DISCRIMINATION OF EMOTIONAL AND LINGUISTIC PROSODY WITH COCHLEAR IMPLANT SIMULATIONS

Daan J. van de Velde[a,d], *Arian Khoshchin*[b], *Linda ter Beek*[b], *Niels O. Schiller*[a], *Johan H. M. Frijns*[c,d], *Jeroen J. Briaire*[c]

[a]Leiden University Centre for Linguistics, [b]Leiden University, [c]Leiden University Medical Center, [d]Leiden Institute for Brain and Cognition
d.j.van.de.velde@hum.leidenuniv.nl

ABSTRACT

In cochlear implants (CI), temporal sound features are more successfully transmitted than spectral sound features. This could have consequences for the perception of prosody. This was tested for emotion vs. focus perception in simulated CI hearing, starting from the assumption that for emotional prosody spectral (F0) features are more important than for focus prosody.

Sets of short Dutch phrases were recorded with neutral, emotional (happy and sad) and focused (e.g., 'a BLUE ball' vs. 'a blue BALL') prosody. Temporal or spectral prosody, or both, were cross-spliced from the non-neutral to the neutral utterances, thus controlling for the usable phonetic cues. 17 Dutch subjects identified intended emotions and focus for vocoded (CI-simulated) and unvocoded versions of the phrases.

A benefit of F0 vs. temporal information was found for emotional, but not for focus prosody. This could imply that CI users have more trouble hearing emotional than linguistic (focus) prosody.

Keywords: cochlear implants, prosody, vocoders, pitch, temporal information

1. INTRODUCTION

Cochlear implants (CI) can provide children and adults suffering from sensorineural hearing loss with a sense of hearing. Most users achieve good speech understanding [10]. Nevertheless, hearing is far from normal and problems remain, such as hearing in noise, hearing music and hearing prosody. These problems are partly due to the differential quality of transmission of different acoustic parameters, such as temporal, intensity and spectral information [7].

In the case of prosody, the difference in transmission quality of acoustic parameters is expected to result in more or less perception difficulties with different types or aspects of prosody, since (in a given language) those different types can be conveyed by different acoustic parameters. One distinction of prosody types where this could play a role is between emotional and linguistic prosody. Emotional prosody is the non-segmental information that reflects the emotional state of the speaker; linguistic prosody is the non-segmental information that conveys (certain) pragmatic information about an utterance. Whereas the acoustic realization and paralinguistic meaning of emotional prosody can be of a gradient nature, those of linguistic prosody are discrete. Furthermore, differences have been found on the neural level [9].

For both the prosody of emotions [5] and of focus (accentuation) in Dutch [8], it has been reported that F0 and temporal (durational and rhythmic) information both play a role. The first goal of the present study was therefore to find out if, for the two types of prosody, the cue weightings of F0 and temporal information are different.

The emotional vs. linguistic prosody distinction is one that has (almost) never been investigated in the literature on CI perception. The second goal of this study was therefore to find out if under the degraded acoustic circumstances of CI hearing there would be a difference in discriminability of emotional vs. linguistic prosody in the presence of F0 vs. temporal cues.

2. METHODS

In order to find out if emotions and focus were discriminable with CI simulations, two tests were developed (the emotion test and the focus test, respectively) in which, in each trial, participants were asked to choose which of two emotions (EMOTION TEST) or focus positions (FOCUS TEST), respectively, was perceived for a given stimulus sound. All stimuli were repeated in a variant with only F0, only temporal or both types of information.

2.1. Participants

17 Dutch native speakers participated for credits or as volunteers as part of a larger study. 13 of them were right-handed, 15 were men, and their mean age was 20 years (SD = 3.4 years). None had a hearing loss of larger than 40 dB on any of frequencies 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz or 8 kHz, as tested with the Oscilla AudioConsole 3.3.2 (InMedico, Denmark).

