complex consisted of 16 harmonics. Stimuli with F0s of 104, 147, 208 and 294 Hz for each strategy were individually balanced in loudness against the reference stimulus (ACE, 131 Hz F0) in an adaptive 2I-2AFC procedure. For the remaining frequencies, the adjustment was linearly interpolated. For the pitch-ranking task, five frequency registers were created with reference F0s of 104, 131, 165, 208, and 262 Hz, respectively. Each register contained four comparison tones that were one, two, three, or four semitones higher in F0 than the reference. During each trial two intervals were presented, one containing the reference and the other the comparison tone, in random order. The subject's task was to indicate which interval sounded higher in pitch. All frequency registers were tested separately in one block. Levels were roved by ± 4 dB, drawing from a uniform distribution.

Figure 4: Phoneme score for the Lilliput words in quiet. Error-bars for individual subjects indicate the standard deviation.

Figure 5: Phoneme score for the Lilliput words in stationary noise at an SNR of 10 dB. Error-bars for individual subjects indicate the standard deviation.

4.2 Results

4.2.1 Speech recognition

The speech recognition results are shown in figures 4 to 7. For each speech material a repeated measures analysis of variance (ANOVA) was performed, after checking normality and if necessary performing a Greenhouse-Geisser correction for non-sphericity. All percentage scores were transformed to rationalised arcsine units (RAU) (Studebaker, 1985) before the statistical analysis.

In figures 4 and 5 percentages phonemes correct of the Lilliput CVC words are shown. For S12 no measurement was made for condition ACE OP and for S16 only one measurement was made for condition ACE xPC. A repeated measures ANOVA of RAU score with factor strategy (ACE OP, ACE xPC, F0mod) and noise type (quiet or noise) indicated a significant effect of noise type ($F(1,4) = 10.96, p = 0.03$), but no effect of strategy or the interaction between strategy and noise type.

In quiet the difference in score between ACE xPC and F0mod lies with 90% confidence between -11% and 5% (paired t-test). In noise this difference lies between -10% and 10%.

In figure 6 percentages words correct of the Matrix test are shown. A repeated measures ANOVA of transformed percentage correct with factor strategy (ACE, F0mod) indicated no effect of

Figure 6: Word score for the Matrix sentences in stationary noise at an SNR of 10 dB. Error-bars for individual subjects indicate the standard deviation.

Figure 7: SRT obtained with the LIST sentences in noise. Error-bars for individual subjects indicate the standard deviation.

strategy ($p = 0.91$). The difference in score between ACE xPC and F0mod lies with 90% confidence between -9% and 6% correct.

In figure 7 the SRTs obtained with the LIST speech material are shown. For S12 no measurement was made for condition ACE OP. As the distribution of SRTs was significantly different from normal, we used Wilcoxon signed rank tests to compare between ACE xPC and ACE own processor, and between ACE xPC and F0mod. Neither difference was significant. The difference between ACE xPC and ACE own processor lies with 90% confidence between -0.1 dB and 1.4 dB. The difference between ACE and F0mod lies with 90% confidence between -0.1 dB and 5 dB. Assessing individual differences, there appeared to be a decrement in performance with F0mod for S12, S14 and S17. For the remaining subjects the differences were very small.

4.2.2 Pitch ranking

As expected, there were large inter-subject differences in pitch-ranking performance. Visual assessment of the individual results indicated that subject S13 performed nearly perfectly with ACE, so due to this ceiling effect, no difference with F0mod was found. Subject S12 had average performance with ACE and no improvement with F0mod. The remaining four out of six subjects (S11, S14, S15, S16) showed improvements with F0mod in various conditions. S17 did not take

part in the pitch ranking experiment. Separate ANOVAs were performed for each register. No significant effects were found, except for a significant improvement of F0mod over ACE for register 131 Hz ($F(1, 4) = 16.78, p = 0.01$).

