Children weight dynamic spectral structure more than adults: Evidence from equivalent signals

Joanna H. Lowenstein,[a] Susan Nittrouer, and Eric Tarr
Department of Otolaryngology, The Ohio State University, 915 Olentangy River Road, Suite 4000, Columbus, Ohio 43212
lowenstein.6@osu.edu, nittrouer.1@osu.edu, tarr.18@osu.edu

Abstract: Earlier work using sine-wave and noise-vocoded signals suggests that dynamic spectral structure plays a greater role in speech recognition for children than adults [Nittrouer and Lowenstein. (**2010**). J. Acoust. Soc. Am. **127**, 1624–1635], but questions arise concerning whether outcomes can be compared because sine waves and wide noise bands are different in nature. The current study addressed that question using narrow noise bands for both signals, and applying a difference ratio to index the contribution made by dynamic spectral structure. Results replicated earlier findings, supporting the idea that dynamic spectral structure plays a critical role in speech recognition, especially for children.

PACS numbers: 43.71.Ft, 43.71.An [QJF]
Date Received: August 6, 2012 **Date Accepted:** October 9, 2012

1. Introduction

When it comes to phoneme recognition, it has long been reported that children weight formant transitions more than adults (e.g., Mayo and Turk, 2005; Nittrouer, 1992; Wardrip-Fruin and Peach, 1984). Because intra-syllabic formant transitions are brief bits of longer patterns of spectral change, that finding for phoneme recognition led to the hypothesis that children's recognition of words in sentences should similarly show disproportionately greater effects of this dynamic (i.e., time-varying) spectral structure than that of adults. To test that hypothesis, two earlier studies compared recognition scores for words in sine-wave and noise-vocoded sentences (Nittrouer and Lowenstein, 2010; Nittrouer *et al.*, 2009). Sine-wave synthesis creates signals that preserve dynamic spectral structure better than vocoding, so the prediction was that recognition would be better for sine-wave than for noise-vocoded signals, and disproportionately more so for children than adults. Indeed, these studies showed that both adults and children performed more accurately with sine-wave than with noise-vocoded speech, and the significant age × signal type interaction suggested the effect was greater for children than for adults. As an example of these findings, Table 1 shows scores for adults, 7-year-olds, and 5-year-olds from Nittrouer and Lowenstein (2010). It can be seen that children showed greater differences in scores between the sine-wave and vocoded conditions than adults. Of course, adults' performance was close to ceiling for sine-wave speech, and that may have constrained the magnitude of the effect which could be obtained. Nonetheless, these results at least suggest that the dynamic spectral structure preserved so well in sine-wave speech plays an important role in speech recognition, especially for children.

There are several challenges, however, that militate against accepting these conclusions too readily. The main challenge concerns the nature of the two signals. The sine waves used to track formants in sine-wave synthesis are periodic in quality, but narrow in frequency. Noise-vocoded speech is aperiodic and spectrally broad. Thus the question can be raised as to whether comparing signals with different carriers offers

[a]Author to whom correspondence should be addressed.

Table 1. Mean percent correct words recognized by each group in each condition from Nittrouer and Lowenstein (2010). Standard deviations are in parentheses. VOC refers to noise-vocoded stimuli and SWS refers to sine-wave stimuli.

	Adults	7-year-olds	5-year-olds
VOC	79.5 (8.3)	43.9 (16.6)	30.7 (17.9)
SWS	98.4 (1.4)	91.9 (3.0)	86.3 (6.1)

a valid test of the role of dynamic spectral structure in speech recognition. Perhaps recognition just differs for the two types of carriers.

Another challenge to the conclusion that children show a greater benefit than adults from the enhanced dynamic spectral structure of sine-wave over noise-vocoded signals has to do with how this effect is indexed. Children generally perform poorer than adults. Comparing scores from different regions of a probability function is a tricky business (Boothroyd and Nittrouer, 1988; Wagenmakers *et al.*, 2012). For example, it is rarely the case that the difference between 20% and 30% represents the same magnitude of effect as the difference between 80% and 90%.

To address these concerns, the current study differed from earlier ones in several ways.

1.1 Equivalent carriers

For stimulus generation in this experiment, the same carrier was used in both synthesis that retained formant tracks (comparable to sine-wave speech) and synthesis that preserved temporal envelopes in several channels (comparable to noise-vocoded speech). That carrier consisted of 20-Hz wide bands of noise. The signals resulting from these processing methods will be referred to as narrow-noise dynamic (DYN) and narrow-noise vocoded (VOC) signals.

