Toddlers' recognition of noise-vocoded speech

Rochelle Newman[a] and Monita Chatterjee[b]
Department of Hearing and Speech Sciences, 0100 Lefrak Hall, University of Maryland, College Park, Maryland 20742

(Received 17 July 2011; revised 10 November 2012; accepted 14 November 2012)

Despite their remarkable clinical success, cochlear-implant listeners today still receive spectrally degraded information. Much research has examined normally hearing adult listeners' ability to interpret spectrally degraded signals, primarily using noise-vocoded speech to simulate cochlear implant processing. Far less research has explored infants' and toddlers' ability to interpret spectrally degraded signals, despite the fact that children in this age range are frequently implanted. This study examines 27-month-old typically developing toddlers' recognition of noise-vocoded speech in a language-guided looking study. Children saw two images on each trial and heard a voice instructing them to look at one item ("Find the cat!"). Full-spectrum sentences or their noise-vocoded versions were presented with varying numbers of spectral channels. Toddlers showed equivalent proportions of looking to the target object with full-speech and 24- or 8-channel noise-vocoded speech; they failed to look appropriately with 2-channel noise-vocoded speech and showed variable performance with 4-channel noise-vocoded speech. Despite accurate looking performance for speech with at least eight channels, children were slower to respond appropriately as the number of channels decreased. These results indicate that 2-yr-olds have developed the ability to interpret vocoded speech, even without practice, but that doing so requires additional processing. These findings have important implications for pediatric cochlear implantation. © *2013 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4770241]

PACS number(s): 43.71.Ft, 43.66.Ts [RYL] Pages: 483–494

I. INTRODUCTION

Cochlear implants (or CIs) are prosthetic devices that provide auditory perception to severely hearing-impaired listeners by bypassing the damaged cochlea and electrically stimulating the auditory nerve. These devices have had substantial impact on thousands of hearing-impaired listeners across the world. Present-day cochlear implants, however, are not able to transmit the entire speech signal to the listener; because of both technological and biological limitations, the signal provided by these devices is considerably degraded in terms of its spectrotemporal resolution. Given these limitations, one primary area of research involves determining which components of the signal are the most important for accurate speech perception.

Much of our knowledge in this domain has come from research on normal-hearing (NH) adults listening to simulated CI speech. Noise-vocoded speech is a type of signal that is thought to simulate, for a normal-hearing listener, the perception of speech as heard through a cochlear implant. This signal is created by dividing the speech signal into a set of separate frequency bands, taking the overall amplitude envelope from each band, and using these envelopes to modulate bands of noise that are centered over the same frequency regions. These noise bands, when combined together, can be perceived as speech, albeit of a very unnatural form (Shannon *et al.*, 1995). This process of dividing the signal into frequency bands and transmitting the amplitude envelope of each band is conceptually similar to the way in which a cochlear implant processes the speech signal. As a result, noise-vocoded speech sentences are often used as simulations of CI speech, although a true CI listener would also have a variety of other perceptual decrements, caused by loss of auditory neurons, reorganization of central auditory pathways following deafness, etc.

Noise-vocoded speech has been used in numerous groundbreaking studies with adult listeners, exploring such issues as the minimum number of bands required for accurate perception (Friesen *et al.*, 2001), the role of subject factors (such as age; Eisenberg *et al.*, 2000; Sheldon *et al.*, 2008a), and the role of top-down contextual and lexical effects on performance (Davis *et al.*, 2005; Hervais-Adelman *et al.*, 2008; Sheldon *et al.*, 2008b). Studies with vocoded speech have helped to explain patterns of performance seen in cochlear-implant listeners with respect to the effect of different types and levels of background noise on performance (Fu *et al.*, 1998; Ihlefeld *et al.*, 2010). Studies have also looked at the effect of frequency shifts (as might be caused by a shallow insertion depth) and frequency compression/expansion (Baskent and Shannon, 2003, 2007; Fu and Shannon, 1999; Shannon *et al.*, 1998) as well as the effect of frequency transposition (as might be used to avoid cochlear dead regions; Baskent and Shannon, 2006). Thus vocoded speech has paved the way for a wide array of research investigations, and these studies have had important influences on the design of CI processors.

Clearly, noise-vocoded speech is not identical to speech heard through a CI. Studies with normally hearing individuals attending to these "CI simulations," however, offer a

[a]Author to whom correspondence should be addressed. Electronic mail: rnewman1@umd.edu
[b]Present address: Boys Town National Research Hospital, 555 N 30th St., Omaha, NE 68131.

unique advantage. CI patients tend to demonstrate large intersubject variability in performance, possibly owing to variations in difficult-to-control factors such as the duration-of deafness prior to implantation, etiology of hearing loss, auditory history, etc., which can influence both the peripheral and central pathways in the auditory system. Moreover, because these patients are difficult to recruit, studies testing CI patients tend to have relatively small numbers of participants; as a result, random differences among participants in these other factors can potentially have large effects on study outcomes. The primary advantage of studying the normally hearing system's processing of CI-simulated speech is that these factors are eliminated from consideration. The *disadvantage* is simply the flip side of the same coin: These studies do not provide sufficient information about the likely scenario in actual CI patients. Perhaps it would be reasonable to think of studies with CI simulations as presenting the "best-case" scenario. That said, the NH listeners in these experiments usually do not have much experience listening to spectrally degraded speech, while the CI patients that their performance is compared against have everyday experience with their device. Given these caveats, it is important to note that CI simulations using noiseband vocoder techniques have shown patterns of results in NH listeners that generally parallel results obtained with full-spectrum speech in CI patients; any divergences have also been useful and resulted in new insights (e.g., Fu and Shannon, 1999; Friesen *et al.*, 2001; Chatterjee and Peng, 2008).

As important as such studies are, they face a critical limitation: To date, all such studies have been run with adult listeners, or, in a few cases, with school-aged children (Eisenberg *et al.*, 2000; Nittrouer and Lowenstein, 2010; Nittrouer *et al.*, 2009). Yet many cochlear implants are being fitted on much younger children. FDA guidelines suggest implantation as young as 12 months, and some implantation is happening even younger. Although children implanted early clearly can benefit from their CIs (Gilley *et al.*, 2008; Sharma *et al.*, 2005; Tajudeen *et al.*, 2010), the processing strategies used with such children are based on our knowledge gained from adults. A long history of research in developmental auditory perception demonstrates that infants and children are not merely small adults—they attend to different aspects of the signal and process sound differently than do older listeners (see, for example, Elliott *et al.*, 1989; Nittrouer, 1992, 1996; Polka *et al.*, 2008; Werker and Tees, 1984; Werner and Bargones, 1992). It is likely, then, that the same would hold for CI users and for spectrally degraded speech. If so, children might receive greater benefit from processing strategies tailored to their listening abilities or strategies. Examining young children's performance with noise-vocoded speech could illuminate areas of commonality and difference between adult and child listeners. Such work would have both applied and theoretical benefits: First, such research could provide critical information that would be relevant to improving processing strategies for implant devices aimed at young children. Second, noise-vocoded speech is a particularly interesting form of degraded speech in that it primarily preserves temporal envelope information, while severely reducing spectral information. Examining young children's performance with this type of a signal provides information regarding the types of acoustic cues that they can process appropriately (see, for example, Nittrouer *et al.*, 2009). Understanding children's ability to interpret such a signal therefore provides an important backdrop to many theories of developmental auditory neuroscience.

Children and adults differ in their processing of speech in a number of ways; this might be relevant to predictions for their processing of degraded speech. Most importantly, adult listeners are able to use top-down knowledge of the language to help them "restore" missing or degraded information; these top-down effects have been shown in a wide range of tasks and studies (e.g., Samuel, 1981; Warren and Obusek, 1971; Warren *et al.*, 1997). Young listeners have had less experience with their language and thus may not have as much top-down knowledge upon which to rely. While young children can use prior knowledge to help them interpret partial information (Newman, 2006; Swingley *et al.*, 1999), they require more of the signal to be intact to be successful, and top-down knowledge does not serve to "fill in the gaps" in toddlers' perception (Newman, 2006). Children also have slower processing than adults, and some studies suggest that children aged 4 yr and younger may be unable to use constraining contextual information quickly enough to influence perception (Cole and Perfetti, 1980; Walley, 1987; but see Nittrouer and Boothroyd, 1990 for findings with older children). Children in general require larger acoustic differences to discriminate speech sounds (Elliott *et al.*, 1986; Elliott *et al.*, 1989); these findings likewise suggest that children might be affected by spectrally degraded signals to a greater extent than are adults. All of these differences suggest that young children may have more difficulty with degraded speech than do older children or adults. (See Newman, 2006 for a further discussion of these differences.)

