Toddlers' comprehension of degraded signals: Noise-vocoded versus sine-wave analogs

Rochelle S. Newman
Department of Hearing and Speech Sciences, 0100 Lefrak Hall, University of Maryland, College Park, Maryland 20742, USA
Rnewman1@umd.edu

Monita Chatterjee
Boys Town National Research Hospital, 555 North 30th Street, Omaha, Nebraska 68131, USA
monita.chatterjee@boystown.org

Giovanna Morini
Department of Hearing and Speech Sciences, 0100 Lefrak Hall, University of Maryland, College Park, Maryland 20742, USA
gmorini@udel.edu

Robert E. Remez
Barnard College, Columbia University, 3009 Broadway, New York, New York 10027, USA
remez@columbia.edu

Abstract: Recent findings suggest that development changes the ability to comprehend degraded speech. Preschool children showed greater difficulties perceiving noise-vocoded speech (a signal that integrates amplitude over broad frequency bands) than sine-wave speech (which maintains the spectral peaks without the spectrum envelope). In contrast, 27-month-old children in the present study could recognize speech with either type of degradation and performed slightly better with eight-channel vocoded speech than with sine-wave speech. This suggests that children's identification performance depends critically on the degree of degradation and that their success in recognizing unfamiliar speech encodings is encouraging overall.

© 2015 Acoustical Society of America
[AC]
Date Received: April 18, 2015 **Date Accepted:** August 17, 2015

1. Introduction

Much of the speech we hear around us is not perfectly clear; for example, it may occur in the presence of reverberation or be masked by background noise. It may also be degraded, such as speech over a poor cell phone connection, or through a cochlear implant. A great deal of research has investigated listeners' ability to compensate for these types of degradations. Much of this research has relied on signals that have been artificially degraded in some manner. Two such examples are noise-vocoded speech and sine-wave analogs to speech.

Noise-vocoded (henceforth NV) speech typically provides a coarse grain of spectral detail due to the summation of spectrotemporal variation in broad frequency bands (Shannon *et al.*, 1995). This method is used to simulate, for a normal-hearing listener, speech spectra delivered through a cochlear implant. To create a noise-vocoded analog of a speech signal, the signal is divided into a set of separate frequency bands; the narrower and more numerous the frequency bands, the more the noise-vocoded signal represents the spectral detail of the original incident speech. The overall amplitude envelope of the signal is then extracted from each band, and these extracted amplitude envelopes are then used to modulate bands of (initially equal-amplitude) noise that are centered over the same frequency regions. These noise bands are then combined to create the complete signal. Normal-hearing adults can tolerate extensive frequency blur imposed by the technique of vocoding and perceive the spoken message of the original speech in noise-vocoded signals with as few as four 1-kHz wide bands or channels spanning the speech spectrum (Shannon *et al.*, 1995).

In contrast to NV speech, sine-wave analogs to speech (SWS) maintain the global dynamic spectral structure of the peaks but not the valleys of the power spectrum. To create sine-wave analogs, the first three or four formants (or resonant energy bands) in the original speech signal are each replaced with a time-varying sinewave

(Remez *et al.*, 1981). This results in a signal that lacks the acoustic signature of the resonant properties of the human vocal tract but maintains the time-varying spectral properties of the peaks in the signal. Despite this degradation, these signals can also be comprehended quite well by adult listeners.

In some sense, these two signals are degraded in complementary ways, one blurring the spectral details and the other sharpening them. Yet while neither sounds like normal speech, both can be understood quite well by adult listeners after minimal exposure. School-age children also succeed at listening to degraded signals. Eisenberg *et al.* (2000) found that children aged 10–12 yr performed quite similarly to adult listeners when listening to NV speech. In contrast, children aged 5–7 yr required more spectral bands than did older listeners to reach the same level of comprehension. Newman and Chatterjee (2013) found that toddlers aged 27 months showed recognition of eight-channel NV speech at levels approaching that for full speech, although they were more variable with four-channel NV speech.

These latter two developmental studies examined NV speech in particular as this form of degradation seems most relevant for children listening through a cochlear implant. Yet Nittrouer and colleagues (Nittrouer and Lowenstein, 2010; Nittrouer *et al.*, 2009) found that the ability to interpret these NV speech and sine-wave analogs had different developmental time courses. Relative to adults, children aged 3–7 yr had much greater difficulty with NV speech than with sine-wave speech. The authors interpreted these results as indicating that children rely on dynamic spectral information to a greater extent than do adults and thus have particular difficulty with NV signals that lack such information. Or, to put it another way, children learn initially to focus on spectral information and thus "learn to extract linguistic form from signals that preserve some spectral structure" earlier in development (Nittrouer *et al.*, 2009, p. 1245). Such results have implications for children hearing speech through cochlear implants, at least if normal-hearing children's performance with NV speech serves as a model for children's performance with an actual implant. Moreover, such comparisons have the potential to inform us about the development of the speech perception mechanism and of the cues that listeners of different ages attend to preferentially.

