Original Article

Trends in Hearing
Volume 22: 1–17

DOI: 10.1177/2331216518804966
journals.sagepub.com/home/tia
SAGE

Multimodal and Spectral Degradation Effects on Speech and Emotion Recognition in Adult Listeners

Chantel Ritter[1] and Tara Vongpaisal[1]

Abstract

For cochlear implant (CI) users, degraded spectral input hampers the understanding of prosodic vocal emotion, especially in difficult listening conditions. Using a vocoder simulation of CI hearing, we examined the extent to which informative multimodal cues in a talker's spoken expressions improve normal hearing (NH) adults' speech and emotion perception under different levels of spectral degradation (two, three, four, and eight spectral bands). Participants repeated the words verbatim and identified emotions (among four alternative options: happy, sad, angry, and neutral) in meaningful sentences that are semantically congruent with the expression of the intended emotion. Sentences were presented in their natural speech form and in speech sampled through a noise-band vocoder in sound (auditory-only) and video (auditory–visual) recordings of a female talker. Visual information had a more pronounced benefit in enhancing speech recognition in the lower spectral band conditions. Spectral degradation, however, did not interfere with emotion recognition performance when dynamic visual cues in a talker's expression are provided as participants scored at ceiling levels across all spectral band conditions. Our use of familiar sentences that contained congruent semantic and prosodic information have high ecological validity, which likely optimized listener performance under simulated CI hearing and may better predict CI users' outcomes in everyday listening contexts.

Keywords

cochlear implant, vocoder, auditory-visual perception, speech, emotion

Date received: 26 September 2017; revised: 16 July 2018; accepted: 30 August 2018

Understanding a spoken message, as well as the emotion in which it is expressed, is an important skill for successful communication. In face-to-face communication, it involves the simultaneous processing of the fine acoustic phonetic details to decode the words (semantic information), the manner in which the sentence is spoken (prosody), as well as visual speech and emotion cues discerned by lip reading and by viewing of a talker's facial expressions. While these processes occur with great ease and efficiency under normal listening conditions, they are more challenging under difficult listening conditions, or under conditions of hearing loss. Gathering speech content and emotion information from sound and vision, individually or in combination, may vary depending on the difficulty and demands of the listening situation.

For profoundly deaf and severely hard of hearing individuals, partial access to hearing sensations are enabled by cochlear implants (CIs)—a sensory prosthesis that is intended to electrically encode gross temporal and amplitude features in the original acoustic–phonetic detail that is sufficient for speech recognition in simple listening conditions (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Many implantees are able to acquire good oral language outcomes by discerning speech features from gross spectral and temporal features (Geers, 2004; Hochmair-Desoyer, Hochmair, Fischer, & Burian, 1980; Svirsky, Robbins, Kirk, Pisoni, & Miyamoto, 2000). Yet some aspects of speech perception, including voice and emotion perception that depend more on pitch and timbre cues, are hampered by the fine spectral detail that is not provided by CI processors (Chatterjee et al., 2015). This is supported by studies showing CI users' disproportionate difficulty in domains

[1]Department of Psychology, MacEwan University, Alberta, Canada

Corresponding author:
Tara Vongpaisal, Room 6-382, City Centre Campus, 10700 – 104 Avenue, Edmonton, Alberta T5J 4S2, Canada.
Email: VongpaisalT@macewan.ca

that depend more on pitch and timbral detail including song identification (Vongpaisal, Trehub, & Schellenberg, 2006) and voice perception (Cleary, Pisoni, & Kirk, 2005; Van Heugten, Volkova, Trehub, & Schellenberg, 2013).

Acoustic simulations of CI hearing demonstrate that spectral degradation interferes with voice recognition making it difficult to identify age and gender characteristics that are otherwise distinctive in acoustic voice samples (Vongpaisal, Trehub, Schellenberg, & Van Lieshout, 2012). In comparison to their speech decoding abilities, younger children have more difficulties with voice recognition at the same levels of degradation and require more spectral resolution than older children and adults (Vongpaisal et al., 2012). The costs of spectral degradation on emotion identification could be exacerbated in more challenging listening conditions (Shannon, Fu, & Galvin, 2004). Providing adequate speech-length samples and optimizing response requirements may overcome some of these challenges in forced choice tasks and ideal listening conditions (Volkova, Trehub, Schellenberg, Papsin, & Gordon, 2013).

The inability to hear distinctive pitch and intonation patterns that mark vocal emotions may have further consequences on CI children's ability to produce expressive emotions in speech (for a comprehensive review on voice perception and production skills of CI users, see Jiam, Caldwell, Deroche, Chatterjee, & Limb, 2017). For instance, it has been observed that CI children's vocal imitations of happy and sad speech notably lack differentiation in comparison with those of their NH peers, especially with regards to the pitch variability that distinguish these basic emotions (Wang, Trehub, Volkova, & Van Lieshout, 2013). As such, optimizing speech emotion perception may lead to more expressive productions of speech in CI children.

What we know is that despite CI users' significant accomplishments in decoding spoken speech from restricted cues (Shannon et al., 1995), discerning the emotional meaning conveyed in the prosody of speech remains challenging by comparison (Tinnemore et al., 2018). While developments to accurately encode a wider array of acoustic features for CI users are underway, enhancing communication outcomes using perceptual and cognitive resources is the focus of the present research. By observing hearing adults listening to vocoder simulations of CI hearing, we can approximate the relative cost of spectral degradation on speech and emotion perception, examine how the integration of cues across and auditory and visual modalities might have a role in mitigating these effects, identify baseline performances that can delineate differences between acoustic and electric hearing in this age demographic, and better anticipate the challenges that CI users may experience in similar tasks.

To date, there have been few studies examining the impact of spectral degradation on emotion recognition using vocoder simulations. In one notable study, Luo, Fu, and Galvin (2007) used acoustic simulations of CI hearing to parametrically vary spectral content (1–16 spectral bands), amplitude cue variation, and filter cutoff frequencies of temporal envelope information affecting NH adults' recognition of angry, anxious, happy, sad, and neutral emotions in comparison with actual CI users' performance. Unlike CI users, NH listeners still managed to identify emotions accurately when critical amplitude cues were removed by amplitude normalization. Their accuracy further improved with increasing spectral bands and temporal envelope information. For CI users, the combination of weak spectral cues and the lack of informative temporal and amplitude cues lead to poor emotion recognition overall. Despite the adverse listening conditions created by these vocoder simulations, NH participants managed to outperform CI users, thereby indicating notable differences between acoustic and electric hearing of emotion in speech.

In another study, Gilbers et al. (2015) compared the perceptual strategies of CI listeners and NH listeners with simulated CI hearing to recognize anger, sadness, joy, and relief. NH listeners relied more on mean pitch differences (i.e., voice fundamental frequency) across emotions, while CI users relied more on wide pitch range contrasts between high-arousal emotions (anger and joy) and low-arousal emotions (sadness and relief). The results suggest that the strong pitch salience available to NH listeners generated a more robust representation of emotion cues enabling them to be more tolerant of the acoustic features that are obscured by spectral degradation. Other factors including better linguistic and cognitive abilities in adults than in children have been shown to increase accuracy in identifying emotions from spectrally degraded speech by (Tinnemore et al., 2018).

Most studies concerning auditory perception with electrical hearing have focused primarily on the auditory modality, largely because of clinical interests in aural/oral communication outcomes (Holt, Kirk, & Hay-McCutcheon, 2011). Sumby and Pollack (1954) were among the first investigators to systematically examine the multimodal benefit to speech perception. The addition of visual speaking cues led to increased intelligibility of speech in adverse listening conditions that included high background noise, reverberation, and competing talkers. Visual cues that complemented auditory speech enabled listeners to perceive similar speech sounds that differed in place of articulation (such as /ba/ vs. /da/). These gains were more pronounced in listening conditions with low acoustic signal-to-noise ratios where performance was observed to increase from near zero to 70% to 80% accuracy with the addition of complementary visual cues.

More recently, Maidment, Kang, Stewart, and Amitay (2015) showed that older children and adults benefited from auditory–visual information to improve word recognition in vocoded speech. In contrast, younger children aged 4 to 5 years old did not experience any gain from the addition of visual information. However, it is possible that the high auditory attention requirements of decoding vocoded speech, coupled with the considerable task demands of selecting among a large set of response items (48 digit–color combinations) contributed to hampering audiovisual integration in these younger children. In this case, it is possible that age-related differences in task demands may have dimished the auditory–visual benefit seen in this younger age group.

Auditory–visual associations may be an important strategy for improving speech decoding and emotion recognition under conditions of spectral degradation. Integration of multisensory information is an automatic process that emerges early in life (Mildner & Koska, 2014). Infants are capable of using multimodal information in a range of perceptual and cognitive tasks (e.g., Bahrick & Lickliter, 2000; Flom & Bahrick, 2010) and are more inclined to perceive speech multimodally than children (Burnham & Dodd, 2004) who later demonstrate age-related improvement in these abilities extending into adulthood (Jerger, Damian, Spence, Tye-Murray, & Abdi, 2009; Sommers, Tye-Murray, & Spehar, 2005). In particular, when auditory and visual cues are complementary, multimodal integration can serve to improve speech understanding (Robbins, Renshaw, & Osberger, 1995; Valkenier, Duyne, Andringa, & Baskent, 2012). For example, visual speech cues from a talker's utterances contribute to improved sentence decoding in background noise, but the gains are notably increased when informative fine structure cues are removed in sinewave vocoded speech, in comparison to acoustic speech with fine structure cues intact (Stacey, Kitterick, Morris, & Sumner, 2016). In short, processing of visual information may become especially important with increasing difficulty in listening conditions.

