Vocal Emotion Recognition with Cochlear Implants

Xin Luo, Qian-Jie Fu, John J. Galvin III

Department of Auditory Implants and Perception, House Ear Institute
2100 West Third Street, Los Angeles, CA 90057, U.S.A.
xluo@hei.org,qfu@hei.org,jgalvin@hei.org

Abstract

Besides conveying linguistic information, spoken language can also transmit important cues regarding the emotion of a talker. These prosodic cues are most strongly coded by changes in amplitude, pitch, speech rate, voice quality and articulation. The present study investigated the ability of cochlear implant (CI) users to recognize vocal emotions, as well as the relative contributions of spectral and temporal cues to vocal emotion recognition. An English sentence database was recorded for the experiment; each test sentence was produced according to five target emotions. Vocal emotion recognition was tested in 6 CI and 6 normal-hearing (NH) subjects. With unprocessed speech, NH listeners' mean vocal emotion recognition performance was 90 % correct, while CI users' mean performance was only 45 % correct. Vocal emotion recognition was also measured in NH subjects while listening to acoustic, sine-wave vocoder CI simulations. To test the contribution of spectral cues to vocal emotion recognition, 1-, 2-, 4-, 8- and 16-channel CI processors were simulated; to test the contribution of temporal cues, the temporal envelope filter cutoff frequency in each channel was either 50 or 500 Hz. Results showed that both spectral and temporal cues significantly contributed to performance. With the 50-Hz envelope filter, performance generally improved as the number of spectral channels was increased. With the 500-Hz envelope filter, performance significantly improved only when the spectral resolution was increased from 1 to 2, and then from 2 to 16 channels. For all but the 16-channel simulations, increasing the envelope filter cutoff frequency from 50 Hz to 500 Hz significantly improved performance. CI users' vocal emotion recognition performance was statistically similar to that of NH subjects listening to 1 - 8 spectral channels with the 50-Hz envelope filter, and to 1 channel with the 500-Hz envelope filter. The results suggest that, while spectral cues may contribute more strongly to recognition of linguistic information, temporal cues may contribute more strongly to recognition of emotional content coded in spoken language.

Index Terms: vocal emotion recognition, cochlear implant.

1. Introduction

The cochlear implant (CI) has restored hearing sensation to many profoundly deafened individuals. Contemporary CI devices generally employ spectrally-based speech-processing strategies, in which the temporal envelope is extracted from a number of frequency analysis bands and used to modulate pulse trains of current delivered to appropriate electrodes. These spectrally-based strategies have generally provided good patient performance in quiet listening conditions. However, CI performance is generally poor for more challenging listening tasks (e.g., speech in noise, music perception, voice gender and speaker recognition, etc.), because of the limited spectral and temporal resolution provided by the implant device [1, 2, 3]. Much recent CI research has been aimed at improving the transmission of spectro-temporal fine structure cues needed to improve patient performance under these difficult listening conditions. For example, Geurts and Wouters proposed increasing the spectral resolution for apical electrodes to better code pitch information [4]. Green et al. proposed sharpening the temporal envelope to enhance periodicity cues transmitted by the speech processor, thereby improving perception of pitch cues [5].

Spoken language conveys not only linguistic information, but also prosodic information (e.g., variations in speech rhythm, intonation, etc.). Prosodic speech cues can convey information regarding the emotion of a talker. Recognition of a talker's emotion can contribute strongly to speech understanding, especially with auditory-only communication (e.g., telephone speech). Acoustic features commonly associated with vocal emotion include pitch (mean value and variability), intensity, speech rate, voice quality and articulation [6]. Using these features, normal-hearing (NH) listeners have been shown to recognize most (70 - 80 % correct) target emotions produced in test [7, 8]. Artificial intelligence systems have also been developed to recognize vocal emotions; neural networks and statistical classifiers using various features as input have been shown to perform similarly to NH listeners in vocal emotion recognition tasks [9, 10, 11, 12].

The present study investigated CI users' ability to recognize vocal emotions in acted emotional speech, given CI patients' limited access to pitch information and spectro-temporal fine structure cues. Vocal emotion recognition was also tested in NH subjects listening to unprocessed speech and speech processed by acoustic CI simulations [13]. In the simulations, different amounts of spectral resolution and temporal information were tested to examine the relative contributions of spectral and temporal cues to vocal emotion recognition.