4.3 Discussion

We investigated a real-time implementation of F0mod, a strategy designed to improve pitch perception with cochlear implants. We evaluated speech recognition in quiet and noise of 7 subjects and pitch ranking performance of 6 subjects.

4.3.1 Speech recognition

Most subjects did not perceive a big difference in sound quality between ACE and F0mod. There were occasional reports of subjects perceiving their own voice as lower, or the sound quality being slightly different but not worse.

For CVC words in quiet and noise, and for the LIST sentences in noise there was no difference between ACE OP and ACE xPC, indicating that the xPC system reliably replicated the function of the clinical processor. Any small individual differences could be caused by differences in pre-processing: in condition ACE OP pre-processing was left unchanged, while in condition ACE xPC all pre-processing was disabled. For CVC words in quiet and noise, and the Matrix sentences in noise there was no difference between ACE xPC and F0mod, neither on average, nor for any individual subject, considering whether the error bars overlapped, indicating that the application of F0mod did not have an effect on speech perception in quiet or noise.

For the LIST sentences in noise there was no significant difference between ACE and F0mod either, but for subjects S12, S14 and S17 the SRT appeared worse. One difference with the Matrix sentences is that the LIST sentences are on purpose slowly spoken and clearly articulated, and therefore easier to understand. This leads to lower SRTs, and therefore in the adaptive procedure lower average SNRs were used. Additionally, the adaptive procedure operated around the 50% correct point, whereas the scores for the Matrix sentences at a fixed SNR of 10 dB were around 80% correct. This effect was further investigated in experiment 2.

4.3.2 Pitch ranking

We found a significant improvement with F0mod over ACE for a reference tone of 131 Hz, while Milczynski et al. (2009) found significant differences for both 131 and 165 Hz. The discrepancy for the 165-Hz condition is probably due to differences in subjects tested. Qualitatively the pitch-ranking results are consistent with those obtained in the more extensive study of Milczynski et al. (2009). The main differences between Milczynski et al. (2009) and the current study are the overall stimulation rate (1800 versus 900 pulses per second) and the implementation (off-line versus on-line). There should not be a major effect of implementation, as we verified using instrumental measures that the electrodiagrams produced by the two implementations were very similar. If there would be an effect of stimulation rate, it would be expected to mainly show up in the higher registers, presumably due to a more detailed representation of the modulation waveform, which was not the case. Remaining differences are probably due to the effect of training: with the application of F0mod, due to interaction with the maxima selection implemented in ACE, spectral cues can change relative to what the subjects are used to in their clinical processor, because for F0mod all channels are modulated synchronously, which leads to different maxima being selected compared to asynchronous envelope modulations.

Figure 8: Training sequence for Matrix sentences with 10 dB SNR for subjects S12 and S17 and in 5 dB SNR for subject S14. The x-axis indicates the presentation order of tests.

5 Experiment 2

5.1 Methods

To explain the difference in speech recognition in noise with the LIST materials for S12, S14 and S17, we first hypothesised that it was an effect of particular SNRs at which performance with F0mod would be worse, which would explain across-subject differences. We measured full performance-intensity functions for ACE and F0mod with S14. The curves were parallel, so no effect of SNR was found.

To further investigate whether this difference for S12, S14 and S17 could have been due to a lack of experience with the F0mod strategy, we designed a training protocol which was run during an extra test session, and evaluated its effect. After switching on the F0mod strategy, two lists of the LIST sentences were conducted, followed by two lists of the Matrix in quiet and in noise (SNR of 10 dB) respectively. Then one list of the Matrix was conducted with visual feedback after each trial and opportunity to replay the sentence. Then a number of Matrix lists was conducted until performance plateaued or test time was used up. Note that the first two lists of the LIST sentences are not considered part of the training. The results of this latter sequence of results are shown in figure 8. Finally, two new lists of the LIST were conducted, first with F0mod, then with ACE xPC. In case the word score with the Matrix sentences at 10 dB SNR was higher than 50%, an SNR of 5 dB was used for the Matrix sentences during the training, otherwise the SNR was kept at 10 dB. This design allowed to monitor the training using the outcomes of the Matrix lists and to validate the training using an independent test-set that was not used during training (the LIST sentences).

