Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 2aSC: Linking Perception and Production (Poster Session)

2aSC33. Toddlers' comprehension of noise-vocoded speech and sine-wave analogs to speech

Rochelle S. Newman*, Monita Chatterjee, Giovanna Morini and Molly Nasuta

***Corresponding author's address: Hearing & Speech Sciences, University of Maryland, 0100 Lefrak Hall, College Park, MD 20742, rnewman1@umd.edu**

A great deal of research has investigated listeners' ability to compensate for degraded speech signals such as noise-vocoded speech (a signal with reduced spectral structure but intact amplitude envelope information) and sine-wave analogs to speech (a signal that maintains the global dynamic spectral structure of the signal at the expense of amplitude envelope information). Nittrouer and colleagues found developmental changes in the ability to comprehend such signals, reporting that while adults perform more accurately with sine-wave analogs than with noise-vocoded speech, school-aged children show the opposite pattern (e.g., Nittrouer Lowenstein & Packer, 2009). In a series of studies, we tested toddler's comprehension of these degraded signals. 27-month-old children saw two images on each trial (e.g., cat, dog), and heard a voice instructing them which image to look at ("Find the cat!"). Sentences were presented either in full speech or were degraded. Toddlers looked at the appropriate object equally long with vocoded speech of 24 channels (60.2%) or 8 channels (62.4%) as with full speech (62.6%), but performed barely above chance with 4 channels (53.6%) and at chance for 2 channels (49.8%). Preliminary results suggest that performance with sine-wave analogs is poorer than 8-channel vocoded speech (56.1%), but testing is ongoing.

Published by the Acoustical Society of America through the American Institute of Physics

Received 21 Jan 2013; published 2 Jun 2013
Proceedings of Meetings on Acoustics, Vol. 19, 060077 (2013)

INTRODUCTION

Much of the speech that we hear is not perfectly clear; for example, it may occur in the presence of reverberation, or be masked by background noise. A great deal of research has investigated listeners' ability to compensate for such degraded speech signals. 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 speech is a signal with reduced spectral structure but intact amplitude envelope information (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). It is thought to simulate the type of signal degradation that is produced by a cochlear implant. Creating noise-vocoded speech involves subdividing a natural speech sample into different frequency ranges, or bands. The overall amplitude of the signal is calculated separately over each band; bands of noise with the same bandwidths and center frequencies as the original bands are then modulated by the envelope extracted from each of those bands. The modulated noisebands are then summed to create the noise-vocoded speech. This manipulation removes the fine spectral structure of the signal within each band, but retains the amplitude envelope. As the number of bands used to create the noise-vocoded speech decreases, the degree of spectral degradation increases. Generally, adult listeners are able to perceive speech well with as few as 4 bands or channels (Shannon et al., 1995).

In contrast to noise-vocoded speech, sine-wave analogs to speech maintain the global dynamic spectral structure of the signal. To create sine-wave analogs, the first three formants (or resonant energy bands) in the original speech signal are each replaced with a time-varying sine-wave tone (Remez, Rubin, Pisoni, & Carrell, 1981). This results in a signal that lacks the resonant properties of the human vocal tract, but maintains the time-varying spectral properties. Despite this degradation, the signal can also be comprehended quite well by adult listeners.

Thus, these two signals are degraded in very different ways. Although the slow time-varying envelope of speech remains in sine-wave speech, rapid envelope cues are removed. Conversely, noise-vocoded speech retains the coarse spectral structure of the signal (to greater or lesser degree depending on the number of channels), but fine spectral information (formant transition trajectories, for instance) is lost. While neither sounds like normal speech, both can be understood quite well by adult listeners, underscoring the remarkable resilience of the speech processing mechanism to severe spectro-temporal degradation of the input and pointing to a necessary and critical role for top-down reconstruction of the intended message under such conditions. Yet Nittrouer and colleagues (Nittrouer & Lowenstein, 2010; Nittrouer, Lowenstein, & Packer, 2009) found that the ability to interpret these two forms of degradation had different developmental time courses. Relative to adults, children across a range of ages had a much greater difficulty with noise-vocoded speech than with sine-wave speech. They 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 noise-vocoded signals which lack such information.

A developmental time course for the ability to process noise-vocoded speech was also indicated by the results of Eisenberg et al. (2000) who found that children aged 10-12 years performed quite similarly to adult listeners when listening to noise-vocoded speech. In contrast, children aged 5-7 years required more spectral bands than did older children or adults in order to reach the same level of comprehension.