In the adult studies discussed in the preceding text, it is normal practice to compare the performance of NH adults to that of post-lingually deaf CI patients. Both groups of listeners would have developed similarly from an auditory/speech/language standpoint. Typically, children receiving CIs at an early age are considered prelingually deaf or prelingually hearing impaired. On the other hand, the normal-hearing children in the present study are developing auditory input and language skills at a normal pace. Thus a direct comparison between early implanted CI children and studies involving NH children of the same age is not justifiable. As with the adult comparisons, here, too, similar precautions must be taken in interpreting the results. Studies with NH children listening to CI simulations can give us some idea of how the normally developing auditory system, presumably naïve to spectrally degraded speech, would handle the task of speech recognition under such conditions. Their CI counterparts would have developed with the device and thus have considerably more expertise with degraded signals. Owing to greater neural plasticity in the early stages of the critical period, early implantation might confer some advantages with degraded signals to CI children that have not as yet been studied. Before such studies can be undertaken or properly considered, however, it is important to know how young

NH children would process such stimuli. Regardless of the caveats discussed in the preceding text, this would provide an important set of reference data, from both a scientific and clinical perspective, against which one can compare the progress of CI children.

Eisenberg *et al*. (2000) first investigated the perception of noise-vocoded speech by children. They explored the number of channels required for accurate speech recognition using a wide variety of speech stimuli, including sentences, words, and nonsense syllables. They found that while children aged 10–12 yr performed similarly to adults, children aged 5–7 yr performed more poorly, requiring more spectral bands to reach levels of performance comparable with their older peers. The authors suggested that their findings have important implications for the amount of speech training provided to young CI users.

Nittrouer and colleagues (Nittrouer and Lowenstein, 2010; Nittrouer *et al*., 2009) tested children's performance on four-channel noise-vocoded speech (which they refer to as amplitude-envelope or AE speech) and sine-wave analog speech. Noise-vocoded speech preserves amplitude structure but has substantially reduced spectral information; sine-wave speech signals preserve spectral information but lack much of the broadband formant structure found in natural speech. They found that children across a range of ages showed much more difficulty with the noise-vocoded signals than with the sine-wave speech, suggesting that children rely to a greater extent on dynamic spectral information than do adults.

Given that children as old as 5–7 yr show reduced ability to recognize spectrally degraded speech, we might expect that infants and toddlers would show even poorer performance. The present study explores this question by testing toddlers aged 27 months with noise-vocoded speech stimuli in a language-guided (or preferential) looking task (also known as a looking-while-listening procedure; Fernald *et al*., 1998). Grieco-Calub *et al*. (2009) used this procedure to test children who use cochlear implants and found their word-recognition to be both slower and less accurate than that of NH children. The source of this difference remains unclear, however; it could be the result of having less prior experience with spoken language, or it could be a result of the reduced information in the speech signal, or both. Testing the same NH children on both degraded and full speech allows for an examination of the effects of the signal degradation, separately from experiential differences.

In the present study, we tested children with normal hearing and normal previous language experience with noise-vocoded stimuli. The children saw two images appear on a television screen in front of them (for example, a car and a ball) and heard a voice telling them which object they should look at (*Can you find the ball?).* On some trials, the speech was presented normally, while on other trials the speech was noise-vocoded. We examined the accuracy of children's looking behavior in these different conditions.

II. EXPERIMENT 1

This first experiment explored whether toddlers could recognize noise-vocoded speech. For this study, we used noise-vocoded speech of 24 and 8 channels as well as full speech. As noted in the preceding text, noise-vocoded speech is created by dividing the original speech signal into a set of separate frequency bands, taking the overall amplitude envelope from each band and using these envelopes to modulate bands of noise that are centered over the same frequency regions. Thus when the signal is divided into more bands, more of the spectral resolution in the original signal is preserved. Adult listeners can perform well with as few as four channels (Shannon *et al*., 1995), although actual performance depends on the specific test conditions, such as whether the task is open- or closed-set, involves phonemes, words, or sentences, etc. But given that toddlers have far less experience with the language than do adult listeners, it seemed reasonable to test these moderate levels of degradation.

The language-guided looking paradigm has proven to be a reliable way of testing young infants (Golinkoff *et al*., 1987) and to be sensitive to a variety of factors that make speech recognition more difficult. For example, prior studies using this method found that toddlers spent a progressively greater proportion of time attending to the target object as the signal-to-noise ratio increased (Newman, 2011). They also spent less time looking appropriately when the target word was mispronounced than when it was pronounced correctly (Swingley and Aslin, 2000). Finally, Fernald *et al*. (1998) reported that the time toddlers required to respond to the correct word decreased progressively with development. In all cases, the task provided a gradient measure of performance that captured the ease of children's language processing. Thus this paradigm seems to be sensitive to a number of factors that influence children's speech recognition, suggesting it would be a good task for measuring toddler's ability to recognize a spectrally degraded signal.

A. Method

1. Participants

Twenty-four toddlers (9 male, 15 female), aged 27 months (range: 26 months, 0 days to 27 months, 25 days) participated. An additional six children participated, but their data were excluded for excessive fussiness/crying ($n = 5$) or having been in a previous version of the study ($n = 1$). The children were assigned to one of six stimulus orders (see Sec. II A 3). An additional three participants were recruited in the event that additional data would be required in one of the stimulus orders, but their data (the last data collected in these orders) were not ultimately needed. Parents reported that their children had normal hearing and were not currently experiencing symptoms indicative of an ear infection.

2. Stimuli

Stimuli consisted of visual images of well-known words and a simultaneous audio signal, presented either in full speech, in 24-channel noise-vocoded speech, or in 8-channel noise-vocoded speech. The visual stimuli consisted of pairs of digital still images (*keys* and *blocks*; *car* and *ball*), matched for approximate size and color. All four objects are generally well-known to children of this age (Fenson *et al*., 1994); if

our particular participants did not know these words, this would become clear from performance in the full-speech condition.

The audio recordings consisted of a single talker producing three sentences, each containing a target word ("Look at the _____! Can you find the _____? See the _____?"). Baseline trials contained similar sentences that did not indicate any particular object ("Look at that! Do you see that? Look over there!"). All sentences were initially recorded in a noise-reducing sound booth, recorded over a Shure SM51 microphone at a 44.1 kHz sampling rate and 16 bits precision. Sentences were isolated and matched for amplitude, and the sentence sequences were then matched for duration by editing the duration of the pauses between sentences; total audio file length was 4.8 s for all trials. These stimuli were then used as the full speech condition.

Noise vocoding was performed using methods akin to published standards (Shannon *et al.*, 1995). The identical sentences as occurred in the full speech condition were vocoded with either 24 or 8 channels, using TigerCIS (Tigerspeech Technology, Qian-Jie Fu, House Ear Institute). The analysis input range was 200–7000 Hz with a 24 dB/octave rolloff. The signal was then split into frequency bands using bandpass filtering (Butterworth filters, 24 dB/oct rolloff), and the envelope of each band was extracted using half-wave rectification and low-pass filtering (400 Hz cutoff frequency). The envelope derived from each band was then used to amplitude-modulate a white noise signal with the same bandwidth as the original signal band. This removed the fine spectro-temporal structure within each frequency band. The resulting modulated noises were combined at equal amplitude ratios to create the final noise-vocoded stimuli.

3. Procedure

Children sat on their caregiver's lap, facing a widescreen TV, and participated in a language-guided looking task. At the start of each of 16 trials, an image of a baby laughing appeared in the center of the screen to attract the participant's attention. Subsequently, participants saw two images, on the left and right sides of the screen, occurring simultaneously, at approximately 20 deg visual angle. The auditory stimulus was presented simultaneously with the two images.