5.2 Results and discussion

In figure 8 the training trajectories for each individual subject are shown. The results after initial acclimatisation with Matrix in quiet and with visual feedback are shown as percentages words correct for each run of the test. There is a clear trend of improving performance. Note that while performance could still have improved with further training, due to testing time constraints we were not able to conduct more experiments.

Figure 9: SRT obtained with the LIST sentences for F0mod, before and after training, and ACE xPC. The error bars indicate standard deviations.

In figure 9 the results for the LIST sentences are shown before and after training. To account for the effect of fatigue in the comparison, the results for ACE xPC were obtained after training for S14 and S17. For S12 we were only able to conduct one list (for which an SRT of 2.3 dB was obtained), instead of the required minimum of two, so this result was discarded and the results before training are shown in the figure. Note that the LIST sentences were not used during the training phase. Any differences in scores with ACE xPC between experiment 1 and 2 are presumably due to measurement noise, subject fatigue and task familiarity. Comparing between F0mod and ACE after training, considering individual results, the error bars always overlap, suggesting that there is no difference in performance. While statistical power is not sufficient to show that there is no relevant difference between the ACE and F0mod after training conditions, it is encouraging that all three subject's scores improved with training and we are hopeful that with chronic exposure to the new strategy, such as in a wearable implementation, performance would be at least the same as with ACE.

6 General discussion and conclusions

We measured the effect of the F0mod strategy on speech perception, and compared with the default clinical ACE strategy. In our acute experiments, we found no significant effect in speech perception for CVC words in quiet, CVC words in noise and Matrix sentences in noise. These results are consistent with those obtained by Milczynski et al. (2012) who found no effect on sentence recognition for Mandarin Chinese.

For the LIST sentences in noise there was no overall effect either, but three out of seven subjects showed a tendency of worse performance with F0mod. In this case an adaptive procedure was used at low average SNRs. In experiment 2 we assessed the effect of brief training with the F0mod strategy and found that performance improved with training.

Vandali & van Hoesel (2012) tested speech perception in quiet and noise (SNRs ranging from 4 to 12 dB, depending on the subject) with the eTone strategy, which similarly to F0mod explicitly modulates the signal according to an estimate of F0. They found no effect on speech perception, but in contrast to the current study only conducted speech tests after nine test sessions in which the subjects could have become accustomed to the new strategy, thus implicitly providing extensive training.

We also evaluated whether our real-time implementation of F0mod yielded the same benefits for pitch perception as found in the much more extensive experiments of Milczynski et al. (2009).

While it is hard to compare the data directly because different subjects were tested and no training was provided, our results were qualitatively consistent with those obtained by Milczynski et al. (2009) and four out of six subjects showed improvements in various conditions. This indicates that the previously obtained positive results hold with the current real-time implementation. Similarly, with a real-time implementation of eTone, Vandali & van Hoesel (2012) administered sung-vowel pitch-ranking tests and after extensive training of rate discrimination they found significantly improved performance for eTone.

Altogether, the results are encouraging, because while F0mod can quite dramatically change the temporal structure of the signal, with no or little training all seven subjects were able to achieve speech scores similar to those obtained with the strategy they use on a daily basis. Considering that in earlier work we found that F0mod can improve F0 discrimination, pitch ranking, familiar melody identification, melodic contour identification (Laneau et al., 2006; Milczynski et al., 2009), and Mandarin tone identification (Milczynski et al., 2012), it would seem beneficial to include F0mod as an option in clinical processors. Given its low computational complexity and fairly straightforward implementation this would be feasible on current clinically available platforms.