1.2 Indexing the contribution of dynamic spectral structure to recognition

A significant problem in comparing effects across groups with overall differences in recognition probabilities concerns how to index the magnitude of the effect. There are a number of ways to handle this problem, but in this experiment the simplest method was used. The difference in scores for the DYN and the VOC stimuli was given as a ratio of the score for VOC stimuli, with the formula

$$\text{difference ratio} = (p_{\text{DYN}} - p_{\text{VOC}})/p_{\text{VOC}}, \tag{1}$$

where p_{DYN} and p_{VOC} are recognition probabilities for the DYN and VOC stimuli, respectively.

1.3 Controlling for linguistic context effects

Finally, one more analysis was used in the current experiment as a control. To make sure any differences among age groups found in this study were not attributable to variation across groups in the contribution of linguistic context to recognition, a factor reported by Boothroyd and Nittrouer (1988) was used. This metric derives from the perspective that the probability of recognizing a whole sentence is related to the probability of recognizing each of its constituent parts, or words, such that

$$p_W = p^n{}_p, \tag{2}$$

where p_W is the probability of recognizing the whole sentence, p_p is the probability of recognizing each part, or word, and n is the number of words in the sentence. However, this relationship holds only if each word must be recognized separately in order to recognize the sentence. With sentence context, that is not the case because context

aids recognition. Thus, j can be substituted for n to represent the number of statistically independent channels of information required for the sentence to be recognized. Now, Eq. (2) can be rewritten as

$$j = \log(p_W)/\log(p_p). \quad (3)$$

Here, j indexes the contribution of sentence context, such that the smaller j is, the greater the effect of that context on recognition. Children as young as three years of age have demonstrated equivalent contributions of sentence context to speech recognition as adults (Nittrouer and Boothroyd, 1990; Nittrouer and Lowenstein, 2010), at least for simple sentences. These j factors were computed in the current experiment to make sure that was the case for these listeners.

1.4 Summary

In summary, the current study was undertaken to test the hypothesis that the apparent benefit observed in earlier experiments for sine-wave over noise-vocoded speech, especially for children, was actually due to the signals having different carriers. That hypothesis would be supported if recognition scores were similar for the two kinds of stimuli used in the current study. The alternative hypothesis going into this experiment was that dynamic spectral structure really does serve an important role in sentence recognition, especially for children, as earlier studies had suggested. That hypothesis would be supported if recognition was better for the DYN than for the VOC stimuli, and disproportionately so for children.

2. Method

2.1 Participants

Sixty-two listeners participated in this experiment: 20 adults between the ages of 18 and 36, 21 7-year-olds (ranging from 7 years, 2 months to 7 years, 11 months), and 21 5-year-olds (ranging from 5 years, 0 months to 5 years, 11 months). All listeners were native speakers of English, and all passed hearing screenings at 25 dB HL for the frequencies 0.5, 1, 2, 4, and 6 kHz. All listeners had histories of normal speech and language skills.

2.2 Equipment

All speech samples were recorded in a sound booth, directly onto the computer hard drive, via an AKG (Vienna, Austria) C535 EB microphone, a Shure (Niles, IL) M268 amplifier, and a Creative Laboratories (Singapore) Soundblaster soundcard. Perceptual testing took place in a sound booth, with the computer that controlled the experiment in an adjacent room. Stimuli were stored on a computer and presented through a Samson (Syosset, NY) headphone amplifier and AKG-K141 headphones. The hearing screening was done with a Welch Allyn (Skaneateles Falls, NY) TM262 audiometer and TDH-39 headphones (Telephonics, Farmingdale, NY).

2.3 Stimuli

The 72 5-word sentences (12 for practice, 60 for testing) used by Nittrouer and Lowenstein (2010) were used in this experiment. These sentences, from HINT-C (Nilsson *et al.*, 1996), are syntactically correct, follow a subject-predicate structure, and are highly predictable semantically. A typical sentence is "Flowers grow in the garden," and Fig. 1 shows spectrograms of this sentence: the unprocessed version is in the top panel, the DYN version in the middle, and the VOC version in the bottom panel. The sentences were recorded at a 44.1-kHz sampling rate with 16-bit digitization by an adult male speaker of American English who is a trained phonetician. Both DYN and VOC versions of all sentences were created.

