Compensating for impaired prosody perception in cochlear implant recipients: A novel approach using speech preprocessing

FELIX KUHNKE[1,2,*], LORENZ JUNG[1,2], AND TAMÁS HARCZOS[1,2,3]

[1] *Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany*

[2] *Institute for Media Technology, Faculty of Electrical Engineering and Information Technology, Ilmenau University of Technology, Ilmenau, Germany*

[3] *Cochlear-Implant Rehabilitationszentrum Thüringen, Erfurt, Germany*

Due to inherent device limitations, cochlear implant (CI) recipients are provided with greatly reduced pitch information. However, detecting changes in pitch is necessary to perceive intonation, a main feature of prosody. Therefore, CI recipients' ability to perceive prosody is typically below that of normal-hearing subjects. We propose a novel preprocessing algorithm to enhance intonation perception by broadening the range of pitch changes in speech signals. To proof this concept, we have developed the pitch range extension (PREX) algorithm. PREX is capable of low-delay pitch modifications to speech signals. In addition, it provides automatic and intonation based amplification of pitch movements. In an evaluation with 23 CI recipients, the proposed algorithm significantly improved intonation perception in a question vs. statement experiment. However, the improved performance of CI subjects was still inferior to the performance of normal-hearing subjects. The results support the idea that preprocessing algorithms can improve the perception of prosodic speech features. Furthermore, we suggest utilizing the PREX algorithm for individualized treatment and rehabilitation.

INTRODUCTION

Over the last decades, speech recognition rates with cochlear implants steadily improved, with the main focus of the cochlear implant (CI) treatment being to improve the perception of words and sentences. However, CI recipients still perform very poorly on pitch-related tasks such as melody recognition (e.g., Wang *et al*., 2011) and the perception of voice pitch information (e.g., Meister *et al*., 2009). The perception of variation of pitch in speech is crucial to perceive intonation, a main aspect of prosody. As a consequence of poor pitch perception, cochlear implantees may not perceive the emotions expressed by a speaker or whether a sentence is meant as a question or a statement.

*Corresponding author: felix.kuhnke@gmail.com

Proceedings of ISAAR 2015: Individual Hearing Loss – Characterization, Modelling, Compensation Strategies. 5th symposium on Auditory and Audiological Research. August 2015, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C. Dalsgaard, L. Tranebjærg, and T. Andersen. ISBN: 978-87-990013-5-4. The Danavox Jubilee Foundation, 2015.

With the steady improvement of cochlear implants, CIs have become powerful computing platforms. Modern CIs employ sophisticated signal processing algorithms that react to the incoming signal and may change their processing parameters to improve sound perception. Apart from that, CI recipients can change their device settings according to the sound environment, applying different sound processing technology at different times. Thus, CIs provide the platform and possibilities for sound-specific and user-specific signal processing algorithms.

We propose to use preprocessing algorithms to enhance features of speech signals that are difficult to perceive for CI recipients, such as pitch. As a proof of concept we developed a method to enhance intonation perception in CI recipients by broadening the range of pitch changes made by speakers.

PITCH RANGE EXTENSION (PREX) ALGORITHM

As already stated, the algorithm should first be implemented as a preprocessing algorithm, meaning that we preprocess the audio signal before it enters the usual CI processing chain. Even though one could imagine modifying the stimulation pattern of the electrodes (speech-processing strategy) directly, the use of a preprocessing algorithm has several advantages. First, the algorithm is independent of implant type and electrode design, which allows usage across different devices. Further, it can be used in devices such as hearing aids, which is advantageous for bimodal fitted patients. Finally, the output quality of the algorithm can easily be evaluated with normal-hearing subjects. However, as a final step the algorithm should be embedded inside the speech-processing strategy for improved performance and lower delays.

