Vocal Emotion Recognition by Normal-Hearing Listeners and Cochlear Implant Users

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

The present study investigated the ability of normal-hearing listeners and cochlear implant users to recognize vocal emotions. Sentences were produced by 1 male and 1 female talker according to 5 target emotions: angry, anxious, happy, sad, and neutral. Overall amplitude differences between the stimuli were either preserved or normalized. In experiment 1, vocal emotion recognition was measured in normal-hearing and cochlear implant listeners; cochlear implant subjects were tested using their clinically assigned processors. When overall amplitude cues were preserved, normal-hearing listeners achieved near-perfect performance, whereas listeners with cochlear implant recognized less than half of the target emotions. Removing the overall amplitude cues significantly worsened mean normal-hearing and cochlear implant performance. In experiment 2, vocal emotion recognition was measured in listeners with cochlear implant as a function of the number of channels

(from 1 to 8) and envelope filter cutoff frequency (50 vs 400 Hz) in experimental speech processors. In experiment 3, vocal emotion recognition was measured in normal-hearing listeners as a function of the number of channels (from 1 to 16) and envelope filter cutoff frequency (50 vs 500 Hz) in acoustic cochlear implant simulations. Results from experiments 2 and 3 showed that both cochlear implant and normal-hearing performance significantly improved as the number of channels or the envelope filter cutoff frequency was increased. The results suggest that spectral, temporal, and overall amplitude cues each contribute to vocal emotion recognition. The poorer cochlear implant performance is most likely attributable to the lack of salient pitch cues and the limited functional spectral resolution.

Keywords: vocal emotion; cochlear implant; normal hearing; spectral resolution; temporal resolution

Speech conveys not only linguistic content but also indexical information about the talker (eg, talker gender, age, identity, accent, and emotional state). Cochlear implant (CI) users' speech recognition performance has been extensively studied, in terms of phoneme, word, and sentence recognition, both in quiet and in noisy listening conditions. In general, contemporary implant devices provide many CI users with good speech understanding under

optimal listening conditions. However, CI users are much more susceptible to interference from noise than are normal-hearing (NH) listeners, because of the limited spectral and temporal resolution provided by the implant device.¹⁻³

More recently, several studies have investigated CI users' perception of indexical information about the talker in speech.^{4,7} For example, Fu et al⁵ found that CI users' voice gender identification was nearly perfect (94% correct) when there was a sufficiently large difference (~100 Hz) in the mean overall fundamental frequency (F0) between male and female talkers. When the mean F0 difference between male and female talkers was small (~10 Hz), performance was significantly poorer (68% correct). In the same study, voice gender identification was also measured in NH subjects listening to acoustic CI simulations,⁸ in

---

From the Department of Auditory Implants and Perception, House Ear Institute, Los Angeles, California.

Address correspondence to: Xin Luo, PhD, Department of Auditory Implants and Perception, House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057; e-mail: xlue@hei.org.

which the amount of available spectral and temporal information was varied. The results for NH subjects listening to CI simulations showed that both spectral and temporal cues contributed to voice gender identification and that temporal cues were especially important when the spectral resolution was greatly reduced. Compared with voice gender identification, speaker recognition is much more difficult for CI users. Vongphoe and Zeng⁷ tested CI and NH subjects' speaker recognition, using vowels produced by 3 men, 3 women, 2 boys, and 2 girls. After extensive training, NH listeners were able to correctly identify 84% of the talkers, whereas CI listeners were able to correctly identify only 20% of the talkers. These results highlight the limited talker identity information transmitted by current CI speech processing strategies.

Another important dimension of speech communication involves recognition of a talker's emotional state, using only acoustic cues. Although facial expressions may be strong indicators of a talker's emotional state, vocal emotion recognition is an important component of auditory-only communication (eg, telephone conversation, listening to the radio, etc). Perception of emotions from nonverbal vocal expression is vital to understanding emotional messages,⁹ which in turn shapes listeners' reactions and subsequent speech production. For children, prosodic cues that signal a talker's emotional state are particularly important. When talking to infants, adults often use "infant-directed speech" that exaggerates prosodic features, compared with "adult-directed speech." Infant-directed speech attracts the attention of infants, provides them with reliable cues regarding the talker's communicative intent, and conveys emotional information that plays an important role in their language and emotional development.¹⁰⁻¹²

Vocal emotion recognition has been studied extensively in NH listeners and with artificial intelligence systems.¹³⁻¹⁹ Emotional speech recorded from naturally occurring situations has been used in some studies^{13,19}; however, these stimuli often suffer from poor recording quality and present some difficulty in terms of defining the nature and number of the underlying emotion types. A more preferred experimental design involves recording professional actors' production of vocal emotion using standard verbal content; a potential difficulty with these stimuli is that actors' production may be overemphasized compared with naturally produced vocal emotion.¹⁵ To make the vocal portrayals resemble

natural emotional speech, actors were usually instructed to imagine corresponding real-life situations before recording with target emotions. Nevertheless, because the simulated emotional productions were reliably recognized by NH listeners (see below), they at least partly reflected natural emotion patterns. The target vocal emotions used in most studies include anger, happiness, sadness, and anxiety. Occasionally, other, more subtle target emotions have been used, including disgust, surprise, sarcasm, and complaint.¹⁴

Acoustic cues that encode vocal emotion can be categorized into 3 main types: speech prosody, voice quality, and vowel articulation.^{9,14,20} Vocal emotion acoustic features have been analyzed in terms of voice pitch (mean F0 value and variability in F0), duration (at the sentence, word, and phoneme level), intensity, the first 3 formant frequencies (and their bandwidths), and the distribution of energy in the frequency spectrum. Relative to neutral speech, angry and happy speech exhibits higher mean pitch, a wider pitch range, greater intensity, and a faster speaking rate, whereas sad speech exhibits lower mean pitch, a narrower pitch range, lower intensity, and a slower speaking rate. Using these acoustic cues, NH listeners have been shown to correctly identify 60% to 70% of target emotions^{15,17,20}; recognition was not equally distributed among the target emotions, with some (eg, angry, sad, and neutral) more easily identified than others (eg, happy). Artificial intelligence systems using neural networks and statistical classifiers with various features as input have been shown to perform similarly to NH listeners in vocal emotion recognition tasks.^{16,17,19}

