Spectral structure across the syllable specifies final-stop voicing for adults and children alike

Susan Nittrouer[a)] and Joanna H. Lowenstein
Department of Speech and Hearing Science, Ohio State University, Columbia, Ohio 43210

(Received 28 March 2007; revised 14 August 2007; accepted 12 October 2007)

Traditional accounts of speech perception generally hold that listeners use isolable acoustic "cues" to label phonemes. For syllable-final stops, duration of the preceding vocalic portion and formant transitions at syllable's end have been considered the primary cues to voicing decisions. The current experiment tried to extend traditional accounts by asking two questions concerning voicing decisions by adults and children: (1) What weight is given to vocalic duration versus spectral structure, both at syllable's end and across the syllable? (2) Does the naturalness of stimuli affect labeling? Adults and children (4, 6, and 8 years old) labeled synthetic stimuli that varied in vocalic duration and spectral structure, either at syllable's end or earlier in the syllable. Results showed that all listeners weighted dynamic spectral structure, both at syllable's end and earlier in the syllable, more than vocalic duration, and listeners performed with these synthetic stimuli as listeners had performed previously with natural stimuli. The conclusion for accounts of human speech perception is that rather than simply gathering acoustic cues and summing them to derive strings of phonemic segments, listeners are able to attend to global spectral structure, and use it to help recover explicitly phonetic structure. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804950]

PACS number(s): 43.71.An, 43.71.Ft, 43.71.Es [MSS] Pages: 377–385

I. INTRODUCTION

Perhaps the trouble all started in 1944 when Frank Cooper and Al Liberman decided to build a reading machine for the blind. At that time they adopted what Liberman would later call the "horizontal view" in his book, *Speech: A Special Code* (1996). According to this view separate segments are aligned in the speech signal in a linear fashion, strictly auditory perceptual processes recover the acoustic character of each segment, and cognitive processes then translate those acoustic descriptors into phonemic units, void of physical attributes. Assuming this much about the acoustic speech signal, Cooper and Liberman turned their attention to what they saw as the truly difficult problem: optically isolating the letters on the page that would need to be converted into acoustic segments. But their own experiments soon revealed the intractable problem that listeners are unable to recognize separate acoustic elements presented at a rate replicating typical speech production. The declassification after World War II of the technology needed to build a sound spectrograph provided a possible clue to the source of the problem: separate segments are not represented in the acoustic speech stream. What ensued were decades of searching for acoustic properties that were at once both invariant correlates of specific phonetic categories as well as robust predictors of listeners' phonetic judgments. Such properties came to be known as acoustic "cues," and were generally defined as portions of the signal that can be isolated visually on the spectrogram, can be manipulated independently in speech synthesis, and can be shown to influence phonetic decisions (Repp, 1982).

Experiments exploring possible acoustic cues were all conducted in essentially the same way: by manipulating one acoustic property along a continuum, most typically in steps of equal (linear) size in the construction of synthetic stimuli, playing those stimuli for listeners in a labeling task, and plotting the probability of a specific phonetic decision as a function of the acoustic setting of the manipulated property. It was not uncommon for experimenters to manipulate one other selected property in a dichotomous manner such that it was set to be appropriate for one or the other phoneme. Under these circumstances two parallel labeling functions are generally derived, with the separation serving as an index of the amount of influence the dichotomously manipulated property has on the phonetic decision.

The voicing of syllable-final stops is one consonantal feature that has been extensively studied in this way. An acoustic difference in many languages between syllables that end in voiceless stops, such as "buck," and those that end in voiced stops, such as "bug," that is clearly apparent on a spectrogram is the duration of the vocalic syllable portion preceding closure, as Fig. 1 shows. The vocalic syllable portion is shorter when the final stop is voiceless than when it is voiced. In a carefully controlled series of experiments done in the 1970s, Raphael and his colleagues demonstrated that adult English speakers show a strong influence of this temporal property on their phonetic decisions (Raphael, 1972; Raphael *et al.* 1975; Raphael *et al.* 1980). These experiments were conducted in the traditional method, using synthetic stimuli in which all acoustic properties were held constant across the set of stimuli, except for the duration of those stimuli and first-formant offsets. Vocalic duration varied in a linear fashion from rather short to rather long. The offset frequency of the first formant (F1), and so the rate and extent of the transition, was manipulated in a dichotomous manner