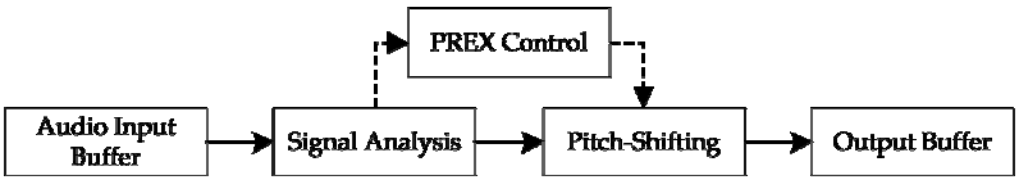

Fig. 1: Overview of the algorithm processing chain.

Figure 1 shows the algorithm overview. The PREX algorithm analyses the incoming speech signal and automatically determines a new pitch value and synthesises the according pitch-shifted signal. In the following we explain the different modules:

Audio samples that have been recorded by the microphone are stored in the Audio Input Buffer. The Signal Analysis module reads a predefined amount of samples from the Audio Input Buffer and estimates fundamental frequency (F0), root-mean-square energy, and zero crossing rate. In the first step these features are used to detect voiced and unvoiced segments of the signal. In the following step the extracted features and the durations between voiced segments are used by the PREX Control module to detect intonational structures. For the PREX protoype, we used

heuristics to classify the change between intonational structures. For example, a new intonational structure was detected after 200 ms of silence, as humans are only able to perceive unconnected F0 contours as one, until the gap between them exceeds 200 ms (Nooteboom, 1997). Based on the intonational structures, the PREX Control module computes the pitch shift scale factors accordingly (see the next section for the calculation of new pitch values). The pitch-shifting is done by a customized implementation of the PSOLA algorithm (Hamon *et al.*, 1989). We enhanced the classical PSOLA for lower delays. Instead of using segments of the size of two (or more) pitch periods we used only a single period of the voiced signal as synthesis segment. By applying an adaptive window for every period we removed signal discontinuities that would otherwise arise during overlap-add synthesis.

The algorithmic delay is dependent on the lowest frequency the algorithm should process. If the F0 value is below this frequency, no pitch-shifting is performed. For a minimum supported frequency of 62.5Hz we get an algorithmic delay of 18.75ms and for 100Hz, 11.7ms, respectively.

Calculating new pitch values (PREX Control)

Based on the detected F0 of the speech signal (f_{in}) the new pitch values (f_{out}) for the synthesized signal are calculated according to Eq. 1:

$$f_{out} = f_{in} + f_{in} \times \log_2\left(\frac{f_{in}}{f_{start}}\right) \times PRSF \qquad \text{(Eq. 1)}$$

PRSF is the pitch range scale factor that allows to modify the global amount of range extension. We found that different factors for up and down pitch range extension are necessary to produce natural sounding results. This can easily be accomplished using separate pitch range scale factors for upward and downward extension. The perception and production of voice pitch is generally not on a linear scale (Nolan, 2003). Therefore, we use a psycho-acoustic logarithmic pitch scale to uniformly describe intonation across different speakers. To preserve natural intonation we use the first F0 of every intonational structure (f_{start}) as reference frequency for pitch range extension. This approach will not alter the pitch register and will produce a more natural sounding result. Finally, the new pitch values are used to compute the corresponding pitch shift scale factor for the pitch shifting.

Figure 2 shows the result of PREX preprocessing on a sentence uttered as a question. The processed F0 curve (dashed line) shows a much higher pitch range.

EVALUATION

We use an intonation hearing experiment based on the question vs. statement paradigm: A number of recorded sentences (stimuli) are presented to the subject. For every stimulus the subject has to decide whether it was spoken as a statement or a question.