Yet the combined use of auditory–visual cues, when paired incongruently, can sometimes result in a differently perceived stimulus as that occurring in the McGurk effect (McGurk & MacDonald, 1976), which has been documented in CI adults (Stropahl, Schellhardt, & Debener, 2017) and CI children (Tona et al., 2015). For instance, older child CI users are more susceptible to the McGurk effect than younger CI and NH children, which likely indicates a greater demand for higher order integration of auditory–visual information with age and with adaption to auditory impairment (Tona et al., 2015). When cues across modalities are congruent, however, the redundancies across the senses can be an important resource to draw upon when one sensory modality is compromised. Over the course of auditory rehabilitation following cochlear implantation, functionality of the visual cortex (Strelnikov et al., 2013), and visual speech perception abilities (Bergeson, Pisoni, & Davis, 2005) and their role in cross-modal plasticity, may be critical in promoting auditory speech recovery and its sustained improvement with CI use.

An auditory–visual benefit has also been observed to improve emotion recognition in conditions of spectral degradation. From clear video recordings of a female talker uttering a single nonsense sentence spoken in happy, sad, angry, and fearful expressions, Most and Michaelis (2012) found that CI children identified emotions more accurately when dynamic expressive cues are presented in audiovisual recordings in comparison to when they are presented in auditory-only and visual-only formats. CI children benefited from an auditory–visual gain that was similar to that of NH children and children fitted with hearing aids, but their identification scores were less accurate overall than their NH peers. Without informative visual cues, which were otherwise reliably available to both hearing and CI children, notable group differences emerged in the confusion patterns across emotions presented in the auditory-only modality (Most & Michaelis, 2012).

While nonsense (e.g., Most & Michaelis, 2012) and semantically neutral (e.g., Tinnemore et al., 2018) sentences are widely used in experimental task conditions to primarily encourage emotion identification by visual and auditory prosodic cues, it remains to be determined whether similar effects can be achieved with semantically meaningful sentences presented in the same formats. That is, we seek to examine emotion identification in sentences that are semantically congruent with the intended emotional expression in a manner that is akin to how speech is often encountered in everyday conversation. Our use of meaningful and familiar sentences will likely increase the ecological validity of CI simulated findings, and will complement the work of other approaches that used experimentally derived linguistic stimuli that were primarily designed to examine specific effects. Recent brain imaging studies on speech perception using ecologically valid linguistic contexts such as spontaneous speech accompanied by meaningful gestures (Weisberg, Hubbard, & Emmorey, 2017), coherent narratives extracted from fables (Xu, Kemeny, Park, Frattali, & Braun, 2005), and figurative language in short stories (Nagels et al., 2013) have not only corroborated findings from research using experimentally derived stimuli, but also revealed unique brain responses that were associated with the sensory and semantic integration that occurred exclusively in natural speech perception. Thus, linguistic materials with high ecological

validity, as those employed in this study, can extend insights from vocoder-simulated speech perception beyond those gained from experimentally-derived speech materials used widely in previous research.

The primary aim of the current study is to expand previous research findings on speech and emotion recognition by investigating the effects of multimodal information on hearing adults' speech and emotion recognition in vocoded simulations of spectrally degraded speech. Because discerning speech content from temporal cues is largely possible under these conditions, we predict that speech perception will improve with increasing spectral content in the auditory-only condition. Additional visual cues will improve speech decoding further, but will confer less advantage than in emotion recognition, which is disproportionately worsened by spectral degradation. As speech emotion depends more on fine spectral detail, identification from auditory-only speech information will be poor under the most degraded conditions, but will improve with increasing spectral content. Furthermore, as emotion recognition is a more challenging auditory task than speech decoding in CI listening conditions, the addition of informative visual cues will become more important for emotion recognition under the most degraded conditions. Accordingly, we expect to observe a larger auditory–visual benefit for emotion recognition accuracy than for speech decoding.

Method

Participants

A total of 30 adults ($M = 21.8$, standard deviation [SD] = 3.1 years) were recruited from the undergraduate research participant pool in the Department of Psychology at MacEwan University. All reported that they had NH and normal (or corrected) vision, and all spoke English as their first language. Motor ability was assessed prior to the experiment requiring participants to demonstrate facility in using a computer mouse for making response selections. All participants demonstrated normal manual skills in using the computer mouse and demonstrated no physical challenges in making timely responses using this device.

Stimuli

We produced audiovisual recordings of a single female actor generating expressive sentence-length speech in four emotions: Happy, sad, angry, and neutral (see Table 1 for the set of sentences and Supplemental Materials for auditory–visual examples). To create conditions that would best represent the naturalistic conditions encountered in conversational speech, we constructed sentences in which the semantic content is consistent with the prosodic expression of the emotions. The final corpus of audiovisual stimuli was selected based on a validation study that was conducted in our

Table 1. Stimuli Sentences.

Happy sentences
Speech decoding condition:
1. "It's so good to see you"
2. "The puppy is coming home today"
3. "It's beautiful outside today"
4. "I love you so much"
5. "My favorite movie is on"
Emotion recognition:
1. "Let's go play outside!"
2. "We're going on a trip today!"
3. "We're going to Grandma's today"
4. "We won our game today"
5. "I am so proud of you"
Angry sentences
Speech decoding condition:
1. "Why did you do that?"
2. "Don't talk to me like that!"
3. "Give that back, it's mine!"
4. "Don't be so rude!"
5. "Turn that game off right now!"
Emotion recognition:
1. "Stop yelling at me!"
2. "I can't believe you broke the toy!"
3. "You never share anything!"
4. "I don't like this toy!"
5. "It's my turn on the computer!"
Neutral sentences
Speech decoding condition:
1. "I'm going to the store"
2. "The pencil is on the table"
3. "He can run very fast"
4. "My neighbor is outside"
5. "I bought chips for the movie"
Emotion recognition:
1. "The movie is playing now"
2. "I have read many books"
3. "The store opens in ten minutes"
4. "The glass is beside you"
5. "Dinner is at five o'clock"
Sad sentences
Speech decoding condition:
1. "We had a fight yesterday"
2. "I am feeling sick today"
3. "I wish it would stop raining"
4. "I'm too tired to play"
5. "I can't go, I'm grounded"
Emotion recognition:
1. "I miss my mom very much"
2. "I just want to go home"
3. "My goldfish died yesterday"
4. "I can't find my friends"
5. "My arm is really sore"

laboratory in which undergraduate students categorized the emotions of each recording. The final set was selected among those that were most accurately (at least 85% correct) and reliably categorized. Sentences ranged from four to six words in length and were an average of 2.5 s in duration. Table 2 lists the average mean pitch and speech rate of the talker.

The audio tracks were extracted from the videos and processed through a noise-band vocoder script that was implemented in MATLAB according to the algorithm described in Shannon et al. (1995), Eisenberg, Shannon, Martinez, Wygonski, and Boothroyd (2000), and previously implemented in Vongpaisal et al. (2012). The audio data were processed through a series of bandpass filters (two, three, four, and eight bands with increasing spectral resolution), spanning a 300- to 6000-Hz frequency range. The cross-over frequencies for the spectral band conditions are reported in Table 3. The original sound files (digitized at 22 kHz, 16-bit resolution) were processed through a preemphasis Butterworth filter with cutoff frequency of 24 kHz and subsequently passed through a series of fourth-order elliptical band-pass filters (passband peak to peak ripple of 0.5 dB and minimum stopband attenuation of 40 dB) corresponding to the number of spectral bands. The temporal envelope, extracted from the original input signal by the Hilbert transform and low-passed filtered (cutoff 160 Hz), was used to modulate narrow-band Gaussian white noise. The product was then processed through the original input filters after which the outputs were then summed to form a noise-band signal. The result is a vocoded sample that preserves the temporal envelope and amplitude profile of the original acoustic signal, but its fine structure is replaced with white noise. Figure 1 shows spectrograms of an original and vocoded speech sample where it can be seen that formant spectral details are less distinct with increasing spectral degradation. The spectral band conditions were chosen based on a preliminary pilot study, in which 4, 8, 16, and 32 bands were used on an adult NH population that included 38 participants. Participants reached ceiling level for the higher resolution conditions, thus additional lower spectral band conditions were included in order to observe greater variability in performance for the current study.

While keeping the amplitude profile intact, the overall sound levels for each vocoded sound file was adjusted (using Audacity audio software) to match those of the natural unprocessed versions. To create auditory–visual stimuli for the spectral band conditions, the original soundtrack from the videos was extracted and replaced with its vocoded version using the Apple iMovie application. Each vocoded sound file was time-matched and dubbed on to the silent videos and saved to create integrated audiovisual stimuli.

For the emotion recognition task, a custom computer program was developed to present sound and video stimuli and to record participants' selections by computer mouse. Sound files were presented at 65 dB (A) as measured by sound level meter (Check Mate Galaxy Audio SPL Meter, CM-130) positioned at the ear-level of listeners. Sentences were blocked by spectral band condition and by modality. The order of the blocks, and the sentences within them, were randomized for each participant. For each spectral band and unprocessed speech condition, five sentences were randomly selected with the condition that no sentence was repeated across spectral band condition, modality condition, and across speech and emotion tasks.

Table 2. Speech Feature Characteristics of the Female Talker.

	Emotion			
Speech Feature	Neutral	Angry	Sad	Happy
Average voice pitch (F0), Hz	235.0	280.4	217.1	284.5
SD (Hz)	13.6	38.8	16.9	28.5
Average speech rate (words/s)	2.24	1.95	1.94	1.99
SD (words/s)	0.35	0.46	0.50	0.32
Average intensity range (dB)	31.9	39.0	29.8	35.4
SD (dB)	3.2	5.8	2.3	3.7

Note. *SD* = standard deviation.

Procedure

All procedures were approved by MacEwan University's Institutional Review Board and carried out in full accordance with the ethical standards of the Canadian

Table 3. Cutoff Frequencies (Hz) of Filterbanks for Each Spectral Band Condition.