The study began with two practice trials, which were not included in the data analysis. These two trials consisted of images of a cat and dog, and a voice telling the child to find one of the two objects. On one trial the correct answer was on the left, and on the other trial the correct answer was on the right. These trials were intended to familiarize the children with the general task and setting.

This was followed by 14 test trials. There were four trials for each condition (full-speech, 24-channel, 8-channel) and two baseline trials used to measure general looking preferences for the object pairs. The four trials for each condition each instructed the child to look at a different object. That is, the child was told to look at the car, ball, blocks, and keys, one time each in each of the three conditions. Objects were always presented in the same pairs (car and ball; keys and blocks), and the correct answer was always one of the two choices. That is, when children were told to look at the car, one of the two objects on the screen was in fact the car. Baseline trials were similar, but the voice simply told the infant to "Look at that! Do you see that! Look over there!" On these trials, the child was not told *which* of the two objects to attend to, and thus their percentage of looking to one object vs the other can be taken as their general looking preference among the two choices. This is used as a comparison – if children comprehend when told to "look at the car," for example, they should spend a greater proportion of time looking at the car vs the ball in that situation than when simply told to look more generally. The two baseline trials were presented in full speech, and one occurred with each object pair (car/ball, keys/blocks).

The first 600 ms (18 frames) of each trial occurred prior to the first presentation of the target word. Because children could not know which object to look at until the word was first produced, these initial 18 frames were ignored in all coding data. Some studies have used these initial portions of trials as a measure of baseline rather than including full-length baseline trials (c.f., Meints *et al.*, 1999), particularly in designs entailing a longer period of time before target word onset. We chose not to use this approach because looking over such a short time window may not be a good measure of baseline preference. Our experience has been that when two images initially appear, the child's natural reaction is to look back and forth to identify the two choices before settling on a particular object; as a result, short stretches of time at the start of trials tend to result in looking time that is artificially closer to 50% looking at each object.

Participants were presented with one of six different trial orders; across these orders, we counterbalanced which image of the pair appeared on the left (vs right) side of the screen. Within each order, trial order was pseudo-randomized with the restriction that the correct response did not occur on the same side (left vs right) more than three trials in a row. Across the full set of trials, each side contained the correct response an equal number of trials. As there were 24 participants, four heard each of the six orders.

The caregiver listened to masking music over headphones throughout the study to prevent any biasing of the child's behavior. In addition to participating in the experimental session, parents were asked to complete the Language Development Survey (Rescorla, 1989) for their children. This is a screening checklist for estimating productive vocabulary; it consists of 310 words, and parents were asked to indicate which words their child produced.

4. Coding

A digital camera recorded each child's eye gaze throughout the study at a rate of 30 frames per second. Two experimenters, blind to condition, individually coded each child's looking behaviors on a frame-by-frame basis using SUPERCODER coding software (Hollich, 2005). From this, the infants' total duration of looking at each of the two images on each trial was calculated.

If the two coders disagreed on any trial by more than 15 frames (0.5 s), a third coder was used. The averages of the

two closest codings were used as the final data. This occurred on a total of 23 of the 336 trials (14 test trials per child × 24 children) or just under 7% of the time. The final data were extremely reliable; correlations on the percentage of left (vs right) looking for each individual participant ranged from 0.95881 to 0.99993 with an average correlation of 0.99440. Such a high correlation is important for ensuring that the results accurately reflect the children's looking behavior.

There are a number of different measures that have been taken from preferential looking studies. Looking time can be calculated based on the single longest look (Schafer and Plunkett, 1998) or based on total looking over a trial or a over a specific window (e.g., Grieco-Calub *et al.*, 2009). It can be measured in seconds or in the proportion of time spent looking at the appropriate vs inappropriate picture; these can lead to different results if children spend time looking at neither image, because a measure based on raw seconds will be reduced if the child spends half the trial looking at his or her feet, whereas a measure based on the proportion of time looking at the target vs nontarget object will not be. Finally, looking time to the target object can be compared either to looking to the alternate object on that trial[1] or to the same object on baseline trials (i.e., how long the child would presumably have looked based on chance alone) or on trials in which an alternative object is named or to a putative baseline of 50% We chose to base all measures on overall proportions of looking to the correct object. We present these data as the proportion of looking time to the target object when named minus the looking time to that object on baseline trials; this is a method that we and others have used successfully in the past (e.g., Naigles and Gelman, 1995; Newman, 2011). We presume that if children can understand the speech, despite any vocoding present, they will look longer to each image when it is named than in the baseline condition. That is, if a child recognizes the word "car," then he or she should look longer at the car when they hear, "Look at the car!" than when they hear, "Look at that!"; this comparison to baseline looking accounts for the fact that children may have preexisting biases to attend longer to some images than others. These measures are based on the full trial durations, rather than on a particular temporal window, as we were unsure whether the time course for responding would be similar for noise-vocoded speech as for full speech. We used two-tailed tests for all measures despite having a directional prediction. We also report the overall proportion looking time to the target object as a more intuitive measure of how often children looked correctly; this should be above 50% if children understand the word.

The use of frame-by-frame coding allows not only an analysis of a child's overall looking per trial but also an analysis of the amount of time it takes children to turn in the appropriate direction. We conducted two types of temporal analyses. First, we determined each child's direction of looking at each point in time (each frame) and averaged these data for all trials in a given condition. It is important to note that this measure of the average time course of looking is quite different from the individual-participant reaction time measures typically used with this paradigm and originally developed by Fernald and colleagues (Fernald and Hurtado, 2006; Fernald *et al.*, 2006; Fernald *et al.*, 1998). The use of a within-subjects design with multiple stimulus conditions resulted in some concerns regarding the calculation of a true RT measure in the present study. In particular, calculating RTs using Fernald's method requires limiting the analysis to only those trials in which children were looking at the distractor at word onset (approximately 40% of the trials in Fernald *et al.*, 1998). Reaction times are then measured on those trials only when they occurred within a 300–1800 ms window after word onset (a further subset). To avoid over-weighting individual trials, Fernald included a participant's data only for conditions with at least two trials remaining. As Fernald notes, this approach often results in the exclusion of a substantial proportion of trials; as a result, it works best in study designs in which children participate in many trials of the same type. The within-subjects design used in the current study resulted in only four trials per condition; thus, on average, infants would be expected to be looking incorrectly at word onset in only one or two trials, and not all of these would contain a shift in the critical window. In essence, then, this approach results in very few data points in the current design and could result in over-weighting single anomalous trials. We therefore elected to pursue both approaches. We first calculated an average looking time response across participants, in which we recorded looking time at all frames, and averaged across all trials (regardless of where the infant happened to be looking at target word onset). We then averaged across participants to find the time at which children, in general, begin looking at the appropriate object more often than chance. This measure includes data from more trials, but averages across individuals who may be responding differently. We also collected participant RTs based on Fernald's methodology with the exception that we included data even if there was only a single trial.

B. Results and discussion

1. Accuracy

We examined children's looking for each of the three speech conditions individually; for each condition, we calculated the proportion of time the child spent looking at each object when named and subtracted from that the proportion of time the child spent looking at the object on baseline trials. This difference was then averaged across the four objects in the study (car, ball, blocks, and keys) and compared to zero using a single-sample *t*-test; we used a P value of 0.05 as the critical value in all cases. We refer to this value as the increase over baseline looking. We also report the proportion of time overall that the children looked toward the target object, but this value is merely illustrative and was not part of the statistical analysis.

For the full-speech condition, children looked toward the target object 62.6% of the time, a 13.2% increase over their baseline looking [SD = 11.2; $t(23) = 5.77$, $P < 0.0001$]. For the 24-channel speech, children looked toward the target object 60.2% of the time, a 10.2% increase over their baseline looking [SD = 16.4; $t(23) = 3.05$, $P < 0.006$]. Finally, for the eight-channel speech, children looked toward the target object 62.4% of the time, a 12.4% increase over baseline

looking [SD = 14.8; $t(23) = 4.10$, $P < 0.0005$].[2] Thus in all three conditions, children looked significantly longer at the named object than would be expected by chance, demonstrating their ability to recognize the appropriate word. (See Fig. 5 for individual participants' looking proportions.)

