Children's Recognition of Spectrally Degraded Cartoon Voices

Marieke van Heugten,[1] Anna Volkova,[2] Sandra E. Trehub,[2] and E. Glenn Schellenberg[2]

Objectives: Although the spectrally degraded input provided by cochlear implants (CIs) is sufficient for speech perception in quiet, it poses problems for talker identification. The present study examined the ability of normally hearing (NH) children and child CI users to recognize cartoon voices while listening to spectrally degraded speech.

Design: In Experiment 1, 5- to 6-year-old NH children were required to identify familiar cartoon characters in a three-alternative, forced-choice task without feedback. Children heard sentence-length utterances at six levels of spectral degradation (noise-vocoded utterances with 4, 8, 12, 16, and 24 frequency bands and the original or unprocessed stimuli). In Experiment 2, child CI users 4 to 7 years of age and a control sample of 4- to 5-year-old NH children were required to identify the unprocessed stimuli from Experiment 1.

Results: NH children in Experiment 1 identified the voices significantly above chance levels, and they performed more accurately with increasing spectral information. Practice with stimuli that had greater spectral information facilitated performance on subsequent stimuli with lesser spectral information. In Experiment 2, child CI users successfully recognized the cartoon voices with slightly lower accuracy (0.90 proportion correct) than NH peers who listened to unprocessed utterances (0.97 proportion correct).

Conclusions: The findings indicate that both NH children and child CI users can identify cartoon voices under conditions of severe spectral degradation. In such circumstances, children may rely on talker-specific phonetic detail to distinguish one talker from another.

Key words: Talker identification, Children, Cochlear implants, Degraded speech.

(Ear & Hearing 2014;35;118–125)

INTRODUCTION

Successful interpretation of verbal messages depends on the integration of linguistic and talker-specific cues. For example, the sentence, "I borrowed my parents' car," has very different implications when spoken by an 18-year-old neighbor or his 15-year-old sibling. In more commonly occurring circumstances, listeners treat information differently depending upon whether the source of that information is a recognized authority or novice. Accordingly, understanding voice identification can enhance our understanding of speech comprehension more generally. There has been considerable research on adults' recognition of voices (e.g., Van Lancker & Kreiman 1987; Loebach et al. 2008) but despite its importance for speech perception, there is very limited information about voice recognition in children. In the present article, we ask whether children can identify familiar voices under conditions of spectral degradation. In Experiment 1, we evaluate the ability of children with normal hearing (NH) to identify familiar voices from auditory signals that are spectrally degraded by means of noise vocoding. In Experiment 2, we evaluate comparable abilities in deaf children who use cochlear implants (CIs), which provide them with spectrally degraded signals. These experiments, taken together, provide new insights into children's representation of familiar voices.

Speech conveys information about the speaker's physical (e.g., age, sex), psychological (e.g., mood, stress level), and social (e.g., education, regional origin; see Kreiman et al. 2005) characteristics. Such indexical information contributes to rapid impressions of unknown talkers, underlies the identification of familiar talkers, and facilitates speech perception. Adults generally experience little difficulty identifying familiar talkers under optimal conditions (e.g., quiet), especially when response alternatives are provided (Van Lancker et al. 1985). In less than optimal circumstances, however, talker familiarity facilitates speech perception, both for adults (Nygaard et al. 1994; Nygaard & Pisoni 1998; Clarke & Garrett, 2004; Bradlow & Bent 2008; Maye et al. 2008) and young children (White & Aslin 2011; Schmale et al. 2012; Van Heugten & Johnson Reference Note 1).

Attention to talker-specific cues is evident early in development. Within days of birth, infants recognize their mother's voice (DeCasper & Fifer 1980; Spence & Freeman 1996). By 7 months of age, they can segregate two women's voices but only when one of the voices is familiar (Barker & Newman 2004). They also differentiate talkers who speak a familiar language but not those who speak an unfamiliar language (Johnson et al. 2011). By the preschool period, children recognize the voices of their kindergarten classmates (Bartholomeus 1973). They also identify the voices of familiar cartoon characters in a six-alternative forced-choice task, with 4- and 5-year olds performing better (81% and 86% correct, respectively) than 3-year olds (61%; Spence et al. 2002). Moreover, children 3 to 6 years of age can learn to identify two previously unfamiliar talkers in the course of a laboratory test session, but their accuracy remains modest even after 32 training trials (Creel & Jimenez 2012).

Despite the ubiquity of talker recognition, there has been little success in identifying the critical cues underlying this ability (for a review, see Kreiman et al. 2005). Pitch and timbre (i.e., voice quality) are undoubtedly relevant to talker identification (Van Lancker & Kreiman 1987), but familiar talkers can be identified to some extent from sine-wave analogs of speech that lack timbre and fundamental frequency cues, presumably by means of talker-specific timing of consonant and vowel articulation (Fellowes et al. 1997; Remez et al. 1997; Sheffert et al. 2002). Adults can also identify sex of talker from vocoded speech that simulates the spectral degradation of cochlear implants (CIs) (Fu et al. 2004; Vongphoe & Zeng 2005), but they have difficulty differentiating same-sex voices in the context of such degradation (Fu et al. 2004).

Despite the spectrally degraded information provided by cochlear prostheses, CI users often achieve excellent levels of speech recognition in quiet (e.g., Vongphoe & Zeng 2005), and

[1] Laboratoire de Sciences Cognitives et Psycholinguistique, CNRS/EHESS/DEC-ENS, Paris, France; and [2] Department of Psychology, University of Toronto, Mississauga, ON, Canada.