[a)] Author to whom correspondence should be addressed. Electronic mail: nittrouer.1@osu.edu

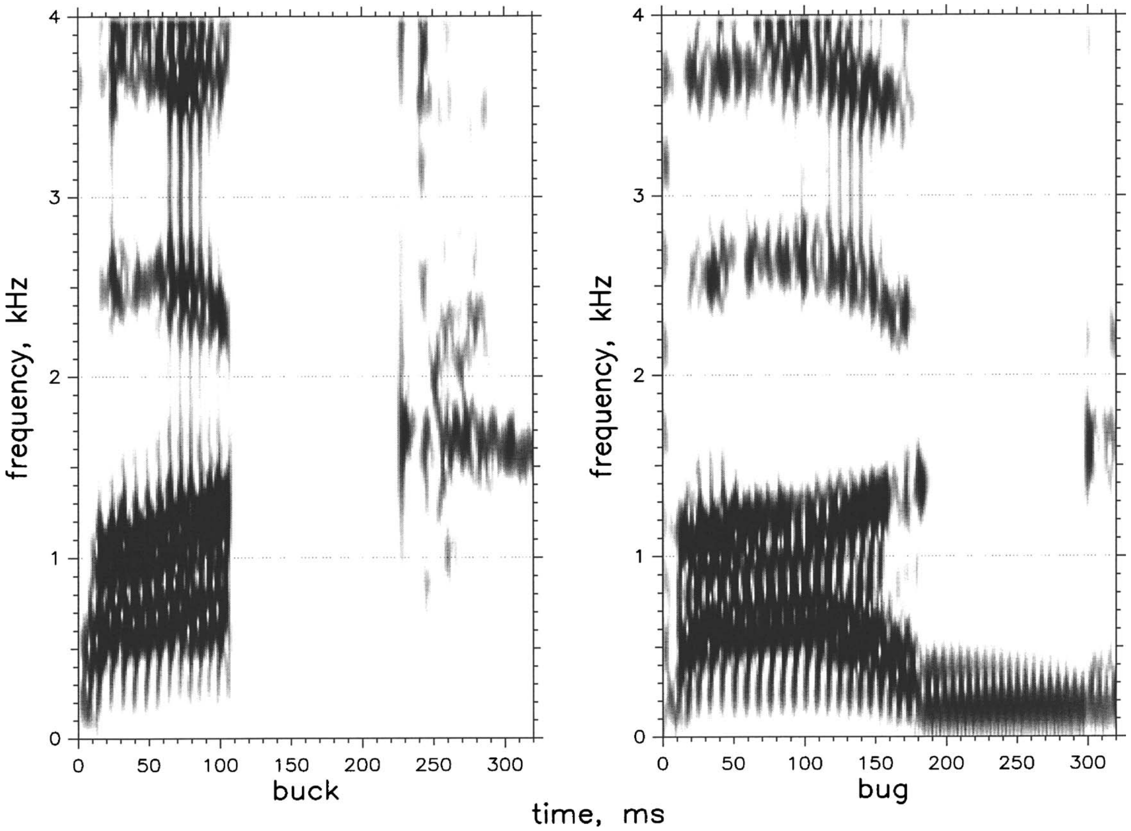

FIG. 1. Spectrogram of *buck* and *bug* spoken by a male, adult speaker.

in these experiments to signal either a voiced or voiceless final stop: higher F1 offsets more strongly support voiceless decisions, and lower F1 offsets support voiced decisions. Results consistently showed that both vocalic duration and F1 offset frequency contributed to voicing decisions.

Investigators interested in language acquisition became interested in the question of whether children use acoustic cues in the same way as adults to reach decisions about phonetic identity, and voicing decisions for syllable-final stops served as one focus of this work. Several investigators constructed synthetic stimuli similar to those used in the experiments of Raphael and colleagues, with vocalic duration varying in a continuous manner and F1 set dichotomously to be appropriate for either a final voiced or voiceless stop. The common conclusions of these experiments were that children did not base their voicing decisions as strongly on vocalic duration as adults did, but that the frequency of F1 at stimulus offset exerted a stronger influence on children's responding than on that of adults (e.g., Greenlee, 1980; Krause, 1982; Wardrip-Fruin and Peach, 1984). Those findings fit the larger picture that was emerging concerning differences between adults' and children's speech perception. Although not without its detractors, experiments with several kinds of stimuli were generally revealing that children's phonetic decisions appeared more strongly dependent on formant transitions, and less dependent on other sorts of acoustic cues (e.g., Morrongiello *et al.* 1984; Nittrouer and Studdert-Kennedy, 1987; Parnell and Amerman, 1978). The collective findings of these experiments led to the proposal that children modify the amount of perceptual weight they assign to individual acoustic cues as they get older and gain experience with a first language (e.g., Nittrouer, 1996). Of course, this model fit the general theoretical perspective of the day, which again was that separate acoustic cues are recovered during speech perception and translated by a cognitive processor to derive phonemic labels.