Severe to profound hearing loss not only reduces the functional bandwidth of hearing but also limits hearing impaired (HI) listeners' frequency, temporal, and intensity resolution,²¹ making detection of the subtle acoustic features contained in vocal emotion difficult. Several studies have investigated vocal emotion recognition in HI adults and children wearing hearing aids (HAs) or CIs.²²⁻²⁷ In the House study,²³ 2 semantically neutral utterances ("Now I am going to move" and "2510") were produced in Swedish by a female talker according to 4 target emotions (angry, happy, sad, and neutral). Mean vocal emotion recognition performance for 17 CI users increased slightly from 44% correct at 2 weeks following processor activation to 51% correct at 1 year after processor activation. More recently, Pereira²⁴ measured vocal emotion recognition in 20 Nucleus-22 CI patients, using the

same utterances and 4 target emotions as used in the House study,²³ produced in English by 1 female and 1 male actor. Mean CI performance was 51% correct, whereas mean NH performance was 84% correct. When the overall amplitude of the sentences was normalized, NH performance was unaffected, whereas CI performance was reduced to 38% correct, indicating that CI listeners depended strongly on intensity cues for vocal emotion recognition. Shinall²⁵ and Peters²⁷ recorded a different emotional speech database to measure both identification and discrimination of vocal emotions in NH and CI listeners, including adult and child subjects. The database consisted of 3 utterances (“It’s time to go,” “Give me your hand,” and “Take what you want”) produced by 3 female speakers according to 4 target emotions (angry, happy, sad, and fearful); instead of the neutral emotion target used by House²³ and Pereira,²⁴ the fearful emotion target was used because it was suspected that child subjects might not understand neutrality. In the Shinall²⁵ and Peters²⁷ studies, talker and sentence variability significantly affected vocal emotion recognition performance of CI users but not of NH listeners. For some children who were still developing their emotion perception, age seemed to affect vocal emotion recognition performance.²⁵ However, Schorr²⁶ found somewhat different results in a group of 39 pediatric CI users. Using nonlinguistic sounds as stimuli (eg, crying, giggling), the investigator found that the age at implantation and duration of implant use did not significantly affect pediatric CI users’ vocal emotion recognition performance.

These previous studies demonstrate CI users’ difficulties in recognizing vocal emotions. In general, CI users more easily identified the sad, angry, and neutral target emotions than the happy and fearful targets. Similar to NH listeners, CI users tended to confuse happiness with anger, happiness with fear, and sadness with neutrality. Considering these results, House²³ suggested that CI patients primarily used intensity cues to perceive vocal emotions and that F0, spectral profile, and voice source characteristics provided weaker, secondary cues. However, in these previous studies, only a small number of utterances were tested (at most 3), and detailed acoustic analyses were not performed on the emotional speech stimuli, making it difficult to interpret the perceptual results.

In the present study, an emotional speech database was recorded, consisting of 50 semantically neutral, everyday English sentences produced by 1 female

and 1 male talker according to 5 primary emotions (angry, happy, sad, anxious, and neutral). After initial evaluations with NH English-speaking listeners, the 10 sentences that produced the highest vocal emotion recognition scores were selected for experimental testing. Acoustic features such as F0, overall root mean square (RMS) amplitude, sentence duration, and the first 3 formant frequencies (F1, F2, and F3) were analyzed for each sentence and for each target emotion. In experiment 1, vocal emotion recognition was measured in adult NH listeners and adult postlingually deafened CI subjects (using their clinically assigned speech processors); the overall amplitude cues across sentence stimuli were either preserved or normalized to 65 dB. In experiment 2, the relative contributions of spectral and temporal cues to vocal emotion recognition were investigated in CI users via experimental speech processors. Experimental speech processors were used not to improve performance over the clinically assigned processors but rather to allow direct manipulation of the number of spectral channels and the temporal envelope filter cutoff frequency. In experiment 3, the relative contributions of spectral and temporal cues to vocal emotion recognition were investigated in NH subjects listening to acoustic CI simulations. In experiments 2 and 3, speech was similarly processed for CI and NH subjects using the Continuous Interleaved Sampling (CIS)²⁸ strategy, varying the number of spectral channels and the temporal envelope filter cutoff frequency.

The House Ear Institute Emotional Speech Database (HEI-ESD)

Methods

The House Ear Institute emotional speech database (HEI-ESD) was recorded for the present study. One female and 1 male talker (both with some acting experience) each produced 50 semantically neutral, everyday English sentences according to 5 target emotions (angry, happy, sad, anxious, and neutral). The same 50 sentences were used to convey the 5 target emotions in order to minimize the contextual and discourse cues, thereby focusing listeners’ attention on the acoustic cues associated with the target emotions. For the recordings, talkers were seated 12 inches in front of a microphone (AKG 414B-ULS, AKG Acoustics GmbH, Vienna, Austria) and were instructed to perform according to the target emotions.

Table 1. Sentences Used in the Present Study

Sentences
1 It takes 2 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.

The recording technician would roughly judge whether the target emotions seemed to be met. Speech samples were digitized using a 16-bit A/D converter at a 22050-Hz sampling rate, without high-frequency pre-emphasis. During recording, the relative intensity cues were preserved for each utterance. After recording was completed, the database was evaluated by 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 talkers $\times$ 5 emotions $\times$ 10 sentences). Table 1 lists the sentences used for experimental testing.

Acoustic Analyses of the HEI-ESD

The F0, F1, F2, and F3 contours for each sentence produced with each target emotion were calculated using Wavesurfer speech processing software (under development at the Center for Speech Technology (CTT), KTH, Stockholm, Sweden).²⁹ Figure 1 shows the mean F0 values (across the 10 test sentences) for the 5 target emotions. A 2-way repeated measures analysis of variance (RM ANOVA) showed that mean F0 values significantly differed between the 2 talkers ($F_{1,36} = 1183.8$, $P < .001$) and across the 5 target emotions ($F_{4,36} = 127.8$, $P < .001$); there was a significant interaction between talker and target emotion ($F_{4,36} = 17.4$, $P < .001$). Post hoc Bonferroni t tests showed that for the female talker, target emotions were ordered in terms of mean F0 values (from high to low) as happy, anxious, angry, neutral, and sad; mean F0 values were significantly different between most target emotions ($P < .03$), except for neutral versus sad. For the male talker, target emotions were ordered in terms of mean F0 values (from high to low) as anxious, happy, angry, neutral, and

Figure 1. Mean F0 values of test sentences for the 5 target emotions. The white boxes show the data for the female talker, and the gray boxes show the data for the male talker. The lines within the boxes indicate the median; the upper and lower boundaries of the boxes indicate the 75th and 25th percentiles. The error bars above and below the boxes indicate the 90th and 10th percentiles. The symbols show the outlying data.

sad; mean F0 values were significantly different between most target emotions ($P < .03$), except for anxious versus happy, happy versus angry, and neutral versus sad.