2.2. Stimuli

All stimuli were based on natural recordings of utterances made by a professional linguist. For the utterances of the EMOTION TEST, she was asked to pronounce 12 phrases of the format ARTICLE-COLOUR-NOUN (e.g., 'een blauwe bal' – 'a blue ball') (1) once without a specific emotion (neutral), (2) once with a happy and (3) once with a sad emotion, all with more or less the same pace. This last instruction was included because it was believed that any large phrase-level temporal differences would yield ceiling-level performance in discrimination.

For the utterances of the FOCUS TEST, the speaker was asked to pronounce 12 phrases of the format ARTICLE-COLOUR-NOUN-'en een' (e.g., 'een gele bloem en een' – 'a yellow flower and a') (1) once without focus on any of the words (neutral), (2) half of them once with narrow focus on the colour and (3) the other half once with narrow focus on the noun. The two trailing words avoided phrase-final prosody on the noun. Note that the colour and noun combinations chosen, although highly comparable, were not identical to those in the emotion test.

For all emotion and focus stimuli, three new variants (PHONETIC PARAMETERS) were resynthesized using Praat software [1], after segmenting them into allophones. (1) The pitch contour of the neutral utterance was per segment replaced with that of the non-neutral variant shortened or lengthened to match the neutral phrases' segment durations (F0 condition). (2) Every segment of the neutral utterances was elongated or shortened using PSOLA (Pitch-Synchronous Overlap Add) to match the duration of the corresponding segment in the non-neutral variant of that utterance (TEMPORAL condition). (3) Both procedures 1 and 2 were applied to every neutral phrase (BOTH condition).

This yielded variants in which only F0, only durational, or both types of information were available for listeners, respectively. All these stimuli were subsequently processed such that they mimicked the signal received by (some) cochlear implant users. This was done by applying an 8-channel sinewave vocoder based on Continuous Interleaved Sampling (CIS). This signal is band-passed for 200 to 7000 Hz with 24 dB/octave filter slopes and then detected for envelopes with a cut-off frequency of 240 Hz (24 dB/octave). This procedure was carried out using the AngelSim^{TM} vocoder software (Emily Shannon Fu Foundation, http://angelsim.tigerspeech.com/).

Stimuli were presented in two forms (PROCESSING conditions), (1) CI simulated (VOCODED), (2) non-CI simulated (UNVOCODED), allowing to find the size of the penalty of vocoding for each of the acoustic conditions (pitch vs. temporal vs. both). There were thus 144 (12 phrases × 2 emotions × 3 phonetic parameters × 2 processing types) stimuli for the emotion test and 72 for the focus test (6 instead of 12 phrases per focus position). Stimuli were approximately 1.5 s long. The focus test also contained 24 extra stimuli that had no processing whatsoever, but these were not analysed for the present study.

2.3. Procedure

Both the EMOTION TEST and the FOCUS TEST were run on a computer with the E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). Participants were seated in a sound-proof booth approximately 70 cm. in front of the screen and heard the stimuli through headphones. After a short training session, three (EMOTION TEST) or two (FOCUS TEST) blocks of trials were run, between which participants could choose to pause.

Trials consisted of a fixation point (1250 ms), presentation of the stimulus sound and time to respond (4000 ms), and finally an inter-stimulus interval (200 ms). When the stimulus was played, a picture of the object described in the stimulus (e.g., a blue ball) was shown on the screen, as well as the two response options on the left and right side of the screen ('happy' and 'sad', for the EMOTION TEST; the colour and the object, for the FOCUS TEST). Participants chose by button-press which emotion or focus position they had heard. The order of conditions (F0, TEMPORAL, BOTH), processing forms (VOCODED, UNVOCODED) and stimuli was randomized. The order of tests (EMOTION TEST, FOCUS TEST) was counterbalanced across participants. The EMOTION TEST lasted for around 8 minutes, the FOCUS TEST around 6 minutes. Accuracy and reaction time data were registered.

2.3. Analysis

The analysis was carried out on the accuracy data. Differences in means were tested separately for the two tests of the experiment by Repeated Measures (RM) ANOVA in SPSS 21 (IBM Corp., 2012). We adopted a *p*-value of 0.05 as a significance threshold.