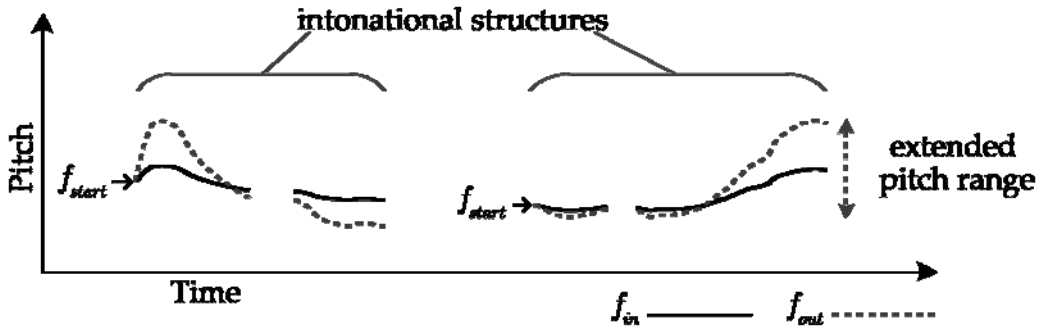

Fig. 2: Pitch range extension based on intonational structures. F0 contour of a german sentence uttered as question, before (solid line) and after PREX preprocessing (dashed line).

Stimuli

We recorded 36 sentences from 3 female and 3 male speakers, once as statement and once as question. To provide no lexical cues, the sentences were the same for questions and statements, e.g., “She will arrive at ten o’clock.” vs. “She will arrive at ten o’clock?”. This produced 72 test stimuli. In addition, all 72 stimuli were processed with the PREX algorithm, resulting in a total of 144 stimuli with a total length of 5 minutes and 24 seconds.

Subjects

23 CI recipients from the Cochlear-Implant Rehabilitationszentrum Thüringen were asked to perform the test. The subjects had the following characteristics: Subjects were aged 17 to 77 years, with a mean of 54 years. The duration of the subjects’ CI-experience was ranging from 1 month to over 11 years. 12 female and 11 male subjects participated. Furthermore, subjects were using Cochlear and MED-EL implants.

Procedure

A single loudspeaker (YAMAHA Monitor Speaker MS101 II) was positioned in front of the subject. The maximum presentation level at the subject’s position, about 50 cm in front of the loudspeaker, was set to 70 dB(A) (measured using stimuli of the first male speaker). All stimuli were played from a laptop. Furthermore, a small program was developed to play the stimuli in random order.

The tests were carried out with one CI subject at a time. First, the task was explained. At this point, unilaterally implanted subjects were asked to put on a single-sided headphone to mask the contralateral ear with noise. We used uncompressed OLSA noise (Wagener *et al.*, 1999) at 81.9 dB(A). Afterwards, every subject completed a training phase, where feedback was given for every response.

The length and intensity of the training phase was dependent on the subject. This was required, because some subjects easily identified the different sentences and were quickly ready for the main test, while others needed repetitive presentation of stimuli and even visual cues to learn what they had to listen for. However, all subjects received the same training stimuli, which were not used in the main test. The main test usually took between 25 to 35 minutes. Every stimulus was presented only once and no feedback was provided during testing. The testing was conducted in an anechoic room at the Cochlear-Implant Rehabilitationszentrum Thüringen. Afterwards, the percentage of correct question/statement identifications (score) was measured for every subject. Furthermore, the subjects' responses were sorted to provide scores for natural stimuli and PREX modified stimuli.

Test verification with normal-hearing subjects

A verification test was conducted to assess whether or not the PREX preprocessing harms the recognition of intonation and that the natural stimuli can be correctly identified by normal-hearing (NH) subjects. Five NH subjects participated in the test. The average identification performance for both types of stimuli was 99.7%. Surprisingly, one subject achieved 98.6% (142 of 144 correct), whereas all others reached 100%. The erroneous identifications could be caused by a lack of concentration as every stimulus was only played once. However, the results showed that NH subjects can perform the test with near perfect results.

RESULTS

Because of the small sample size, scores were not assumed to be normally distributed. The box plot in Fig. 3 shows the scores for both stimulus groups. It can be seen that the identification of questions and statements for natural stimuli and modified stimuli was worse compared to the near perfect score achieved by NH subjects. Furthermore, the median of the PREX stimuli score is about 10% higher than the corresponding natural stimuli score median. To analyse the results in more detail, a scatter plot was used.