The F0 variation range (ie, maximum F0 minus minimum F0) was calculated for each sentence produced with each target emotion. Figure 2 shows the range of F0 variation (across the 10 test sentences) for the 5 target emotions. A 2-way RM ANOVA showed that the range of F0 variation significantly differed between the 2 talkers ($F_{1,36} = 132.5$, $P < .001$) and across the 5 target emotions ($F_{4,36} = 53.3$, $P < .001$); there was a significant interaction between talker and target emotion ($F_{4,36} = 16.7$, $P < .001$). Post hoc Bonferroni t tests showed that the female talker had a significantly larger range of F0 variation than the male talker ($P < .03$) for all target emotions. For the female talker, target emotions were ordered in terms of range of F0 variation (from large to small) as happy, anxious, angry, neutral, and sad; the range of F0 variation was significantly different between most target emotions ($P < .001$), except for anxious versus angry and neutral versus sad. For the male talker, target emotions were ordered in terms of range of F0 variation (from large to small) as happy, angry, anxious, neutral, and sad; the range of F0 variation was significantly larger for happy, angry, and anxious than for neutral and sad.

Figure 2. Range of F0 variation of test sentences for the 5 target emotions. The white boxes show the data for the female talker, and the gray boxes show the data for the male talker. The lines within the boxes indicate the median; the upper and lower boundaries of the boxes indicate the 75th and 25th percentiles. The error bars above and below the boxes indicate the 90th and 10th percentiles. The symbols show the outlying data.

($P < .01$) and was not significantly different within the 2 groups of emotions.

The mean F1, F2, and F3 values were calculated for each sentence produced with each target emotion. Figure 3 shows the mean F1 values (across the 10 test sentences) for the 5 target emotions. A 2-way RM ANOVA showed that mean F1 values significantly differed between the 2 talkers ($F_{1,36} = 197.4$, $P < .001$) and across the 5 target emotions ($F_{4,36} = 17.0$, $P < .001$); there was a significant interaction between talker and target emotion ($F_{4,36} = 10.7$, $P < .001$). Post hoc Bonferroni t tests showed that for the female talker, target emotions were ordered in terms of mean F1 values (from high to low) as happy, anxious, angry, sad, and neutral; mean F1 values were significantly different between happy and any other target emotion ($P < .001$) but did not significantly differ among the remaining 4 target emotions. For the male talker, target emotions were ordered in terms of mean F1 values (from high to low) as angry, anxious, happy, neutral, and sad; mean F1 values significantly differed only between angry and sad ($P = .01$). A 2-way RM ANOVA showed that mean F2 values significantly differed between the 2 talkers ($F_{1,36} = 15.7$, $P = .003$) but not across the 5 target emotions ($F_{4,36} = 1.8$, $P = .15$). Similarly, mean F3 values significantly differed between the 2 talkers ($F_{1,36} = 53.2$, $P < .001$) but not across the 5 target emotions ($F_{4,36} = 1.3$, $P = .28$).

Figure 3. Mean F1 values of test sentences for the 5 target emotions. The white boxes show the data for the female talker, and the gray boxes show the data for the male talker. The lines within the boxes indicate the median; the upper and lower boundaries of the boxes indicate the 75th and 25th percentiles. The error bars above and below the boxes indicate the 90th and 10th percentiles. The symbols show the outlying data.

The RMS amplitude was calculated for each sentence produced with each target emotion. Figure 4 shows the overall RMS amplitude (across the 10 test sentences) for the 5 target emotions. A 2-way RM ANOVA showed that RMS amplitudes significantly differed between the 2 talkers ($F_{1,36} = 69.0$, $P < .001$) and across the 5 target emotions ($F_{4,36} = 150.8$, $P < .001$); there was a significant interaction between talker and target emotion ($F_{4,36} = 39.5$, $P < .001$). Post hoc Bonferroni t tests showed that RMS amplitudes for anxious, neutral, and sad were significantly higher for the male talker than for the female talker ($P < .001$); the RMS amplitude for happy was significantly higher for the female talker than for the male talker ($P < .001$). For the female talker, target emotions were ordered in terms of mean RMS amplitudes (from high to low) as happy, angry, anxious, neutral, and sad; RMS amplitudes were significantly different between most target emotions ($P < .001$), except for happy versus angry and neutral versus sad. For the male talker, target emotions were ordered in terms of mean RMS amplitudes (from high to low) as angry, anxious, happy, neutral, and sad; RMS amplitudes were significantly different between most target emotions ($P < .02$), except for anxious versus happy and neutral versus sad.

The overall duration was calculated for each sentence produced with each target emotion, which was inversely proportional to the speaking rate. Figure 5

Figure 4. Overall root mean square (RMS) amplitudes of test sentences for the 5 target emotions. The white boxes show the data for the female talker, and the gray boxes show the data for the male talker. The lines within the boxes indicate the median; the upper and lower boundaries of the boxes indicate the 75th and 25th percentiles. The error bars above and below the boxes indicate the 90th and 10th percentiles. The symbols show the outlying data.

Figure 5. Overall duration of test sentences for the 5 target emotions. The white boxes show the data for the female talker, and the gray boxes show the data for the male talker. The lines within the boxes indicate the median; the upper and lower boundaries of the boxes indicate the 75th and 25th percentiles. The error bars above and below the boxes indicate the 90th and 10th percentiles. The symbols show the outlying data.

shows the overall duration (across the 10 test sentences) for the 5 target emotions. A 2-way RM ANOVA showed that overall duration significantly differed across the 5 target emotions ($F_{4,36} = 8.3$, $P < .001$) but not between the 2 talkers ($F_{1,36} = 3.0$, $P = .11$); there was a significant interaction between

talker and target emotion ($F_{4,36} = 9.8$, $P < .001$). Post hoc Bonferroni t tests showed that for the female talker, overall duration was significantly shorter for neutral, compared with the remaining 4 target emotions ($P < .003$), and was significantly longer for sad than for angry and anxious ($P < .005$). However, for the male talker, there was no significant difference in overall duration across the 5 target emotions.

In summary, among these analyzed acoustic features, mean pitch values, the ranges of pitch variation, and overall RMS amplitudes are the most different across the 5 target emotions. Because of the limited spectrotemporal resolution provided by the implant device, CI users may have only limited access to these acoustic features. The following perceptual experiments were conducted to investigate the overall and relative contributions of intensity, spectral, and temporal cues to vocal emotion recognition by NH and CI listeners.

Experiment 1: Vocal Emotion Recognition by NH and CI Listeners

Subjects

Eight adult NH subjects (5 women and 3 men; age range 22-40 years, with a median age of 28) and 8 postlingually deafened adult CI users (4 women and 4 men; age range 41-73 years, with a median age of 58 years) 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. Table 2 shows the relevant demographic details for the participating CI subjects. Informed consent was obtained from all subjects, all of whom were paid for their participation.