0196/0202/14/3501-0118/0 • Ear and Hearing • Copyright © 2013 by Lippincott Williams & Wilkins • Printed in the U.S.A.

many acquire the spoken language of their community from electrical input alone (Svirsky et al. 2000). Spectral degradation, however, poses substantial difficulties for *talker differentiation*, especially for children (Cleary & Pisoni 2002; Vongphoe & Zeng 2005; Vongpaisal et al. 2012). For example, school-age CI users with 4 or more years of implant experience achieve only 57% correct on a talker differentiation task with variable linguistic content within and across talkers (Cleary & Pisoni 2002). At times, however, children's difficulty is probably exacerbated by laboratory stimuli or tasks that are insufficiently engaging. For example, NH infants sometimes perform well on speech perception tasks with natural, sentence-length utterances produced by parents of same-age infants (Van Heugten & Johnson 2012), but they often experience difficulty on similar tasks with isolated syllables produced by actors (Houston & Jusczyk 2000). The facilitative effects of engaging tasks have also been observed under conditions of spectral degradation. For example, child CI users can differentiate their mother's natural-sounding utterances from those of other female talkers in a game-like task with closed-set responding (Vongpaisal et al. 2010). In a similar task with three response alternatives (chance level of 33%), NH children classify the unfamiliar talker of noise-vocoded utterances as a man, woman, or child with over 60% accuracy (Vongpaisal et al. 2012). In the context of limited spectral information, children are likely to capitalize on individual differences in consonant and vowel articulation and speaking rate, especially when sentence-length stimuli are provided (Vongpaisal et al. 2010, 2012).

Noise-vocoded speech was first introduced by Shannon et al. (1995) to simulate the coding of speech in CIs and to demonstrate that excellent speech recognition is possible under conditions of severe spectral degradation. This vocoding scheme involves filtering speech into a number of frequency bands, using the amplitude envelope of each band to modulate Gaussian noise, and summing the modulated bands. The resulting signal preserves temporal and amplitude cues but no spectral detail within each band. Similar speech intelligibility is achieved with sine-vocoded speech, which uses amplitude-modulated sine waves rather than noise bands (Dorman et al. 1997). Both types of vocoding have become common means of examining the influence of various spectral and temporal parameters on speech intelligibility (e.g., Shannon et al. 1995; Faulkner et al. 2000) and voice discrimination (e.g., Fu et al. 2004).

Practice with vocoded speech increases the subsequent comprehensibility of such speech (e.g., Davis et al. 2005; Hervais-Adelman et al. 2008, 2011). Adaptation is relatively rapid, especially when feedback is provided or when listeners have the opportunity of hearing clear versions before hearing the distorted ones (Davis et al. 2005; Hervais-Adelman et al. 2008). Even without prior exposure to noise-vocoded speech, NH children can decode sentence-length utterances at levels of spectral degradation that pose difficulty for identifying talker age or sex (Vongpaisal et al. 2012; Newman & Chatterjee 2013). Obviously, NH children do not have long-term exposure to spectrally degraded input, as is the case for child CI users, so their performance in a single test session may underestimate the performance that can be achieved after additional exposure or training.

In principle, the speech perception benefits attributable to talker familiarity could be available to child CI users, as they are to NH children (White & Aslin 2011; Schmale et al. 2012; Van Heugten & Johnson Reference Note 1). To date, however, studies of familiar talker identification by child CI users have been restricted to the maternal voice (Vongpaisal et al. 2010), which is more familiar than any other voice. Moreover, there has been no evaluation of NH children's identification of familiar talkers from samples of noise-vocoded speech. Here we ask whether NH children can identify the voices of familiar cartoon characters under conditions of spectral degradation and whether child CI users can do so under their usual conditions of listening. Cartoon voices are of particular interest and relevance to young children, constituting a suitable test case of the recognition of voices that are familiar but much less familiar than the voice of the mother or other immediate family members. Cartoon voices are designed to be highly distinctive and engaging, and children often have enduring, affectionate ties with the characters. In the present study, we examined the ability of NH children to identify the voices of familiar cartoon characters at several levels of spectral degradation (Experiment 1) and the ability of young CI users to identify the same voices (Experiment 2).

EXPERIMENT 1

The goal of the present experiment was to examine the impact of spectral degradation on the identification of cartoon voices by 5- to 6-year-old children with normal hearing. Preschool children recognize cartoon voices in the context of a six-alternative, forced-choice task and ideal listening conditions (Spence et al. 2002). The focus here was on children's identification of such voices from primarily temporal cues that are available in noise-vocoded sentences. To maintain children's interest in the spectrally degraded stimuli, the voice identification task was designed as an engaging game. After hearing each utterance, children responded by selecting one of three colorful images of different cartoon characters on a touch-sensitive monitor. The availability of three alternatives rather than six (Spence et al. 2002) reduced task difficulty, a change that was warranted by the degraded stimuli. In fact, the use of two or three response alternatives is common in research with young children (Morton & Trehub 2001; Volkova et al. 2013), including research on talker recognition (Creel & Jimenez 2012; Vongpaisal et al. 2012). Even on simple tasks with only two alternatives, young children can be distracted by irrelevant cues (e.g., utterance content) and often perseverate on unsuccessful response strategies (Morton & Trehub 2001; Morton et al. 2003). Children were expected to perform at or near ceiling on intact versions, but their performance was expected to decrease progressively with increasing levels of spectral degradation.

Because NH children have no experience with spectrally degraded speech, exposure to less-degraded versions is likely to assist them subsequently on more degraded versions, as it does for adults (Hervais-Adelman et al. 2008). A secondary manipulation involving the presentation of utterances examined this possibility. The order of presentation of stimuli at various levels of degradation was randomized in one condition and blocked in another condition by level of degradation, proceeding from greatest to least degradation. The highly degraded stimuli in early trials of the blocked presentation would provide little opportunity for transfer of training (e.g., Hervais-Adelman et al. 2011). By contrast, exposure to some stimuli with lesser degradation in early trials of the randomized version could facilitate adaptation to the mode of distortion or enhance subsequent performance by prior exposure to utterance content (Church & Fisher 1998). Unpredictable distortion in the randomized condition could also have negative consequences, especially for children.