In 2004, a series of related experiments were conducted involving adults' and children's labeling of words ending in voiced or voiceless final stops (Nittrouer, 2004). In the first of these experiments, synthetic stimuli were constructed in the manner of earlier experiments so that vocalic duration varied across a continuum and F1 at stimulus offset was set to each of two frequencies appropriate for either a voiced or voiceless stop. These experiments replicated the findings of those earlier developmental studies: children's responses were less dependent on duration of the vocalic signal portion and more dependent on the frequency of F1 at stimulus offset than were responses of adult listeners. In the second kind of stimulus preparation, natural productions of words ending with voiced or voiceless final stops were edited so that vocalic duration varied from short to long. In those experiments it was observed that responses of children and adults alike were strongly related to whether the stimulus had been derived from a word ending with a voiceless or a voiced stop. For children this meant that labeling functions resembled those obtained for the children who had heard the synthetic stimuli, leading to the easily supportable conclusion that children were basing their responses strongly on formant transitions near stimulus offsets, as they had with the synthetic stimuli. For adults, however, this pattern of responding meant there was a substantial change from the pattern of responding seen in numerous earlier experiments with synthetic stimuli. This difference in response patterns led to the conclusion that perhaps experiments done with synthetic stimuli had constrained our ability to determine which cues listeners use in phonetic decisions with natural speech. It may be that even adults rely primarily on formant transitions near a syllable's end to make decisions about the voicing of final stops.

The problem with that conclusion, however, is that there are numerous cues that vary in natural stimuli as a function of phonetic structure, and so it is hard to determine which one accounts for most of the variability in phonetic decisions. It is this very fact that has always made natural stimuli so undesirable for empirical study: With all those uncontrolled acoustic properties it is difficult to know how much each one explains about perceptual responding. The finding that adults and children showed similar labeling functions for edited natural stimuli ending in voiced and voiceless final stops might not mean that adults and children were basing decisions on the same acoustic property. Perhaps adults and children actually rely on different acoustic attributes, both of which happen to vary across voicing conditions in natural stimuli. The current experiment was originally undertaken to address this possibility.

One difference between synthetic stimuli generally used in experiments exploring the perception of syllable-final stop voicing and natural tokens was that synthetic stimuli preserved voicing-related differences in F1 offset transitions only. In fact, all formant transitions at the syllable's end are affected by the voicing category of the final stop. Because the vocal folds are abducted before the vocal tract achieves complete closure in the production of a syllable ending in a voiceless final stop, formants have not attained their final frequencies at voicing offset. In the production of words with voiced final stops, on the other hand, speakers continue voicing through closure, and so those final frequency destinations are reached. The difference in F1 depending on voicing is always that F1 is higher at voicing offset for words with voiceless, rather than voiced, final stops because F1 frequency is tightly linked to the degree of vocal tract opening. For higher formants, however, the relation between the formant's frequency at voicing offset and the voicing feature of the final stop depends on the place of closure for the stop. But regardless of the exact nature of that relation, dynamic spectral information at the ends of syllables was clearly richer in the natural stimuli than in the earlier synthetic stimuli, and Nittrouer (2004) suggested that findings for edited natural tokens showed that all listeners were basing their voicing decisions on those final formant transitions.