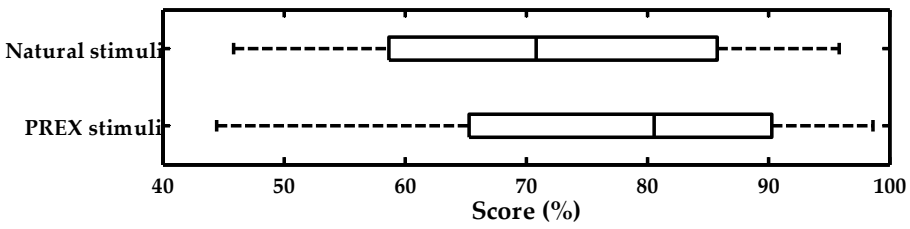

Fig. 3: Box plot of the results (percent correct scores) for the two stimulus groups.

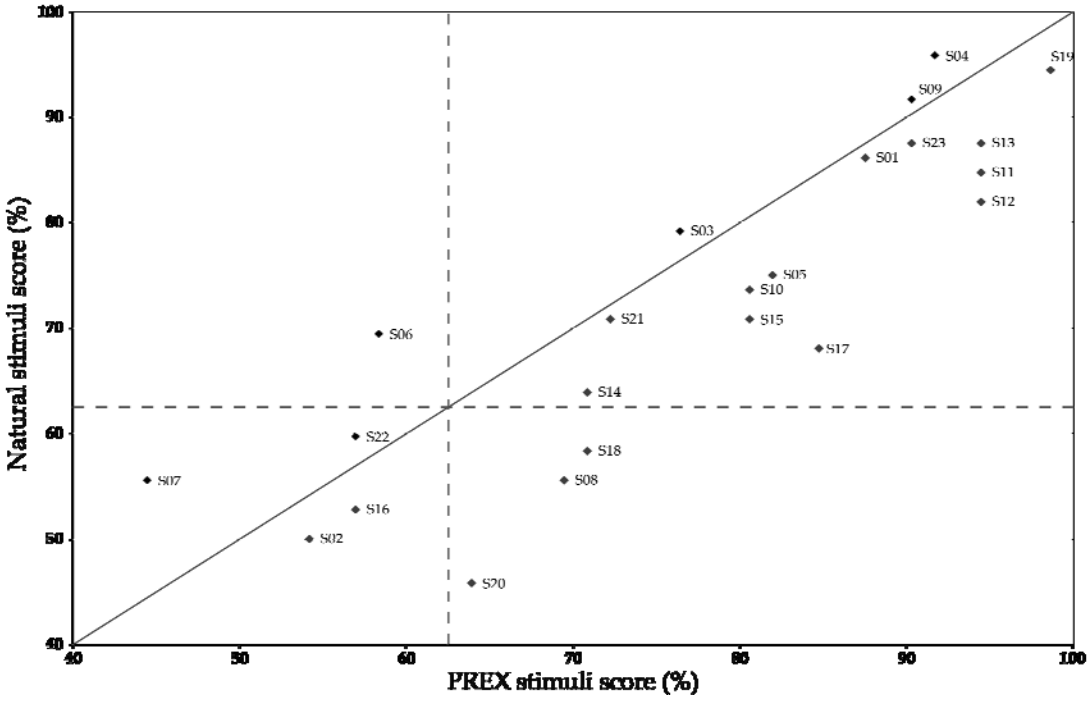

Fig. 4: Percentages of correct question/statement identifications (score) for all CI subjects. Performance for the natural stimuli is plotted against the performance for the PREX processed stimuli. The diagonal line represents equal performance for both types of stimuli. Subjects had to achieve more than 62.5% (shown with a dashed line) for each condition to perform better than chance.

The scatter plot (Fig. 4) uncovers the scores of every subject. The plot shows that some CI subjects (S02, S07, S16, and S22) had huge problems in perceiving the difference between question and statement stimuli. The binomial test revealed that these subjects did not perform significantly better than chance (p = 0.5), which would require at least 45 of 72 (62.5%) correct identifications for either case (p = 0.0444), two sided binomial test). Interestingly, subjects S06, S08, S18, and S20 performed only above chance level for one stimuli group. The other subjects were above chance level but identification scores were widely distributed and subjects showed large inter-individual scatter. However, scores were mostly close to the diagonal, indicating only small differences between the stimuli groups.