2. Methods

2.1. Subjects

Six NH subjects (3 males and 3 females) and six CI users (3 males and 3 females) participated in the present study. All participants were native English speakers. All NH subjects had pure-tone thresholds better than 20 dB HL at octave frequencies from 125 to 8000 Hz in both ears. All CI users were post-lingually deafened; 5 of the 6 CI subjects had at least one-year experience with their device (the sixth subject, a user of Cochlear's Freedom device, had 4 months' experience with the device). CI subjects included 3 Nucleus-22 users, 2 Nucleus-24 users, and one Freedom user. CI subjects were tested using their clinically assigned speech

processors. All subjects were paid for their participation.

2.2. Stimuli and speech processing

An emotional speech database (HEI-ESD) was recorded for the present study. One male and one female talker each produced 50 simple, everyday English sentences according to 5 target emotional qualities (i.e., neutral, anxious, happy, sad, and angry). The same sentences were used to convey the different target emotions in order to minimize the contextual and discourse cues, thereby focusing listeners' attention on the acoustic cues associated with the target emotions. Speech samples were digitized using a 16-bit A/D converter at a 22,050 Hz sampling rate, without high-frequency pre-emphasis. The relative intensity cues were preserved for each emotional quality; speech samples were not normalized. The database was first evaluated with 3 NH English-speaking listeners; the 10 sentences that produced the highest vocal emotion recognition scores were selected for experimental testing, resulting in a total of 100 tokens (2 speakers $\times$ 5 emotions $\times$ 10 sentences). Table 1 lists the sentences used in experimental testing.

Table 1: *Sentences used in the present study*

List	Sentences
1	It takes two days.
2	I am flying tomorrow.
3	Who is coming?
4	Where are you from?
5	Meet me there later.
6	Why did you do it?
7	It will be the same.
8	Come with me.
9	The year is almost over.
10	It is snowing outside.

CI subjects' vocal emotion recognition was tested using unprocessed speech. NH subjects' vocal emotion recognition was tested using unprocessed speech, as well as speech processed by acoustic, sine-wave vocoder CI simulations. The Continuous Interleaved Sampling (CIS) strategy [14] (used in many CI devices), was simulated as follows. After pre-emphasis (1st-order Butterworth high-pass filter at 1200 Hz), the input speech signal was band-pass filtered into 1, 2, 4, 8, or 16 frequency bands (4th-order Butterworth filters). The overall input acoustic frequency range was from 100 to 7000 Hz; for each spectral resolution condition, the corner frequencies of the analysis bands were calculated according to Greenwood's formula [15]. The temporal envelope from each band was extracted by half-wave rectification and low-pass filtering (4th-order Butterworth) at either 50 or 500 Hz (according to the experimental condition). The temporal envelope from each band was used to modulate a sine wave generated at center frequency of the analysis band. Finally, the amplitude-modulated sine waves from all frequency bands were summed and then normalized to have the same long-term root-mean-square (RMS) amplitude as the input speech signal. Figure 1 shows a simplified block diagram of the CI simulation speech processing.

2.3. Procedure

Subjects were seated in a double-walled sound-treated booth and listened to the stimuli presented in free field over a single loudspeaker (Tannoy Reveal). The presentation level (65 dBA) was calibrated according to the average power of the "angry" emotion sentences produced by the male talker (which had the highest long-term RMS amplitude among the 5 emotions produced by the 2 speakers). A closed-set, 5-alternative identification task was used to measure vocal emotion recognition. In each trial, a sentence was randomly selected (without replacement) from the stimulus set and presented to the subject; subjects responded by clicking on one of the five response choices shown on screen (labeled "neutral," "anxious," "happy," "sad," and "angry"). No feedback or training was provided. Responses were collected and scored in terms of percent correct. There were at least two runs for each experimental condition. For the CI simulations, the test order of speech processing conditions was randomized across subjects, and different between the two runs.