Speech Processing

Vocal emotion recognition was tested with NH and CI subjects using stimuli in which the relative overall amplitude cues between sentences were either preserved or normalized to 65 dB. Beyond this amplitude normalization, there was no further processing of the speech stimuli. CI subjects were tested using their clinically assigned speech processors (see Table 2 for individual speech processing strategies). For the originally recorded speech stimuli, the relative intensity cues were preserved for each emotional quality. The highest average RMS

Table 2. Relevant Demographic Information for the Cochlear Implant Subjects Who Participated in the Present Study

Subject	Age	Gender	Etiology	Prosthesis	Strategy	Duration of Hearing Loss, y	Years With Prosthesis
S1	63	F	Genetic	Nucleus-24	ACE	30	3
S2	60	F	Congenital	Nucleus-24	SPEAK	17	8
S3	53	F	Congenital	Nucleus-24	SPEAK	Unknown	5
S4	49	M	Trauma	Nucleus-22	SPEAK	0.66	14
S5	65	M	Trauma	Nucleus-22	SPEAK	4	16
S6	56	M	Genetic	Freedom	ACE	30	0.33
S7	73	F	Unknown	Nucleus-24	ACE	30	6
S8	41	M	Measles	Nucleus-24	SPEAK	Unknown	7

Note: ACE = Advanced Combination Encoding; SPEAK = Spectral Peak.

amplitude (across the 10 test sentences) was for the angry target emotion with the male talker (ie, 65.3 dB). For the amplitude-normalized speech, the overall RMS amplitude of each sentence (produced by each talker with each target emotion) was normalized to 65 dB.

Testing Procedure

Subjects were seated in a double-walled sound-treated booth and listened to the stimuli presented in sound field over a single loudspeaker (Tannoy Reveal, Tannoy Ltd, North Lanarkshire, Scotland, UK); subjects were tested individually. For the original speech stimuli (which preserved the relative overall amplitude cues among sentences), 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 RMS amplitude among the 5 target emotions produced by the 2 talkers). For the amplitude-normalized speech, the presentation level was fixed at 65 dBA. CI subjects were tested using their clinically assigned speech processors. The microphone sensitivity and volume controls were set to their clinically recommended values for conversational speech level. CI subjects were instructed to not change this setting during the course of the experiment.

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 1 of the 5 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 2 runs for each experimental condition. The test order of speech processing conditions was randomized across subjects and was different between the 2 runs.

Results

Figure 6 shows vocal emotion recognition scores for NH listeners and for CI subjects using their clinically assigned speech processors, obtained with originally recorded and amplitude-normalized speech. Mean NH performance was 89.8% correct with originally recorded speech and 87.1% correct with amplitude-normalized speech; a paired *t* test showed that NH performance was significantly different between the 2 stimulus sets ($t_7 = 3.2$, $P = .01$). Mean CI performance was 44.9% correct with originally recorded speech and 37.4% correct with amplitude-normalized speech; a paired *t* test showed that CI performance was significantly different between the 2 stimulus sets ($t_7 = 3.2$, $P = .01$). Although much poorer than NH performance, CI performance was significantly better than chance performance level (ie, 20 % correct); (1-sample *t* tests: $P \leq .004$ for CI performance both with and without amplitude cues). Amplitude normalization had a much greater impact on CI performance than on NH performance.

There was great intersubject variability within each subject group. Among CI subjects, mean performance was higher for the 3 subjects using the Advanced Combination Encoding (ACE) strategy (52.5% correct with originally recorded speech) than for the 5 subjects using the Spectral Peak (SPEAK) strategy (40.3% correct with originally recorded speech). However, there were not enough subjects

Figure 6. Mean vocal emotion recognition scores (averaged across subjects) for normal-hearing (NH) listeners and for cochlear implant (CI) subjects using their clinically assigned speech processors, obtained with originally recorded (white bars) and amplitude-normalized speech (gray bars). The error bars represent 1 SD. The dashed horizontal line indicates chance performance level (ie, 20% correct).

to conduct a correlation analysis between CI performance and speech processing strategy. Interestingly, the best CI performer (S6: 56.5% correct with originally recorded speech) had only 4 months' experience with the implant, while the poorest CI performer (S5: 27.5% correct with originally recorded speech) had 16 years' experience with the implant. Thus, longer experience with the implant did not predict performance in the vocal emotion recognition task. Also, the best CI performer (S6) had a long history of hearing loss before implantation (30 years), whereas CI subject S4, whose performance was much lower (41.5% correct with originally recorded speech), had only 8 months of hearing loss before implantation. Thus, shorter duration of hearing loss did not predict performance in the vocal emotion recognition task either.

Table 3 shows the confusion matrix for CI data (collapsed across 8 subjects) with originally recorded speech. Recognition performance was highest for neutral and sad (mean: 65% correct) and poorest for anxious and happy (mean: 25% correct). The most frequent response was neutral, and the least frequent response was happy. Aside from the tendency toward neutral responses, there were 2 main groups of confusion: angry/happy/anxious (which had relatively higher RMS amplitudes, higher mean F0 values, and wider ranges of F0 variation) and sad/neutral (which

had relatively lower RMS amplitudes, lower mean F0 values, and smaller ranges of F0 variation). Recognition performance was higher for the female talker (mean: 50.0% correct) than for the male talker (mean: 39.8% correct), consistent with the acoustic analysis results, which showed that the acoustic characteristics of the target emotions were more distinctive for the female talker than for the male talker (see the previous section).

Table 4 shows the confusion matrix for CI data (collapsed across 8 subjects) with amplitude-normalized speech. Compared with CI performance with originally recorded speech (see Table 3), amplitude normalization greatly worsened recognition of the angry, sad, and neutral target emotions and had only small effects on recognition of the happy and anxious target emotions. After amplitude normalization, there was greater confusion between target emotions with higher original overall amplitudes (ie, angry, happy, and anxious) and those with lower original overall amplitudes (ie, neutral and sad). When overall amplitude cues were removed, performance dropped 10.7 percentage points with the female talker and 4.2 percentage points with the male talker.

Taken together, CI vocal emotion recognition performance was significantly poorer than that of NH listeners, suggesting that CI subjects perceived only limited amounts of vocal emotion information via their clinically assigned speech processors. Both CI and NH listeners used overall RMS amplitude cues to recognize vocal emotions. In experiments 2 and 3, the relative contributions of spectral and temporal cues to vocal emotion recognition were investigated in CI and NH listeners, respectively.

Experiment 2: Vocal Emotion Recognition by CI Subjects Listening to Experimental Speech Processors

Subjects

Four of the 8 CI subjects who participated in experiment 1 participated in experiment 2 (S1, S2, S6, and S7 in Table 2).