At first glance, these looking proportions appear somewhat low. However, there is no standard proportion of looking that is considered typical across studies. For example, Meints *et al*. (1999) reported correct looking time in their best condition hovering around 55% for 24-month-olds; in contrast, Fernald *et al*. (2006) reported accuracies as high as 76% with participants of the same age. Accuracy depends on a wide variety of factors, including the age of the participants, how well the particular words are known, the confusability or discriminability of the acoustic forms of the words, the typicality of the visual representations of those objects, the conceptual similarity between the objects, etc. Our decision to only use single-syllable, stop-consonant-initial words, which are potentially more confusable, is likely one factor resulting in our looking times being lower than those reported by Fernald *et al*. Still, it is not clear what proportion of time we should expect infants to look at the correct object; what is most important is that in all three conditions, children successfully looked at the target object when named, implying recognition of the word. Moreover, a one-way repeated-measures ANOVA showed no significant differences across these three conditions, $F(2,46) < 1$, $P > 0.65$, suggesting that children behaved similarly with vocoded speech as with full speech (see Fig. 1).

Such robust performance on the vocoded speech is somewhat surprising and raises the question as to whether children might have learned to recognize vocoded speech over the course of the experiment via explicit comparison with the full-speech trials. Although this seemed unlikely, we decided to examine only those vocoded trials that had not been preceded by a full-speech token of the same sentence. We again found what appears to be strong but variable performance: Children looked appropriately 63.7% of the time for the 24-channel speech, and 62.6% appropriately for the 8-channel speech. However, only the latter actually demonstrated a significant increase in looking over baseline performance [for the 24-channel speech, there was a nonsignificant 9.5% increase over baseline, SD = 25.5, $t(23) = 1.83$, $P = 0.08$, whereas for the 8-channel speech there was a 12.2% increase over baseline performance, SD = 18.8, $t(23) = 3.17$, $P < 0.005$]. Because the eight-channel speech is putatively the more difficult, the significant result on those trials suggests that the good performance with vocoded speech was not merely an artifact of having already heard the words in a full-speech context.

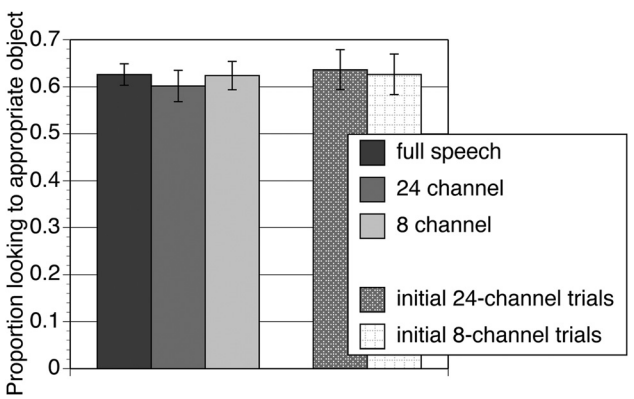

FIG. 1. Proportion of time spent looking to the target object in full speech, 24-channel noise-vocoded speech, and 8-channel noise-vocoded speech (left side of graph) in Experiment 1. Bars to the right show performance on noise vocoded trials not preceded by a full-speech sentence. Baseline performance was approximately 50% in all cases.

One limitation of the current study is that the children only had to distinguish between two choices at any given time. This is a very limited set-size, and children might not require access to the full phonological form of the word to make the distinction. In that sense, this task may be overestimating children's abilities. This is perhaps a greater concern for the keys/blocks pair, which have more dissimilar phonetic forms, than for the ball/car pair, which may be more difficult to discriminate acoustically (at least in full speech, although perhaps not in vocoded speech). We decided to examine these two pairs separately. In the full-speech condition, children actually did perform better with the blocks/keys pair [18.2% (SD = 12.5) increase from baseline, vs 8.2% (SD = 15.2) increase from baseline for car/ball, $t(23) = 3.01$, $P < 0.007$]. But this was not the case in either the 24-channel vocoded speech [8.3% (SD = 21.4) vs 12.1% (SD = 17.6), $t(23) = -0.85$, $P > 0.40$] or 8-channel vocoded speech [11.5% (SD = 23.2) vs 13.3% (SD = 14.7), $t(23) = -0.34$, $P > 0.70$]. This suggests that the children's excellent performance in the spectrally degraded stimuli was not being driven solely or primarily by one of the word pairs in particular.

2. Average time course of looking and reaction time measures

Frame-by-frame looking coding began with the onset of the visual stimulus (the pair of objects). Because this occurred immediately after the disappearance of the central attention-getter, children generally began each trial continuing to look at the center of the screen; once the two objects appeared, they shifted their gaze toward one or the other object. As a result, the average looking time to the target object across participants begins at roughly 0%. Children then looked somewhat randomly between the two images, while the voice was saying, "Look at the"; this is shown by accuracy in the 40% to 50% range. At some point after the onset of the first repetition of the target word, children began looking to the target object more frequently than to the nontarget object. As can be seen in Fig. 2, this occurred at different average time points for the three different conditions. Across the participants, looking time was significantly greater than chance (based on a one-tailed *t*-test for a minimum of three sequential frames) by 567 ms after the onset of the first repetition of the target word in the full-speech condition. This same reference point occurred at 800 ms in the 24-channel condition but not until 1300 ms for the 8-channel condition. Thus while children showed accurate responding overall in all three conditions, based on the disparate

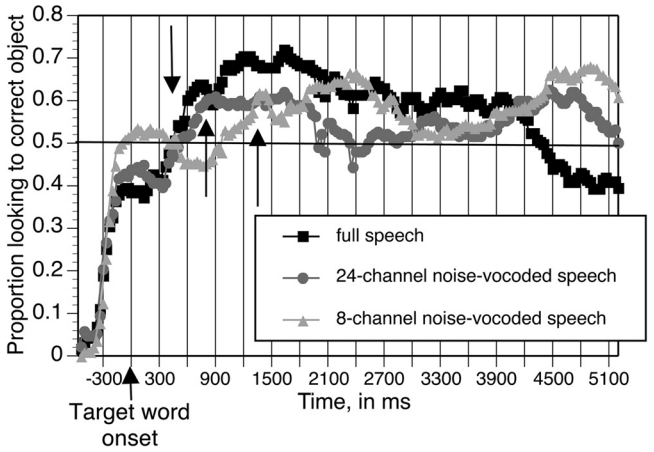

FIG. 2. Looking time to the target object as a function of time, for full speech, 24-channel noise-vocoded speech, and 8-channel noise-vocoded speech in Experiment 1. Arrows indicate the points at which each curve shows looking time significantly greater than chance.

amounts of time it took for children to reach above-chance performance, we infer that they were somewhat slower to identify the correct word when the speech was more degraded. This mirrors effects with adults, who also show slower reaction times to vocoded speech as the number of channels declines (Drgas and Blaszak, 2009).

Results using Fernald's method of reaction time analysis similarly found slower RTs for the eight-channel condition, although there were very few trials per condition and, as a result, quite variable effects. On average, the time for the first shift to the target object occurred at 641 ms (SD = 291) in the full-speech condition, 605 ms (SD = 279) in the 24-channel condition, and 767 ms (SD = 383) in the 8-channel condition. Eighteen participants (of 24) provided RT data for all three conditions; across these children, we found a significant overall difference, $F(2,34) = 4.94$, $P = 0.013$. Follow-up t-tests showed a significant difference between the 24- and 8-channel conditions only [full vs 24: $t(20) = 0.65$, $P > 0.50$; full vs 8: $t(18) = 1.55$, $P = 0.14$; 24 vs 8: $t(17) = 3.53$, $P < 0.005$; the differing df is the result of a number of participants who had no shift at all in a given condition].