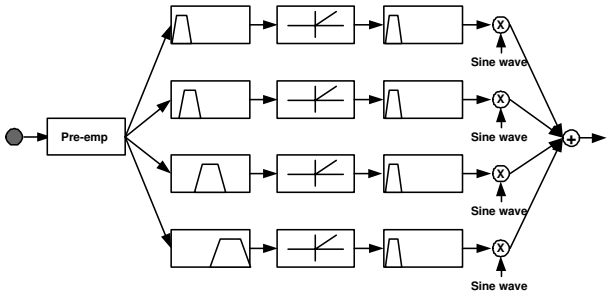

Figure 1: *Example of 4-channel speech processing for the CI simulation.*

3. Results

Figure 2 shows vocal emotion recognition performance for CI users (filled squares) and NH subjects (filled triangles) listening to unprocessed speech. Mean NH performance (across subjects) with the CI simulations is shown as a function of the number of spectral channels; the error bars show 1 standard deviation. The open circles show data with the 50-Hz envelope filter and the filled circles show data with the 500-Hz envelope filter. The dashed horizontal line shows chance performance level (20 % correct).

With unprocessed speech, mean NH performance was 90 % correct, while mean CI performance was only 45 % correct. Note that there was large inter-subject variability in each subject group. While much lower than NH performance, CI performance was significantly better than chance performance level [paired t-test: $t(5) = 6.1, p = 0.002$].

With the acoustic CI simulations, NH subjects' vocal emotion recognition performance was significantly affected by both the number of spectral channels and the temporal envelope filter cutoff frequency [one-way, repeated measures analysis of variance (ANOVA): $F(10, 50) = 67.5, p < 0.001$]. Post-hoc Bonferroni pair-wise comparisons were performed for all experimental processors. With the 50-Hz envelope filter, performance significantly improved when the number of spectral channels was increased ($p < 0.05$), except from 1 to 2, and from 4 to 8. With the 500-Hz envelope filter, performance significantly improved only when the number of spectral channels was increased from 1 to more than 1, and from 2 to 16 ($p < 0.05$). NH Performance with unprocessed speech was significantly better than performance in any of the CI simulations ($p < 0.05$), except the 16-channel simulation with the

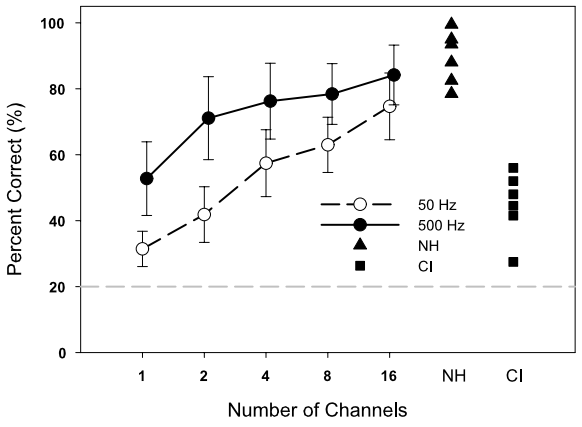

Figure 2: *Vocal emotion recognition by CI users and NH subjects (listening to unprocessed speech and CI simulations).*

500-Hz envelope filter. For all but the 16-channel processors, performance was significantly better with the 500-Hz envelope filter than with the 50-Hz envelope filter ($p < 0.05$). There was no significant difference in performance between the 1-channel processor with the 500-Hz envelope filter and the 2-, 4-, and 8-channel processors with the 50-Hz envelope filter. Similarly, there was no significant difference in performance between the 2-channel processor with the 500-Hz envelope filter and the 8- and 16-channel processors with the 50-Hz envelope filter. Finally, there was no significant difference in performance between the 16-channel processor with the 50-Hz envelope filter and the 2-, 4-, 8- and 16-channel processors with the 500-Hz envelope filter.

NH performance with the CI simulations was also compared to CI performance with unprocessed speech using a one-way ANOVA [$F(10, 55) = 18.4, p < 0.001$]. Post-hoc Bonferroni pair-wise comparisons showed no significant difference in performance between CI users and NH subjects listening to 1, 2, 4, or 8 spectral channels with the 50-Hz envelope filter, or to 1 spectral channel with the 500-Hz envelope filter ($p > 0.05$). CI subject performance was significantly worse than that of NH subjects listening to 16 channels with the 50-Hz envelope filter, or to 2, 4, 8 or 16 channels with the 500-Hz envelope filter ($p < 0.05$).