Participants and Methods

Participants • The participants were 24 NH children who were 5 to 6 years of age (M = 6.1 years, SD = 0.3 years) from the Greater Toronto Area. All children were estimated to be from middle- or upper-income families, as determined from census data linked to their residential address. The children were native speakers of English and had no family or personal history of hearing problems, according to parental report. They regularly watched television (TV) programs featuring at least three cartoon characters from the present stimulus set. An additional 4 children were excluded from the sample because of equipment error (2) or failure to complete the test session (2).

Apparatus and Stimuli • All testing was conducted in a double-walled sound-attenuating booth (Industrial Acoustics Co., Bronx, NY), with the child seated facing a touch-screen monitor (ELO LCD Touch Systems, Milpitas, CA). Auditory stimuli, consisting of utterances from cartoon characters, were presented by means of an amplifier (Harmon-Kardon HK3380, Woodbury, NY) and two loudspeakers (Electro-Medical Instruments, Mississauga, ON, Canada), each located at a 45-degree angle from the participant. Visual stimuli consisting of brightly colored images of the cartoon characters were presented on the monitor. Stimulus delivery and response recording were controlled by a custom program on a Windows XP workstation outside the booth. The auditory stimulus set consisted of utterances from 11 cartoon characters in popular TV shows for young children (see Table 1). For each cartoon character, five noise-free utterances were selected from TV episodes. Because the content varied across characters, utterances with content cues to the identity of the character (i.e., stereotyped expressions) were avoided.

The noise-band vocoder used in the present study (Sheldon et al. 2008) was implemented as described by Shannon et al. (1995) and Eisenberg et al. (2000). Digital sound files were passed through a series of bandpass filters to create conditions with 4, 8, 12, 16, or 24 frequency bands spanning a frequency range of 300 to 6000 Hz. The temporal envelope of the signal was extracted by means of the Hilbert transform and used to modulate narrowband Gaussian white noise. These filtered bands were then added to generate auditory signals that preserved the original temporal envelope and amplitude profile without the original fine structure. Each child received 90 trials, consisting of five utterances presented at six levels of degradation (vocoded stimuli with 4, 8, 12, 16, and 24 frequency bands and the original unprocessed stimuli) for each of three characters. In other words, the same 15 utterances were presented at all levels of degradation. All stimuli were equated for amplitude and presented at approximately 70 dB (A level).

Table 1 shows mean syllable duration (a rough index of speaking rate), mean sentence duration (and range of sentence duration), and fundamental frequency (f0; and f0 range) for cartoon characters selected by children. It is apparent that Dora and Elmo speak considerably more slowly and use a wider f0 range than most of the other characters. Overall, the relatively slow speaking rate, high pitch, and large pitch range relative to typical adult-directed speech corroborate the child-directed speech style (e.g., Fernald et al. 1989; Creel & Jimenez 2012) of these cartoon characters.

Procedure • Children were tested in a three-alternative, forced-choice task. First, they selected the three most familiar cartoon characters from the set of 12, and parents confirmed the familiarity of their choices. Then the child and experimenter entered the test booth. The experimenter sat behind the child

TABLE 1. Mean syllable duration and fundamental frequency (f0) of the cartoon characters selected by children

TV Program	Cartoon Character	Mean Syllable Duration (msec)	Mean Utterance Duration (range, sec)	Mean f0 (range, Hz)	Sample Utterance
Dora the Explorer	Dora	536	1.90 (1.24–2.62)	384 (180–819)	I miss him so much.
Dora the Explorer	Boots	453	1.92 (1.55–2.13)	287 (225–699)	That wouldn't be too good.
SpongeBob	SpongeBob	259	1.71 (1.07–2.11)	196 (117–413)	But it doesn't make any sense.
SpongeBob	Patrick	245	1.50 (0.84–2.02)	226 (125–395)	I thought I was doing a pretty good job.
SpongeBob	Squidward	197	1.82 (1.37–2.76)	210 (93–285)	Now go spread the word.
SpongeBob	Sandy	389	1.32 (0.92–1.63)	398 (175–632)	That's gotta hurt.
Bob the Builder	Bob	236	1.52 (1.05–2.17)	276 (126–603)	I can't wait to see it.
Bob the Builder	Wendy	316	2.18 (1.77–2.75)	333 (178–498)	We had a bit of a slow start.
Sesame Street	Elmo	462	1.59 (1.03–2.84)	495 (194–684)	Tell us about yourself.
Sesame Street	Ernie	279	1.38 (1.11–1.74)	316 (186–418)	These are all our friends out there.
Max and Ruby	Ruby	305	2.01 (1.58–2.63)	346 (172–574)	He's just getting dressed.

f0 was measured from the middle 50% of six vowels that occurred at least once for each character (a, æ, ε, ʌ, ɪ, and i). Verbal content of sample utterances is provided.

to discourage interaction during the task. Children were told that they would see pictures of the three cartoon characters on the monitor and that they would hear them talk. They were also told that the characters would sometimes sound funny or hard to understand. Their task was to indicate who was talking by touching the corresponding picture on the monitor. Children were told that they could ask to have an utterance repeated. After the children expressed their understanding of the instructions, the three characters appeared on the screen and the presentation of auditory stimuli began. Touch responses were recorded only after the entire utterance was presented. Some children required occasional encouragement (e.g., *You're really trying hard. That's great!*), but they received no feedback regarding response accuracy. After each response, the experimenter or the child pressed a button to proceed to the next trial. Testing continued until all 90 trials were completed.