Yet these findings seem somewhat at odds with results from Newman and Chatterjee, who showed that toddlers were successful at perceiving NV speech without prior experience. This difference may be the result of the number of channels used in the different studies; Nittrouer *et al.* (2009) tested children with four-channel NV speech, whereas Newman and Chatterjee (2013) found much stronger performance with eight-channel NV speech. But the argument of Nittrouer suggests the need for directly comparing toddlers on their ability to recognize NV speech and sine-wave analogs to speech, the focus of the current paper. The goal of this investigation was to examine both the developmental time-course for recognizing degraded speech signals more generally and to estimate the extent to which toddlers succeed using the sharply resolved spectral detail in SWS in contrast to the frequency blur imposed on the speech spectrum by NV speech.

2. Experiment

2.1 Method

2.1.1 Participants

Twenty-four toddlers (14 male, 10 female), aged 27 months (range: 25 months, 25 days to 28 months, 2 days) participated. Parents reported that their children had normal hearing, were not currently experiencing symptoms indicative of an ear infection, and heard at least 90% English in the home. An additional seven children participated, but their data were excluded for excessive fussiness ($n = 5$), low English percentage ($n = 1$), or equipment difficulties ($n = 1$).

2.1.2 Test materials

Test items included an audible component (sentences that were either nondegraded, noise-vocoded at eight channels or a sine-wave analog) and a visible component (still pictures of well-known objects). Four objects were presented in pairs on test trials (keys/blocks; car/ball) with an additional pair used for practice trials (cat/dog); all were matched for size and color-scheme.

The nondegraded audio test items were spoken by a single female talker, recorded over a Shure SM51 microphone at a 44.1 kHz sampling rate and 16 bits precision. Sentences instructed the child to attend to a particular object ("Look at the ___! Can you find the ___? See the ___?") or on baseline trials told the child to look more generally ("Look at that! Do you see that? Look over there!"). Sentences were matched for duration (4.8 s) and average root-mean-square amplitude.

Noise vocoding was performed using methods akin to published standards (Shannon *et al.,* 1995) using ANGELSIM v. 8.1 (Tigerspeech Technology, Qian-Jie Fu, House Ear Institute) with eight channels. The nondegraded audio stimuli were first bandpassed to limit the input range to that between 0.2 and 7.0 kHz and were then divided into eight bands (Butterworth filters, 24 dB/octave rolloff); 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; these bands were then combined at equal amplitude ratios to make the noise-vocoded stimuli.

Sine-wave analogs to speech were created by hand using the procedures described in Remez *et al.* (2011). This approach was selected because linear prediction estimates of formant frequency, bandwidth, and amplitude often are erroneous when the spectrum changes rapidly or discontinuously and are not suitable as a basis for speech synthesis.

2.1.3 Procedure

A child sat on a caregiver's lap facing a widescreen television. At the start of each trial, an image of a laughing baby appeared in the center of the screen to attract the child's attention. Subsequently, two images appeared, separated by approximately 20° visual angle, along with a simultaneous audio sequence.

The first two trials, using the words "cat" and "dog," were considered practice and were presented in nondegraded speech. On one of these two trials, the correct answer appeared on the left, and on the other, it appeared on the right. This was followed by 14 test trials: 4 test trials in the nondegraded condition (one requesting that the child look at each of the four test objects), 4 in the eight-channel noise-vocoded condition, 4 in the sine-wave analog condition, and 2 baseline nondegraded trials for comparison purposes (see procedure); baseline trials just told the children to look at the objects in general, but did not name them (see Sec. 2.1.2), and one such baseline trial occurred for each pair of objects. These 14 trials were presented in pseudo-random order with the restriction that the correct response did not occur on the same side more than three trials in a row. All trial types were intermixed rather than blocked. Participants were presented with one of six different trial orders, which counterbalanced for which image appeared on the left vs right. Audible sentences began simultaneously with the visible items with the target word first appearing 600 ms (18 frames) into the sentence; looking during these initial 18 frames was excluded from data analysis.