4. Discussion

Although the present results may not be readily generalized to the wide range of conversation scenarios, these data provide useful comparisons between NH and CI listeners' perception of vocal emotion with unprocessed speech, as well as interesting observations regarding NH performance with the different CI simulations. With unprocessed speech, NH subjects' vocal emotion recognition performance was comparable to results reported in previous studies [9, 10]. Not surprisingly, CI performance was much poorer than NH performance. Note that, because of the limited number of CI subjects, it was not possible to correlate CI performance with implant device type; however, there was no clear trend in the pattern of results. While CI users may be able to perceive intensity and speech rate cues coded in emotionally produced speech, they have only limited access to pitch, voice quality, and articulation cues, due to the reduced spectral and temporal resolution provided by the implant device. Consequently, mean CI performance was only half of that of NH subjects. However, mean CI performance was well above chance performance level, suggesting that CI users may have perceived at least some of the vocal emotion cues. The relative contributions of intensity, speech rate, pitch, voice quality, and articulation cues to CI users' vocal emotion recognition are presently unclear.

The CI simulation results obtained with NH subjects suggest that temporal cues may contribute more strongly to vocal emotion recognition than spectral cues, especially when 8 or fewer spectral channels were available. For all but the 16-channel processors, increasing the temporal envelope filter cutoff frequency significantly improved performance. A similar result for voice gender recognition, which also depends strongly on fundamental frequency (F0) and periodicity cues, was reported for NH subjects listening to acoustic CI simulations [2]. Temporal cues have also been shown to significantly contribute to speaker identification in acoustic and electric hearing [3]. In terms of spectral resolution with the 500-Hz envelope filter, the greatest improvements in performance occurred when the number of channels was increased from 1 to 2, beyond which performance did not significantly improve until 16 channels were available. Improved recognition of linguistic information as the spectral resolution was increased from 1 to 2 channels may have contributed to the improved vocal emotion recognition. With the 50-Hz envelope filter, performance gradually improved as the number of channels was increased, suggesting that spectral cues, in the absence of periodicity cues, may contribute strongly to vocal emotion recognition.

The apparent trade-off between spectral and temporal cues in the CI simulations is somewhat difficult to interpret. For example, 1 - 2 spectral channels with the 500-Hz envelope filter produced similar performance as 2 - 8 spectral channels with the 50-Hz envelope filter. It is possible that, while the 500-Hz envelope filter provided important periodicity cues, spectral sidebands around the sine-wave carriers may have provided salient spectral pitch cues. While with the 50-Hz envelope filter, the sidebands would have provided much weaker spectral cues. In the present study, with 16 spectral channels, the temporal envelope filter cutoff frequency did not significantly affect performance. It is possible that pitch cues were adequately preserved with 16 spectral channels, and that any additional cues due to spectral sidebands and/or periodicity fluctuations did not contribute to performance.

There was no significant difference in performance between CI users and NH subjects listening to CI simulations with 1 - 8 spectral channels and the 50-Hz envelope filter, or with 1 spectral channel and the 500-Hz envelope filter. Again, it is somewhat difficult to interpret these results. While most CI users have between 16 and 22 channels available in their device, their functional spectral resolution is generally limited to around 8 channels. Also, given the stimulation rates used in CI subjects' clinically assigned speech processors (250 - 1800 Hz/channel), CI subjects would have most likely received some periodicity cues (certainly more than those transmitted by the simulations with the 50-Hz envelope filter). Previous studies have suggested that sine-wave (rather than noise-band) CI simulations are most comparable to CI users in performance, because of the better representation of the temporal envelope and/or improved channel selectivity. However, given the wide range of spectral resolution with the 50-Hz envelope filter simulations that produced similar performance to CI users, it is difficult to know which simulation condition best reflected CI

performance. The comparable performance between CI users and the 1-channel simulation with the 500-Hz envelope filter suggests that temporal periodicity cues may compensate for reduced spectral resolution. Future studies with CI users, in which the spectral resolution and amount of temporal cues are directly controlled (i.e., via research interface rather than via clinical speech processors), may shed further light on the relative contributions of spectral and temporal cues to vocal emotion recognition, as well as on the most appropriate CI simulation with which to test NH listeners.