Spectral band condition	Cutoff frequencies (Hz)								
Two bands	300	1528	6000						
Three bands	300	814	2210	6000					
Four bands	300	722	1528	3066	6000				
Eight bands	300	477	722	1061	1528	2174	3066	4298	6000

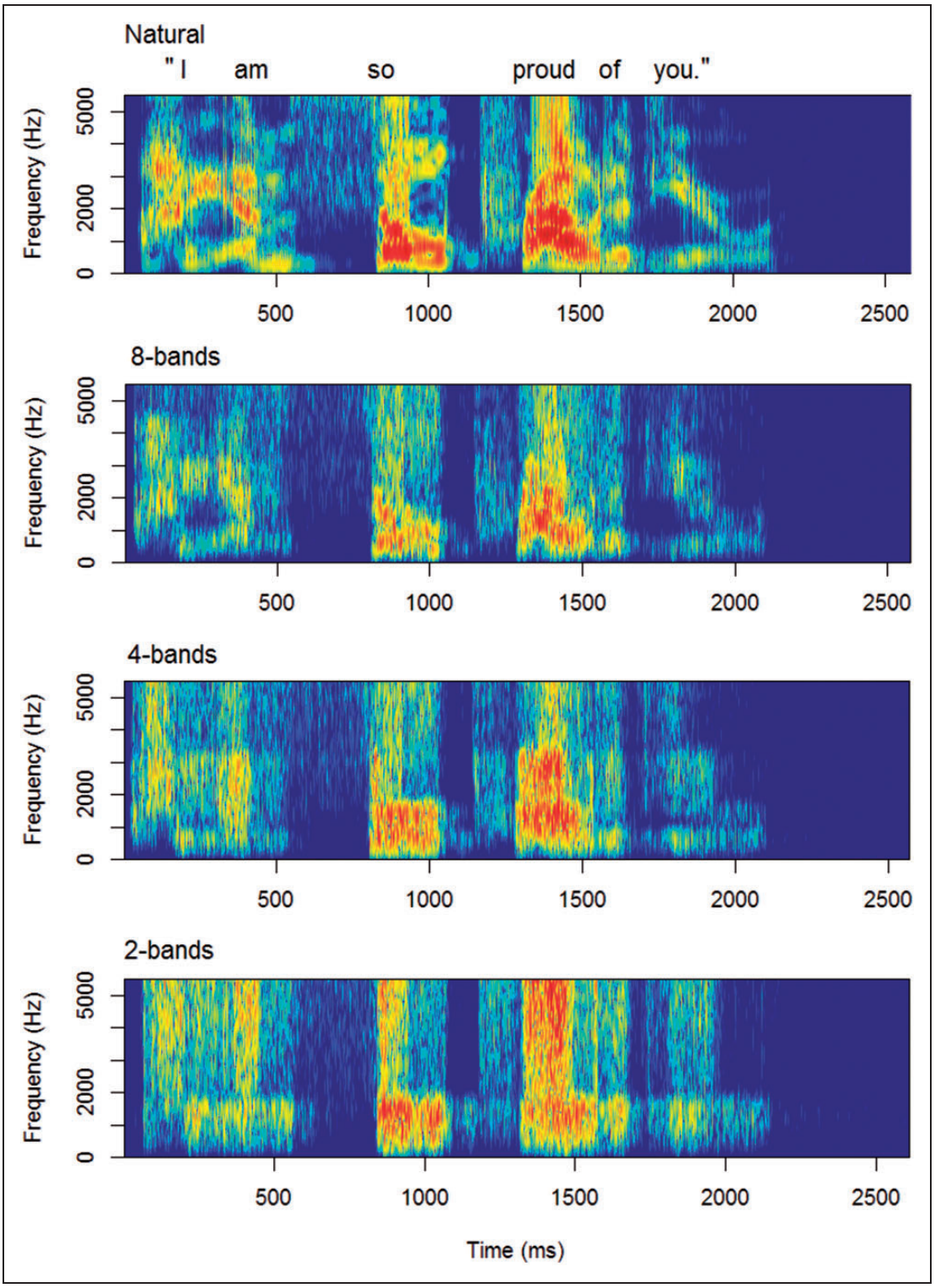

Figure 1. Spectrogram of an original and noise-band vocoded sentence.

Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. All participants provided written consent and completed a questionnaire of age, sex, language, and hearing status information. Course credit was given for their participation. All testing was conducted in a quiet room and was completed in a single experimental session lasting approximately 30 min in duration.

A practice session consisting of two auditory-only trials was administered to each participant in order to familiarize them with the sound of vocoded speech. For this purpose, sentences vocoded with 16 spectral bands were generated. This short exposure to the sounds of vocoded speech was intended to facilitate participants' transition to the task of listening to more distorted versions rather than to significantly increase the intelligibility of vocoded sentences as created by a perceptual *pop-out* effect (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005).

Participants were instructed to repeat the words that they heard and identify the emotion conveyed in the speech sample. In addition, two practice audiovisual samples of the same sentences were presented with the vocoded speech was dubbed onto the original visual recordings. Participants were instructed to repeat the words that they heard and identify the emotion in the video sample. Following their familiarization with vocoded sounds, they were informed that the subsequent task would involve listening to samples with greater levels of spectral degradation. Once participants were assured of the task requirements, they proceeded with the experiment.

The order of the speech decoding and emotion recognition task, modality, and spectral degradation conditions were counterbalanced for each participant. In the auditory-only speech decoding task, participants listened to sentence-length speech samples presented through loudspeakers (Harman/Kardon HK 195 multimedia speaker system) placed 60 cm apart on either side of the computer monitor that was positioned 60 cm in front of the listener. Following the presentation of the sentence, participants were asked to repeat the words verbatim. They proceeded to the next trial by clicking on the *Play* button presented on the computer monitor. In the auditory–visual version of this task, participants watched video samples of the talker speaking the sentences. Following each trial, participants' verbal responses were recorded with a microphone (Yeti LE Professional USB microphone) connected directly to a laptop computer and were saved for offline analysis.

In the auditory-only version of the emotion recognition task, participants listened to sound files with a static picture of a loudspeaker displayed onscreen for the duration of the trial. In the auditory–visual version, participants watched a video of the talker onscreen. Following each trial, participants were asked to identify the emotion by selecting among four icons presented on the computer monitor depicting *happy*, *sad*, *angry*, and *neutral* faces. No feedback was provided after each trial. All sound samples were presented at approximately 65 dB SPL through loudspeakers positioned at each side of a central computer monitor on a desktop surface.

Accuracy and response times were measured to assess speech and emotion recognition performance. While both tasks did not occur under speeded task conditions as commonly employed in paradigms that assess cognitive load, self-paced response time measurements as used in the current study can be a useful index of cognitive processing and cognitive effort when intelligibility is affected by noise (Pals, Sarampalis, Van Rijn, & Başkent, 2015) and by spectral degradation as modeled by CI simulated hearing (Pals, Sarampalis, & Başkent, 2013).

Results

Speech Decoding Analysis

The recordings were assessed offline by one experimenter and a speech decoding score was calculated as the total number of correctly repeated whole words per sentence. Accuracy per spectral band condition was reported as the percentage of correctly repeated words across all sentences in that condition. Similar to a previous method used evaluate performance in a sentence repetition task (Hudgins & Cullinan, 1978), we measured response latency as the interval between the offset of the prerecorded sentence and the onset of the participant's vocal repetition of that sentence.

Figure 2 depicts the speech decoding accuracy per spectral band condition. To examine whether accuracies achieved in both auditory and auditory–visual speech decoding of the original (unprocessed) recordings were at ceiling, one-sample t tests were conducted against perfect accuracy and confirmed that ceiling accuracies were achieved in these conditions ($ps > .05$). Further, all 30 participants achieved 100% accuracy in these conditions indicating that our recorded speech samples were readily intelligible when clear speech is presented in both modality conditions. Inspection of the score in the eight-band auditory–visual condition revealed near-ceiling scores. However, upon a closer examination of the scores between the eight-band and the original auditory–visual conditions, they were significantly different from each other, $t(29) = -3.29$, $p = .003$. In contrast to the original speech condition where all participants achieved perfect accuracy, 20 of 30 participants achieved 100% accuracy in the eight-band auditory–visual condition with the remaining participants scoring in the 86% to 95% accuracy range. Taken together, ceiling performance was not achieved in the eight-band auditory–visual condition.

To account for the ceiling, or near-ceiling effects, in the auditory–visual condition and its nonlinear and nonadditive effect on the proportion correct scale, statistical analyses were performed on the rationalized arcsine transformation of these scores (Studebaker, 1985). The

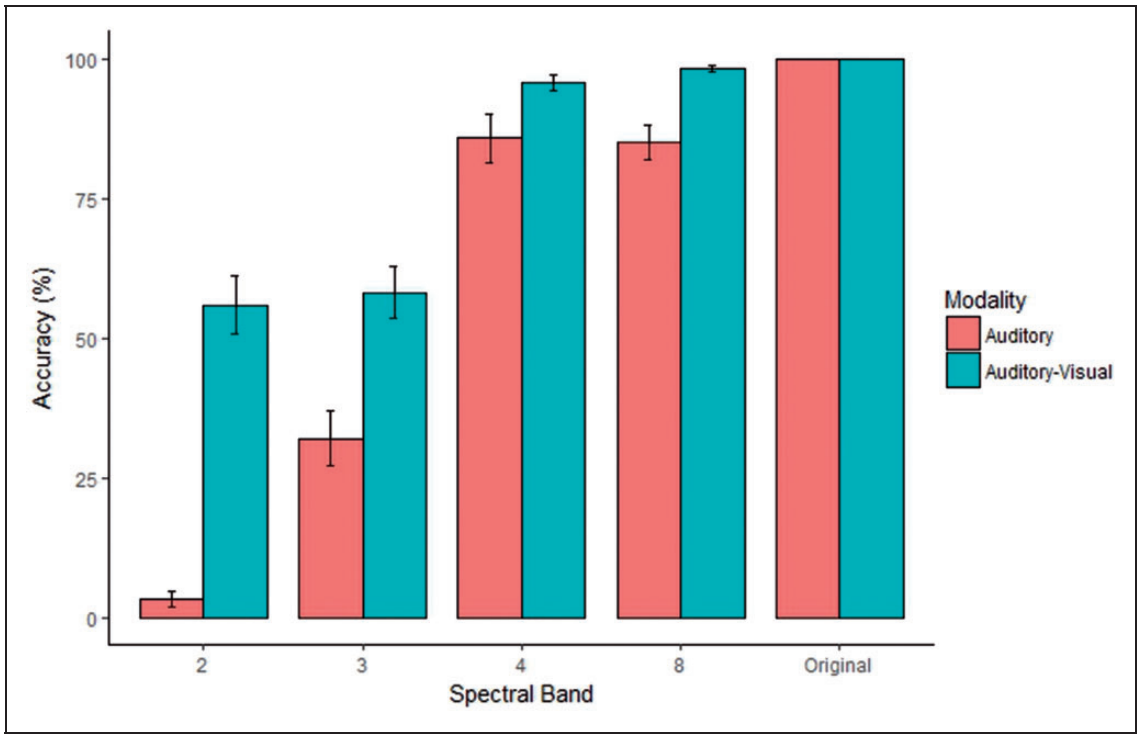

Figure 2. Mean speech decoding accuracy per spectral band condition for auditory and auditory–visual modalities. Error bars represent standard error of the mean.