Design • Children were randomly assigned to one of two conditions: (1) the 90 utterances in random order or (2) blocked presentation, with each block consisting of utterances from the three characters at the same level of degradation, and the order of blocks proceeding from the highest level of degradation (utterances with 4 frequency bands) to lesser and lesser degradation, ending with the intact or unprocessed utterances. Within each block, the order of utterances was randomized. The same utterance never occurred on successive trials in either condition. A separate random order was generated for each participant.

Results and Discussion

Figure 1 (left panel) shows the proportion of correct responses in the random and blocked orders for each of the six levels of degradation. Performance was at or near ceiling for the unprocessed stimuli, with only 3 children making errors in this condition. Two children (1 in the random order, 1 in the blocked order) selected an incorrect cartoon character on one of the 15 trials. The third child (blocked order) selected an incorrect character twice. Because of the lack of variability in this condition, data from the intact stimuli were excluded from the between-group comparisons. We used a two-way mixed-design analysis of variance on the remaining conditions, with Order (random, blocked) as a between-subjects factor and Degradation (4, 8, 12, 16, 24 bands) as a within-subjects factor. All analyses were conducted with SPSS. Because sphericity was violated ($\chi^2(9) = 24.31$; $p = 0.004$), the degrees of freedom for within-subjects comparisons were corrected with Greenhouse–Geisser estimates. The analyses revealed a main effect of Order ($F(1, 22) = 18.86$, $p < 0.001$), indicating better performance in the random order (M = 0.81, SD = 0.11) than in the blocked order (M = 0.63, SD = 0.08). There was also a main effect of Degradation ($F(2.7, 60.2) = 35.73$, $p < 0.001$), indicating better performance with lesser levels of degradation. These effects were qualified by an interaction between Order and Degradation ($F(2.7, 60.2) = 4.99$, $p = 0.005$), stemming from an advantage for the random order at eight bands ($t(22) = 4.52$, $p < 0.001$) and 12 bands ($t(22) = 4.53$, $p < 0.001$), but not at 16 bands ($t(22) = 2.05$, $p = 0.052$) and 24 bands ($p = 0.384$). The difference between orders at four bands was marginal ($t(22) = 2.76$, $p = 0.012$; all comparisons used a Bonferroni-corrected α-level of 0.01).

We also evaluated the levels of degradation at which children in random and blocked conditions identified the voices at better than chance levels (0.33 proportion correct). Performance exceeded chance at all levels of spectral degradation in both the random and blocked conditions, as revealed by one-sample *t* tests and nonparametric, one-sample Wilcoxon signed-rank tests with a Bonferroni-corrected α-level of 0.0083 (all $p \leq 0.006$, one-tailed). Thus, despite the fact that greater spectral degradation reduced children's accuracy of identifying cartoon voices, the greatest levels of spectral degradation did not prevent children from recognizing the voices. Finally, the variable selection of cartoon characters across children precluded specification of the cues used to distinguish one character from another. However, because several children (17 of 24) selected Dora as one of their three characters, and Dora's speaking rate and f0 range were distinctly different from those of many other characters, we asked whether those cues facilitated identification at the highest level of spectral degradation. This did not appear to be the case. Identification of Dora (average recognition scores of 0.36 in the blocked order and 0.58 in the random order) was no better than that of other characters. As can be seen in Table 1, moreover, within-talker sentence duration was highly

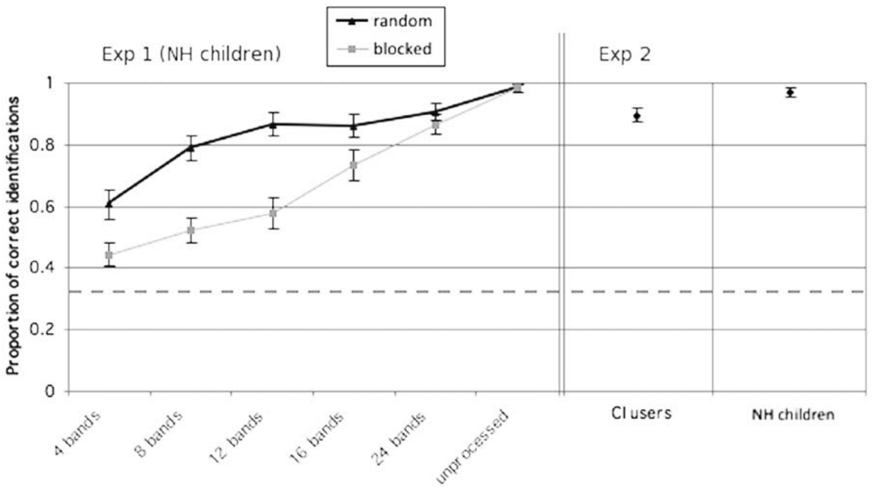

Fig. 1. Proportion of correct responses as a function of number of spectral bands for NH children in Experiment 1 (left panel) and proportion of correct responses for the CI users (middle panel) and NH children (right panel) in Experiment 2. Error bars indicate standard errors of the mean, and the interrupted line indicates chance performance. CI indicates cochlear implant; NH, normally hearing.

variable and therefore unlikely to be a reliable cue to identity. This suggests that voice identification was probably driven by differences in talker-specific phonetic realization.