3. RESULTS

All participants completed the two tests. Table 1 and Table 2 show means and standard deviations of accuracy scores in all cells of both tests, respectively.

One-sample *t*-tests showed that scores in all cells of both the EMOTION TEST and the FOCUS TEST

was above chance ($p < .001$). Figure 1 and Figure 2 show scores and 95% confidence intervals for all cells of both tests, respectively.

Table 1: Mean accuracy (and standard deviation) scores of the EMOTION TEST per processing and per phonetic parameter condition, plus total values.

PROCESSING	PHONETIC PARAMETER F0	temporal	both	total
	Mean accuracy (SD)			
unvocoded	98,0 (13,9)	57,3 (49,5)	98,5 (12,3)	84,6 (36,2)
vocoded	61,8 (48,7)	60,9 (48,9)	75,1 (43,3)	66,1 (47,4)
total	80,0 (40,0)	59,0 (49,2)	86,7 (33,9)	75,5 (43,0)

Table 2: Mean accuracy (and standard deviation) scores of the FOCUS TEST per processing and per phonetic parameter condition, plus total values.

PROCESSING	PHONETIC PARAMETER F0	temporal	both	total
	Mean accuracy (SD)			
unvocoded	68,8 (46,4)	58,3 (49,4)	72,1 (44,9)	66,6 (47,2)
vocoded	58,4 (49,4)	55,0 (49,9)	69,7 (46,1)	61,2 (48,8)
total	63,6 (48,2)	56,7 (49,6)	71,0 (45,5)	63,9 (48,0)

In the EMOTION TEST, by RM-ANOVA, the main effect of PROCESSING ($F(1,16) = 117,1$, $p < .001$) was significant. Post-hoc Bonferroni corrected tests showed that this effect was significant for the F0 ($p < .001$) and the BOTH condition ($p < .001$), but not for the TEMPORAL condition ($p = .310$). The main effect of PHONETIC PARAMETER ($F(2,15) = 71,5$, $p < .001$) was significant, as well as Bonferroni corrected post-hoc tests for all three pairs of PHONETIC PARAMETERS (F0 vs. TEMPORAL: $p < .001$; F0 vs. BOTH: $p = .005$; TEMPORAL vs. BOTH: $p < .001$). The interaction between PROCESSING and PHONETIC PARAMETER ($F(2,15) = 22,1$, $p < .001$) was significant, as well as the three subcontrasts (F0 vs. TEMPORAL: $F(1,16) = 44,2$, $p < .001$; F0 vs. BOTH: $F(1,16) = 13,4$, $p = .002$; TEMPORAL vs. BOTH: $F(1,16) = 21,0$, $p < .001$).

In the FOCUS TEST, the main effect of PROCESSING ($F(1,16) = 5,9$, $p < .027$) was significant, but post-hoc tests ($p = 0.016$ threshold) showed an insignificant effect for each of the three separate PHONETIC PARAMETERS (F0: $p = .023$; TEMPORAL: $p = .493$; BOTH: $p = .571$). The main effect of PHONETIC PARAMETER ($F(2,15) = 8,0$, $p < .004$) was significant. Post-hoc Bonferroni corrected tests showed that the effect was significant for F0 vs. TEMPORAL ($p < 0.018$) and for F0 vs. BOTH ($p < 0.018$) but not for F0 vs. TEMPORAL ($p = 0.483$). The interaction between PROCESSING and PHONETIC PARAMETER ($F(2,15) = 0.86$, $p = .445$) was not significant.

4. DISCUSSION

In the present study, we sought to find out if under the degraded conditions of simulated cochlear implant hearing, there would be a difference in discriminability of emotional and focus position prosody when only F0 cues, only temporal cues or both cues (control condition) would be present in the signal. In other words, the goal was to discover (1) if for emotional vs. linguistic (focus) prosody, listeners would rely to a different degree on F0 and temporal cues, and consequently (2) if CI users would therefore probably have different degrees of access to the two types of prosody.