rationalized arcsine transformed units (R) ranged from −1.45 to 30.37 rau, corresponding to accuracy scores of 0% to 100%, respectively. For ease of interpretation, however, the associated figures for speech decoding performance are displayed as percentage correct.

Because the speech decoding accuracy for the original speech samples was at ceiling ($M = 100\%$, $SD = 0\%$; or $R = 30.37$ rau, $SD = 0$ rau), it was not included in the analysis. A 2 modality (auditory-only, auditory-visual) × 4 spectral degradation (two, three, four, and eight spectral bands) repeated-measures analysis of variance (ANOVA; with Greenhouse–Geisser correction for sphericity) on speech decoding accuracy revealed a main effect of modality—$F(1.0, 29.0) = 51.37$, $p < .001$, $\eta^2 = .14$—with participants achieving higher accuracies in the auditory–visual condition ($M = 76.9\%$, $SD = 16.4\%$; $R = 24.87$ rau, $SD = 7.07$ rau) than in the auditory-only condition ($M = 51.5\%$, $SD = 19.0\%$; $R = 16.71$ rau, $SD = 12.68$ rau). In addition, the analysis revealed a main effect of spectral degradation: $F(2.0, 58.7) = 192.93$, $p < .001$, $\eta^2 = 0.54$. However, a two-way interaction between modality and spectral degradation—$F(1.6, 45.5) = 45.90$, $p < .001$, $\eta^2 = .10$—revealed that the loss of spectral detail affected speech decoding differently depending on whether speech was presented with auditory-only or auditory–visual information. To analyze this interaction, separate one-way repeated measures ANOVAs were conducted for each modality followed by post hoc multiple comparisons. A significant simple main effect of auditory modality—$F(2.5, 71.2) = 204.42$, $p < .001$, $\eta^2 = .88$—followed by pairwise t tests (Holm–Bonferroni correction) revealed that accuracy in the two-band condition ($M = 3.2\%$, $SD = 8.2\%$; $R = .41$ rau, $SD = 4.45$ rau), while significantly different from floor-level performance ($p = .03$), improved monotonically ($ps < .001$) across the three- and four-band conditions ($Ms = 31.9\%$, 85.8%; $SDs = 26.8\%$, 24.1%, respectively; or $Rs = 12.58$, 26.84 rau; $SDs = 8.39$, 6.78 rau, respectively). Finally, accuracy in the four-band condition matched that attained in the eight-band condition ($M = 84.9\%$, $SD = 16.7\%$; or $R = 27.0$ rau, $SD = 3.9$ rau, $p > .05$).

A significant simple main effect in the auditory–visual modality—$F(1.9, 55.0) = 44.32$, $p < .001$, $\eta^2 = .60$—revealed that accuracies in the lower two- and three-band conditions ($Ms = 55.9\%$, 58.1%; $SDs = 28.7\%$, 25.5%, respectively; or $Rs = 19.6$, 20.5 rau, $SD = 7.9$, 6.5 rau, respectively) were similar ($p > .05$) yet significantly poorer ($ps < .001$) than those attained in the higher four- and eight-band conditions ($Ms = 95.7\%$, 98.1%; $SDs = 8.2\%$, 3.1%, respectively; or $Rs = 29.4$, 30.0 rau; $SDs = 1.8$, $.7$ rau, respectively), which were also similar in accuracy ($p > .05$). Paired t tests comparing accuracies between the auditory and auditory–visual modalities at each spectral band condition revealed that speech decoding improved significantly in the

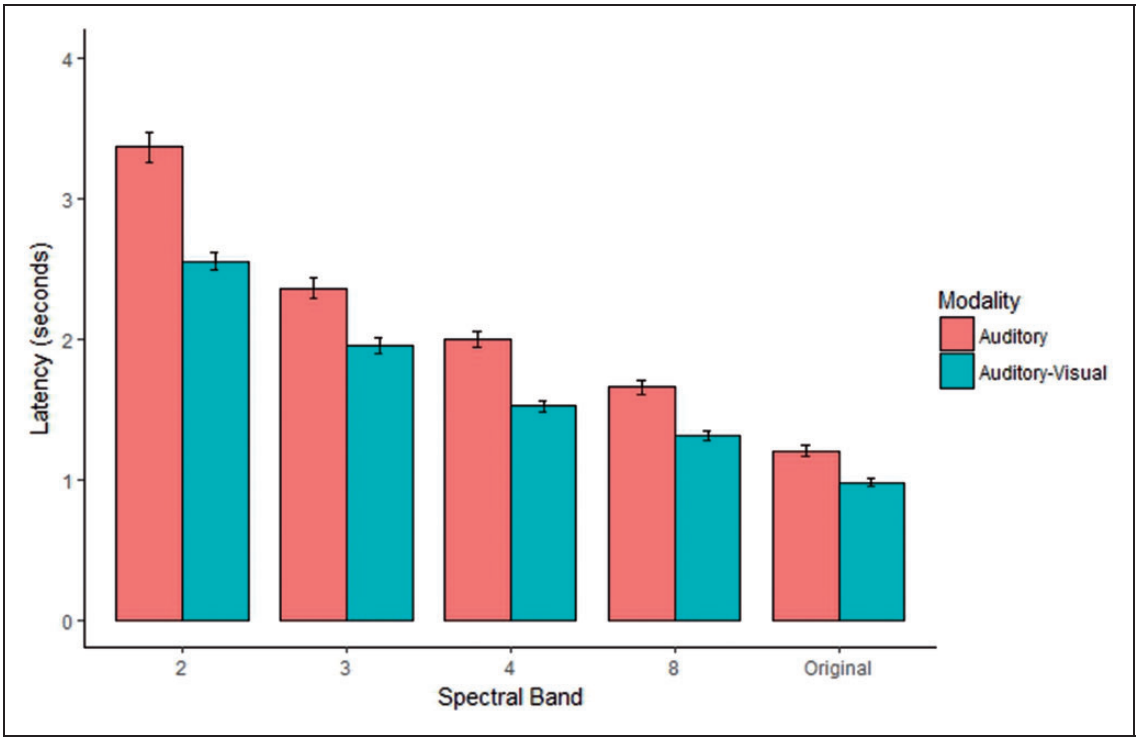

Figure 3. Latency of speech decoding per spectral band condition for auditory and auditory–visual modalities. Error bars represent standard error of the mean.

multimodal condition across all levels of spectral degradation ($ps < .05$).

To examine how modality and spectral degradation affected participants' response latencies (see Figure 3), response times were submitted to a 2 modality (auditory-only, auditory–visual) × 5 spectral degradation (two, three, four, and eight spectral bands, Original) ANOVA with Greenhouse–Geisser correction for sphericity. A main effect of modality—$F(1,29) = 37.46$, $p < .001$, $\eta^2 = .08$—revealed that speech decoding took longer in the auditory condition ($M = 2.11$ s, $SD = .93$ s) than in the auditory–visual condition ($M = 1.66$ s, $SD = 0.67$ s). The main effect of spectral degradation was also significant: $F(2.3, 67.9) = 172.33$, $p < .001$, $\eta^2 = .66$. Response times decreased monotonically (Ms = 2.96, 2.15, 1.76, 1.48, 1.09; SDs = .81, .58, .52, .42, .27 s) with increasing spectral information, as confirmed by paired t tests ($ps < .001$) conducted between successive spectral band conditions (two, three, four, eight, and Original, respectively).

More importantly, as seen in Figure 3, the significant two-way interaction between modality and spectral degradation condition—$F(2.9, 84.1) = 7.50$, $p < .001$, $\eta^2 = .02$—indicated that spectral degradation affected response latencies across modality conditions differently. An analysis of the simple main effects revealed that the difference in response latencies between the auditory-only and auditory–visual modality were significant at each spectral band condition (paired t tests, $ps < .001$), with spectral degradation having a more pronounced effect on increasing the response time of auditory-only speech decoding at the lowest spectral band conditions.