The DYN stimuli were synthesized in MATLAB using formant tracks obtained from Praat (Boersma and Weenink, 2009). The center frequencies of $F1$, $F2$, and $F3$ were obtained for 6-ms windows, and were hand-corrected as needed so that outputs

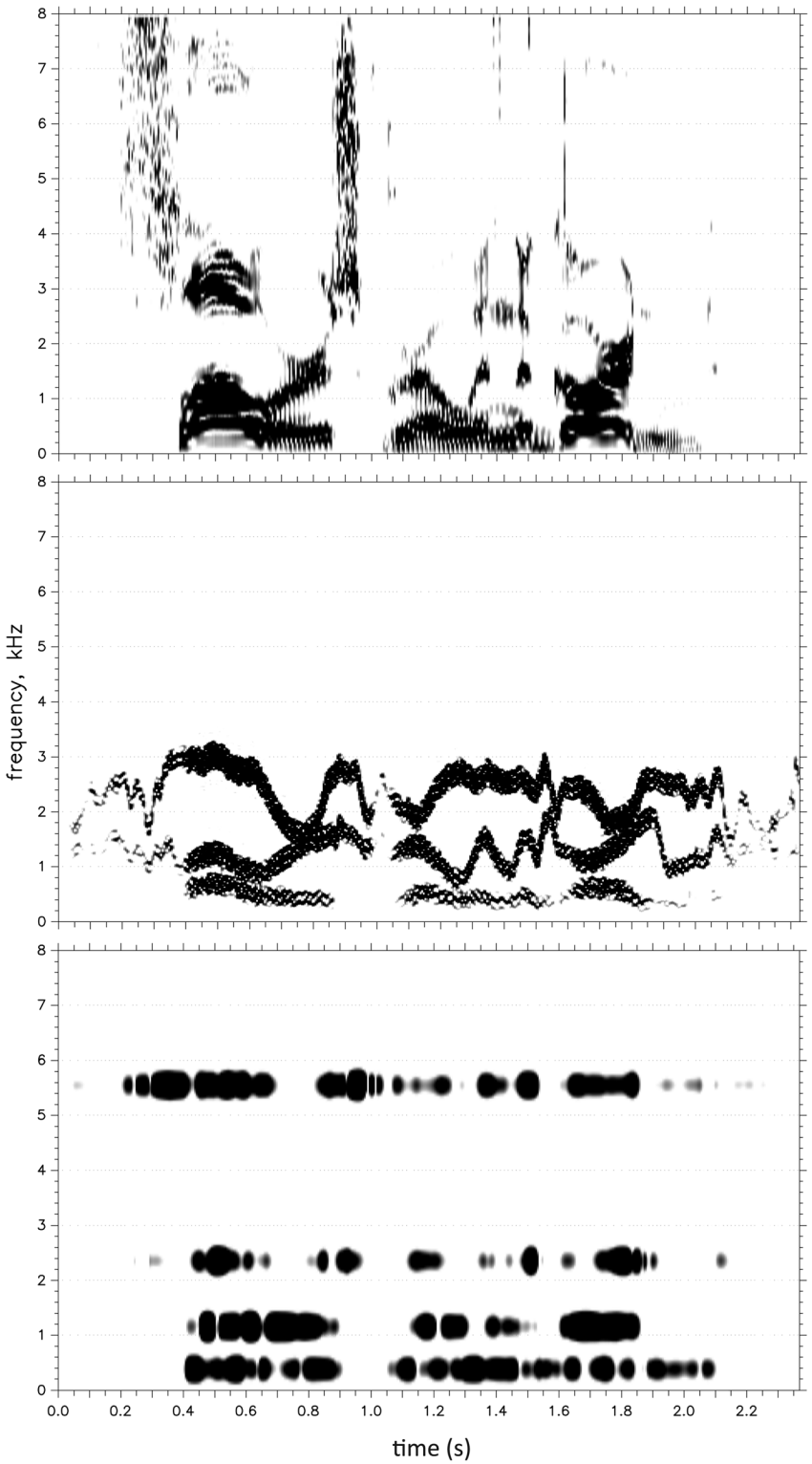

Fig. 1. Spectrograms of the sentence "Flowers grow in the garden" in its unprocessed form (top), as a narrow-noise wave signal (middle) and as a narrow-noise vocoded signal (bottom).

corresponded accurately to spectrograms of the sentence. Derived formant frequencies were imported into MATLAB, and a local-averaging filter was applied to the formant frequencies to reduce transients that can arise from LPC analysis. Three white noise signals, one for each formant, were synthesized in MATLAB using a random number generator. The length of the noise signals matched the length of the corresponding sentence. Each noise signal was filtered in 6-ms windows following the analysis windows used in PRAAT. A band-pass filter, centered at the frequency of each formant, was applied in each time window. The band-pass filters had cut-off frequencies 10 Hz higher and lower than the center frequency. After filtering, the three noise signals were added together and the gross temporal envelope of the original sentence was applied.