Acknowledgements

We are grateful to our test subjects for their inexhaustible and enthusiastic collaboration. We thank the staff of the audiology team of the University Hospital Leuven for their support. This study was performed as part of IWT Flanders project number 110722 in cooperation with Cochlear Ltd.

References

- Bagshaw P.C., Hiller S.M. & Jack M.A., 1993. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Proc. Eur. Conf. Speech Commun.*, pp. 1003–1006. Berlin.
- Boersma P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc Inst phonetic Sci.*
- Boersma P. & Weenink D., 2001. Praat, a system for doing phonetics by computer. *Glott Int*, 5(9/10), 341–345.
- de Cheveigne A. & Kawahara H., 2002. YIN, a fundamental frequency estimator for speech and music. *J Acoust Soc Am*, 111(4), 1917.
- Donaldson G.S., Dawson P.K. & Borden L.Z., 2011. Within-subjects comparison of the HiRes and Fidelity120 speech processing strategies: speech perception and its relation to place-pitch sensitivity. *Ear Hear*, 32(2), 238–50.
- Francart T. & McDermott H.J., 2013. Psychophysics, fitting, and signal processing for combined hearing aid and cochlear implant stimulation. *Ear Hear*, 34(6), 685–700.
- Francart T., van Wieringen A. & Wouters J., 2008. APEX 3: a multi-purpose test platform for auditory psychophysical experiments. *J Neurosci Methods*, 172(2), 283–93.
- Geurts L. & Wouters J., 2001. Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants. *J Acoust Soc Am*, 109(2), 713–26.

- Geurts L. & Wouters J., 2004. Better place-coding of the fundamental frequency in cochlear implants. *J Acoust Soc Am*, 115(2), 844–52.
- Green T., Faulkner A. & Rosen S., 2004. Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. *J Acoust Soc Am*, 116(4 Pt 1), 2298–2310.
- Green T., Faulkner A., Rosen S. & Macherey O., 2005. Enhancement of temporal periodicity cues in cochlear implants: effects on prosodic perception and vowel identification. *J Acoust Soc Am*, 118(1), 375–85.
- Hochmuth S., Brand T., Zokoll M.a., Castro F.Z., Wardenga N. & Kollmeier B., 2012. A Spanish matrix sentence test for assessing speech reception thresholds in noise. *Int J Audiol*, 51(7), 536–44.
- Kasturi K. & Loizou P.C., 2007. Effect of filter spacing on melody recognition: acoustic and electric hearing. *J Acoust Soc Am*, 122(2), EL29–34.
- Laneau J., Wouters J. & Moonen M., 2004. Relative contributions of temporal and place pitch cues to fundamental frequency discrimination in cochlear implantees. *J Acoust Soc Am*, 116(6), 3606–19.
- Laneau J., Wouters J. & Moonen M., 2006. Improved music perception with explicit pitch coding in cochlear implants. *Audiol Neurotol*, 11(1), 38–52.
- Luo X., Landsberger D.M., Padilla M. & Srinivasan A.G., 2010. Encoding pitch contours using current steering. *J Acoust Soc Am*, 128(3), 1215–23.
- McKay C.M., McDermott H. & Clark G.M., 1995. Pitch matching of amplitude-modulated current pulse trains by cochlear implantees: the effect of modulation depth. *J Acoust Soc Am*, 97(3), 1777–1785.
- Milczynski M., Chang J.E., Wouters J. & van Wieringen A., 2012. Perception of Mandarin Chinese with cochlear implants using enhanced temporal pitch cues. *Hear Res*, 285(1-2), 1–12.
- Milczynski M., Wouters J. & van Wieringen A., 2009. Improved fundamental frequency coding in cochlear implant signal processing. *J Acoust Soc Am*, 125(4), 2260–2271.
- Plomp R. & Mimpen A.M., 1979. Speech-reception threshold for sentences as a function of age and noise level. *J Acoust Soc Am*, 66(5), 1333–42.
- Rabiner L.R., Cheng M.J., Member S. & Rosenberg A.E., 1976. Detection Algorithms. *IEEE Trans Acoust speech signal Process*, ASSP-24(5), 399–418.
- Shannon R.V., 1992. Temporal modulation transfer functions in patients with cochlear implants. *J Acoust Soc Am*, 91(4 Pt 1), 2156–64.
- Shannon R.V., Zeng F.G., Kamath V., Wygonski J. & Ekelid M., 1995. Speech recognition with primarily temporal cues. *Science (80-)*, 270(5234), 303.
- Studebaker G.A., 1985. A "rationalized" arcsine transform. *J Speech Hear Res*, 28(3), 455–62.
- Van Immerseel L.M. & Martens J.P., 1992. Pitch and voiced/unvoiced determination with an auditory model. *J Acoust Soc Am*, 91(6), 3511–26.
- van Wieringen A. & Wouters J., 2008. LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands. *Int J Audiol*, 47(6), 348–55.