Emotion Recognition Analysis

Figure 4 displays the emotion recognition accuracy per spectral band condition. Exceptionally, accuracy scores were at ceiling across all spectral band conditions in the auditory–visual modality ($M = 99.6\%$, $SD = 1.6\%$) and were thus subsequently omitted from further analyses. While the assumption of normality was not met in the current data set, the assumption of sphericity was satisfied. Ceiling accuracies achieved in auditory–visual emotion recognition indicate that spectral degradation had no effect on emotion recognition when more reliable visual cues were provided. By contrast, emotion recognition in the auditory-only modality ($M = 70.0\%$, $SD = 20.1\%$) was affected by spectral degradation—$F(4, 116) = 56.85$, $p < .001$, $\eta^2 = .66$—yet scores in each spectral band condition were reliably above chance performance (where chance is 25%, $ps < .05$). Emotion recognition accuracy in the two-band condition ($M = 30.8\%$, $SD = 24.2\%$) was significantly poorer ($ps < .001$) than in three- and four-band conditions (Ms = 61.7%, 68.3%; SDs = 27.6%, 25.4%, respectively), which were not significantly different

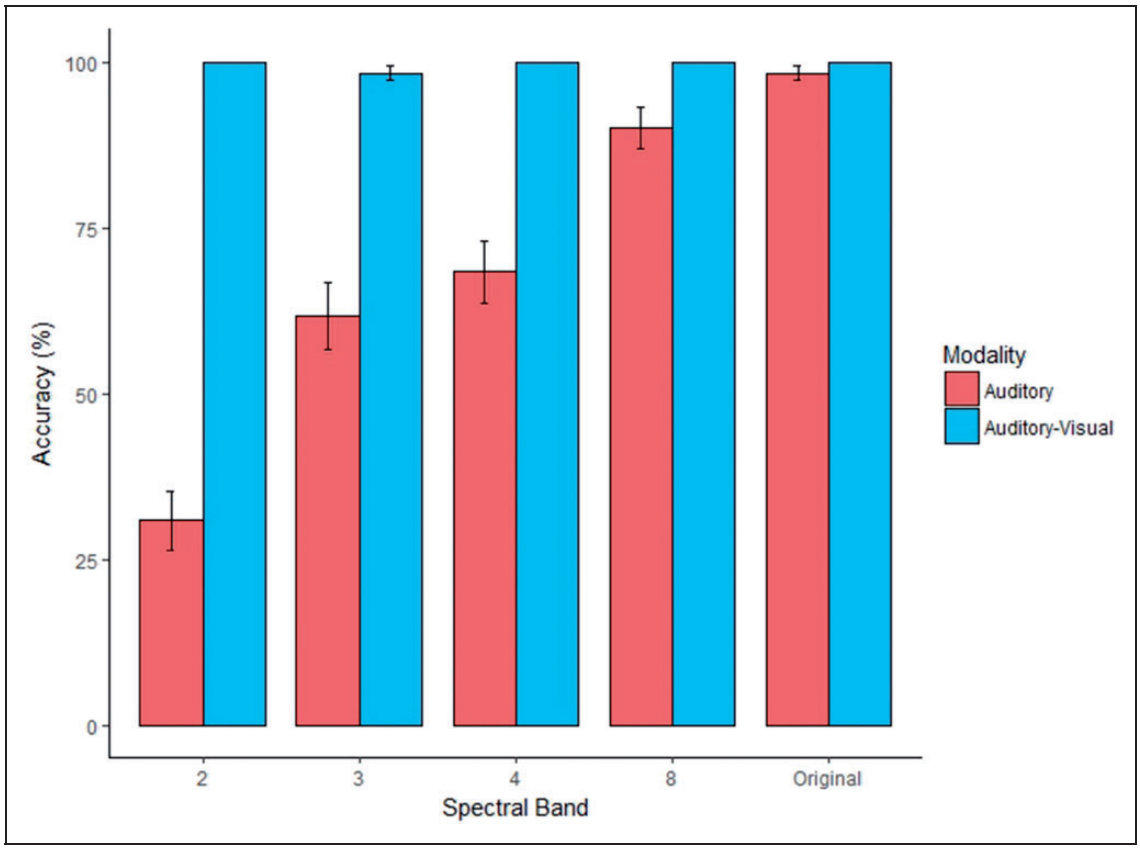

Figure 4. Mean emotion recognition accuracy per spectral band condition for auditory and auditory–visual modalities. Error bars represent standard error of the mean.

from each other. However, accuracy scores in these intermediate spectral degradation conditions were significantly poorer ($ps < .001$) than scores achieved in the eight-band condition ($M = 90.0\%$; $SD = 16.9\%$) and the original speech condition ($M = 98.3\%$, $SD = 6.3\%$), which was at ceiling levels.

Response latency in each condition was calculated as the average interval between mouse responses used to select an emotion icon onscreen. An analysis of latencies using the same ANOVA test revealed no significant main effects or two-way interaction between spectral degradation and modality (see Figure 5). Thus, unlike the response times used to discern the words in the sentence repetition task, the latency responses were not sensitive to spectral degradation and modality effects in the judgment of emotion in this context. In the auditory-only modality, the mean latencies across two-, three-, four-, eight-band, and Original conditions were 4.40, 4.19, 4.26, 3.62, and 4.27 s ($SDs = 1.57$, 2.21, 1.99, .65, and 1.53 s), respectively. In the auditory–visual modality, the mean latencies across conditions with increasing spectral resolution were 4.35, 4.38, 4.30, 4.43, and 4.34 s ($SDs = 0.36$, 1.36, .64, 1.80, and .39 s), respectively.

To examine the pattern of errors across emotion categories, we analyzed the frequency of confusions between the target emotion and responses. Figure 6 shows the confusion matrices for each spectral band condition in the auditory-only emotion recognition task. Each confusion matrix was submitted to a Kappa analysis to determine the agreement between the observed and expected responses (Congalton & Green, 2009). The greatest number of errors occurred in the most degraded two-band condition resulting in poor agreement between observed and expected responses, $\hat{K} = .10$, standard error (SE) = .07. While there were considerably fewer errors in the three- and eight-band conditions, agreement between observed and expected responses increased from poor to fair, $\hat{K} = .15$, .23 ($SEs = .09$, .10, respectively). The fewest errors occurred in the eight-band condition with a moderate agreement between observed and expected responses, $\hat{K} = .46$, $SE = .16$.

In the most degraded two-band condition, confusions were distributed across a wide range of emotion categories with the most frequent occurring among sad, happy, and neutral emotions. The frequency of errors among emotions decreased with increasing spectral

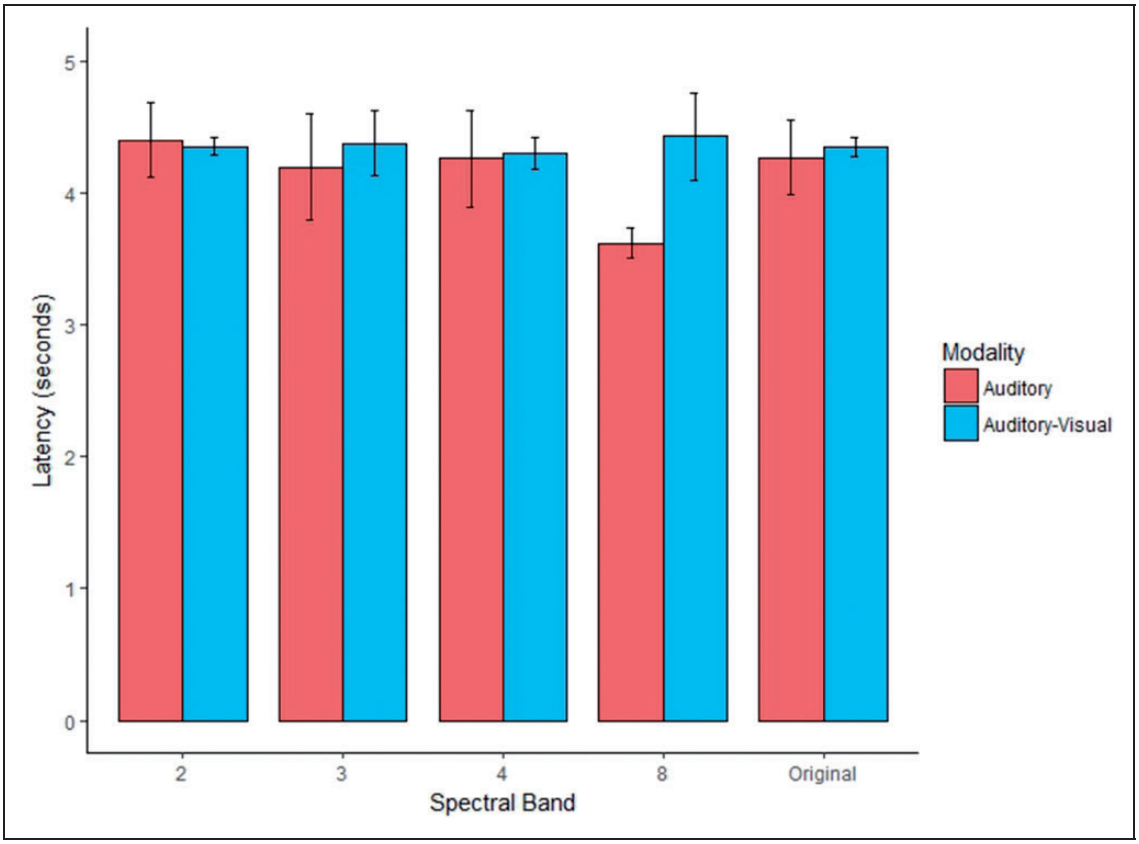

Figure 5. Latency of emotion recognition per spectral band condition for auditory and auditory–visual modalities. Error bars represent standard error of the mean.

resolution, and there was no discernible systematic pattern of errors occurring across spectral band conditions. Thus, increasing spectral resolution increased emotion perception across categories used in the current task.

Discussion

The goal of the present research was to ascertain the impact of spectral degradation on speech decoding and emotion recognition, and to determine whether the provision of congruent dynamic visual cues in a talker's facial expression bolsters performance in these abilities in comparison to unimodal auditory input. We assessed the impact of these factors on participants' accuracy and latency responses in tasks requiring them to report the words and identify the emotion in spoken sentences presented in auditory-only and audiovisual presentations of a female talker. The results from this study demonstrated that listeners capitalized on multimodal auditory–visual speech presentations to improve both speech and emotion recognition accuracies. While the addition of visual speech cues increased speech decoding accuracies incrementally with increasing spectral detail, the addition of visual cues bolstered emotion recognition performance to ceiling levels across all spectral degradation conditions. That is, when auditory information was compromised and made unreliable by spectral degradation, listeners relied primarily on visual prosody to achieve perfect accuracy in emotion identification. While we did not instruct participants to adopt a particular strategy, the forced choice nature of the response format likely promoted an attentional strategy wherein the participant could rely exclusively on the dynamic visual information and not on the interpretation of the semantic content of sentences to identify emotions accurately.