The accuracy results showed for both the EMOTION TEST and the FOCUS TEST that in CI simulation the prosody was more difficult to distinguish than in the unprocessed condition. This cost was higher for the EMOTION TEST (84,6% vs. 66,1%) than for the FOCUS TEST (66,6% vs. 61,2%), but the latter was more difficult in general. Split by PHONETIC PARAMETER, this effect was only significant for the F0 and the BOTH conditions in the EMOTION TEST but not in the FOCUS TEST. This implies that the benefit of F0 information relative to temporal information is greater in the EMOTION TEST than in the FOCUS TEST.

In both tests, there was a cost of making only one cue available as opposed to two, suggesting that both cues contribute to the prosody discrimination. Importantly, however, in the EMOTION TEST but not in the FOCUS TEST, there was a benefit of F0 over TEMPORAL information, as well as an interaction between the PHONETIC PARAMETER and PROCESSING conditions. This suggests that the vocoder processing leaves the temporal information relatively intact (performance in all cells was above chance), whereas it significantly affects the F0 information.

Taken together, the results could be taken to indicate that for emotional prosody, F0 information is a more important cue than temporal information, whereas for linguistic (focus) prosody, there is no benefit of one cue over the other. CI processing affects emotional prosody more than focus prosody, as for emotional prosody listeners have to rely relatively more on F0 than on temporal information.

These results are in accordance with

literature stating that temporal information is technically better preserved than F0 information in CI users [e.g., 7]. The heavier reliance in prosody perception on temporal as opposed to F0 information has also been reported [6].

Figure 1: Mean accuracy and 95% confidence intervals of the EMOTION TEST for UNVOCODED (light grey bars), VOCODED (dark grey bars) and the three PHONETIC PARAMETERS.

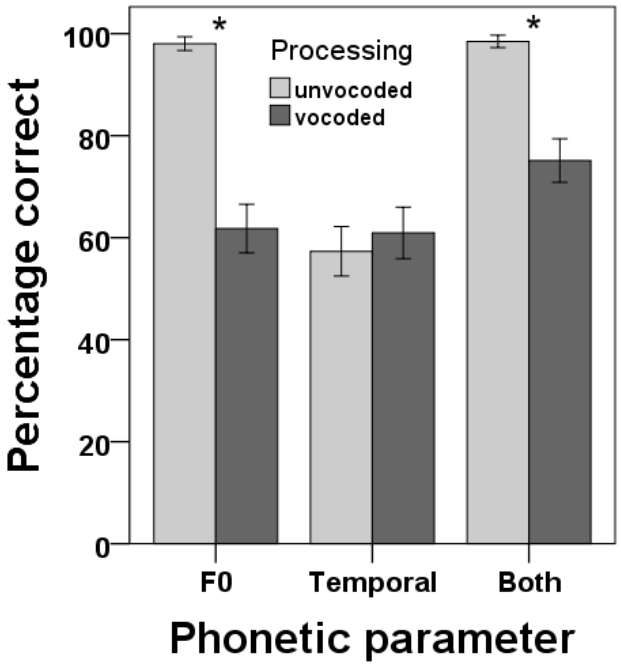
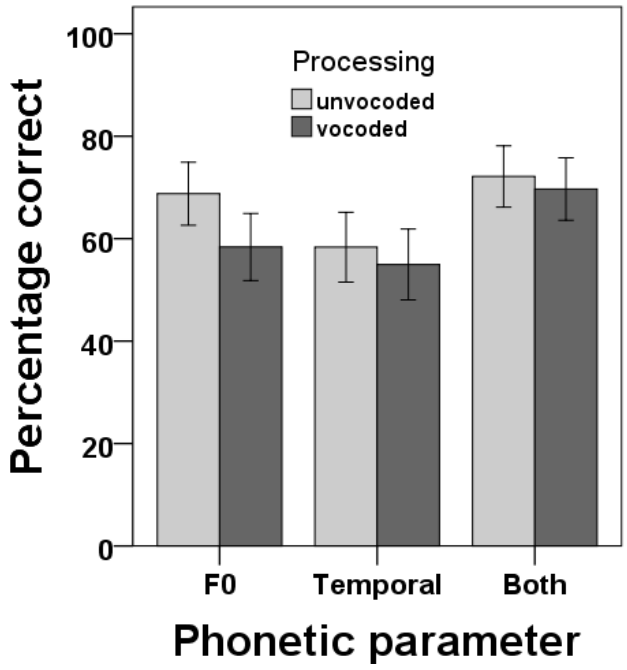


Figure 2: Mean accuracy and 95% confidence intervals of the FOCUS TEST for UNVOCODED (light grey bars), VOCODED (dark grey bars) and the three PHONETIC PARAMETERS.