Speech Processing

Vocal emotion recognition was tested using the amplitude-normalized speech from experiment 1. Stimuli were processed and presented via custom research interface (HEINRI^{30,31}), bypassing subjects'

Table 3. Confusion Matrix (for Originally Recorded Speech) for All Cochlear Implant Subjects in Experiment 1

Originally Recorded Speech		Response				
		Angry	Happy	Anxious	Sad	Neutral
Target emotion	Angry	140 (43.8)	50 (15.6)	39 (12.2)	11 (3.4)	80 (25.0)
	Happy	65 (20.3)	78 (24.4)	82 (25.6)	9 (2.8)	86 (26.9)
	Anxious	49 (15.3)	47 (14.7)	81 (25.3)	13 (4.1)	130 (40.6)
	Sad	4 (1.2)	1 (0.3)	16 (5.0)	209 (65.3)	90 (28.1)
	Neutral	13 (4.1)	15 (4.7)	44 (13.8)	38 (11.9)	210 (65.6)

Note: The corresponding percentages of responses to target emotions are listed in parentheses.

Table 4. Confusion Matrix (for Amplitude-Normalized Speech) for All Cochlear Implant Subjects in Experiment 1

Amplitude-Normalized Speech		Response				
		Angry	Happy	Anxious	Sad	Neutral
Target emotion	Angry	98 (30.6)	47 (14.7)	65 (20.3)	10 (3.1)	100 (31.2)
	Happy	60 (18.8)	74 (23.1)	78 (24.4)	10 (3.1)	98 (30.6)
	Anxious	56 (17.5)	56 (17.5)	74 (23.1)	15 (4.7)	119 (37.2)
	Sad	9 (2.8)	7 (2.2)	23 (7.2)	174 (54.4)	107 (33.4)
	Neutral	40 (12.5)	28 (8.8)	59 (18.4)	14 (4.4)	179 (55.9)

Note: The corresponding percentages of responses to target emotions are listed in parentheses.

clinically assigned speech processors. The research interface allowed direct manipulation of the spectral and temporal resolution provided by the experimental speech processor. The CIS strategy²⁸ was implemented as follows. To transmit temporal envelope information as high as 400 Hz without aliasing, the stimulation rate on each electrode was fixed at 1200 pulses per second (pps), that is, 3 times the highest experimental temporal envelope filter cutoff frequency (400 Hz). Given the hardware limitations of the research interface in terms of overall stimulation rate and the 1200 pps/electrode stimulation rate, a maximum of 8 electrodes could be interleaved in the CIS strategy. The threshold and maximum comfortable loudness levels on each selected electrode were measured for each subject. The input speech signal was pre-emphasized (first-order Butterworth high-pass filter at 1200 Hz) and then band-pass filtered into 1, 2, 4, or 8 frequency bands (fourth-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.³² The temporal envelope from each band was extracted by half-wave rectification and low-pass filtering (fourth-order Butterworth filters) at either 50 or 400 Hz (depending on the experimental condition).

After extracting the acoustic envelopes, an amplitude histogram was calculated for each sentence (produced by each talker according to each target emotion). To avoid introducing any additional spectral distortion to the speech signal, the histogram was computed for the combined frequency bands rather than for each band separately. After computing the histogram, the peak level of the acoustic input range of the CIS processors was set to be equal to the 99th percentile of all amplitudes measured in the histogram; the top 1% of amplitudes were peak-clipped. The minimum amplitude of the acoustic input range was set to be 40 dB below the peak level; all amplitudes below the minimum level were center-clipped. This 40-dB acoustic input dynamic range was mapped to the electrode

Table 5. Selected Electrodes and Stimulation Modes for the Cochlear Implant Subjects in Experiment 2

Subject	Selected Electrodes	Stimulation Mode
S1	22, 19, 16, 13, 10, 7, 5, 3	MP1+2
S2	21, 19, 16, 13, 10, 7, 5, 3	MP1+2
S6	22, 19, 16, 13, 10, 7, 4, 1	MP1
S7	22, 20, 18, 16, 14, 12, 10, 8	MP1+2

Note: MP1+2 stimulation mode refers to monopolar stimulation with 1 active electrode and 2 extracochlear return electrodes. MP1 stimulation mode refers to monopolar stimulation with 1 active electrode and 1 extracochlear return electrode.

dynamic range by a power function, which was constrained so that the minimal acoustic amplitude was always mapped to the minimum electric stimulation level and the maximal acoustic amplitude was mapped to the maximum electric stimulation level. The current level of electric stimulation in each band was set to the acoustic envelope value raised to a power³³ (exponent = 0.2). This transformed amplitude envelope was used to modulate the amplitude of a continuous, 1200 pps, 25 μ s/phase biphasic pulse train. The stimulating electrode pairs depended on individual subjects' available electrodes. For the 8-channel processor, the 8 electrodes were distributed equally in terms of nominal distance along the electrode array. The electrodes for the 4-, 2-, and 1-channel processors were selected from among the 8 electrodes used in the 8-channel processor and again distributed equally in terms of nominal distance along the electrode array. Table 5 shows the experimental electrodes and corresponding stimulation mode for each subject. In summary, there were 4 spectral resolution conditions (1, 2, 4, and 8 channels) $\times$ 2 temporal envelope filter cutoff frequencies (50 and 400 Hz), resulting in a total of 8 experimental conditions.

The acoustic amplitude dynamic range of each input speech signal was mapped to the dynamic range of electric stimulation in current levels; the presentation level was comfortably loud. If there were channel summation effects for the different multichannel processors, most comfortable loudness levels on individual electrodes were globally adjusted to achieve comfortable loudness.

Testing Procedure

The same testing procedure used in experiment 1 was used in experiment 2, except that vocal emotion

recognition was tested while CI subjects were directly connected to the research interface, rather than the sound-field testing used in experiment 1.

Results

Figure 7 shows vocal emotion recognition scores for 4 CI subjects listening to amplitude-normalized speech via experimental processors, as a function of the number of channels. The open downward triangles show data with the 50-Hz temporal envelope filter, and the filled upward triangles show data with the 400-Hz temporal envelope filter. For comparison purposes, the filled circle shows mean performance for the 4 CI subjects listening to amplitude-normalized speech via their clinically assigned speech processors. Note that via clinically assigned speech processors, the 4 CI subjects in experiment 2 had higher mean vocal emotion recognition scores (45.6% correct) than the 8 CI subjects in experiment 1. Mean performance ranged from 24.5% correct (1-channel/50-Hz envelope filter) to 42.4% correct (4-channel/400-Hz envelope filter). A 2-way RM ANOVA showed that performance was significantly affected by both the number of channels ($F_{3,9} = 12.2$, $P = .002$) and the temporal envelope filter cutoff frequency ($F_{1,9} = 14.6$, $P = .03$); there was no significant interaction between the number of channels and envelope filter cutoff frequency ($F_{3,9} = 0.3$, $P = .80$). Post hoc Bonferroni t tests showed that performance significantly improved when the number of channels was increased from 1 to more than 1 (ie, 2-8 channels); ($P < .05$); there was no significant difference in performance between other spectral resolution conditions. Post hoc analyses also showed that performance was significantly better with the 400-Hz envelope filter than with the 50-Hz envelope filter ($P = .03$).