We speculate that response latencies were less informative on emotion recognition performance as the self-paced and forced choice nature of the task placed less working memory demands on participants. Thus, this measure may not have been sensitive to task difficulty as it was for the sentence repetition task, which demanded greater memory and cognitive resources under conditions of spectral degradation (Hudgins & Cullinan, 1978).

The current findings are consistent with those reported in a recent investigation examining adolescent and adult CI users' closed set emotion recognition from multisensory talker information (Fengler et al.,

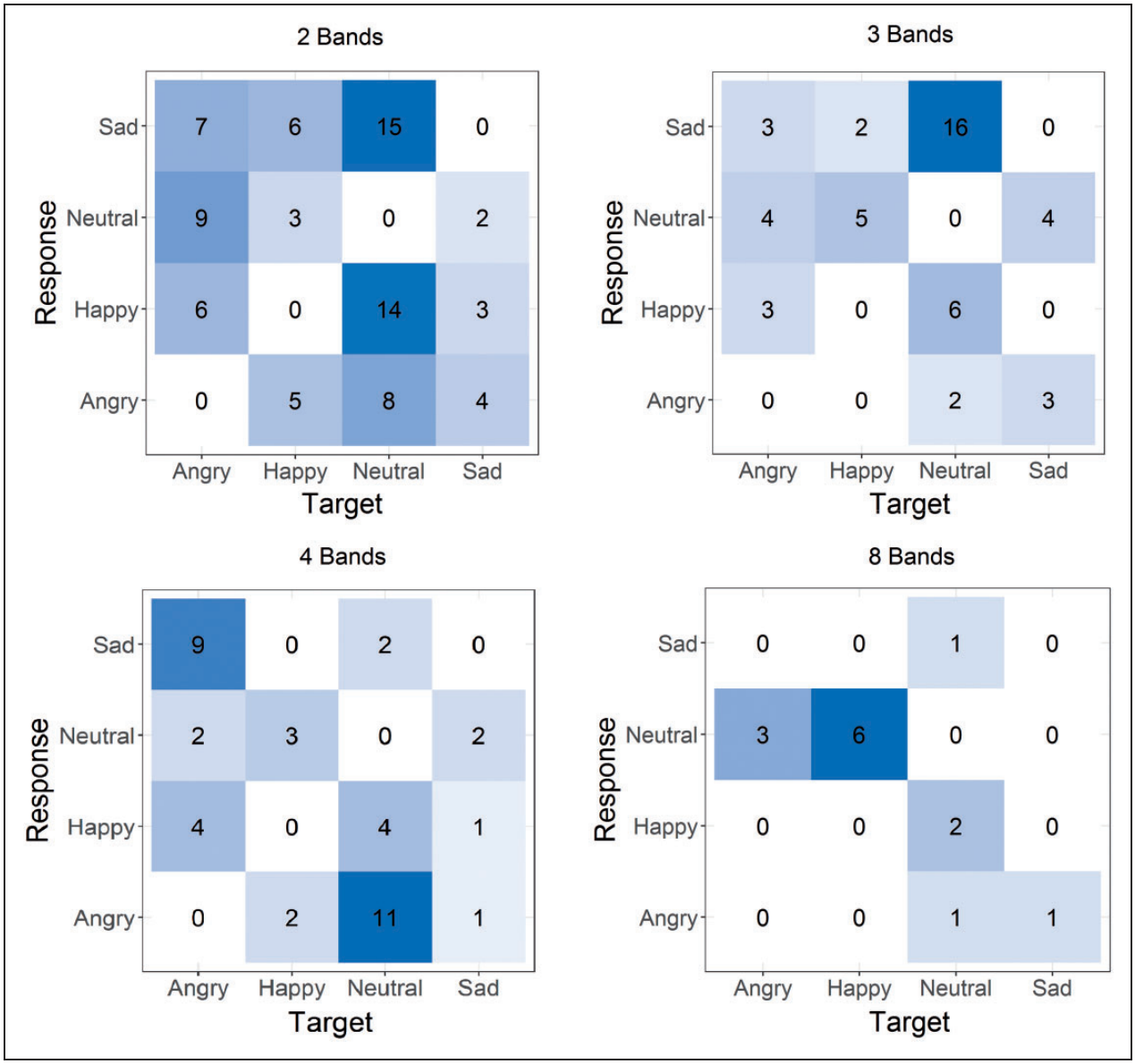

Figure 6. Confusion matrices, auditory-only emotion recognition. Darker colors represent more errors, while lighter colors represent fewer errors. The observed frequency of errors for each response–target emotion combination is reported in each cell.

2017). While CI users were less accurate than hearing controls in recognizing emotions from prosodic vocal cues alone, they relied more strongly on visual facial expressions as indicated by a cost to performance when visual facial cues were incongruent with the vocal expressive cues.

Our findings are also consistent with those observed by Fengler et al. (2017) in that the ceiling performance seen across spectral band conditions is indicative that multisensory emotion perception is dominated by the visual modality in actual CI users. That is, when the lack of spectrotemporal fine structure interferes with the perception of auditory prosodic information, observers can flexibly adapt to attend to the more reliable sensory mode presented in the task at hand. A limitation of the current study is the absence of a visual-only emotion identification condition that would enable us to ascertain the relative contribution of auditory information in the multimodal condition. However, an analysis of the distribution of scores in the auditory-only emotion recognition condition indicates that few participants were able to achieve perfect scores on the basis of degraded auditory input alone (see Figure 7), especially in the most degraded conditions. Notably, in the two-band condition, only one participant was able to achieve perfect emotion identification from auditory cues alone. Thus, the ceiling performance observed across spectral degradation conditions in auditory–visual emotion

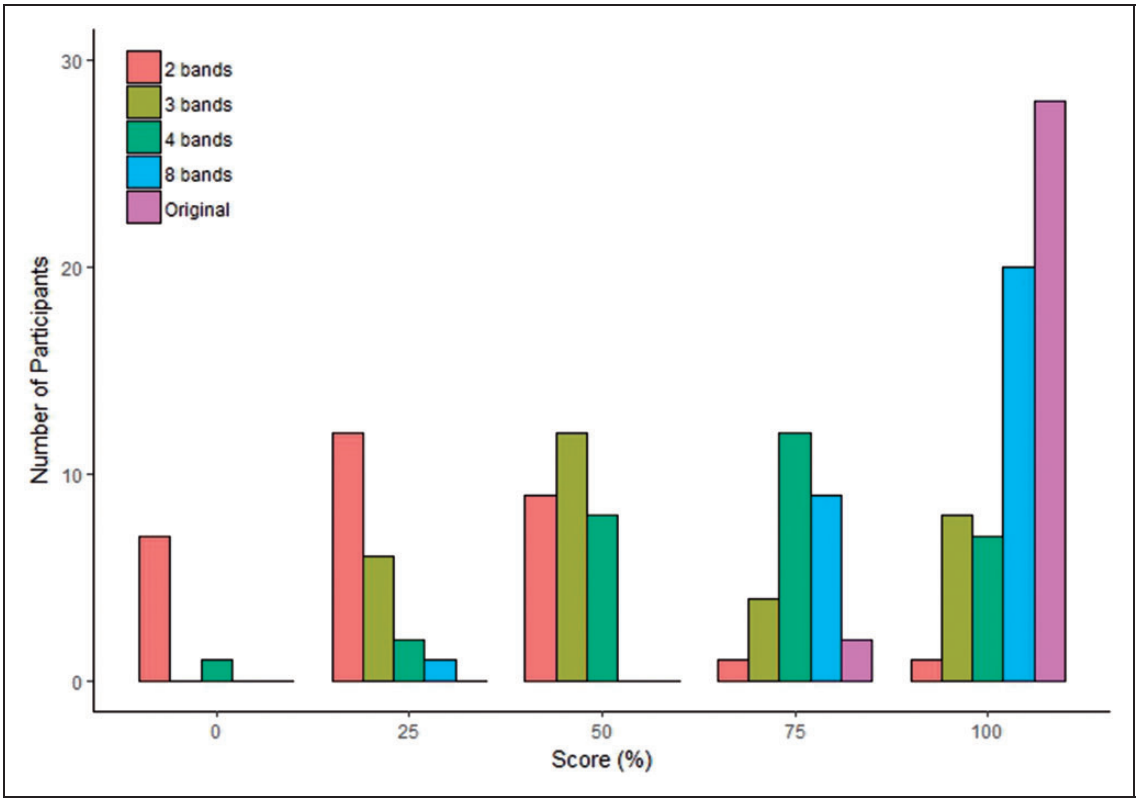

Figure 7. Number of participants per score across spectral band conditions in the auditory-only emotion recognition task.

recognition is likely attributed to informative visual cues to a greater extent, and to auditory cues to a lesser extent.

An analysis of the errors in the auditory-only condition revealed that confusions between happy, sad, and neutral emotions were more frequent under greater spectral degradation and they diminished with increasing spectral detail. The angry emotion was least confused, which is likely attributed to its contrasting and wide intensity range in comparison to the other emotions. While the frequency of errors decreased with increasing spectral resolution, systematic confusions attributed to emotional intensity or valence were not apparent across spectral degradation conditions. The poor agreement between observed and expected responses in participants' confusions patterns is indicative of this. We found, however, that informative visual cues enabled listeners to overcome all confusions occurring in spectrally degraded auditory-only presentations and identify emotions with perfect accuracy.