The current study explores these developmental changes more fully by examining still younger children: toddlers. In Experiment 1, children aged 27 months were tested on their ability to recognize noise-vocoded (henceforth NV-speech) of either 8 or 4 channels. In Experiment 2, children of the same age were compared on their ability to recognize 8-channel NV-speech vs. sine-wave analogs to speech (henceforth, SWS). We therefore examine both the developmental time-course for recognizing degraded speech signals more generally, and the extent to which toddlers depend on spectral cues vs. amplitude-envelope cues in their speech recognition.

EXPERIMENT 1: NOISE-VOCODED SPEECH[1]

This first experiment explored whether toddlers could recognize noise-vocoded speech; we compared NV-speech of 8 and 4 channels, as well as unprocessed (original) speech. We use the language-guided looking paradigm, in which children are presented with two visual images of familiar objects and a speech sample that matches one of the two images. If children are able to understand the speech, despite the degradation of the signal, they should look longer to (and shift their gaze faster to) the image that matches the speech sample they are hearing. This has proven to be a reliable and sensitive method for testing young children (Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987).

[1] This Experiment appeared in Newman and Chatterjee (2013)

Method

Participants

Eighteen children (11 m, 7 f), aged 27 months (range: 26 months, 5 days - 27 months, 24 days) participated. Data from an additional 4 children were excluded for hearing problems (n=2) or fussiness (n=2).

Stimuli

Stimuli consisted of both an auditory component (sentences that were either nondegraded or noise-vocoded at 8 or 4 channels) and a video component (still pictures of well-known objects). Four objects presented in pairs were used for test trials (keys and blocks; car and ball) and an additional pair was used for practice trials (cat and dog); all were matched for size and color-scheme.

The nondegraded audio stimuli were spoken by a single female talker, recorded over a Shure SM51 microphone at a 44.1 kHz sampling rate and 16 bits precision. They consisted of sentences instructing the child to attend to a particular object("Look at the _____! Can you find the _____? See the _____?") or telling the child to look more generally ("Look at that! Do you see that? Look over there!"). Sentences were matched for duration (4.8 sec) and average r.m.s. amplitude.

Noise vocoding was performed using methods akin to published standards (Shannon et al., 1995) using TigerCIS (Tigerspeech Technology, Qian-Jie Fu, House Ear Institute), with either 4 or 8 channels. The nondegradad audio stimuli were first bandpassed to limit the input range to that between 0.2 and 7.0 kHz, and was then split into 4 or 8 bands (Butterworth filters, 24 dB/oct 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.

Procedure

Children sat on their caregiver's lap facing a widescreen TV. 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 degrees visual angle, along with a simultaneous audio sequence.

The first two trials were considered practice; on one of these two trials the correct answer appeared on the left, and on the other it appeared on the right. Practice stimuli were presented in nondegraded speech.

This was followed by 14 test trials: 4 test trials in the nondegraded condition (one for each test object), 4 in the 8-channel noise-vocoded condition, 4 in the 4-channel noise-vocoded condition, and 2 baseline trials. The baseline trials (which used nondegraded speech) measured infants' general looking preferences for one object of a pair over the other when children were not instructed which way to look. Sentences began simultaneously with the video stimuli, with the target word first appearing 600 ms (18 frames) into the sentence; looking during these initial 18 frames prior to the word were excluded from data analysis.

Children were tested with one of 6 different trial orders; which image appeared on the left vs. right was counterbalanced across orders, and trial order was pseudo-randomized within each order (with the restriction that the correct response did not occur on the same side more than 3 trials in a row).

The caregiver listened to masking music over headphones throughout the study to prevent any biasing of the child's behavior, and completed the Language Development Survey (Rescorla, 1989) as a measure of their child's productive vocabulary.

Children's eye gaze was recorded by a digital camera at 30 frames per second; two researchers (blind to condition) coded these videos on a frame-by-frame basis using Supercoder coding software (Hollich, 2005). A third coder coded any trial on which the two researchers disagreed by more than 15 frames (0.5 sec); this occurred on 23 trials (9% of the trials). The final data was extremely reliable; correlations on the percentage of left (vs. right) looking ranged from 0.9886 to 0.9997 per participant, with an average correlation of 0.9962.