A 1-way RM ANOVA was performed to compare performance with the experimental and clinically assigned speech processors ($F_{8,24} = 7.9$, $P < .001$). Post hoc Bonferroni t tests showed that performance with the clinically assigned speech processors was significantly better than that with the 1- or 2-channel processors with the 50-Hz envelope filter ($P < .01$). Performance with the 1-channel processor with the 50-Hz envelope filter was significantly worse than that with the 2-, 4-, or 8-channel processors with the 400-Hz envelope filter ($P < .01$). Finally, performance with the 2-channel processor with the 50-Hz envelope filter was significantly worse than that with

Figure 7. Mean vocal emotion recognition scores for 4 cochlear implant subjects listening to amplitude-normalized speech via experimental processors, as a function of the number of channels. The open downward triangles show data with the 50-Hz temporal envelope filter, and the filled upward triangles show data with the 400-Hz temporal envelope filter. The filled circle shows mean performance for the 4 cochlear implant subjects listening to amplitude-normalized speech via clinically assigned speech processors (experiment 1). The error bars represent 1 SD. The dashed horizontal line indicates chance performance level (ie, 20% correct).

the 4-channel processor with the 400-Hz envelope filter ($P < .05$). There was no significant difference in performance between the remaining experimental speech processors.

In summary, CI subjects used both spectral and temporal cues to recognize vocal emotions. However, the contribution of spectral cues plateaued at 2 frequency channels.

Experiment 3: Vocal Emotion Recognition by NH Subjects Listening to Acoustic CI Simulations

Subjects

Six of the 8 NH subjects who participated in experiment 1 participated in experiment 3.

Speech Processing

NH subjects were tested using the amplitude-normalized speech from experiment 1, processed by acoustic CI simulations.⁸ The CIS strategy²⁸ was simulated as follows. After pre-emphasis (first-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 (fourth-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.³² The temporal envelope from each band was extracted by half-wave rectification and low-pass filtering (fourth-order Butterworth filters) 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. The starting phase of the sine wave carriers was fixed at 0°. Finally, the amplitude-modulated sine waves from all frequency bands were summed and then normalized to have the same RMS amplitude as the input speech signal. Thus, there were 5 spectral resolution conditions (1, 2, 4, 8, and 16 channels) $\times$ 2 temporal envelope filter cutoff frequencies (50 and 500 Hz), resulting in a total of 10 experimental conditions.

Testing Procedure

The same testing procedure used in experiment 1 was used in experiment 3, except that vocal emotion recognition was tested while NH subjects listened to acoustic CI simulations.

Results

Figure 8 shows vocal emotion recognition scores for 6 NH subjects listening to amplitude-normalized speech processed by acoustic CI simulations, as a function of the number of channels. The open downward triangles show data with the 50-Hz temporal envelope filter, and the filled upward triangles show data with the 500-Hz temporal envelope filter. For comparison purposes, the filled circle shows mean performance for the 6 NH subjects listening to unprocessed amplitude-normalized speech (experiment 1). Mean performance ranged from 22.6% correct (1-channel/50-Hz envelope filter) to 80.2% correct (16-channel/500-Hz envelope filter). A 2-way RM ANOVA showed that performance was significantly affected by both the number of channels ($F_{4,20} = 80.1$, $P < .001$) and the temporal envelope filter cutoff frequency ($F_{1,20} = 70.5$, $P < .001$); there was a significant interaction between the number of channels and envelope filter cutoff frequency ($F_{4,20} = 6.1$, $P = .002$). Post hoc Bonferroni t tests

Figure 8. Mean vocal emotion recognition scores for 6 normal-hearing subjects listening to amplitude-normalized speech via acoustic CI simulations, as a function of the number of channels. The open downward triangles show data with the 50-Hz temporal envelope filter, and the filled upward triangles show data with the 500-Hz temporal envelope filter. The filled circle shows mean performance for the 6 normal-hearing subjects listening to unprocessed amplitude-normalized speech (experiment 1). The error bars represent 1 SD. The dashed horizontal line indicates chance performance level (ie, 20% correct).

showed that with the 50-Hz envelope filter, performance significantly improved as the number of channels was increased ($P < .03$), except from 2 to 4 channels. With the 500-Hz envelope filter, performance significantly improved only when the number of channels was increased from 1 to more than 1 (ie, 2-16 channels) or from 2 to 16 ($P < .01$). Post hoc analyses also showed that for all spectral resolution conditions, performance was significantly better with the 500-Hz envelope filter than with the 50-Hz envelope filter ($P < .02$).

A 1-way RM ANOVA was performed to compare performance with simulated CI processors and unprocessed speech ($F_{10,50} = 65.0$, $P < .001$). Post hoc Bonferroni t tests showed that performance with unprocessed speech was significantly better than that with most acoustic CI simulations ($P < .002$) except the 4-, 8-, or 16-channel processors with the 500-Hz envelope filter. Performance was significantly different between most simulated CI processors ($P < .05$). There was no significant difference in performance between the 1-channel processor with the 500-Hz envelope filter and the 2- or 4-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- or 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 4- or 8-channel processors with the 500-Hz envelope filter.

In summary, when listening to acoustic CI simulations, NH subjects used both spectral and temporal cues to recognize vocal emotions. The contribution of spectral and temporal cues was stronger in NH subjects listening to the acoustic CI simulations than in CI subjects (experiment 2).

General Discussion

Although the present results for vocal emotion recognition may not be readily generalized to the wide range of conversation scenarios, these data provide comparisons between NH and CI listeners' perception of vocal emotion with acted speech and show the relative contributions of various acoustic features to vocal emotion recognition by CI users.