The importance of visual cues in emotion perception, whether provided unimodally or in combination with auditory information, have been noted previously in the performance of young CI users. For instance, Most and Aviner (2009) observed that CI users' identification of emotion from a closed set was significantly poorer in the auditory-only modality than in the visual-only and auditory–visual modalities—the latter two conditions eliciting equal effectiveness in emotion perception suggesting that auditory information added no benefit when combined with visual cues. However, it is noted that when emotional expressions are placed on neutral sentences, as those used by Most and Aviner, CI listeners may direct more attention toward the visual modality where the most informative cues are present instead of the auditory modality where the speech content has no meaningful information. By contrast, when congruent sentence content and prosodic information are matched as employed in this study, a condition is created such that reliable emotion cues are distributed across both auditory and visual modalities thereby requiring the listener to attend to both modalities. Thus, studies that employ neutral sentences (Tinnemore et al., 2018), semantically anomalous phrases (e.g., Dorman et al., 2016), or pseudowords (e.g., Fengler et al., 2017) in order to disentangle the segmental and prosodic contributions to emotion perception may underestimate the actual emotion perception abilities of CI users in everyday contexts. Paradigms employing ecologically valid linguistic stimuli, such as familiar and meaningful sentences, have shown a pattern of sensory-semantic integration that is distinct from that occurring in

experimentally derived stimuli that do not have a meaningful linguistic context. For instance, the brain's response varies distinctly across the presentation of random words, unrelated sentences, and coherent narratives with activation expanding from left-lateralized core language areas to bilateral activation reflecting a greater demand for sensory-semantic integration to process increasingly meaningful and coherent language (Xu et al., 2005). The congruency between the semantic meaning and emotion expression in a linguistic context will likely yield a more unified mental representation of speech and emotion content. Future research could determine more precisely the relative contribution of semantic and prosodic information on these mental representations by varying the congruency of these factors and observing its impact on auditory–visual integration in speech and emotion perception.

The current study's approach of using semantically meaningful sentence materials that are congruent with the intended emotion provides an important benchmark of CI simulated performance to be considered alongside those from neutral, pseudowords, and semantically neutral sentence materials. While we observed progressively better speech recognition performance in the multimodal condition than in the unimodal auditory-only condition, a striking contrast occurred in the multimodal emotion recognition task where ceiling performance was achieved across all levels spectral degradation. This result suggests that participants relied on the available visual speech and facial emotion cues to overcome degraded emotion cues presented in the auditory channel, especially in the lowest spectral band conditions. The greater importance of nonverbal emotion information (available in facial expression and vocal prosody) over verbal emotion information (sentence meaning) has been observed a study examining subjective emotion judgments and cortical responses to audiovisual recordings of sentence-length stimuli spoken by actors (Jacob et al., 2012). These audiovisual recordings captured natural vocal and facial expressions—deemed by the investigators to have high ecological validity—that were either congruent or incongruent with the sentence meaning that indicated an actor's current emotional state. The behavior index used to measure the relative importance of verbal versus nonverbal cues indicated that, for the most part, nonverbal cues played a dominant role in participants' judgments of emotions even when the semantic meaning of the verbal content contradicted these nonverbal cues. Notably, a ceiling effect in this behavior index was observed in a third of their participants, which the investigators interpreted as a clear indicator of the non-verbal cues' dominance over verbal cues in emotion judgments. These findings corroborate the occurrence of the ceiling effect observed in the current audiovisual emotion recognition task where the addition of nonverbal visual cues could have exerted a strong influence to raise performance to ceiling levels. To deter this ceiling effect, we would need to employ finer grained manipulations in the auditory–visual features or task conditions, with a possible negative consequence of reducing the external validity of our findings.

Furthermore, the greater social relevance of emotionally meaningful information confers greater cognitive advantages than those with less emotion meaning and is likely to be favorable to the performance of CI users. This is supported by neuroimaging studies that documented stronger activation for non-verbal emotional stimuli at the level of the whole brain (Jacob et al., 2012) and heightened activation involved in the processing of affective states that have social relevance (Lamm & Singer, 2010). While we have optimized our task conditions using semantically meaningful speech materials that match the intended emotional expression, future studies could vary the congruence of these cues to examine whether this ceiling effect still occurs in auditory–visual presentations.

Our report on hearing listeners' performance in decoding speech and emotion from spectrally degraded acoustic signals are consistent with observations on CI users' difficulties perceiving the auditory pitch and timing cues that cue emotion (Volkova et al., 2013; Wang et al., 2013). In addition, our reported accuracies in auditory-only speech decoding are generally consistent with those in previous reports using vocoder simulations to examine speech perception at similar levels of spectral degradation. For instance, the generally high speech decoding performance at four- and eight-band spectral resolution (i.e., 85%) observed in the current study are within the range of those reported by Loizou, Dorman, and Tu (1999) who reported speech perception accuracies as high as 90% with a five-channel vocoder simulation and asymptotic performance with eight-channel simulations. It is noteworthy that in our four-spectral band condition, speech decoding accuracies greatly outpaced emotion recognition accuracy at the same level even when task conditions were optimized in a forced choice response format. While these outcomes are encouraging, these speech perception scores are still well below those achieved in their intact acoustic form. Given the vocoder frequency range spans 300 to 6000 Hz and that our female talker's average voice pitch falls just below the lower limit, it is likely that the weak voice pitch cues conveyed in the temporal envelope of CI hearing figured less prominently in listeners' perception of emotion (Green, Faulkner, & Rosen, 2004). Instead, speech rate and intensity change are more reliable and secondary prosodic cues that differentiate emotions when the signal is degraded (Huang, Newman, Catalano, & Goupell, 2017). In the current study, happy is spoken at a faster rate in comparison to neutral, angry, and sad, which were spoken at slower rates. The angry

emotion spanned a wider intensity range than the remaining three emotions (see Table 2).

Limitations of CI users' performance in more adverse real-world listening conditions will likely require training strategies that capitalize on informative cues from other sensory modalities. For instance, in the lowest spectral resolution conditions, our results indicate that speech decoding on the basis of auditory temporal cues alone is poor. With the addition of visual speech cues, however, performance accuracies increased considerably from near zero to over 50% accuracy in the two-band condition. While the multimodal advantage was seen across all levels of spectral degradation, the greatest gains were observed in conditions with the poorest spectral resolution.

Enhancing the communication outcomes of CI users through multimodal means has considerable implications given the documented challenges that extend generally to speech communication and emotion reasoning. For instance, difficulties with emotion understanding are associated with greater risk for developing symptoms of psychopathology or poor social functioning (Eisenberg, Spinrad, & Eggum, 2010). Because emotion understanding is a multidimensional percept involving the integration of sensory and linguistic information, hearing loss may place individuals at risk for developing ineffective skills to decode the expression of discrete emotions that are critical for effective communication. Visual input alone is ineffective for a comprehensive understanding of emotion as there are some indications that young CI children have difficulties discriminating facial expressive cues from static images, which may be exacerbated by late auditory input (Wiefferink, Rieffe, Ketelaar, De Raeve, & Frinjns, 2013).

Results from the current vocoder simulations demonstrating gains from multisensory input are consistent with similarly documented effects in candidate and current CI users. For instance, there are indications that visual skill in speech reading can form the foundation for good speech outcomes following cochlear implantation in prelingually deaf children. In addition, speechreading and speech perception of audiovisual information prior to implantation are reliable predictors of postimplantation success in speech and language development (Bergeson & Pisoni, 2004; Bergeson, Pisoni, & Davis, 2003; Bergeson et al., 2005). Further, spoken language accompanied by signing has been found to improve CI users' speech recognition, comprehension, and learning in a college classroom setting (Blom, Marschark, & Machmer, 2017).

While vocoder CI simulations have been used previously to examine auditory training effects and signal processing (Bernstein, Demorest, Coulter, & O'Connell, 1991) and to examine speech feature manipulation (Li & Fu, 2007), the current findings extend the scope of these simulations by identifying some parameters in which multimodal auditory–visual integration can improve speech and emotion perception under conditions of poor spectral resolution. They can inform new directions in rehabilitation schemes that enable CI listeners to capitalize on multimodal cues early on in order to accelerate acclimatization to their devices and optimize auditory learning following implantation.

Declaration of Conflicting Interests

The author(s) declared no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research study was supported by a SSHRC Insight Development Grant and a Project Grant from MacEwan University's Research Office to T. Vongpaisal.

Supplemental Material

Supplemental material for this article is available online.

References

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*, 190–201. doi: 10.1037//0012-1649.36.2.190

Bergeson, T. R., & Pisoni, D. B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 749–772). Cambridge: MIT Press.

Bergeson, T. R., Pisoni, D. B., & Davis, R. A. (2003). A longitudinal study of audiovisual speech perception by children with hearing loss who have cochlear implants. *Volta Review*, *103*, 347–370.

Bergeson, T. R., Pisoni, D. B., & Davis, R. A. O. (2005). Development of audiovisual comprehension skills in prelingually deaf children with cochlear implants. *Ear and Hearing*, *26*, 149–164.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & O'Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing impaired subjects. *The Journal of the Acoustical Society of America*, *90*, 2971–2984.

Blom, H., Marschark, M., & Machmer, E. (2017). Simultaneous communication supports learning in noise by cochlear implant users. *Cochlear Implants International*, *18*, 49–56. doi: 10.1080/14670100.2016.1265188

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychology*, *44*, 209–220. doi: 10.1002/dev.20032

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., . . . Christensen, J. A. (2015). Voice emotion recognition by cochlear-implanted children

and their normally-hearing peers. *Hearing Research*, *332*, 151–162. doi: 10.1016/j.heares.2014.10.003

Cleary, M., Pisoni, D. B., & Kirk, K. I. (2005). Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *48*, 204–223.