The present study revealed a robust recognition of cartoon voices among 5- to 6-year-old children, consistent with findings obtained previously with somewhat younger children under optimal listening conditions (Spence et al. 2002). Children's performance on the intact stimuli approached ceiling levels, but even with as few as four bands of spectral information, performance was well above chance. In general, four spectral bands are considered adequate for adults' recognition of speech in quiet (Shannon et al. 1995) but inadequate for voice sex recognition, although performance also depends on the available temporal cues (Fu et al. 2005). It is likely that the ecological validity of the present materials, as exemplified by the familiar talkers (Vongpaisal et al. 2010), sentence-length utterances (Vongpaisal et al. 2012), and other iconic qualities of cartoon voices contributed to the high levels of performance in the present study. In addition, the limited variety of talkers and response options minimized the cognitive demands on children, which may have contributed to the high levels of performance.

The method of stimulus presentation had differential consequences, with randomized trials generating better performance than blocked trials, especially at high levels of spectral degradation. Prior exposure to the sentences at low levels of degradation may have allowed children to adapt to the manner of distortion (Hervais-Adelman et al. 2008), resulting in enhancement on subsequent trials with greater degradation. Specifically, exposure to intermediate levels of degradation might promote adaptation to noise-vocoded signals, with subsequent transfer to situations involving increased degradation, in line with adults' generalization across frequency regions after exposure to spectrally distorted speech (Hervais-Adelman et al. 2011). Children in the random order may also have profited from the opportunity to map utterance content onto specific talkers, but this would have required remembering five different utterances for each of the three characters. They could have benefited as well from increased practice on the task. For high levels of spectral degradation, children in the random order had completed more prior test trials, on average, than children in the blocked order. This increased practice may have enhanced their performance independent of or in addition to perceptual adaptation. Note, however, that children received no feedback, so the benefits of such exposure would involve implicit learning.

Overall, the findings suggest that NH children can identify familiar cartoon voices from spectrally degraded speech. However, they do not address the question of whether child CI users, who lack exposure to spectrally rich versions of those voices and for whom listening in general involves more effort, would be capable of recognizing familiar cartoon voices. The ability of CI users to do so was examined in Experiment 2.

EXPERIMENT 2

The goal of the present experiment was to examine the ability of child CI users to recognize the voices of familiar TV cartoon characters. Previous research confirmed that CI users can identify their mother's voice (Vongpaisal et al. 2010). As noted, the voices of cartoon characters are much less familiar than the mother's voice. Moreover, exposure to cartoon voices occurs in noninteractive contexts with competing visual cues. It is possible, then, that child CI users accord less attention to auditory than to visual cues from TV than is the case for their NH counterparts. That situation could result in less stable representations of the voices of TV characters and, consequently, poorer voice recognition. Even if child CI users allocate comparable attention to auditory cues in audiovisual contexts, the decreased availability of spectral cues could make their representations less distinctive or robust. One potential consequence would be an inability or reduced ability to identify the voices of cartoon characters relative to their NH peers.

Despite child CI users' difficulty with talker differentiation in some circumstances (Cleary & Pisoni 2002; Cleary et al. 2005; Vongpaisal et al. 2010), their experience with spectrally degraded speech provides one potential advantage over NH children tested with noise-vocoded (i.e., spectrally degraded) speech. For child CI users, talker recognition strategies used in everyday contexts would be applicable to the test context. In the present experiment, the test stimuli for CI users were the same as those available while watching TV except for superior listening conditions afforded by a sound-attenuating booth and high-quality loudspeakers. Thus, even if child CI users are less accurate at identifying the cartoon voices than NH peers tested with unmodified voices, they might still recognize the voices of the cartoon characters at better than chance levels. Adult CI users' ability to differentiate male from female voices is roughly equivalent to the performance of NH adults on vocoded stimuli with four to eight spectral bands (Fu et al. 2005). One might therefore expect child CI users' performance on unmodified stimuli to be poorer than the performance of NH children in Experiment 1 on stimuli with four to eight spectral bands.

Participants and Methods

Participants • The participants were 15 deaf children who were 4 to 7 years of age (M = 5.5 years; SD = 0.7 years) with bilateral CIs. Information about implant type, age at testing, and age at activation is shown in Table 2. The children had a minimum of 2 years of CI experience and, except for 3 participants with progressive hearing loss, the others were implanted in infancy. Children's absolute thresholds in the speech range were within normal limits (10 to 30 dB HL). All children participated in auditory–verbal therapy for a minimum of 2 years after implantation, they communicated exclusively by auditory–oral means, and they were in age-appropriate school classes with their NH peers. The present sample of child CI users had advantages relative to the general population of child CI users because they were considered successful CI users, they had access to a wide range of professional services, they had parents who were highly educated and committed to their children's progress, and they had the same language at home, at school, and for therapeutic interventions. An additional 15 NH children who were 4 to 5 years of age (M = 4.7 years; SD = 0.3 years) and presumably from middle- or upper-income families (determined by census information based on their residential address) were tested with the same stimuli and methods as the CI users. As in Experiment 1, all participants were familiar with at least three cartoon characters in the stimulus set.