The caregiver listened to masking music over headphones throughout the study to prevent any biasing of a child's behavior, and each caregiver completed a Language Development Survey (Rescorla, 1989) as a measure of the child's productive vocabulary. A digital camera recorded each child at a rate of 30 frames per second. Two experimenters, blind to condition, individually coded each child's gaze direction on a frame-by-frame basis using SUPERCODER coding software (Hollich, 2005). From this, an infant's total duration of looking at each of the two images on each trial was calculated. A third coder coded any trial on which the two researchers disagreed by more than 15 frames (0.5 s); this occurred on 28 trials across the 24 participants (or 8.3% of trials). These average data were used to calculate an infant's total duration of looking at each of the two images; we expect greater looking to the correct (named) image than the opposite image if children understand the speech despite any signal degradation.

2.2 Results and discussion

We examined children's looking for each of the three conditions individually, calculating the proportion of time a child spent looking at each object when named, from target word onset, and subtracting the proportion of time the child spent looking at the object on baseline trials. This difference was averaged across the four objects in the study (car, ball, blocks, keys), and compared to zero using a single-sample *t*-test, with a critical *p*-value of 0.05. We refer to the difference in proportion looking time as the increase over baseline looking (see Fig. 1).

For the full-speech condition, children looked toward the target object 21.7% over their baseline looking [SD = 9.6; $t(23) = 11.05$, $p < 0.0001$]; see Fig. 1. For the eight-channel NV speech, children showed an 18% increase over baseline looking [SD = 10.1; $t(23) = 8.75$, $p < 0.0001$]. These two conditions were not significantly different from one another [$t(23) = 1.47$, $p = 0.16$], replicating the pattern of results found in

Newman and Chatterjee (2013). However, the trend was for children to show slightly better performance in full-speech than NV speech.

The critical condition, however, is the SWS. Here children showed a 12% increase over baseline looking [SD = 11.4; $t(23) = 5.16$, $p < 0.0001$]. Although this was significantly above chance, it was not as strong as children's performance in either of the other two conditions [vs full-speech, $t(23) = 4.03$, $p = 0.0005$; vs NV, $t(23) = 3.50$, $p = 0.002$].

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. However, the children performed significantly more poorly with the SWS than with the NV speech, a pattern opposite than predicted by Nittrouer *et al.* (2009).

One concern is that given the random order of the trials, children's success with NV and/or SWS may have been the result of direct comparison across trials, particularly if they had heard the target words in the full speech condition first. In each of the trial orders, one of each of the four target words (keys, blocks, car, or ball) occurred for the first time in SWS, another occurred for the first time in NV speech, and a third occurred for the first time in full speech; the fourth varied across the orders. We therefore decided to look separately at the first occurrence of each target word—this instance could not benefit from a cross-trial comparison of the particular recording. Here we found a somewhat surprising pattern of results. Children were still significantly above chance in all three conditions [full speech: 14.6% increase in looking, $t(23) = 4.32$, $p < 0.0005$; SWS: 14.5% increase in looking, $t(23) = 2.75$, $p = 0.011$; NV: 25.5% increase in looking, $t(23) = 9.41$, $p < 0.0001$]. But performance was strongest in the first NV trial, and the only significant difference was between that and the full speech condition [full vs NV: $t(23) = 2.48$, $p = 0.02$; full vs SWS: $t(23) = 0.01$, $p > 0.90$; SWS vs NV: $t(23) = 1.76$, $p = 0.092$]. While this pattern of results is unexpected, it does suggest two inferences: first, children's successful identification in these NV and SWS trials suggests that they were able to recognize the appropriate referent for these degraded signals even when they had not previously heard the sentence in a full-speech condition. Second, even without practice, children at this age find it easier to recognize novel words in NV speech than in SWS.

Our final analysis examined whether there was a difference across items that might be informative. To do this, we calculated the proportion of time the child spent looking to the correct object when it was named minus that spent looking to that same object on the baseline trials; we measured this individually for each word, in each condition, for each participant and then examined the effect that degradation had on the varying words. Results from this analysis are shown in Table 1. Degradation had the least impact (relative to full speech) on the word car: children looked to the correct object 18.9% of the time in the full speech, 15.7% of the time in SWS, and 19.4% of