3. Vocabulary correlations

We examined whether performance on this task correlated with children's vocabulary level. Children with more advanced lexical skills might conceivably also be better equipped to deal with degraded stimuli as a result of stronger lexical knowledge; alternatively, children who are better able to interpret degraded stimuli might have more opportunities to learn vocabulary, as they might be able to acquire useful lexical information in noisy or other difficult listening conditions. Indeed, Swingley and Aslin (2000) found that children with larger productive vocabularies were better able to identify words that had been mispronounced, although vocabulary did not correlate with performance on correctly pronounced words. One might predict that the ability to identify mispronounced words would be related to the ability to identify degraded words and thus that vocabulary would likewise play a role in the present study. Similarly, Grieco-Calub and colleagues found significant correlations between accuracy in their looking-while-listening task and productive vocabulary in young children with CIs.

Children in the current study varied tremendously in the number of words they were reported to produce; their vocabularies ranged from 18 words to 309 words (mean = 209, standard deviation = 105). However, while the increase in looking time in the full-speech condition did show a marginal correlation with vocabulary [$r(20) = 0.33$, $P < 0.10$], performance in the two vocoded conditions did not ($r = -0.04$ and $r = -0.03$). Apparently, vocabulary size is not a strong predictor of the ability to recognize spectrally degraded signals in these normal-hearing participants.

III. EXPERIMENT 2

The results from Experiment 1 suggest that toddlers can identify noise-vocoded speech, even with very little practice, at least in this closed-set (two-choice) task. This is the case for both 24- and 8-channel noise-vocoded speech. We decided to explore whether they would continue to be successful with further signal degradations. Thus the current experiment is identical to that in Experiment 1 except that we presented the sentences in full speech, in four-channel noise-vocoded speech, or in two-channel noise-vocoded speech.

A. Method

1. Participants

Twenty-four toddlers (11 male, 13 female), aged 27 months (range: 25 months, 23 days to 28 months, 0 days) participated. An additional two children participated, but their data were excluded for fussiness; four more participants were replaced after an error was identified in the order file they received.

2. Stimuli

Stimuli were created similarly to those in Experiment 1 except that the vocoded stimuli consisted of either 2 or 4 channels rather than 8 or 24 channels.

3. Procedure

Procedure was identical to that in Experiment 1. Coding reliability was similar with 22 trials requiring a third coder (compared to 24 in Experiment 1), and correlations on looking time ranging from 0.9473 to 0.9998 per participant, with an average correlation of 0.99540.

B. Results and discussion

For the full-speech condition, children looked toward the target object 69.2% of the time, an 18.9% increase over their baseline looking [$t(23) = 8.34$, $P < 0.0001$]. There were some hints that they looked correctly in the four-channel condition, but this was a *very* small, statistically marginal effect [53.6%, a 3.6% increase over baseline, $t(23) = 2.03$, $P = 0.054$]. They did not appear to look correctly in the two-channel condition [49.8%, a < 1% decrease from baseline, $t(23) = -0.07$ $P > 0.90$]. Thus while the children did seem to be performing

the task in general, they did not show evidence of an ability to interpret these more spectrally degraded signals as well as those from Experiment 1. A one-way repeated measures ANOVA identified an effect of condition [$F(2,46) = 23.45$, $P < 0.0001$]; follow-up t-tests showed a significant difference between the full-speech condition and both the four-channel [$t(23) = 6.25$, $P < 0.0001$] and two-channel [$t(23) = 5.69$, $P < 0.0001$] conditions, which did not differ from one another [$t(23) = 1.32$, $P > 0.20$]. These results are shown in Fig. 3. (See Fig. 5 for individual participants' looking proportions.)

Thus although a single-sample t-test demonstrated a trend toward longer looking in the four-channel condition than in the baseline condition, this performance difference was very small, with less than 54% correct looking, and was also not significantly better than in the two-channel condition (in which children performed at chance levels). While there are some hints of success in this condition, it does not appear that toddlers are performing at the level shown in Experiment 1. Looking at the scores from individual participants, there were only three children (of 24) who showed >60% accurate looking in the four-channel condition. This is in comparison to 22 children who did so in the full speech condition in this experiment, and 12 who had done so in the eight-channel condition in Experiment 1.

Although it certainly seems as though children performed less well in this four-channel condition than in the eight-channel condition from Experiment 1, such a comparison is difficult to make across different participant groups. Perhaps the participants in this study were simply less adept at interpreting degraded speech in general than those in Experiment 1, and thus the difference in performance with noise-vocoded speech had more to do with the participants than with the stimuli. We therefore tested this issue explicitly in Experiment 3, where we compared performance in the four- and eight-channel conditions in the same group of children.

Because most children did not look appropriately in the vocoded-speech conditions in this study, examining their reaction times did not seem appropriate. However, we again examined whether performance on this task correlated with children's vocabulary level and found no such correlations (full speech, $r = -0.03$; four-channel, $r = 0.07$; two-channel, $r = 0.18$) despite finding substantial variation in reported vocabulary (range 21 – 309, average 212 words, standard deviation = 106).

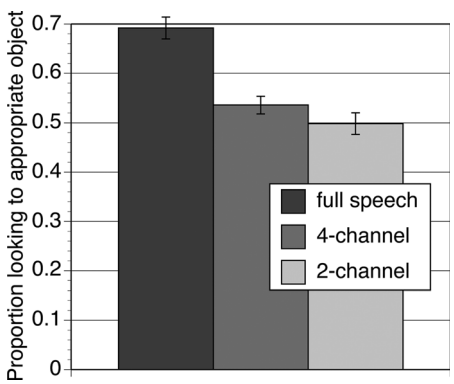

FIG. 3. Proportion of time spent looking to the target object in full speech, 4-channel noise-vocoded speech, and 2-channel noise-vocoded speech in Experiment 2. Baseline performance was approximately 50% in all cases.

IV. EXPERIMENT 3

The results from Experiments 1 and 2 suggest that toddlers can identify eight-channel noise-vocoded speech but not four-channel speech. However, this interpretation requires a comparison across the different groups of children who participated in the two studies. The current experiment was identical to the prior two experiments with the sentences presented to the same children in full speech, in eight-channel noise-vocoded speech (as in Experiment 1), or in four-channel noise-vocoded speech (as in Experiment 2).

A. Method

1. Participants

Eighteen toddlers (11 male, 7 female), aged 27 months (range: 26 months, 5 days to 27 months, 24 days) participated. An additional four children participated, but their data were excluded for hearing problems ($n = 2$) or fussiness ($n = 2$).

2. Stimuli

Stimuli consisted of the eight-channel noise-vocoded speech from Experiment 1, the four-channel noise-vocoded speech from Experiment 2, and the full-speech stimuli, which had been used in both prior experiments.

3. Procedure

The procedure was identical to that in Experiments 1 and 2. Coding reliability was similar, with 23 trials requiring a third coder, and correlations on looking time ranging from 0.9886 to 0.9997 per participant with an average correlation of 0.9962.

B. Results and discussion

1. Accuracy

For the full-speech condition, children looked toward the target object 69.1% of the time, a 19.1% increase over their baseline looking [$t(17) = 5.43$, $P < 0.001$]. For the eight-channel speech, children looked toward the target object 63.8% of the time, a 13.8% increase over their baseline looking [$t(17) = 4.06$, $P < 0.001$]. Finally, for the four-channel speech, children looked toward the target object 57.7% of the time, a 7.7% increase over baseline looking [$t(17) = 2.63$, $P = 0.018$]; this is similar to that in Experiment 2, albeit significant rather than marginal. A repeated-measures one-way ANOVA showed that the effect of condition was significant, $F(2,34) = 4.50$, $P < 0.02$, suggesting that the children's looking performance differed across the three conditions. Follow-up paired comparisons show no difference between the full-speech and eight-channel speech conditions, replicating the pattern in Experiment 1 [$t(17) = 1.19$, $P > 0.20$]. There was a significant difference between the full-speech and four-channel speech condition [$t(17) = 3.15$, $P < 0.006$], as shown in Fig. 4. However, the difference between the

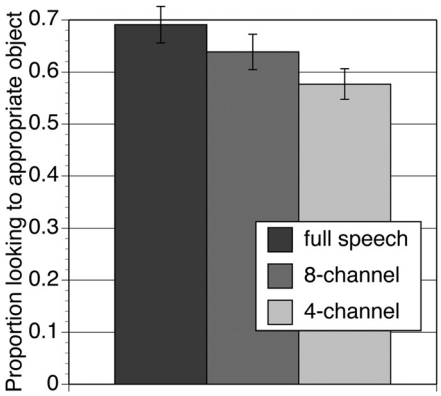

FIG. 4. Proportion of time spent looking to the target object in full speech, 8-channel noise-vocoded speech, and 4-channel noise-vocoded speech in Experiment 3. Baseline performance was approximately 50% in all cases.

eight- and four-channel conditions was only marginal, $t(17) = 1.85$, $P = 0.08$ (two-tailed).