Visual inspection of Fig. 4 also suggests a trend towards the lower right side of the scatter plot, as points are more often found below the diagonal. This finding indicates a better performance with the PREX stimuli. The non-parametric Wilcoxon matched-pairs signed rank test (two sided) was employed to test for a significant difference of medians between PREX stimuli scores and natural stimuli scores. In the indicated case, the test rejected the null hypothesis of equal medians at the 5%

significance level with p = 0.0103. Therefore, the score for PREX stimuli is significantly higher than the score for natural stimuli.

In addition, we analysed for relationships between scores and subject characteristics. Astonishingly, no significant relationship was found for experience, processing strategy, age, or residual hearing. This surprising finding may be associated with the small number of subjects.

DISCUSSION

Based on the results, it can be concluded that the PREX preprocessing significantly improves question vs. statement identification in CI recipients. The increased ability to identify sentences as questions or statements suggests that PREX preprocessing improves the overall perception of intonation and prosody. However, these results must be interpreted with caution, as the question vs. statement test cannot be considered representative for all forms of intonation perception.

While a significant difference was found, it seems to be very small when the median values are taken as references. On the other hand, the subjects heard PREX processed stimuli in the test situation for the first time and they had no time to get accustomed to the new stimuli. Using PREX on a daily basis might reveal additional improvements.

The main weakness of the evaluation is that speech intelligibility of PREX stimuli was not measured. Even though NH subjects did not report any difficulties to understand the processed sentences, the same cannot be assumed for CI recipients. However, CI subjects did not report any problems with intelligibility. Often, they stated that they did not have any problem understanding the sentences but did not know whether it was spoken as a question or a statement.

CONCLUSION AND OUTLOOK

We presented a preprocessing algorithm that enhances intonation perception by broadening the range of pitch changes in speech signals. It provides automatic and intonation based amplification of pitch movements. In an evaluation with 23 CI recipients, the proposed algorithm significantly improved intonation recognition, likely caused by the fact that pitch movements became more easily identifiable by CI subjects. Based on these findings, it would be very interesting to see if PREX processing could improve speech intelligibility for tonal languages such as Mandarin.

The results support the idea that the perception of a variety of speech features that are difficult to perceive for CI recipients or hearing aid users can be improved by speech preprocessing algorithms. These additional speech features include loudness, vowel quality, and duration. Speech preprocessing methods could be used in CI rehabilitation to specifically exercise and improve individual weaknesses. The areas of application range from speech perception to speech production. Similar to music students learning musical pieces at lower tempi, CI recipients could use time stretching to learn voice recognition at a lower speech tempo. Furthermore, PREX

preprocessing could be used while training speech production. Implantees may achieve an improved recognition of their own sound production and subsequently improve their prosody production. Finally, signal modifications do not need to be fixed to a certain intensity, but could be set to meet individual needs.

REFERENCES

Hamon, C., Mouline, E., and Charpentier, F. (**1989**), "A diphone synthesis system based on time-domain prosodic modifications of speech," Proc. International Conference on Acoustics, Speech, and Signal Processing, 1989, 238-241.

Meister, H., Landwehr, M., Pyschny, V., Walger, M., and Wedel, H.v. (**2009**), "The perception of prosody and speaker gender in normal-hearing listeners and cochlear implant recipients," Int. J. Audiol., **48**, 38-48.

Nolan, F. (**2003**), "Intonational equivalence: An experimental evaluation of pitch scales," Proc. 15th International Congress of Phonetic Sciences, 771-774.

Nooteboom, S. (**1997**), "The prosody of speech: Melody and rhythm," in *The Handbook of Phonetic Sciences*. Eds. Hardcastle, W.J. and Laver, J. (Oxford, UK: Blackwell), pp. 640-673.

Wagener, K., Kühnel, V., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests," Zeitschrift für Audiologie/Audiological Acoustics, **38**, 4-15.

Wang, W., Zhou, N., and Xu, L. (**2011**), "Musical pitch and lexical tone perception with cochlear implants," Int. J. Audiol., **50**, 270-278.