The near-perfect NH performance with originally recorded speech (~90% correct) was comparable to or better than results reported in previous studies,^{15,20,27} suggesting that the 2 talkers used in the present study accurately conveyed the 5 target emotions within the 10 test sentences. Thus, the HEI-ESD used in the present study appears to contain appropriate stimuli with which to measure CI users' vocal emotion recognition. Also consistent with previous studies,^{9,20} the HEI-ESD stimuli contained the requisite prosodic features needed to support the acoustic basis for vocal emotion recognition. According to acoustic analyses, mean pitch values, the ranges of pitch variation, and overall RMS amplitudes were strong emotion indicators for the 2 talkers. Happy, anxious, and angry speech had higher RMS amplitudes, higher mean F0 values, and wider ranges of F0 variation, whereas neutral and sad speech had lower RMS amplitudes, lower mean F0 values, and smaller ranges of F0 variation. Acoustic analyses showed no apparent distribution pattern in terms of overall sentence duration or speaking rate across the target emotions; if anything, neutral speech tended to be shorter in duration, whereas sad and happy speech tended to be longer in duration, at least for the female talker. Similarly, there was no apparent distribution pattern in terms of the first 3 formant frequencies across the target emotions, except that happy speech had higher mean F1 values than those for the remaining 4

target emotions, at least for the female talker. Other acoustic features (eg, formant bandwidth, spectral tilt) may be analyzed to more precisely describe variations in vowel articulation and spectral energy distribution associated with the target emotions. In terms of intertalker variability, the acoustic characteristics of the target emotions were more distinctive for the female talker than for the male talker. However, it is unclear whether this gender difference may be reflected in other talkers.

CI users' vocal emotion recognition performance via clinically assigned speech processors (~45% correct with originally recorded speech) was much poorer than that of NH listeners, consistent with previous studies.^{23-24,27} Although CI users may perceive intensity and speaking rate cues coded in emotionally produced speech, they have only limited access to pitch, voice quality, vowel articulation, and spectral envelope cues, attributable to the reduced spectral and temporal resolution provided by the implant device. Consequently, mean CI performance was only half that of NH listeners. However, mean CI performance with originally recorded speech was well above chance performance level, suggesting that CI users perceived at least some of the vocal emotion information. The confusion patterns for CI subjects among the target emotions were largely consistent with those found in previous studies.²³⁻²⁴ The neutral and sad target emotions were most easily recognized by CI subjects but possibly for different reasons. CI subjects may have failed to detect emotional acoustic features in many sentence stimuli and thus tended to respond with "neutral" (the most popular response choice). The speech produced according to the sad target emotion had the lowest RMS amplitudes, the lowest mean F0 values, and the smallest ranges of F0 variation and thus was likely to be quite distinct from the happy, angry, and anxious targets. As shown in Table 3, CI subjects were able to differentiate between target emotions with higher amplitudes (ie, angry, happy, and anxious) and those with lower amplitudes (ie, neutral and sad), suggesting that CI subjects used the relative intensity cues. The poorest performance was with the happy and anxious target emotions, which were often confused with each other and with the angry target emotion. Acoustic analyses revealed that these 3 target emotions differed mainly in mean F0 values and ranges of F0 variation. Most likely, CI subjects' limited access to pitch cues and other spectral details contributed to the confusion among these 3 target emotions.

Because of the greatly reduced spectral and temporal resolution, overall amplitude cues contributed more strongly to vocal emotion recognition for CI users than for NH listeners. Different from the Pereira study,²⁴ the present study found that overall amplitude cues significantly contributed to vocal emotion recognition not only for CI users but also for NH listeners, even though NH listeners had full access to other emotion features such as pitch cues and spectral details.

CI performance with the experimental speech processors (experiment 2) suggested that both spectral and temporal cues contributed strongly to vocal emotion recognition. With the 1-channel/50-Hz envelope filter processor, CI performance (24.5% correct) was only slightly above chance performance level, suggesting that the slowly varying amplitude envelope of the wide-band speech signal contained little vocal emotion information. For all spectral resolution conditions, increasing the envelope filter cutoff frequency from 50 to 400 Hz significantly improved CI performance, because the periodicity fluctuations added to the temporal envelopes contained temporal pitch cues. For both envelope filter cutoff frequencies, CI performance significantly improved when the number of channels was increased from 1 to 2, beyond which performance saturated. Channel interactions may explain why CI subjects were unable to perceive the additional spectral envelope cues provided by 4 or 8 spectral channels. A trade-off between spectral (ie, the number of channels) and temporal cues (ie, the temporal envelope filter cutoff frequency) was observed for CI users. For example, the 1-channel/400-Hz envelope filter processor produced the same mean performance as the 4-channel/50-Hz envelope filter processor, suggesting that temporal periodicity cues may compensate for reduced spectral resolution to some extent. CI performance with the clinically assigned speech processors was comparable to that with the 4- or 8-channel processors with the 400-Hz envelope filter, which is not surprising given that the ACE strategy used in the clinically assigned speech processors used a similar number of electrodes per stimulation cycle (8) and a similar stimulation rate (900, 1200, or 1800 pps).

The contribution of spectral and temporal cues was stronger in the acoustic CI simulations with NH listeners (experiment 3) than in the experimental processors with CI subjects (experiment 2). For all spectral resolution conditions, increasing the envelope filter cutoff frequency from 50 to 500 Hz

significantly improved NH performance. Although the maximum temporal envelope filter cutoff frequency used in the experimental CI processors (400 Hz) was slightly lower than that used in the acoustic CI simulations (500 Hz), it is unlikely that this difference accounts for the better NH performance under similar spectral resolution conditions, because the F0 values (both long-term and fluctuation range) were below 400 Hz for both talkers. Although the 500-Hz temporal envelopes may have provided important periodicity cues, it is possible that the spectral sidebands around the sine-wave carriers may have provided additional spectral pitch cues, which would not be available to real CI users. With the 50-Hz envelope filter, NH performance gradually improved as the number of channels was increased. Compared with CI users, NH listeners may have experienced less channel interaction and thus were able to perceive the additional spectral envelope cues associated with 16 spectral channels. The present results are consistent with many previous studies showing that CI users cannot access the additional spectral cues beyond 6 to 8 channels, whereas NH listeners are able to use these additional spectral cues in acoustic CI simulations, up to 16 spectral channels. With the 500-Hz envelope filter, the greatest improvement in NH performance occurred when the number of channels was increased from 1 to 2, beyond which performance did not significantly improve until 16 channels were provided. The temporal periodicity and/or spectral sideband cues with the 500-Hz envelope filter may have contributed to the lesser effects of increased spectral resolution. In summary, sine-wave acoustic CI simulations may have provided speech cues that are not available in the real implant case and thus may have overestimated the effects of spectral and temporal cues to vocal emotion recognition by CI users.

Nonetheless, the present results demonstrate the importance of overall amplitude, pitch, and other spectral cues to vocal emotion recognition by CI users. Future implant devices and their speech processing strategies must increase the functional spectral resolution and enhance the reception of salient voice pitch cues to improve CI users' vocal emotion recognition. Such strategies may also provide better talker identification, music perception, tone recognition, and speech understanding in noise, because these listening conditions all require good reception of voice pitch cues.