Method • Testing was conducted as in Experiment 1 except that only the unprocessed stimuli were used for all children, and some of the child CI users were tested in the same manner (comparable apparatus, stimuli, and procedure) at a local

TABLE 2. Background of child CI users in Experiment 2

Child	Age (yrs) at Test	Age (yrs) at 1st and 2nd CI Activation	Etiology	Device L/R	Strategy
1*	5.8	3.4; 3.4	Genetic	Contour/Freedom	ACE
2	4.8	0.8; 1.7	Genetic	Contour/Freedom	ACE
3	5.3	1.1; 1.1	Genetic	Contour/Contour	ACE
4	5.8	1.0; 3.6	Genetic	Contour/Freedom	ACE
5*	6.6	2.5; 4.0	Unknown	Contour/Contour	ACE
6	5.0	1.0; 4.6	Genetic	Contour/Contour	ACE
7	5.1	0.9; 1.8	Genetic	Contour/Freedom	ACE
8	6.1	0.8; 1.5	Genetic	Contour/Freedom	ACE
9	5.4	1.7; 1.7	Unknown	Contour/Freedom	ACE
10*	6.3	3.1; 6.3	Mondini dysplasia	Freedom/Freedom	ACE
11	4.8	1.1; 1.1	Genetic	Contour/Freedom	ACE
12	6.1	1.0; 3.5	Unknown	Freedom/Freedom	ACE
13	4.1	1.1; 1.1	Unknown	Freedom/Freedom	ACE
14	6.4	1.3; 2.3	Genetic	Freedom/Freedom	ACE
15	5.5	1.7; 2.7	Genetic	Freedom/Freedom	ACE

**Progressive hearing loss from birth.*
CI, cochlear implant; L, left, R, right.

hospital facility. There were 15 different trials for each participant: five utterances for each of the three characters. All utterances were presented twice, for a total of 30 trials, and the order of trials was randomized with the constraint that the same utterance could not appear on successive trials.

Results and Discussion

Figure 1 shows the performance of CI users (middle panel) and NH children (right panel). Mean proportion of correct responses was 0.90 (SD = 0.09) for CI users and 0.97 (SD = 0.05) for NH children. As was the case for unprocessed stimuli in Experiment 1, NH children's performance was at ceiling. Accordingly, nonparametric statistics were used for data analysis. A Mann–Whitney test revealed that the performance of CI users was significantly less accurate than that of NH children (U = 71.5; $p = 0.009$). One-sample Wilcoxon tests indicated, however, that the proportion of correct responses substantially exceeded chance levels for both groups of listeners (CI users: $z = -3.42$; $p < 0.001$; NH children: $z = -3.91$; $p < 0.001$, respectively).

Obviously, the child CI users had to contend with spectrally degraded input, but the NH children did not. We attempted to compare the performance of CI users in the present experiment with that of NH children tested under conditions of spectral degradation (Experiment 1). For these purposes, performance in the blocked condition (equivalent level of degradation within a block) provided a more suitable basis of comparison than performance in the random condition where children derived benefit from stimuli heard previously at lesser levels of degradation. Child CI users' performance was no different from that of NH children tested at 24 frequency bands (U = 75; $p = 0.460$) and they outperformed NH listeners at greater levels of degradation (all $p \leq 0.005$; all two-tailed comparisons with a Bonferroni-corrected α-level of 0.005). To ensure that NH children's performance at the highest level of spectral degradation was not unduly affected by the first few trials in the blocked condition with very unusual-sounding stimuli, their performance was examined with the first four trials excluded. Overall performance on stimuli with four frequency bands remained unchanged. According to Fu et al. (2005), adult CI users' accuracy of differentiating speaker sex is comparable with the performance of NH listeners tested at four or eight frequency bands, at least when tested on syllables in isolation. In the context of sentence-level utterances, child CI users exceeded this performance by a considerable margin.

The present experiment revealed that child CI users, or at least those who share the health and demographic advantages of the present sample, succeed in identifying cartoon voices, but they do so less accurately than their NH peers who listen to unprocessed voices. By contrast, child CI users performed more accurately than NH children tested under conditions of moderate to severe levels of spectral degradation. Even in the absence of rich spectral information at the encoding phase (i.e., at home), child CI users were able to develop robust representations of the voices of cartoon characters, which supported subsequent recognition of those voices. It will be important to establish whether comparable voice recognition can be achieved by the wider population of child CI users in addition to the advantaged sample tested here.

The NH children had 5 to 6 years of experience with speech in general and with a variety of talkers. Their hours of exposure to the cartoon voices may have been comparable with that of child CI users, but the quality of the input obviously differed for the two groups. The importance of spectral cues in the long-term representations of NH children was reflected in sharply reduced performance with increasing levels of degradation, suggesting that their representations of voices depend heavily on fine-grained frequency information that is largely unavailable to child CI users. Tellingly, only at relatively modest levels of degradation was their performance equivalent to that of child CI users. This may indicate that the stored spectral information is of limited assistance to NH children when speech is distorted and a shift in listening strategies is essential.

Unlike their NH peers, child CI listeners had considerable experience with degraded speech. Additional exposure to vocoded speech is likely to improve NH children's ability to map the degraded input onto their long-term representations of the voices, in line with gains in speech perception observed in

NH adults after exposure to noise-vocoded speech (Davis et al. 2005; Hervais-Adelman et al. 2008), and with the superior performance in the present sample of NH children who were tested in the random condition.

GENERAL DISCUSSION

We demonstrated that children identify familiar cartoon voices in the absence of spectral cues that are often considered critical to talker recognition (cf. Van Lancker & Kreiman 1987). In Experiment 1, NH children were tested on their ability to identify such voices from intact speech samples and from speech samples with varying levels of spectral degradation. Children identified the voices at all levels of degradation, but performance accuracy decreased with increasing degradation, highlighting the importance of spectral cues to talker identification (Fu et al. 2004). Limited practice with less distorted versions appeared to facilitate subsequent performance with more distorted versions, as reflected in superior performance when levels of degradation were randomized rather than presented in order of decreasing difficulty.

In Experiment 2, child CI users and another sample of NH children were evaluated on their ability to identify the same cartoon voices. Although CI users performed less accurately than their NH peers, they were successful in identifying the familiar cartoon voices. The findings of both experiments, taken together, confirm that talker recognition is possible in the absence of temporal fine structure although the availability of such cues clearly enhances talker recognition (Vongphoe & Zeng 2005; Vongpaisal et al. 2010, 2012).