There are differences in acoustic structure earlier in the syllable as a function of whether the final stop is voiced or voiceless, as well. In particular, F1 rises more rapidly at voicing onset and achieves a higher frequency when the final stop is voiceless than when it is voiced (Summers, 1987). Summers (1988) showed that this acoustic difference alone influences voicing decisions for adult listeners. Consequently, the possibility needed to be considered that perhaps it was this acoustic characteristic that was influencing decisions of some listeners for natural tokens in the Nittrouer (2004) study, and adults were seen as more likely to use this perceptual strategy. That is, perhaps adults are able to take advantage of dynamic spectral patterns across the lengths of syllables in making phonetic decisions whereas children might be restricted to using only that information in the signal region generally affiliated with the segment in question. To examine this possibility we presented adults and children in the current experiment with synthetic stimuli in which formants replicated the patterns of natural syllables over their entirety, and also with stimuli in which formant transitions were deliberately held constant at syllable offsets, but allowed to vary in a natural manner earlier in the syllables.

In summary, this experiment was conducted in order to extend the work of Nittrouer (2004) in two ways. First, by presenting synthetic stimuli that replicated natural tokens of words ending in voiced and voiceless final stops we would be able to examine whether the difference in adults' response patterns observed for synthetic and natural stimuli had something to do with the naturalness of the stimuli. Would adults and children show the same labeling functions for stimuli replicating the acoustic structure of the natural stimuli in the 2004 study, but possessing a synthetic nature? Generally concern exists that children may show decrements in performance for synthetic compared to natural speech (e.g., Mirenda and Beukelman, 1987; Reynolds and Jefferson, 1999), but in this case the question may be asked of adults' responding, as well. In the second stimulus manipulation we asked what it was in the natural stimuli that supported adults' and children's voicing decisions. In particular the hypothesis was explored that perhaps children based their voicing decisions on formant transitions going into vocal tract closure, but adults based their decisions on formant characteristics earlier in the stimulus. This hypothesis would be supported if children showed greater differences in their response patterns between the two stimulus types used in this experiment than adults showed.

II. METHOD

A. Listeners

Adults between the ages of 18 and 39 years participated, as well as 8 year olds, 6 year olds, and 4 year olds. Children were between −1 and +5 months of their birthdays. All listeners had to meet certain criteria to participate. All participants were native English speakers with no histories of speech, language, or hearing problems. All were required to pass hearing screenings of the frequencies 0.5, 1, 2, 4, and 6 kHz presented at 25 dB hearing level to each ear separately. Children could have had no more than five episodes of otitis media before their second birthdays. Children needed to perform at or better than the 30th percentile on the Goldman-Fristoe Test of Articulation 2, Sounds-in-Words subtest (Goldman and Fristoe, 2000), and adults needed to read at or better than an 11th grade reading level on the Wide Range Achievement Test – Revised (Jastak and Wilkinson, 1984). Meeting these criteria were 24 adults, 24 8 year olds, 32 6 year olds, and 30 4 year olds.

B. Equipment and materials

Perceptual testing took place in a soundproof booth, with the computer that controlled the experiment in an adjacent room. The hearing screening was done with a Welch Allen TM 262 audiometer and TDH-39 earphones. Stimuli were stored on a computer and presented through a Creative Labs Soundblaster card, a Samson headphone amplifier, and AKG-K141 headphones. The experimenter recorded responses with a keyboard connected to the computer. Two pictures on 8 in. × 8 in. cards were used to represent the response labels of *buck* (a male deer) and *bug* (a ladybug). Game boards with ten steps were also used with children: they moved a marker to the next number on the board after each block of test stimuli. Cartoon pictures were used as reinforcement and were presented on a color monitor after completion of each block of stimuli. A bell sounded while the pictures were being shown and served as additional reinforcement.

C. Stimuli

Two sets of stimuli were created for this experiment: synthetic *buck/bug* that copied the patterns of frequency change across the first three formants of natural *buck* and *bug* tokens ("natural-formant" stimuli), and synthetic *buck/bug* that had a relatively high or low F1 frequency at syllable center ("high/low F1" stimuli), but ambiguous formant transitions near the syllable's end. Both sets of stimuli were created using the Sensyn Laboratory Speech Synthesizer. In both sets of stimuli, f0 started at 130 Hz and fell linearly throughout the stimulus to an offset frequency of 100 Hz. All stimuli were completely voiced, with no stimulus portion replicating voicing during closure for the final stop or a release burst. Each stimulus that was created was subsequently manipulated so that individual tokens varied along a continuum from 100 to 260 ms, in nine 20-ms steps. So, vocalic duration served as a nondynamic (i.e., static) property that has been shown to have a large influence on voicing decisions of English-speaking adults in experiments with synthetic stimuli that vary only vocalic duration and F1-offset transitions. All stimuli were presented as isolated words in a labeling task.