Conclusions

An emotional speech database was recorded for the present study that included 50 sentences produced by 2 talkers according to 5 target emotions. Acoustic analyses showed that mean pitch values, the ranges of pitch variation, and overall amplitudes were strong acoustic indicators for the target emotions. Consistent with the acoustic analyses, perceptual tests revealed that spectral envelope, temporal pitch, and overall amplitude cues significantly contributed to vocal emotion recognition by both NH listeners and CI subjects. Although CI subjects were able to use some of these cues to achieve moderate levels of vocal emotion recognition, CI performance remained much poorer than that of NH listeners. The results suggest that salient voice pitch cues may be the most important cues for vocal emotion recognition, and future speech processing strategies must improve the transmission of temporal fine structure and voice pitch cues to provide better vocal emotion recognition for CI patients.

Acknowledgments

We are grateful to all subjects for their participation in these experiments. We also thank 3 anonymous reviewers for their constructive comments on an earlier version of this article. This research was supported in part by grants R01-DC004993 and R03-DC008192 from the National Institutes of Health, Bethesda, Maryland. Portions of this work were presented at the 9th International Conference on Spoken Language Processing, Pittsburgh, Pennsylvania, 2006, and at the 30th Midwinter Meeting of the Association for Research in Otolaryngology, Denver, Colorado, 2007.

References

1. Fu QJ, Shannon RV, Wang X. Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. *J Acoust Soc Am.* 1998; 104:3586-3596.
2. Friesen LM, Shannon RV, Baskent D, Wang X. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am.* 2001;110:1150-1163.
3. Fu QJ, Nogaki G. Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. *J Assoc Res Otolaryngol.* 2005;6:19-27.
4. Fu QJ, Chinchilla S, Galvin JJ III. The role of spectral and temporal cues in voice gender discrimination by

- normal-hearing listeners and cochlear implant users. *J Assoc Res Otolaryngol*. 2004;5:253-260.
5. Fu QJ, Chinchilla S, Nogaki G, Galvin JJ III. Voice gender identification by cochlear implant users: the role of spectral and temporal resolution. *J Acoust Soc Am*. 2005;118:1711-1718.
  6. Gonzalez J, Oliver JC. Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *J Acoust Soc Am*. 2005;118:461-470.
  7. Vongphoe M, Zeng FG. Speaker recognition with temporal cues in acoustic and electric hearing. *J Acoust Soc Am*. 2005;118:1055-1061.
  8. Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. *Science*. 1995;270:303-304.
  9. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*. 1996;70:614-636.
  10. Fernald A. Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Dev*. 1989;60:1497-1510.
  11. Cooper RP, Aslin RN. Preference for infant-directed speech in the first month after birth. *Child Dev*. 1990;61:1584-1595.
  12. Trainor LJ, Austin CM, Desjardins RN. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychol Sci*. 2000;11:188-195.
  13. Williams CE, Stevens KN. Emotions and speech: some acoustical correlates. *J Acoust Soc Am*. 1972;52:1238-1250.
  14. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J Acoust Soc Am*. 1993;93:1097-1108.
  15. Scherer KR. Vocal communication of emotion: A review of research paradigms. *Speech Communication*. 2003;40:227-256.
  16. Dellaert F, Polzin T, Waibel A. Recognizing emotion in speech. In: *Proceedings of the International Conference on Spoken Language Processing*. 1996;1970-1973. Online archive of International Speech Communication Association proceedings available at: <http://www.isca-speech.org>.
  17. Petrushin VA. Emotion recognition in speech signal: experimental study, development, and application. In: *Proceedings of the International Conference on Spoken Language Processing*. 2000;2:222-225. Online archive of International Speech Communication Association proceedings available at: <http://www.isca-speech.org>.
  18. Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction. *IEEE Signal Process Mag* 2001;18:32-80.
  19. Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. *IEEE Trans Speech Audio Process*. 2005;13:293-303.
  20. Yildirim S, Bulut M, Lee CM, et al. An acoustic study of emotions expressed in speech. In: *Proceedings of the International Conference on Spoken Language Processing*. 2004;2193-2196. Online archive of International Speech Communication Association proceedings available at: <http://www.isca-speech.org>.
  21. Moore BC. Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear Hear*. 1996;17:133-161.
  22. Most T, Weisel A, Zaychik A. Auditory, visual and auditory-visual identification of emotions by hearing and hearing-impaired adolescents. *Br J Audiol*. 1993;27:247-253.
  23. House D. Perception and production of mood in speech by cochlear implant users. In: *Proceedings of the International Conference on Spoken Language Processing*. 1994;2051-2054. Online archive of International Speech Communication Association proceedings available at: <http://www.isca-speech.org>.
  24. Pereira C. The perception of vocal affect by cochlear implantees. In: Waltzman SB, Cohen NL, eds. *Cochlear Implants*. New York, NY: Thieme Medical; 2000:343-345.
  25. Shinall AR. *Emotion Perception in Pre-kindergarten School Children at Central Institute for the Deaf* [master's thesis]. St Louis, Mo: Washington University School of Medicine; 2005.
  26. Schorr EA. *Social and Emotional Functioning of Children With Cochlear Implants* [dissertation]. College Park: University of Maryland; 2005.
  27. Peters KP. *Emotion Perception in Speech: Discrimination, Identification, and the Effects of Talker and Sentence Variability* [dissertation]. St Louis, Mo: Washington University School of Medicine; 2006.
  28. Wilson BS, Finley CC, Lawson DT, Wolford RD, Eddington DK, Rabinowitz WM. Better speech recognition with cochlear implants. *Nature*. 1991;352:236-238.
  29. Sjölander K, Beskow J. *Wavesurfer speech processing software*. Department of Speech, Music and Hearing, KTH, Sweden, 2006. Available at: <http://www.speech.kth.se/wavesurfer/>. Accessed November 1, 2005.
  30. Shannon RV, Adams DD, Ferrel RL, Palumbo RL, Grandgenett M. A computer interface for psychophysical and speech research with the Nucleus cochlear implant. *J Acoust Soc Am*. 1990;87:905-907.
  31. Wygonski J, Robert ME. *HEI Nucleus Research Interface Specification*. Los Angeles, Calif: House Ear Institute; 2001.
  32. Greenwood DD. A cochlear frequency-position function for several species—29 years later. *J Acoust Soc Am*. 1990;87:2592-2605.
  33. Fu QJ, Shannon RV. Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners. *J Acoust Soc Am*. 1998;104:2570-2577.