- Vandali A.E., Sucher C.M., Tsang D.J., McKay C.M., Chew J.W. & McDermott H., 2005. Pitch ranking ability of cochlear implant recipients: a comparison of sound-processing strategies. *J Acoust Soc Am*, 117(5), 3126–3138.
- Vandali A.E. & van Hoesel R.J.M., 2011. Development of a temporal fundamental frequency coding strategy for cochlear implants. *J Acoust Soc Am*, 129(6), 4023–36.
- Vandali A.E. & van Hoesel R.J.M., 2012. Enhancement of temporal cues to pitch in cochlear implants: Effects on pitch ranking. *J Acoust Soc Am*, 132(1), 392–402.
- Vandali A.E., Whitford L.A., Plant K.L. & Clark G.M., 2000. Speech perception as a function of electrical stimulation rate: using the Nucleus 24 cochlear implant system. *Ear Hear*, 21(6), 608–624.
- Wagener K.C. & Brand T., 2005. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. *Int J Audiol*, 44(3), 144–156.
- Wu C.C. & Luo X., 2013. Current steering with partial tripolar stimulation mode in cochlear implants. *J Assoc Res Otolaryngol*, 14(2), 213–31.

Subject	Age	CI Use	CI side	Aetiology	Implant
S11	63	15	Left	congenital	Nucleus 24 Contour
S12	76	6	Right	otosclerosis	Contour Advance
S13	51	2	Right	progressive	Contour Advance
S14	62	11	Left	infection	Nucleus 24 Contour
S15	18	17	Left	meningitis	Nucleus 24
S16	78	3	Right	genetic, progressive	CI512
S17	76	1	Left	unknown	Contour Advance

Table I: Subject details: Age is in years at the time of testing. CI use is the number of years of implant use at the time of testing.

List of Figures

1	Schematic representation of the test setup	6
2	Error measures: vErr, uvErr, gErr for the evaluated speech materials. Error bars indicate the standard deviation of the average across sentences.	10
3	Error rate (Err) measure for complete set of PB and LIST sentences.	10
4	Phoneme score for the Lilliput words in quiet. Error-bars for individual subjects indicate the standard deviation.	12
5	Phoneme score for the Lilliput words in stationary noise at an SNR of 10 dB. Error-bars for individual subjects indicate the standard deviation.	12
6	Word score for the Matrix sentences in stationary noise at an SNR of 10 dB. Error-bars for individual subjects indicate the standard deviation.	13
7	SRT obtained with the LIST sentences in noise. Error-bars for individual subjects indicate the standard deviation.	13
8	Training sequence for Matrix sentences with 10 dB SNR for subjects S12 and S17 and in 5 dB SNR for subject S14. The x-axis indicates the presentation order of tests.	15
9	SRT obtained with the LIST sentences for F0mod, before and after training, and ACE xPC. The error bars indicate standard deviations.	16