In the present study, there was significant inter-subject variability in both CI and NH performance. Interestingly, there was significant inter-subject variability in NH performance with unprocessed speech. This variability suggests that individual NH subjects may have interpreted the five target emotions (and their production by the two talkers) somewhat differently. Nonetheless, mean NH performance was double that of CI users. In light of the simulation results, future implant devices and speech processing strategies must increase the functional spectral resolution and/or enhance the reception of temporal pitch cues to improve CI users' vocal emotion recognition. Such strategies may also provide better talker recognition, music perception, and recognition of speech in noise, as all these listening conditions require good reception of pitch cues.

5. Conclusions

With unprocessed speech, mean NH performance in a vocal emotion recognition task was 90 % correct, while mean CI performance was only 45 % correct. With acoustic, sine-wave vocoder CI simulations, NH performance was significantly affected by both the number of spectral channels and the temporal envelope filter cutoff frequency. With the 50-Hz envelope filter, NH performance gradually improved as the spectral resolution was increased from 1 to 16 channels. However, with the 500-Hz envelope filter, performance significantly improved only when the spectral resolution was increased from 1 to 2, and then from 2 to 16 channels. For all but the 16-channel processors, NH performance was significantly better with the 500-Hz envelope filter than with the 50-Hz envelope filter. The best mean NH performance among the CI simulations (16 channels, 500-Hz envelope filter: 84 % correct) was statistically similar to that with unprocessed speech. There was no significant difference in performance between CI users and NH subjects listening to CI simulations with 1 - 8 channels and the 50-Hz envelope filter, or with 1 channel and the 500-Hz envelope filter. These results suggest a potential trade-off between spectral resolution and periodicity cues when performing a vocal emotion recognition task. The deficit in CI performance suggests that future speech processing strategies must improve access to spectral and temporal fine structure cues to enhance CI users' recognition of pitch cues, which contribute strongly to vocal emotion recognition.

6. Acknowledgements

We are grateful to all subjects for their participation in these experiments. Research was supported in part by NIH (R01-DC-004993 and R03-DC-008192).

7. References

[1] Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X.-S., "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants", J. Acoust. Soc. Amer., 110(2): 1150–1163, 2001.

[2] Fu, Q.-J., Chinchilla, S., Nogaki, G., and Galvin, J. J. III, "Voice gender identification by cochlear implant users: The role of spectral and temporal resolution", J. Acoust. Soc. Amer., 118(3): 1711–1718, 2005.

[3] Vongphoe, M. and Zeng, F.-G., "Speaker recognition with temporal cues in acoustic and electric hearing", J. Acoust. Soc. Amer., 118(2): 1055–1061, 2005.

[4] Geurts, L. and Wouters, J., "Better place-coding of the fundamental frequency in cochlear implants", J. Acoust. Soc. Amer., 115: 844–852, 2004.

[5] Green, T., Faulkner, A., Rosen, S., and Macherey, O., "Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification", J. Acoust. Soc. Amer., 118: 375–385, 2005.

[6] Williams, C. E. and Stevens, K. N., "Emotions and speech: some acoustical correlates", J. Acoust. Soc. Amer., 52(4): 1238–1250, 1972.

[7] Murray, I. R. and Arnott, J. L., "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", J. Acoust. Soc. Amer., 93(2): 1097–1108, 1993.

[8] Scherer, K. R., "Vocal communication of emotion: A review of research paradigms", Speech Commun., 40: 227–256, 2003.

[9] Dellaert, F., Polzin, T., and Waibel, A., "Recognizing emotion in speech", in Proc. ICSLP, 1996, 1970–1973.

[10] Petrushin, V. A., "Emotion recognition in speech signal: experimental study, development, and application", in Proc. ICSLP, 2000.

[11] Zhou, G., Hansen, J. H. L., and Kaiser, J. F., "Nonlinear feature based classification of speech under stress", IEEE Trans. Speech Audio Process., 9(3): 201–216, 2001.

[12] Lee, C.-M. and Narayanan, S. S., "Toward detecting emotions in spoken dialogs", IEEE Trans. Speech Audio Process., 13(2): 293–303, 2005.

[13] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues", Science, 270: 303–304, 1995.

[14] Wilson, B. S., Finley, C. C., and Lawson, D. T., "Better speech recognition with cochlear implants", Nature, 352: 236–238, 1991.

[15] Greenwood, D. D., "A cochlear frequency-position function for several species - 29 years later", J. Acoust. Soc. Amer., 87: 2592–2605, 1990.