These findings are the first to indicate that spectrally degraded versions of cartoon voices activate long-term representations of such voices in NH children. Undoubtedly, children's usual talker-identification strategies prioritize features such as vocal timbre and pitch patterning, which were less accessible under conditions of severe spectral degradation. In other contexts, young NH children do not always display flexible, or situation-appropriate, use of cues in speech processing (e.g., Eisenberg et al. 2000; but see Newman & Chatterjee 2013), even with limited response options (Morton & Trehub 2001; Morton et al. 2003), which makes it all the more impressive that they succeeded in the present talker identification task.

It has become clear that native speakers of a language exhibit subtle differences in the pronunciation of phonemes and allophones (e.g., Allen et al. 2003; Smith & Hawkins 2012). What is also clear is that listeners use such phonetic variability to identify specific talkers (e.g., Allen & Miller 2004). In fact, NH adults are capable of identifying familiar talkers at better than chance levels from spectrally impoverished, sine-wave replicas of their speech (Remez et al. 1997), even when the acoustic correlates of differences in vocal tract size are neutralized (Fellowes et al. 1997). Child CI users and NH children listening to vocoded speech are presumed to use differences in articulatory timing as well as global differences in speech rhythm and rate to identify talkers or classify them by sex and age level (Vongpaisal et al. 2010, 2012). The presumption is that children in the present study, CI users as well NH children presented with spectrally degraded speech, capitalized on articulatory timing differences to identify the voices of cartoon characters even though other cues were available. Given that individual differences in intonation and speech rate only add minimally to child CI users' accuracy of talker identification (Vongpaisal et al. 2010), it is likely that phonetic timing played a more important role.

To date, the only evidence of familiar talker recognition by child CI users involved the highly familiar maternal voice (Vongpaisal et al. 2010). The present research establishes definitively that children with CIs have stable, long-term representations of other voices. Unquestionably, the voices of cartoon characters are much less familiar than the maternal voice. Nonetheless, such voices have other characteristics that may make them salient and memorable for children and for CI users in particular. The speech style of cartoon characters has much in common with maternal speech to infants and young children, which is marked by considerable pitch and amplitude modulation, slow speaking rate, and hyperarticulation (Fernald et al. 1989; Kuhl et al. 1997; Lam & Kitamura 2012). This speech register increases the transparency of the speaker's expressive intentions (Fernald 1989; Bryant & Barrett 2007) and may make those intentions more accessible to child CI users (Nakata et al. 2012; Volkova et al. 2013). Adults also make various speech adjustments, including hyperarticulation, in their interactions with listeners with known hearing loss (Ferguson 2004; Smiljanić & Bradlow 2005), resulting in increased speech intelligibility (Picheny et al. 1985). It is possible, indeed likely, that exaggerated cues that increase the transparency of expressive intentions and the intelligibility of speech also magnify cues to talker identity. It remains to be determined whether children can recognize talkers from spectrally degraded utterances without the characteristic exaggeration of child-directed speech and whether some voices, like those of parents, siblings, or beloved cartoon characters, are privileged because of their emotional value. Finally, because phonological knowledge affects talker recognition (Perrachione et al. 2011), one would expect the course of learning to recognize specific voices to be more protracted for child CI users than for their NH peers. Future research could address these questions.

In sum, the present study indicates that young NH children as well as CI users can recognize talkers under conditions of severe spectral degradation. Like their adult counterparts, children use idiosyncratic aspects of speech to identify talkers when limited spectral information is available.

ACKNOWLEDGMENTS

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (to S.E.T. and E.G.S.).

The authors declare no conflict of interest.

Address for correspondence: Sandra E. Trehub, Department of Psychology, University of Toronto Mississauga, 3359 Mississauga Road North, Mississauga, ON, Canada L5L 1C6. E-mail: sandra.trehub@utoronto.ca

Received November 19, 2012; accepted June 21, 2013.

REFERENCES

Allen, J. S., Miller, J. L., DeSteno, D. (2003). Individual talker differences in voice-onset-time. *J Acoust Soc Am*, *113*, 544–552.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *J Acoust Soc Am*, *115*, 3171–3183.

Barker, B. A., & Newman, R. S. (2004). Listen to your mother! The role of talker familiarity in infant streaming. *Cognition*, *94*, B45–B53.

Bartholomeus, B. (1973). Voice identification by nursery school children. *Can J Psychol*, *27*, 464–472.

Bryant, G. A., & Barrett, H. C. (2007). Recognizing intentions in infant-directed speech: Evidence for universals. *Psychol Sci*, *18*, 746–751.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729.

Church, B. A., & Fisher, C. (1998). Long-term auditory word priming in preschoolers: Implicit memory support for language acquisition. *J Mem Lang*, *39*, 523–542.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *J Acoust Soc Am*, *116*, 3647–3658.

Cleary, M., & Pisoni, D. B. (2002). Talker discrimination by prelingually deaf children with cochlear implants: Preliminary results. *Ann Otol Rhinol Laryngol Suppl*, *189*, 113–118.

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. *J Speech Lang Hear Res*, *48*, 204–223.

Creel, S. C., & Jimenez, S. R. (2012). Differences in talker recognition by preschoolers and adults. *J Exp Child Psychol*, *113*, 487–509.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., et al. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *J Exp Psychol Gen*, *134*, 222–241.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*, 1174–1176.

Dorman, M. F., Loizou, P. C., Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *J Acoust Soc Am*, *102*, 2403–2411.

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., et al. (2000). Speech recognition with reduced spectral cues as a function of age. *J Acoust Soc Am*, *107*(5 Pt 1), 2704–2710.