Studies on CI users found that they have much difficulty distinguishing stimuli based on prosody. The overall accuracy results in the BOTH condition of the current study are however still low compared to some other studies on comparable tasks by CI users [4]. This could, among other factors, be due to (1) the fact that normally hearing listeners are not used to the processed signal, or (2) the fact that the stimuli were doubled processed (i.e., cross-spliced prosody and vocoding).

Regarding the first goal of our study, the distinction of the current study between emotional and linguistic prosody and their differing results sheds light on the discussion about cues used for the two types. The current results support models claiming that F0 information is relatively important for emotional prosody [5] but do not support studies claiming that F0 information is paramount for (accentual) focus in Dutch, as we didn't find a difference between the F0 and the TEMPORAL condition for the FOCUS TEST.

Regarding the second goal, extending our findings with CI simulations to actual CI users, emotional prosody discrimination seems to be relatively difficult as compared to focus discrimination for that clinical population. This is because, although the mean level of performance was higher for the emotional than for the focus stimuli (this could be attributed to intrinsic difficulty level differences between the two studies), the cost of vocoder processing was higher for the former than for the latter.

5. CONCLUSION

To our knowledge, the current study was the first to study cue reliance in emotional vs. linguistic (focus) prosody in (simulated) CI perception. Emotional, but not focus prosody was found to rely more heavily on F0 than on temporal information. The cost of CI processing, in which temporal information is more preserved, is therefore higher for emotional than for focus perception. Based on our results, we recommend that future studies take into account (1) a possible interaction between prosody type (emotional vs. linguistic) and the cues by which they can be distinguished by CI users, and (2) the availability of cues, for instance by synthetically removing one or more cues. Future studies on this subject could be performed using utterances from more than one speaker, distinguishing more than two different emotions and investigating more than one type of linguistic prosody.

7. REFERENCES

[1] Boersma, P., Weenink, D. 2012. Praat: doing phonetics by computer [Computer program].

[2] Eefting, W. 1991. The Effect of Information Value and Accentuation on the Duration of Dutch Words, Syllables, and Segments. *J. Acoust. Soc. Am.* 89(1), 412-424.

[3] IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

[4] Meister, H., Landwehr, M., Pyschny, V., Walger, M., von Wedel, H. 2009. The perception of prosody and speaker gender in normal-hearing listeners and cochlear implant recipients. *Int J Audiol* 48(1), 38-48.

[5] Murray, I. R., Arnott, J. L. 1993. Toward the Simulation of Emotion in Synthetic Speech - a Review of the Literature on Human Vocal Emotion. *J. Acoust. Soc. Am.* 93(2), 1097-1108.

[6] O'Halpin, R. 2009. *The perception and production of stress and intonation by children with cochlear implants*. Doctoral dissertation, UCL, London.

[7] Shannon, R. V. 2002. The relative importance of amplitude, temporal, and spectral cues for cochlear implant processor design. *Am. J. Audiol.* 11(2), 124-127.

[8] 't Hart, J., Cohen, A. 1973. Intonation by rule: a perceptual quest. *J. Phon.*, 309-327.

[9] Witteman, J., van Ijzendoorn, M. H., van de Velde, D., van Heuven, V. J. J. P., Schiller, N. O. 2011. The nature of hemispheric specialization for linguistic and emotional prosodic perception: A meta-analysis of the lesion literature. *Neuropsychologia* 49(13), 3722-3738.

[10] Zeng, F. G. 2004. Trends in cochlear implants. *Trends Amplif.* 8(1), 1-34.