For the VOC signals, a MATLAB routine was used. All signals were first filtered with an upper cut-off frequency of 8000 Hz and a lower cut-off frequency of 50 Hz. Cut-off frequencies for the analysis bands were 800, 1600, and 3200 Hz, which created 4-channel stimuli. This number of channels was selected with a view to trying to make the strongest test of the hypotheses: Although the VOC stimuli had one more physical channel than the DYN stimuli, 4-channel vocoding is common. Also, it was thought that the additional, high-frequency channel would provide the most equitable match across conditions because those VOC signals lacked dynamic spectral structure. Thus, if an advantage was found nonetheless for the DYN stimuli, just that much stronger of a case could be made that dynamic spectral structure is important.

Next the temporal envelope was extracted for each of the channels by half-wave rectification followed by low-pass filtering at 20 Hz, using a Butterworth filter with a transition band to 25 Hz and a 40-dB stop band above that. Four carrier signals, one for each channel, were created using a random number generator to make white noise. Each noise was filtered in a band centered at the middle frequency of the corresponding channel with a cut-off frequency of 10 Hz higher and lower than the center frequency. The temporal envelope of each channel was used to modulate the corresponding narrow-noise carrier signal. Finally, the synthesized signals for all channels were recombined.

All stimuli (natural, DYN, and VOC) were equalized for root mean square amplitude across sentences after they were created.

2.4 Procedures

All stimuli were presented at a peak intensity of 68 dB SPL under headphones. Before testing with each participant, the software randomly selected 30 sentences to present as DYN stimuli and 30 to present as VOC stimuli. Half of the participants heard all 30 DYN sentences first, and then the VOC sentences; the other half of the participants heard sentences in the opposite order. Training for each condition consisted of six practice sentences. For training, listeners were told they would first hear the sentence in a man's voice, and they should repeat it. They then heard the same sentence in its processed form, and they again repeated it. None of the listeners had any difficulty hearing the DYN or VOC sentences as speech.

During testing, the order of presentation of the sentences within each condition was randomized independently for each listener. Each processed sentence was played once, and the listener repeated it as best as possible. The number of incorrect words for each sentence was entered into the program interface during testing. Five- and seven-year-olds moved a game piece along a game board after every ten sentences to help maintain interest.

After hearing all sentences in their processed forms, all sentences were played to listeners in their unprocessed forms. Listeners could get no more than 10% of the words wrong on this post test, or their data would be eliminated from analysis.

3. Results

One 7-year-old and one 5-year-old consistently responded "I don't know" for the VOC sentences, so their data were not included. Data are thus included for 20 listeners of each age.

Table 2. Mean percent correct words recognized by each group in each condition. Standard deviations are in parentheses. VOC refers to narrow-noise vocoded stimuli and DYN refers to narrow-noise dynamic stimuli.

	Adults	7-year-olds	5-year-olds
VOC	49.8 (19.2)	29.9 (9.9)	12.0 (6.1)
DYN	78.9 (5.1)	59.2 (8.7)	45.2 (13.4)

3.1 Word recognition

Word recognition for unprocessed signals was above 99% correct for all listeners. Table 2 shows mean correct word recognition for each group for each kind of processed stimulus. A two-way analysis of variance (ANOVA) was performed on these scores, with age as the between-subjects factor and signal type as the within-subjects factor. The main effect of age was significant, $F(2,57) = 93.71$, $p < 0.001$, as were all *post hoc* contrasts among age groups ($p < 0.001$ in all cases). The main effect of signal type was also significant, $F(1,57) = 322.26$, $p < 0.001$. The age $\times$ signal type interaction was not significant. Based on these outcomes it can be concluded that recognition scores were better for DYN than for VOC stimuli, and recognition improved with increasing age.