Congalton, R. G., & Green, K. (2009). *Assessing the accuracy of remotely sensed data: Principles and practices, Second edition*. Boca Raton, FL: CRC Press.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology*, *134*, 22–241. doi: 10.1037/0096-3445.134.2.222

Dorman, M. F., Liss, J., Wang, S., Berisha, V., Ludwig, C., & Natale, S. C. (2016). Experiments on auditory-visual perception of sentences by users of unilateral, bimodal, and bilateral cochlear implants. *Journal of Speech, Language, and Hearing Research*, *59*, 1505–1519. doi: 10.1044/2016_JSLHR-H-15-0312

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., & Boothroyd, A. (2000). Speech recognition with reduced spectral cues as a function of age. *Journal of the Acoustical Society of America*, *107*, 2704–2710.

Eisenberg, N., Spinrad, T. L., & Eggum, N. D. (2010). Emotion-related self-regulation and its relation to children's maladjustment. *Annual Review of Clinical Psychology*, *6*, 495–525. doi: 10.1146/annurev.clinpsy.121208.131208

Fengler, I., Nava, E., Villwock, A. K., Büchner, A., Lenarz, T., Röder, B. (2017). Multisensory emotion perception in congenitally, early, and late deaf CI users. *PLoS One*, *12*(10), e0185821doi: 10.1371/journal.pone.0185821

Flom, R., & Bahrick, L. E. (2010). The effects of intersensory redundancy on attention and memory: Infants' long-term memory for orientation in audiovisual events. *Developmental Psychology*, *46*, 428–436. doi: 10.1037/a0018410

Gilbers, S., Fuller, C., Gilbers, D., Broersma, M., Goudbeek, M., Free, R., & Başkent, D. (2015). Normal-hearing listeners and cochlear implants users perception of pitch cues in emotional speech. *i-Perception*, *6*, 1–19. doi: 10.1177/0301006615599139

Geers, A. E. (2004). Speech, language, and reading skills after early cochlear implantation. *Archives of Otolaryngology—Head & Neck Surgery*, *130*, 634–638. doi: 10.1001/archotol.130.5.634

Green, T., Faulkner, A., & Rosen, S. (2014). Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. *The Journal of the Acoustical Society of America*, *116*, 2298doi: 10.1121/1.1785611

Hochmair-Desoyer, I. J., Hochmair, E. S., Fischer, R. E., & Burian, K. (1980). Cochlear prostheses in use: Recent speech comprehension results. *Archives of Otorhinolaryngology*, *229*, 81–98.

Holt, R. F., Kirk, K. I., & Hay-McCutcheon, M. (2011). Assessing multimodal spoken word-in-sentence recognition in children with normal hearing and children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *54*, 632–657. doi: 10.1044/1092-4388(2010/09-0148)

Huang, Y. T., Newman, R., Catalano, A., & Goupell, M. J. (2017). Using prosody to infer discourse status in cochlear-implant and normal-hearing listeners. *Cognition*, *166*, 184–200.

Hudgins, J. C., & Cullinan, W. L. (1978). Effects of sentence structure on sentence elicited imitation responses. *Journal of Speech and Hearing Research*, *21*, 809–819.

Jacob, H., Kreifelts, B., Brück, C., Erb, M., Hösl, F., & Wildgruber, D. (2012). Cerebral integration of verbal and nonverbal emotional cues: Impact of individual nonverbal dominance. *Neuroimage*, *61*, 738–747. doi: 10.1016/j.neuroimage.2012.03.085

Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: A new multimodal picture-word task. *Journal of Experimental Child Psychology*, *102*, 40–59. doi: 10.1016/j.jecp.2008.08.002

Jiam, N. T., Caldwell, M., Deroche, M. L., Chatterjee, M., & Limb, C. J. (2017). Voice emotion perception and production in cochlear implant users. *Hearing Research*, *352*, 30–39.

Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Structure and Function*, *214*, 579–591.

Li, T., & Fu, Q. J. (2007). Perceptual adaptation to spectrally shifted vowels: Training with nonlexical labels. *Journal of the Association for Research in Otolaryngology*, *8*, 32–41.

Luo, X., Fu, Q. J., & Galvin, J. J. (2007). Vocal emotion recognition by normal-hearing listeners and cochlear implant users. *Trends in Amplification*, *11*, 301–215. doi: 10.1177/1084713807305301

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *The Journal of the Acoustical Society of America*, *106*, 2097–2103.

Maidment, D. W., Kang, H. J., Stewart, H. J., & Amitay, S. (2015). Audiovisual integration in children listening to spectrally degraded speech. *Journal of Speech, Language, and Hearing Research*, *58*, 61–68. doi: 10.1044/2014_JSLHR-S-14-0044

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. doi: 10.1038/264746a0

Mildner, V., & Koska, T. (2014). Recognition and production of emotions in children with cochlear implants. *Clinical Linguistics & Phonetics*, *28*, 543–554. doi: 10.3109/02699206.2014.927000

Most, T., & Aviner, C. (2009). Auditory, visual, and auditory-visual perception of emotions by individual with cochlear implants, hearing aids, and normal hearing. *Journal of Deaf Studies and Deaf Education*, *14*, 449–464.

Most, T., & Michaelis, H. (2012). Auditory, visual, and auditory-visual perceptions of emotions by young children with hearing loss versus children with normal hearing. *Journal of Speech, Language, and Hearing Research*, *55*, 1148–1162. doi: 10.1044/10924388(2011/11-0060)

Nagels, A., Kauschke, C., Schrauf, J., Whitney, C., Straube, B., & Kircher, T. (2013). Neural substrates of figurative language during natural speech perception: An fMRI

study. *Frontiers in Behavioral Neuroscience*, *7*. doi: 10.3389/fnbeh.2013.00121

Pals, C., Sarampalis, A., & Başkent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech, Language, and Hearing Research*, *56*, 1075–1084. doi:10.1044/1092-4388(2012/12-0074)

Pals, C., Sarampalis, A., van Rijn, H., & Başkent, D. (2015). Validation of a simple response time measure of listening effort. *The Journal of the Acoustical Society of America*, *138*, 187–192. doi: 10.1121/1.4929614

Robbins, A. M., Renshaw, J. J., & Osberger, M. J. (1995). *Common Phrases Test*. Indianapolis: Indiana University School of Medicine.

Shannon, R. V., Fu, Q. J., & Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngologica*, *552*, 50–54.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear & Hearing*, *26*, 263–275. doi: 0196/0202/05/2603-0263/0

Sumby, W. H., & Pollack, L. (1954). Visual contribution of speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212–215. doi: 10.1121/1.1907309

Stacey, P. C., Kitterick, P. T., Morris, S. D., & Sumner, C. J. (2016). The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure. *Hearing Research*, *336*, 17–28. doi:10.1016/j.heares.2016.04.002

Strelnikov, K., Rouger, J., Demonet, J. F., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (2013). Visual activity predicts auditory recovery from deafness after adult cochlear implantation. *Brain*, *136*, 3682–3695. doi: 10.1093/brain/awt274

Stropahl, M., Schellhardt, S., & Debener, S. (2017). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: The Oldenburg Audio Visual Speech Stimuli (OLAVS). *Psychonomic Bulleting & Review*, *24*, 863–872. doi: 10.3758/s13423 016-1148-9

Studebaker, G. A. (1985). A ''rationalized'' arcsine transform. *Journal of Speech and Hearing Research*, *28*, 455–462.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological Science*, *11*, 153–158.

Tona, R., Naito, Y., Moroto, S., Yamamoto, R., Fujiwara, K., Yamazaki, H., . . . Kikuchi, M. (2015). Audio–visual integration during speech perception in prelingually deafened Japanese children revealed by the McGurk effect. *International Journal of Pediatric Otorhinolaryngology*, *79*, 2072–2078. doi: 10.1016/j.ijporl.2015.09.016

Tinnemore, A. R., Zion, D. J., Kulkarni, A. M., & Chatterjee, M. (2018). Children's recognition of emotional prosody in spectrally degraded speech is predicted by their age and cognitive status. *Ear & Hearing*, *39*(5), 874–880.

Valkenier, B., Duyne, J. Y., Andringa, T. C., & Baskent, D. (2012). Audiovisual perception of congruent and incongruent Dutch front vowels. *Journal of Speech Language and Hearing Research*, *55*, 1788–1801. doi: 10.1044/1092-4388(2012/11-0227)

van Heugten, M., Volkova, A., Trehub, S. E., & Schellenberg, G. (2013). Children's recognition of spectrally degraded cartoon voices. *Ear & Hearing*, *35*, 118–125. doi: 10.1097/AUD.0b013e3182a468d0

Volkova, A., Trehub, S. E., Schellenberg, E. G., Papsin, B. C., & Gordon, K. A. (2013). Children with bilateral cochlear implants identify emotion in speech and music. *Cochlear Implants International*, *14*(2), 80–90. doi: 10.1179/1754762812Y.0000000004

Vongpaisal, T., Trehub, S. E., & Schellenberg, E. G. (2006). Song recognition by children and adolescents with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *49*, 1091–1103.

Vongpaisal, T., Trehub, S. E., Schellenberg, E. G., & van Lieshout, P. (2012). Age-related changes in talker recognition with reduced spectral cues. *Journal of the Acoustical Society of America*, *131*, 501–508. doi: 10.1121/1.3669978

Wang, D. J., Trehub, S. E., Volkova, A., & van Lieshout, P. (2013). Child implant users' imitation of happy-and sad-sounding speech. *Frontiers in Psychology*, *4*, 1–8. doi: 10.3389/fpsyg.2013.00351

Weisberg, J., Hubbard, A. L., & Emmorey, K. (2017). Multimodal integration of spontaneously produced presentational co-speech gestures: An fMRI study. *Language, Cognition, and Neuroscience*, *32*, 158–174. doi: 10.1080/23273798.2016.1245426

Wiefferink, C. H., Rieffe, C., Ketelaar, L., De Raeve, L., & Frijns, J. H. (2013). Emotion understanding in deaf children with a cochlear implant. *Journal of Deaf Studies and Deaf Education*, *18*, 175–186. doi: 10.1093/deafed/ens042

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: Emergent features of word, sentence, and narrative comprehension. *Neuroimage*, *25*, 1002–1015. doi:10.1016/j.neuroimage.2004.12.013.