1. Natural-formant

The natural-formant stimuli were based on natural tokens of an adult, male speaker producing the words *buck* and *bug* in isolation. Nittrouer *et al.* (2005) provide complete results of acoustic analyses on these and similar words, and measures from that study served as a guide for creating these stimuli. In both kinds of stimuli, F3 was held constant at 2700 Hz until 50 ms before offset. For the more *buck*-like stimulus, F3 fell to an ending frequency of 2500 Hz, while in the more *bug*-like stimulus it fell to 2400 Hz. For the more *buck*-like stimulus, F2 started at 1000 Hz and rose linearly throughout the stimulus to 1200 Hz at stimulus offset. For the more *bug*-like stimulus, F2 was constant at 1000 Hz until 50 ms before offset, at which time it rose linearly to 1400 Hz. Regarding F1, it started at 400 Hz, rose linearly to 800 Hz over the first 50 ms, and remained at 800 Hz until stimulus offset for the more *buck*-like stimulus. For the more *bug*-like stimulus, F1 started at 400 Hz, rose linearly to 625 Hz over the first 50 ms, and remained at 625 Hz until 50 ms before offset, at which time it fell linearly to 250 Hz. Thus, all formants varied across the voicing conditions as they do in natural syllables: F2 and F3 more closely approximated a "velar pinch" and F1 was lower in frequency at stimulus offset in the voiced condition. Furthermore, F1 differed across the length of the stimuli as natural tokens of *buck* and *bug* do. There were 18 natural-formant stimuli: two formant patterns x nine vocalic durations.

2. High/low F1

For both of the high/low F1 stimuli, F3 was constant at 2700 Hz until 50 ms before offset, at which time it fell to its ending frequency of 2450 Hz (midpoint of settings for the voiced and voiceless conditions in the natural-formant stimuli). F2 was constant at 1000 Hz until 50 ms before offset, at which time it rose linearly to 1300 Hz (again, midpoint of the settings for the voiced and voiceless conditions in the natural-formant stimuli). Thus, both F2 and F3 were set to ambiguously signal voicing for the final stop. For the most *buck*-like of these high/low F1 stimuli, F1 started at 400 Hz, rose linearly to 800 Hz over the first 50 ms, and stayed at 800 Hz until 50 ms before offset. At that time it fell by 175 Hz to its ending frequency of 625 Hz. For the most *bug*-like of the stimuli, F1 started at 400 Hz, rose linearly to 625 Hz over the first 50 ms, and stayed at 625 Hz until 50 ms before offset. At that time it fell by 175 Hz to its ending frequency of 450 Hz. So F1 differed at syllable center across stimuli, but fell by the same amount at offset. There were 18 of these high/low F1 stimuli: two F1 patterns X nine vocalic durations.

D. Procedures

Adults attended one test session and children attended two. Screening procedures were completed first. Next, two sets of stimuli were used in pretesting. A set of completely natural tokens of *buck* and *bug* was presented, with whatever voicing during closure was in the signal and release bursts. These tokens were taken from three different adult, male speakers. Two tokens each of *buck* and *bug* were used from each speaker, making a total of 12 stimuli for the pretest. Listeners had to respond correctly to 11 of them to proceed to the next pretest. The next pretest consisted of the same 12 stimuli, only with the release bursts and voicing during closure removed. Listeners had to respond correctly to 11 of these stimuli in order to proceed to testing.

In testing, all listeners were presented with both sets of stimuli. The order of presentation of the natural-formant and the high/low F1 stimuli was randomized across listeners. The same procedures were followed for each set of stimuli. Practice items were presented before the testing began. Practice items consisted of *buck* and *bug* "best exemplars," which were the stimuli that should most strongly evoke the correct words. For example, for the natural-formant stimuli, the stimulus with the *buck*-like spectral settings that was 100 ms long and the stimulus with the *bug*-like spectral settings that