3.2 Top-down effects

Using the formula from Boothroyd and Nittrouer (1988), *j* factors were computed for individual listeners using word and sentence recognition scores. As this computation requires the use of sentence recognition scores, mean percentages of sentences recognized correctly for each group are shown in Table 3 for VOC and DYN stimuli separately. Because both 5-year-olds and 7-year-olds had less than 5% correct mean sentence recognition for the VOC sentences, *j* factors were only calculated for the DYN sentences. Those *j* factors could be calculated for all adults and 7-year-olds, but for only 19 of the 20 5-year-olds because one 5-year-old had less than 5% correct recognition for DYN sentences. Mean *j* factors [and standard deviations (SDs)] were 2.97 (0.44) for adults, 2.78 (0.43) for 7-year-olds, and 2.70 (0.71) for 5-year-olds. A one way ANOVA computed on those individual *j* factors was not significant ($p > 0.10$), so it may be concluded that all listeners used sentence context to a similar extent.

3.3 Age-related differences in the effects of dynamic spectral structure

Finally, difference ratios were examined. These values (and SDs) were 0.86 (1.04) for adults, 1.25 (0.99) for 7-year-olds, and 3.77 (3.01) for 5-year-olds. A one-way ANOVA performed on these scores showed a significant effect of age, $F(2,57) = 13.38$, $p < 0.001$. *Post hoc* contrasts between 5-year-olds and each of the other groups were significant ($p < 0.001$), but the contrast between adults and 7-year-olds was not ($p > 0.10$). Consequently it may be concluded that dynamic spectral structure contributed to word recognition for all listeners, but that contribution was greatest for 5-year-olds.

4. Discussion

This study was conducted to examine the role of dynamic spectral structure to speech recognition using 20-Hz wide bands of noise for formant-tracking synthesis and vocoding in order to address concerns that there were differences in the nature of the stimuli

Table 3. Mean percent correct sentences recognized by each group in each condition. Standard deviations are in parentheses.

	Adults	7-year-olds	5-year-olds
VOC	24.0 (12.9)	4.5 (4.7)	0.1 (1.8)
DYN	50.5 (8.6)	24.8 (8.5)	13.7 (7.9)

used in earlier studies. When scores for this experiment are compared to those from Nittrouer and Lowenstein (2010) with the same sentences, but with traditional sine-wave and noise-vocoded signals, it can be seen that performance was diminished in both conditions for all three age groups (Table 1 versus Table 2). Nonetheless, performance was better for the signals that preserved dynamic spectral structure than for those that did not. For the youngest children in this study, an enhanced effect of dynamic spectral structure was observed, compared to the other two groups.

In summary, the current experiment demonstrated that robust dynamic spectral structure across long stretches of the signal serves an important function in speech recognition for listeners, especially for children. This effect was observed for signals matched as closely as possible in their fundamental quality.

Acknowledgments

This work was supported by a grant from the National Institutes of Health, National Institute on Deafness and Other Communication Disorders, Grant No. R01 DC000633.

References and links

Boersma, P., and Weenink, D. (**2009**). "Praat: Doing phonetics by computer (version 5.1.1) [computer program]," htpp://www.praat.org (Last viewed 7/19/2009).

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.

Mayo, C., and Turk, A. (**2005**). "The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception," J. Acoust. Soc. Am. **118**, 1730–1741.

Nilsson, M., Soli, S. D., and Gelnett, D. J. (**1996**). *Development and Norming of a Hearing in Noise Test for Children* (House Ear Institute, Los Angeles, CA).

Nittrouer, S. (**1992**). "Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries," J. Phonetics **20**, 351–382.

Nittrouer, S., and Boothroyd, A. (**1990**). "Context effects in phoneme and word recognition by young children and older adults," J. Acoust. Soc. Am. **87**, 2705–2715.

Nittrouer, S., and 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., and Packer, R. (**2009**). "Children discover the spectral skeletons in their native language before the amplitude envelopes," J. Exp. Psychol. Hum. Percept. Perform. **35**, 1245–1253.

Wagenmakers, E. J., Krypotos, A. M., Criss, A. H., and Iverson, G. (**2012**). "On the interpretation of removable interactions: A survey of the field 33 years after Loftus," Mem. Cognit. **40**, 145–160.

Wardrip-Fruin, C., and Peach, S. (**1984**). "Developmental aspects of the perception of acoustic cues in determining the voicing feature of final stop consonants," Lang. Speech **27**, 367–379.