Thus as in Experiment 2, there is some indication that children successfully recognized speech in the four-channel condition, but their performance is quite weak. It is not clear whether this is the result of a consistent pattern of just-above-chance performance shown by most children or whether the significant effect is being driven by a smaller number of participants who are able to successfully recognize speech at this level of spectral degradation. Figure 5 shows individual data points for each participant in the various experiments and conditions. Most of the participants appear to hover around or just over the 50% looking mark in the four-channel condition (both in the present experiment and in Experiment 2); a small number of participants appear to show above-chance looking and could potentially be driving the significant effect. However, they are not statistical outliers and thus could simply reflect the extremes of a normal distribution. Either way, the four-channel condition appears to be on the borderline of what children can do.

This raises the question as to what abilities might be tied to children's success in this condition. Vocabulary does not seem to be the driving factor in children's performance; while we did find positive, but not significant, correlations for both the full-speech condition ($r = 0.23$, $P = 0.18$) and eight-channel condition ($r = 0.36$, $P = 0.07$), we did not find such relationships for the four-channel condition, and indeed the trend went in the opposite direction ($r = -0.23$, $P > 0.50$), replicating the lack of such an effect in Experiment 2. While there is some suggestion across experiments that vocabulary might be a factor in performance more generally, it does not appear to be a factor in performance in the four-channel condition. Identifying what factors allow particular children to succeed at this level of degradation should be an important area for future research.

2. Average time course of looking and reaction time measures

As in Experiment 1, children began looking consistently to the target object after the onset of the first repetition of the target word as shown in Fig. 6. Averaging across all participants, looking time was significantly greater than chance (based on a one-tailed *t*-test across at least three successive frames) by 933 ms post word-onset in the full-speech condition (compared to 567 ms in Experiment 1). This same reference point occurred at 1467 ms in the eight-channel condition (compared to 1300 ms in Experiment 1). That is, similar to Experiment 1, children as a group were slower to look correctly in the eight-channel condition than in the full-speech condition. Accuracy was in general much lower for the four-channel condition, but, surprisingly, reached significance slightly sooner by 1133 ms. However, looking time to the target object did not stay high in this condition, dropping back down to nonsignificant levels after only eight frames, whereas it continued to remain high for both the full speech (remaining significantly above chance for 98 frames) and the eight-channel condition (54 frames).

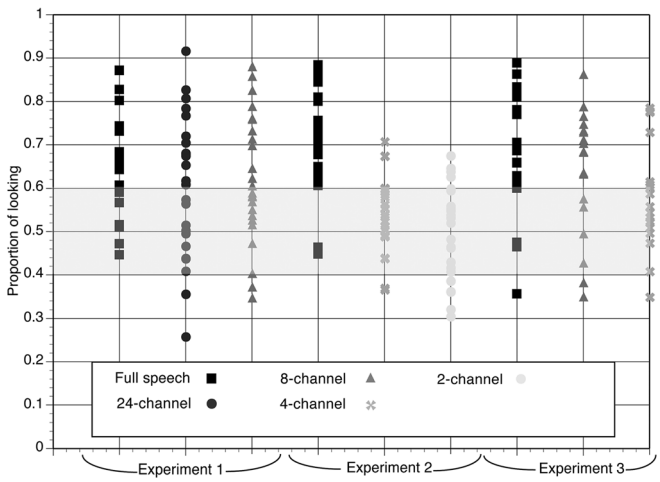

FIG. 5. Individual participants' proportion of looking time to the appropriate object in the three experiments and the different speech conditions. Shaded region represents 40% to 60% looking to the target object or roughly chance performance.

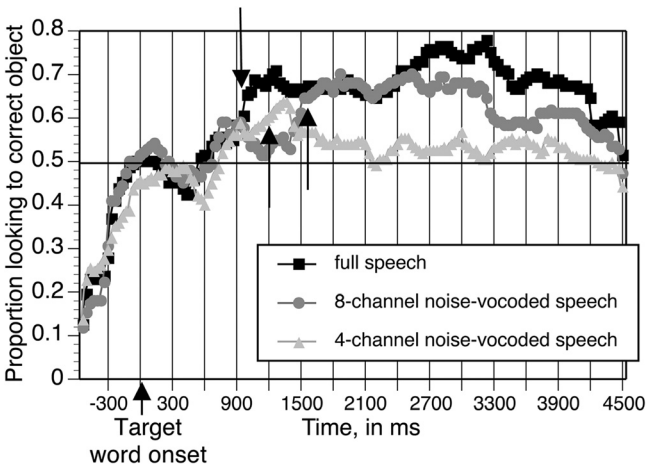

FIG. 6. Looking time to the target object as a function of time, for full speech, 8-channel noise-vocoded speech, and 4-channel noise-vocoded speech in Experiment 3. Arrows indicate the points at which each curve shows looking time significantly greater than chance.

This ambiguity for the four-channel condition may not be surprising because the preceding method of analysis averages across all children, and (as noted earlier) not all children seemed able to accurately perceive the four-channel noise-vocoded speech. Indeed, Fernald's method of individual participant reaction time analysis found somewhat different results. By this method of analysis, the average time for the first shift to the target object occurred at 640 ms in the full-speech condition, quite comparable to the 641 ms time found in Experiment 1. However, in Experiment 1, the shift to the eight-channel condition occurred at 767 ms, whereas it occurred at 625 ms here. Times for the four-channel condition were slower, at 891 ms. As in Experiment 1, this RT difference across the three conditions only approached significance, $F(2,36) = 3.04$, $P = 0.060$. Follow-up t-tests showed that the four-channel condition was significantly slower than either of the less-degraded conditions [full vs 4, $t(13) = 2.17$, $P < 0.05$; eight- vs four-channel, $t(15) = 2.78$, $P < 0.02$], which did not differ from one another [$t(13) = 0.18$, $P > 0.50$].

We decided to look at the RTs for only the four children who showed high proportions of looking in the four-channel condition. This analysis must be treated with caution as it is based on a very limited set of data; from those four children, only two had a trial that could be analyzed in the full-speech condition, and three had a trial that could be analyzed in the four-channel condition, although all contributed data to the eight-channel condition. With so few data points, we are unable to perform a statistical comparison. However, average reaction times were 575 ms for the full-speech condition, but 750 ms for the eight-channel and 717 for the four-channel, suggesting that for these children, reaction times to the two vocoded stimuli may be similar.

Thus in general, there is some evidence to suggest that children's reaction times were generally slower for the four-channel condition than for the other two, although this may have been less so for those children who were most successful at comprehending these items.

V. GENERAL DISCUSSION

In all three studies, toddlers were presented with images of two objects at a time and heard a voice telling them which object to look at. When the voice was presented in normal speech, the children were highly successful, as expected. When the speech was noise-vocoded, they were also highly successful, as long as the vocoded speech included either 8 or 24 channels. When the vocoded speech was based on two channels, the children failed to demonstrate recognition of the signal; when it was based on four channels, children's performance was more variable, showing a significant, but very small, degree of success.

Moreover, this high level of performance with 8- and 24-channel noise-vocoded speech occurred with no prior training on this type of signal and was apparent even on the earliest test trials. Thus our initial set of studies has demonstrated that toddlers have a surprising ability to interpret vocoded speech with very little prior exposure and can do so with speech of as few as eight, or possibly four, channels.