From this, the infants' total duration of looking at each of the two images on each trial was calculated. This was used to calculate the proportion of time the child spent looking to the target object when it was named ("Look at the blocks!") minus the looking time to that same object on baseline trials ("Look at that!"); see Naigles and Gelman (1995) or Newman (2011) for similar methods. 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; thus, positive

values represent longer looking to the correct object, whereas numbers approaching (or below) zero indicate a lack of comprehension.

Results and Discussion

We examined children's looking for each of the three speech conditions individually. For the nondegraded condition, children looked towards the target object 19.1% longer when named than during baseline trials ($t(17)=5.43$, $p<.001$). For the 8-channel NV speech, children looked towards the target object 13.8% longer when named ($t(17)=4.06$, $p<.001$). Finally, for the 4-channel NV speech, children looked towards the target object 7.7% longer ($t(17)=2.63$, $p=.02$). A repeated-measures 1-way ANOVA showed that the effect of condition was significant, $F(2,34)=4.50$, $p<.02$, suggesting that the children's looking performance differed across the three conditions. Follow-up paired comparisons show no difference between the nondegraded and 8-channel NV speech conditions ($t(17)=1.19$, $p>.20$), but a significant difference between the nondegraded and the 4-channel NV speech condition ($t(17)=3.15$, $p<.006$), as shown in Figure 1. The difference between the 8-channel and 4-channel conditions was only marginal, $t(17)=1.85$, $p=.08$ (two-tailed).

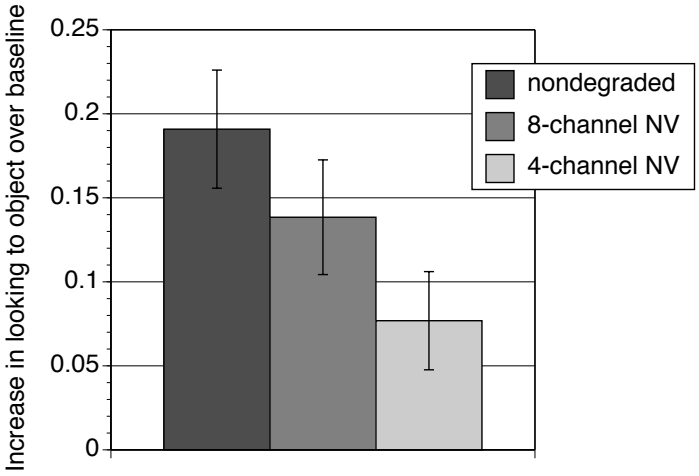

FIGURE 1. Increase in looking to target object when named vs. baseline condition (in proportion) for the three different types of stimuli.

Thus, children successfully recognized speech in the 8-channel noise-vocoded condition, and appeared to do equivalently well in this condition as with nondegraded speech. There was also evidence to suggest that children recognized speech in the 4-channel noise-vocoded condition, but their performance was poorer than for nondegraded speech.

Finally, we examined whether performance on this task correlated with children's vocabulary. We did not find any significant correlations (nondegraded speech: $r=.23$, $p=.18$; 8-channel NV speech: $r=.36$, $p=.07$; 4-channel NV speech, $r=-.23$, $p>.50$), suggesting that vocabulary does not appear to be a factor in children's ability to recognize degraded speech signals.

EXPERIMENT 2: SINE-WAVE ANALOGS TO SPEECH

Experiment 1 demonstrates that children are quite successful at recognizing one form of degraded speech: that produced by noise-vocoding. This form of speech has intact amplitude envelope information, but reduced spectral information. Nittrouer has argued that children rely on dynamic spectral information to a greater extent than do adults; this would imply that they would do as well, if not better, at recognizing a degraded speech signal that maintains that spectral information (such as sine-wave analogs to speech) as they do at recognizing noise-vocoded speech. Experiment 2 tests this prediction by presenting children with both 8-channel NV-speech and SWS.

Method

Participants

This experiment is still ongoing. To date, ten children (7 m, 3 f), aged 27 months (range: 26.1 – 27.9 months) participated and were coded; none needed to be excluded from the study.

Stimuli

Stimuli were identical to those in Experiment 1, except that the 4-channel noise-vocoded speech stimuli were replaced with sine-wave analogs to speech; thus, there were three types of stimuli in this experiment: nondegraded sentences, 8-channel noise-vocoded stimuli, and SWS. The sine-wave analogs were created in Praat (Boersma & Weenink, 2009) using a script written by Chris Darwin (www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS); this script tracks the first three formants of the speech signal and replaces them with sine waves that track the center of each formant.