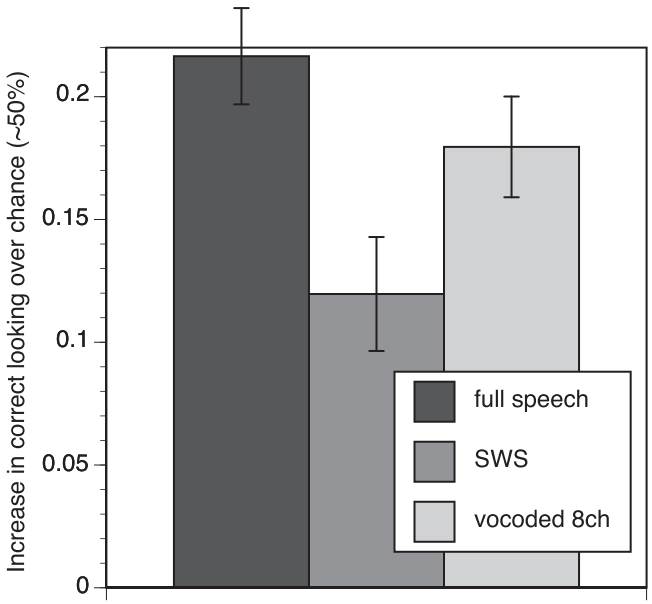

Fig. 1. Increase in looking to target when named, compared to baseline, in full speech, sine-wave analogs to speech, and eight-channel noise-vocoded speech.

Table 1. Accuracy by item in the various conditions; *ball* and *car* are paired in testing as are *blocks* and *keys*.

	Increase over baseline				Effect of degradation	
	Full	SWS	NV	*Average*	Full - SWS	Full - NV
Ball	16.9	10.2	14.5	*13.9*	6.7	2.4
Car	18.9	15.7	19.4	*18.0*	3.2	−0.4
Blocks	30.6	12.6	27.2	*23.5*	18.0	3.4
Keys	21.3	6.7	10.8	*12.9*	14.6	10.5

the time in NV. The change to SWS resulted in decrements for both blocks and keys (difference between SWS and full speech of 18.0% and 14.6%, respectively), suggesting that these two words were harder to distinguish from one another in SWS; in contrast, SWS had far smaller effects on ball and car (3.2% and 6.7%). A similar pattern occurred in NV speech but to a lesser degree. In general, while ball and car showed lower accuracy in the full speech condition, they did not appear to *lose* their discriminability when degraded. In contrast, blocks and keys were highly discriminable in full speech but were more affected by degradation, particularly the degradation in SWS.

In general, children looked significantly longer at the named object than would be expected by chance in all three conditions. However, they performed significantly more poorly with the SWS than with the NV speech; while young children can comprehend SWS with no prior experience, they do less well with this type of degradation than with 8-channel vocoding.

3. Final discussion

In this study, toddlers were presented with two images and heard a voice telling them which object to look at. On some trials, the speech was degraded, and we examined the effect of that degradation on children's looking behavior. This provides an indication of the perceptual standards on which young listeners rely in recognizing spoken words. Not surprisingly, when the speech was not degraded, children appeared to recognize the target words, showing increased looking relative to the baseline trials. We also replicated findings from Newman and Chatterjee (2013), suggesting that children looked nearly as long at the appropriate object with eight-channel NV signals.

Interestingly, children also successfully recognized SWS. This is somewhat surprising because even adults generally need a moment of exposure and instruction to recognize these signals as speech (Remez *et al.,* 1981). While the listening context (and the within-subjects design) likely primed children to interpret the sounds in a speech mode, it is nonetheless surprising that listeners with so little experience with language in general would spontaneously group the three tones in a manner allowing for comprehension of a speech signal. This finding supports the notion that listeners have at their disposal a phonetic mode of perception (see Remez, 2005) and suggests that this mode is in place at a very young age.

Although children were quite successful with sine-wave analogs, they did not perform as well in this condition as in the NV condition. This is also surprising, given prior findings from Nittrouer and Lowenstein (2010) that children (including 3-yr-olds, only roughly 1–1.5 yr older than the toddlers tested here) perform substantially better with SWS than with NV speech. The authors suggest that their findings "could have important implications for how we intervene with individuals, particularly children, who have hearing loss," suggesting that "it may be worth exploiting…residual hearing in order to provide at least the first formant to these individuals through a hearing aid, and this practice could be especially beneficial for young children who rely heavily on dynamic spectral structure" (p. 1633). However, the frequency blur imposed by an eight-band vocoder on a speech spectrum obscures the spectral detail of the formant pattern, yet our findings show that intelligibility of degraded speech is possible despite the absence of an acoustic component presenting a well-resolved first formant (see, also, Remez *et al*., 2013).