Yet it remains unclear how this ability changes as children develop. Our studies used only children of a single age, and the task (selecting the appropriate object from two phonetically dissimilar choices) is difficult to compare directly with the types of tasks typically employed with older children and adults (identification of open-set words and sentences). Yet the results to date suggest that there is some developmental change occurring. Adults typically perform very well with as few as four channels, even in tasks more difficult than the one presented here. Moreover, Eisenberg *et al.* (2000) reported differences between children aged 5–7 yr and those aged 10–12 yr, and Nittrouer and colleagues (Nittrouer and Lowenstein, 2010; Nittrouer *et al.*, 2009) reported that 7-yr-olds were less adept than adults at recognizing words with such signals. These findings suggest that there may be improvements in recognizing spectrally degraded speech that occur as children gain more experience with their language. Thus one direction for future work would be to explore how this ability changes developmentally during early childhood.

Another direction for the future would be to compare children's performance with noise-vocoded speech with that of sine-wave speech (Remez *et al.*, 1981). Sine-wave speech replaces the time-varying resonance bands produced by a human vocal tract with three time-varying sinusoids. As in noise-vocoded speech, this preserves much of the time-varying structure of the speech signal but removes all harmonic structure. However, the resulting quality of the signal is quite different, and in addition, normal-hearing listeners attending to such speech have additional access to spectral cues (for instance, sidebands resulting from the envelope modulation of the sinewaves) that cochlear implant patients would not be able to hear. Regardless, sine-wave speech provides the listener with a different kind of sparse spectral representation of the signal. Comparing performance when different aspects of the speech signal are removed or preserved might provide clues to which properties of the speech signal attract children's attention (see Nittrouer *et al.*, 2009).

Finally, noise-vocoded speech removes some of the acoustic cues that signal aspects of prosody, such as intonation—particularly fundamental frequency cues. While other cues to intonation remain, such as duration and intensity cues, these are typically secondary cues for typical listeners (Denes and Milton-Williams, 1962), and listeners show poorer discrimination of prosodic information in noise-vocoded speech (Peng *et al.*, 2009). While intonational differences are of relatively little importance in the single-syllable English words presented here, they are of immense importance in words in other languages (particularly tone languages) and in interpreting larger units such as sentences. Intonational cues serve to distinguish English questions from statements and to provide information on stress and sentential focus. They also serve to aid in the separation of simultaneous talkers (Brokx and Nooteboom, 1982). It is possible that infants and toddlers may have greater difficulty benefitting from secondary cues to prosodic information than do older children and adults. Examining children's ability to interpret noise-vocoded sentences, particularly sentential

distinctions for which prosodic information plays a critical role, is yet another important area of future study.

The ability to recognize spectrally degraded speech is a critical skill for successfully using a cochlear implant. While children could presumably learn this skill after implantation, having the skill already in place would likely assist a child in learning to use his or her implant appropriately. The present results suggest that children can interpret speech with as few as eight channels but have difficulty with speech that is degraded beyond that point. Although many cochlear implants putatively have more channels than this, cross-channel interference and loss of auditory neurons limits the number of separate channels that listeners can utilize. Indeed, the literature suggests that adult CI listeners are getting at best eight channels of spectral information (Friesen *et al.*, 2001). The average CI listener may be operating at less than this, perhaps around six channels. More critically, some CI listeners may be receiving as few as four channels (Friesen *et al.*); the present results suggest that at least some normally developing young children might have difficulties interpreting speech signals appropriately if the number of channels they could utilize was this low. Considering the fact that children in the present study have had normal auditory/linguistic development since birth, the results presented here may well represent the best-case scenario. As discussed in Sec. I, however, it is also important to remember that early implanted CI children would have the advantage of developing with electric hearing and may be able to overcome some limitations of the device. Results emerging from the population of CI children suggest that *some* children are able to perform at levels comparable to those of their normally hearing peers, both in expressive and receptive language (e.g., Niparko *et al.*, 2010). Moreover, Grieco-Calub and colleagues (2009) recently tested children with CIs on a task (and at an age) quite similar to the one employed here and found that these children were successful at recognizing words in quiet settings, albeit to a lesser degree (and with longer response times) than typically developing chronologically age-matched children.

An important finding of Grieco-Calub *et al.*'s work is that young children with CIs required longer processing times than did their normally hearing peers. This corresponds to the findings here, in which children had longer response times with spectrally degraded speech. These findings suggest that more cognitive resources may be needed to process everyday speech by the pediatric CI population. In demanding environments such as pre-school or classrooms, these children may have to use more resources per task, which might place them at considerable disadvantage relative to their normally hearing peers.

Thus understanding the limits of young children's spectral processing has clear implications for current recommendations regarding implantation of young children. But recognizing such speech is important for many other individuals as well. While noise-vocoded speech, as presented here, is not a naturally occurring signal, spectral degradation in general is a common consequence of hearing impairment, and finding that children can recognize speech despite such degradation is relevant to our understanding of the limitations of this group as well. It appears that even at a very young age, listeners are able to adjust their perception in response to fairly severe signal degradation, demonstrating the impressive flexibility of human auditory processing.

VI. CONCLUSION

Toddlers aged 27 months can accurately recognize noise-vocoded speech of as few as eight channels with virtually no prior training, although they may require longer processing times to do so. They show mixed results with four-channel noise-vocoded speech, however, as many children failed to recognize known words when the number of channels was this low. While such results demonstrate young listeners' remarkable ability to interpret severely degraded speech signals, they also have important implications for decisions regarding implanting young children.

ACKNOWLEDGMENTS

The authors thank George Hollich for the SUPERCODER program, and thank Daniel Eisenberg and Tracy Moskatel for stimulus creation. We thank Amanda Pasquarella, Justine Dombroski, Molly Nasuta, and Lauren Evans for overseeing substantial parts of the coding reported here and particularly thank Giovanna Morini for lab and project oversight. We also thank the following additional students for assistance either in scheduling or testing participants or coding looking time performance: Katrina Ablorh, Candace Ali, Alison Arnold, Alyssa Cook, Jennifer Coon, Sara Dougherty, Sara Edelberg, Lauren Fischer, Arielle Gandee, Laura Horowitz, Megan Janssen, Mina Javid, Amanda Jensen, Esther Kim, Stephanie Lee, Rachel Lieberman, Perri Lieberman, Eileen McLaughlin, Kelly McPherson, Vidda Moussavi, Courtenay O'Connor, Sabrina Panza, Elise Perkins, Rachel Rhodes, Allie Rodriguez, Krista Voelmle, Chelsea Vogel, Amanda Wildman, Kimmie Wilson, and TeHsin Wu. This work was supported by NIH Grant No. R01 DC004786 and by NSF Grant No. BCS0642294 to the University of Maryland.

[1]Comparing looking to the alternative object is an appropriate choice when using raw looking times (in seconds), but doing so when using proportions artificially increases the apparent size of an effect. If a child looks 50% of the time to each object in a baseline period and looks 60% of the time to the object when told to find it, it seems more appropriate to compare 60% to 50% (what they would have done otherwise, a 10% preference) than to compare the 60% target looking to the 40% nontarget looking (resulting in an apparent 20% looking preference). Because baseline trials are paired in the current study, comparing trials to baseline should be roughly comparable to comparing to the hypothetical baseline of 50%; however, there can be slight differences, particularly when an infant does not attend on every trial.

[2]Had we not used our baseline trials as the standard of comparison and instead compared looking to a putative baseline of 50%, we would have found the same pattern of results.

Baskent, D., and Shannon, R. V. (**2003**). "Speech recognition under conditions of frequency-place compression and expansion," J. Acoust. Soc. Am. **113**(4), 2064–2076.

Baskent, D., and Shannon, R. V. (**2006**). "Frequency transposition around dead regions simulated with a noiseband vocoder," J. Acoust. Soc. Am. **119**(2), 1156–1163.