Procedure

The procedure was identical to that in Experiment 1; we used the same orders, but with SWS replacing 4-channel NV speech. Because this study is still ongoing, assignment to orders is not fully-counterbalanced.

Coding was identical to Experiment 1; 20 trials had to be recoded by a third coder (15%); final coding reliability ranged from .983 to .999 per participant, with an average of .995.

Results and Discussion

We again examined children's looking for each of the three speech conditions individually. For the nondegraded condition, children looked towards the target object 15.7% longer when named than during baseline trials ($t(9)=11.28$, $p<.02$), slightly smaller than the 19.1% value found in Experiment 1. For the 8-channel NV speech, children looked towards the target object 18.6% longer when named ($t(9)=5.44$, $p<.05$), compared to 13.8% in Experiment 1. Although the particular values differ from those found in Experiment 1, the general pattern of comparable performance in the nondegraded and 8-channel NV stimuli remains the same, as shown in Figure 2.

Finally, for the SWS, children looked towards the target object 7.4% longer ($t(9)=1.71$, $p<.05$). A repeated-measures 1-way ANOVA showed that the effect of condition was significant, $F(2,18)=4.79$, $p=.022$, suggesting that the children's looking performance differed across the three conditions. Follow-up paired comparisons show no difference between the nondegraded and 8-channel NV speech conditions, as in Experiment 1 ($t(9)=1.13$, $p>.20$), but a significant difference between the NV and SWS ($t(9)=2.41$, $p<.05$), and a marginal difference between the nondegraded speech and SWS ($t(9)=2.21$, $p=.055$, two-tailed).

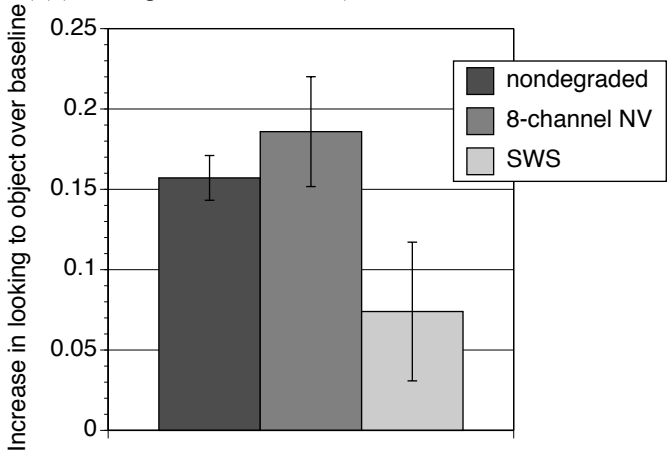

FIGURE 2. Increase in looking to target object when named vs. baseline condition (in proportion) for the three different types of stimuli.

Finally, as in Experiment 1, we examined whether performance on this task correlated with children's vocabulary. We did not find any significant correlations, and indeed all correlations were negative (nondegraded speech: r=-.29; 8-channel NV speech: r=-.08; SWS, r=-.40), suggesting that higher levels of productive vocabulary do not appear to allow children to better interpret these degraded speech signals.

Thus, the general pattern appears to be that SWS is recognized less well than 8-channel NV speech by children of this age. Instead, performance on SWS appears to be more comparable to that of 4-channel NV stimuli tested in Experiment 1.

GENERAL DISCUSSION

In both studies, toddlers were presented with 2 images and heard a voice telling them which object to look at. On some trials, the speech presented by this voice was degraded, and we examined the effect of that degradation on children's looking behavior. This provides an indication of their ability to recognize these degraded signals. Not surprisingly, when the speech was not degraded, children appeared to recognize the target words, looking longer at the appropriate object than during baseline trials. They looked similarly when the speech consisted of 8-channel NV speech: They looked slightly more accurately in the 8-channel NV condition than in the nondegraded condition in one experiment, and slightly less accurately in the other, but neither difference was significant. This suggests that children are able to recognize 8-channel NV speech quite well.

Children were also able to recognize speech in the 4-channel NV and SWS conditions at above-chance levels. However, in both cases they performed more poorly than with 8-channel NV or nondegraded speech.