Faulkner, A., Rosen, S., Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *J Acoust Soc Am*, *108*, 1877–1887.

Fellowes, J. M., Remez, R. E., Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Percept Psychophys*, *59*, 839–849.

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *J Acoust Soc Am*, *116*(4 Pt 1), 2365–2373.

Fernald, A., Taeschner, T., Dunn, J., et al. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J Child Lang*, *16*, 477–501.

Fu, Q. J., Chinchilla, S., Galvin, J. J. (2004). The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users. *J Assoc Res Otolaryngol*, *5*, 253–260.

Fu, Q. J., Chinchilla, S., Nogaki, G., et al. (2005). Voice gender identification by cochlear implant users: The role of spectral and temporal resolution. *J Acoust Soc Am*, *118*(3 Pt 1), 1711–1718.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., et al. (2008). Perceptual learning of noise-vocoded words: Feedback and lexicality. *J Exp Psychol Hum Percept Perform*, *34*, 460–474.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., et al. (2011). Generalization of perceptual learning of vocoded speech. *J Exp Psychol Hum Percept Perform*, *37*, 283–295.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *J Exp Psychol Hum Percept Perform*, *26*, 1570–1582.

Johnson, E. K., Westrek, E., Nazzi, T., et al. (2011). Infant ability to tell voices apart rests on language experience. *Dev Sci*, *14*, 1002–1011.

Kreiman, J., Van Lancker-Sidtis, D., Gerratt, B. R. (2005). Perception of voice quality. In D. Pisoni & R. Remez (Eds.), *Handbook of Speech Perception* (pp. 338–362). Maldon, MA: Blackwell.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*, 684–686.

Lam, C., & Kitamura, C. (2012). Mommy, speak clearly: Induced hearing loss shapes vowel hyperarticulation. *Dev Sci*, *15*, 212–221.

Loebach, J. L., Bent, T., Pisoni, D. B. (2008). Multiple routes to the perceptual learning of speech. *J Acoust Soc Am*, *124*, 552–561.

Maye, J., Aslin, R. N., Tanenhaus, M. (2008). The weckud wetch of the wast: Rapid adaptation to a novel accent. *Cogn Sci*, *32*, 543–562.

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Dev*, *72*, 834–843.

Morton, B. A., Trehub, S. E., Zelazo, P. D. (2003). Sources of inflexibility in 6-year-olds' understanding of emotion in speech. *Child Development, 74*, 1857–1868.

Nakata, T., Trehub, S. E., Kanda, Y. (2012). Effect of cochlear implants on children's perception and production of speech prosody. *J Acoust Soc Am*, *131*, 1307–1314.

Newman, R., & Chatterjee, M. (2013). Toddlers' recognition of noise-vocoded speech. *J Acoust Soc Am*, *133*, 483–494.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Percept Psychophys*, *60*, 355–376.

Nygaard, L. C., Sommers, M. S., Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychol Sci*, *5*, 42–46.

Perrachione, T. K., Del Tufo, S. N., Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, *333*, 595.

Picheny, M. A., Durlach, N. I., Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *J Speech Hear Res*, *28*, 96–103.

Remez, R. E., Fellowes, J. M., Rubin, P. E. (1997). Talker identification based on phonetic information. *J Exp Psychol Hum Percept Perform*, *23*, 651–666.

Schmale, R., Cristia, A., Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Dev Sci*, *15*, 732–738.

Shannon, R. V., Zeng, F. G., Kamath, V., et al. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., et al. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *J Exp Psychol Hum Percept Perform*, *28*, 1447–1469.

Sheldon, S., Pichora-Fuller, M. K., Schneider, B. A. (2008). Effect of age, presentation method, and learning on identification of noise-vocoded words. *J Acoust Soc Am*, *123*, 476–488.

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *J Acoust Soc Am*, *118*(3 Pt 1), 1677–1688.

Smith, R. & Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. *J Phon*, *40*, 213–233.

Spence, M. J. & Freeman, M. S. (1996). Newborn infants prefer the maternal low-pass filtered voice, but not the maternal whispered voice. *Infant Behav Dev*, *19*, 199–212.

Spence, M. J., Rollins, P. R., Jerger, S. (2002). Children's recognition of cartoon voices. *J Speech Lang Hear Res*, *45*, 214–222.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., et al. (2000). Language development in profoundly deaf children with cochlear implants. *Psychol Sci*, *11*, 153–158.

Van Heugten, M., & Johnson, E. K. (2012). Infants exposed to fluent natural speech succeed at cross-gender word recognition. *J Speech Lang Hear Res*, *55*, 554–560.

Van Lancker, D., Kreiman, J., Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I. Recognition of backwards voices. *J Phon*, *13*, 19–38.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*, 829–834.

Volkova, A., Trehub, S. E., Schellenberg, E. G., et al. (2013). Children with bilateral cochlear implants identify emotion in speech and music. *Cochl Impl Int*, *14*, 80–91.

Vongpaisal, T., Trehub, S. E., Glenn Schellenberg, E., et al. (2012). Age-related changes in talker recognition with reduced spectral cues. *J Acoust Soc Am*, *131*, 501–508.

Vongpaisal, T., Trehub, S. E., Schellenberg, E. G., et al. (2010). Children With Cochlear implants recognize their mother's voice. *Ear Hear*, *31*, 555–566.

Vongphoe, M., & Zeng, F. G. (2005). Speaker recognition with temporal cues in acoustic and electric hearing. *J Acoust Soc Am*, *118*, 1055–1061.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Dev Sci*, *14*, 372–384.

REFERENCE NOTE

1. Van Heugten, M., & Johnson, E. K. (in press). Learning to contend with accents in infancy: Benefits of brief speaker exposure. *J Exp Psychol Gen*.