Baskent, D., and Shannon, R. V. (**2007**). "Combined effects of frequency compression-expansion and shift on speech recognition," Ear Hear. **28**(3), 277–289.
Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon. **10**, 23–36.
Chatterjee, M., and Peng, S. C. (**2008**). "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," Hear. Res. **235**, 143–156.
Cole, R. A., and Perfetti, L. A. (**1980**). "Listening for mispronunciations in a children's story: The use of context by children and adults," J. Verbal Learn. Verbal Behav. **19**, 297–315.
Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**(2), 222–241.
Denes, P., and Milton-Williams, J. (**1962**). "Further studies in intonation," Lang. Speech **5**(1), 1–14.
Drgas, S., and Blaszak, M. A. (**2009**). "Perceptual consequences of changes in vocoded speech parameters in various reverberation conditions," J. Speech Lang. Hear. Res. **52**, 945–955.
Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., and Boothroyd, A. (**2000**). "Speech recognition with reduced spectral cues as a function of age," J. Acoust. Soc. Am. **107**(5), 2704–2710.
Elliott, L. L., Busse, L. A., Partridge, R., Rupert, J., and DeGraaff, R. (**1986**). "Adult and child discrimination of CV syllables differing in voicing onset time," Child Dev. **57**, 628–635.
Elliott, L. L., Hammer, M. A., Scholl, M. E., and Wasowicz, J. M. (**1989**). "Age differences in discrimination of simulated single-formant frequency transitions," Percept. Psychophys. **46**, 181–186.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., and Pethick, S. J. (**1994**). "Variability in early communicative development," Monogr. Soc. Res. Child Dev. Ser. **242**(59, 5), 1–173.
Fernald, A., and Hurtado, N. (**2006**). "Names in frames: Infants interpret words in sentence frames faster than words in isolation," Dev. Sci. **9**(3), F33–F40.
Fernald, A., Perfors, A., and Marchman, V. A. (**2006**). "Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year," Dev. Psychol. **42**(1), 98–116.
Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., and McRoberts, G. W. (**1998**). "Rapid gains in speed of verbal processing by infants in the 2nd year," Psychol. Sci. **9**(3), 228–231.
Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**(2), 1150–1163.
Fu, Q. J., and Shannon, R. V. (**1999**). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," J. Acoust. Soc. Am. **105**(3), 1889–1900.
Fu, Q. J., Shannon, R. V., and Wang, X. S. (**1998**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**(6), 3586–3596.
Gilley, P. M., Sharma, A., and Dorman, M. F. (**2008**). "Cortical reorganization in children with cochlear implants," Brain Res. **1239**, 56–65.
Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., and Gordon, L. (**1987**). "The eyes have it: Lexical and syntactic comprehension in a new paradigm," J. Child Lang. **14**, 23–45.
Grieco-Calub, T. M., Saffran, J. R., and Litovsky, R. Y. (**2009**). "Spoken word recognition in toddlers who use cochlear implants," J. Speech Lang. Hear. Res. **52**, 1390–1400.
Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., and Carlyon, R. P. (**2008**). "Perceptual learning of noise vocoded words: Effects of feedback and lexicality," J. Exp. Psychol. Hum. Percept. Perform. **34**(2), 460–474.
Hollich, G. (2005). "SUPERCODER: A program for coding preferential looking," Version 1.5, Purdue University, West Lafayette, IN.
Ihlefeld, A., Deeks, J. M., Axon, P. R., and Carlyon, R. P. (**2010**). "Simulations of cochlear-implant speech perception in modulated and unmodulated noise," J. Acoust. Soc. Am. **128**(2), 870–880.
Meints, K., Plunkett, K., and Harris, P. L. (**1999**). "When does an ostrich become a bird? The role of typicality in early word comprehension," Dev. Psychol. **35**(4), 1072–1078.
Naigles, L. G., and Gelman, S. A. (**1995**). "Overextensions in comprehension and production revisited: Preferential-looking in a study of *dog, cat,* and *cow*," J. Child Lang. **22**, 19–46.
Niparko, J. K., Tobey, E. A., Thal, D. J., Eisenberg, L. S., Wang, N. Y., Quittner, A. L., Fink, N. E., and CDaCI Investigative Team. (**2010**). JAMA **303**(15), 1498–506, Spoken language development in children following cochlear implantation, http://www.ncbi.nlm.nih.gov/pubmed/20407059.
Newman, R. S. (**2006**). "Perceptual restoration in toddlers," Percept. Psychophys. **68**, 625–642.
Newman, R. S. (**2011**). "2-year-olds' speech understanding in multi-talker environments," Infancy **16**(5), 447–470.
Nittrouer, S. (**1992**). "Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries," J. Phonet. **20**(3), 351–382.
Nittrouer, S. (**1996**). "Discriminability and perceptual weighting of some acoustic cues to speech perception by 3-year-olds," J. Speech Hear. Res. **39**, 278–297.
Nittrouer, S., and Boothroyd, A. (**1990**). "Context effects in phoneme and word recognition by young children and older adults," J. Acoust. Soc. Am. **87**(6), 2705–2715.
Nittrouer, S., and Lowenstein, J. H. (**2010**). "Learning to perceptually organize speech signals in native fashion," J. Acoust. Soc. Am. **127**(3), 1624–1635.
Nittrouer, S., Lowenstein, J. H., and Packer, R. R. (**2009**). "Children discover the spectral skeletons in their native language before the amplitude envelopes," J. Exp. Psychol. Hum. Percept. Perform. **35**, 1245–1253.
Peng, S.-C., Lu, N., and Chatterjee, M. (**2009**). "Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners," Audiol. Neurotol. **14**(5), 327–337.
Polka, L., Rvachew, S., and Molnar, M. (**2008**). "Speech perception by 6- to 8-month-olds in the presence of distracting sounds," Infancy **13**(5), 421–439.
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.
Rescorla, L. (**1989**). "The Language Development Survey: A screening tool for delayed language in toddlers," J. Speech Hear. Disord. **54**(4), 587–599.
Samuel, A. G. (**1981**). "Phonemic restoration: Insights from a new methodology," J. Exp. Psychol. Gen. **110**, 474–494.
Schafer, G., and Plunkett, K. (**1998**). "Rapid word learning by fifteen-month-olds under tightly controlled conditions," Child Dev. **69**(2), 309–320.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.
Shannon, R. V., Zeng, F. G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**(4), 2467–2476.
Sharma, A., Dorman, M. F., and Kral, A. (**2005**). "The influence of a sensitive period on central auditory development in children with unilateral and bilateral cochlear implants," Hear. Res. **203**, 134–143.
Sheldon, S., Pichora-Fuller, M. K., and Schneider, B. A. (**2008a**). "Effect of age, presentation method, and learning on identification of noise-vocoded words," J. Acoust. Soc. Am. **123**(1), 476–488.
Sheldon, S., Pichora-Fuller, M. K., and Schneider, B. A. (**2008b**). "Priming and sentence context support listening to noise-vocoded speech by younger and older adults," J. Acoust. Soc. Am. **123**(1), 489–499.
Swingley, D., and Aslin, R. N. (**2000**). "Spoken word recognition and lexical representation in very young children," Cognition **76**, 147–166.
Swingley, D., Pinto, J. P., and Fernald, A. (**1999**). "Continuous processing in word recognition at 24 months," Cognition **71**, 73–108.
Tajudeen, B. A., Waltzman, S. B., Jethanamest, D., and Svirsky, M. A. (**2010**). "Speech perception in congenitally deaf children receiving cochlear implants in the first year of life," Otol. Neurotol. **31**, 1254–1260.
Walley, A. C. (**1987**). "Young children's detections of word-initial and -final mispronunciations in constrained and unconstrained contexts," Cognit. Dev. **2**, 145–167.
Warren, R. M., and Obusek, C. J. (**1971**). "Speech perception and phonemic restorations," Percept. Psychophys. **9**(3-B), 358–362.
Warren, R. M., Reiner Hainsworth, K., Brubaker, B. S., Bashford, J. A., Jr., and Healy, E. W. (**1997**). "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," Percept. Psychophys. **59**(2), 275–283.
Werker, J. F., and Tees, R. C. (**1984**). "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," Infant Behav. Dev. **7**, 49–63.
Werner, L. A., and Bargones, J. Y. (**1992**). "Psychoacoustic development of human infants," Adv. Infancy Res. **7**, 103–145.