Interestingly, we did not find an advantage for SWS over NV speech, based on the data collected so far. Nittrouer and colleagues (Nittrouer & Lowenstein, 2010; Nittrouer et al., 2009) have suggested that children rely more strongly on spectral cues than do adults. They found that while children perform more poorly than adults with all forms of degraded signal, they are proportionately much less hindered by SWS than by noise-vocoding. Nittrouer & Lowenstein (2010) tested children using a sentence repetition task, and found that the 3-year-olds recognized 75% of the words the SWS signals, but only 16% of the words in 4-channel NV speech. Although the 3-year-old children they tested (specific ages not given) are significantly older than the 27-month-old children tested here, the pattern in the current study does not appear to match their findings. One possible explanation may relate to the use of open-class response sets (those where participants can give any answer) vs. closed-class response sets (where there are a limited number of choices for participants to choose among); if children are able to identify some of the phonemes in NV speech, but not all, they might show highly accurate performance in a 2-choice task such as the one here, while not being able to identify words sufficiently accurately to be able to produce them. Another potential issue here is that the sentences in our study did not involve context effects: it is not possible to predict the key word based on the initial portion of the sentence, which was identical for all words. On the other hand, the sentences in Nittrouer and colleagues' study did provide contextual cues. Although Nittrouer et al (2010) found similar context effects across their groups, the effect could not be calculated for the 3-year olds who performed too poorly with NV speech in their study.

To summarize, our results indicate that:

i) 27-month old toddlers are able to recognize 8-channel NV speech about as well as they recognize unprocessed speech;

ii) 4-channel NV speech presents significantly greater difficulty for this age group;

iii) similarly, sine wave analogs to speech also present significant difficulty compared to unprocessed or 8-channel NV speech

iv) toddlers were able to recognize both the 4-channel NV and SW speech at above-chance levels.

Overall, these children exhibit remarkable ability to process severely degraded speech with no previous training or experience.

ACKNOWLEDGMENTS

The authors thank George Hollich for the Supercoder program, and thank Daniel Eisenberg and Tracy Moskatel for assistance with stimulus creation. We thank Amanda Pasquarella, Justine Dombroski, and Lauren Evans for overseeing substantial parts of the coding reported here. We also thank the following additional students for assistance either in scheduling or testing participants, or coding looking time performance: Katrina Ablorh, Mikayka Abrams, Faraz Ahsan, Candace Ali, Alison Arnold, Taryn Bipat, Alyssa Cook, Jennifer Coon, Sara Dougherty, Sara

Edelberg, Lauren Fischer, Andrea Fisher, Arielle Gandee, Laura Horowitz, Megan Janssen, Mina Javid, Amanda Jensen, Esther Kim, Penina Kozlovsky, Stephanie Lee, Rachel Lieberman, Perri Lieberman, Eileen McLaughlin, Kelly McPherson, Vidda Moussavi, Molly Nasuta, Courtenay O'Connor, Maura O'Fallon, Sabrina Panza, Elise Perkins, Rachel Rhodes, Allie Rodriguez, Rebecca Sherman, Katie Shniderman, Veronica Son, Ashley Thomas, Krista Voelmle, Chelsea Vogel, Amanda Wildman, Kimmie Wilson, Catherine Wu, & TeHsin Wu.

This work was supported by NIH grant R01 DC004786 and by NSF grant BCS0642294 to the University of Maryland.

REFERENCES

Boersma, P., & Weenink, D. **(2009)**. *Praat: doing phonetics by computer.* (Version 5.1.05) http://www.praat.org/

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., & Boothroyd, A. (2000). "Speech recognition with reduced spectral cues as a function of age." J. Acoust. Soc. Am., **107**, 2704-2710.

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. **(1987)**. "The eyes have it: Lexical and syntactic comprehension in a new paradigm." J. Child Lang., **14**, 23-45.

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

Naigles, L. G., & 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.

Newman, R. S. **(2011)**. "2-year-olds' speech understanding in multi-talker environments." Infancy, **16**, 447-470.

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

Nittrouer, S., & Lowenstein, J. H. **(2010).** "Learning to perceptually organize speech signals in native fashion." J. Acoust. Soc. Am., **127**, 1624-1635.

Nittrouer, S., Lowenstein, J. H., & Packer, R. R. (2009). "Children discover the spectral skeletons in their native language before the amplitude envelopes." J. Exp. Psychol. - Hum. Percept. Peform., 35, 1245-1253.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & 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. Dis., **54**, 587-599.

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