One possible reason for the difference between studies has to do with methodological choices. For example, Nittrouer and Lowenstein (2010) provided their participants with a brief training session prior to testing in which participants heard the same sentence both in full speech and subsequently in degraded speech; it is possible that this training experience enhanced children's performance with SWS in particular. However, this training period used only six sentences; it does not seem likely to have had a large impact on children's performance. They also asked participants to repeat

back five-word (highly predictable) sentences; this task may encourage reliance on contextual cues in a way different from the two-choice looking task used here, but it is not clear why that would affect SWS differently from NV speech.

Another likely reason for the difference between studies is the different number of channels in the NV speech. Nittrouer and Lowenstein's NV test items consisted of four vocoder channels, presumably because of the apparent similarity between that and the three-tone sine-wave analogs. Yet a direct comparison between the two is difficult: the frequency and amplitude modulation of three tones replicating a formant pattern would seem to preserve a greater proportion of the spectral variation in a typical speech signal than a four channel noise-band vocoded version does. Moreover, it is not clear that four channels are approximate to what listeners with a cochlear implant experience. While work by Friesen and colleagues (Friesen *et al.*, 2001) suggests that adults with implants do not show improvement with increases in the number of electrodes beyond seven or eight (and some did not benefit beyond four), we might expect that part of this limitation would be the result of neural degradation caused by long stretches of time with no acoustic stimulation of the auditory nerve. Generally, four-channel NV speech represents the lower range of performance with a CI, while eight-channel NV speech is generally more representative of the average performance. Further, children needing CIs today are implanted relatively early, within the first year or two of life, and the possibility remains that they might additionally benefit from their experience with the device during the more sensitive period of development. Thus the eight-channel NV stimuli used in this study might be more representative of the type of spectral degradation children might experience. In a recent study of voice emotion recognition, Chatterjee *et al.* (2015) reported that school-aged children with CIs attending to full-spectrum speech showed mean performance similar to hearing adults with eight-channel NV speech. In contrast, the performance of hearing children with eight-channel NV speech was significantly poorer than that of hearing adults and showed a strong developmental effect. This underscores the benefit obtained by CI children from experience with their device and suggests that the poorer performance by hearing children listening to NV speech underestimates CI children's actual performance. Given the success of children at listening to these items, there may be little need to supplement children's CI signals with dynamic spectral information from lower formants at least for speech perception in quiet. As the practical matter of encoding the lower formant frequency dynamics with sufficient resolution in a CI system remains a major technical challenge, this is an important conclusion.

In short, whether children perform better with NV or SWS signals depends critically on the level of degradation found in the NV stimuli. But regardless, the fact that young children do well with both types of degradation is a surprising, yet reassuring, result.

Acknowledgments

This work was supported by NSF Grant No. BCS1152109 to the University of Maryland. M.C. was supported by NIH Grant Nos. R21 DC11905 and R01 DC014233; G.M. was supported by an NRSA T32 and an NSF IGERT to the University of Maryland; R.E.R. was supported by NIH Grant No. NIDCD 000308. We thank George Hollich for his Supercoder program, Kate Francis, Emily Thomas, Andrea Wycoff, Aislinn Crank, and Sara Alice Hanna for the creation of the SWS, and members of the Language Development Lab for assistance with scheduling, testing, and coding participants.

References and links

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., Kulkarni, A. M., and Christensen, J. A. (**2015**). "Voice emotion recognition by cochlear-implanted children and their normally-hearing peers," Hear. Res. **322**, 151–162.

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.

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.

Hollich, G. (**2005**). "SUPERCODER: A program for coding preferential looking," Version 1.5 (Purdue University, West Lafayette, IN).

Newman, R. S., and Chatterjee, M. (**2013**). "Toddlers' recognition of noise-vocoded speech," J. Acoust. Soc. Am. **133**(1), 483–494.

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.

Remez, R. E. (**2005**). "Perceptual organization of speech," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell Publishers, Oxford).

Remez, R. E., Cheimets, C. B., and Thomas, E. F. (**2013**). "On the tolerance of spectral blur in the perception of spoken words," Proc. Meet. Acoust. **19**, 1–6.

Remez, R. E., Dubowski, K. R., Davids, M. L., Thomas, E. F., Paddu, N. U., Grossman, Y. S., and Moskalenko, M. (**2011**). "Estimating speech spectra for copy synthesis by linear prediction and by hand," J. Acoust. Soc. Am. **130**(4), 2173–2178.

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.

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.