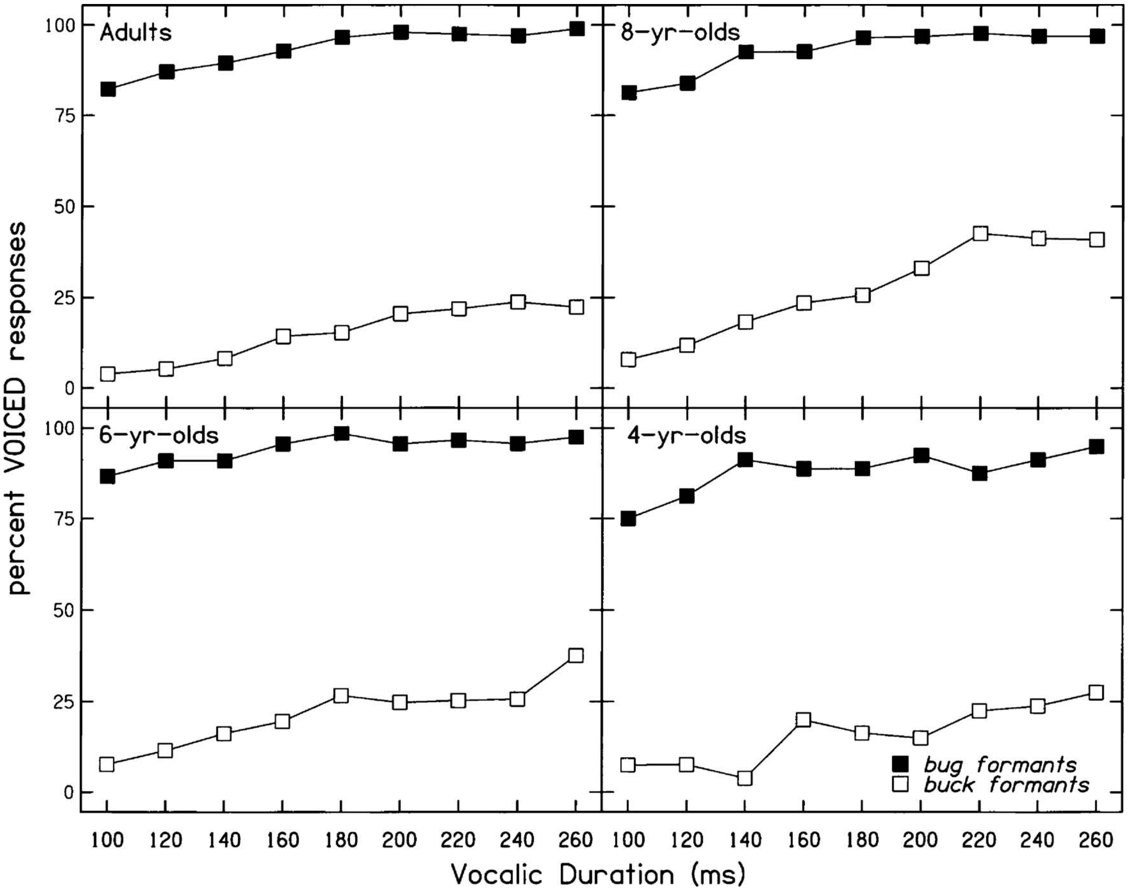

FIG. 2. Labeling functions for the natural-formant stimuli.

was 260 ms long were played, six times each (12 stimuli). The listener had to respond correctly to at least 11 of these stimuli to proceed to testing. During testing, ten blocks of stimuli were presented for both sets. Listeners responded by saying the label and pointing to the picture that represented their selection. To have their data included in the final analysis, participants needed to respond with at least 80% correct responses to these best exemplars during testing. This requirement served as a check that data were analyzed only from participants who maintained attention to the task: because all listeners were required to respond correctly to 90% of these stimuli during practice, they should have been able to do so during testing, if they maintained attention.

For children, cartoon pictures were displayed on the monitor and a bell sounded at the end of each block. They moved a marker to the next space on a gameboard after each block as a way of keeping track of how much more time they had left in the test.

The percentages of *bug* responses to each stimulus were tabulated and subsequently correlated with settings for vocalic duration and spectral structure (appropriate for *buck* or *bug*). The obtained regression coefficients indexed the weights that were assigned to each of these acoustic properties in perception. These correlation coefficients from individual listeners were used for statistical analyses.

III. RESULTS

Of the 30 4 year olds who met the criteria to participate, three did not meet the labeling criterion with unedited, natural *buck/bug* stimuli; that is, those that preserved voicing during closure and burst releases. Another ten did not meet criterion with natural *buck/bug* stimuli that had the voicing during closure and burst releases edited out. So, a total of 17 4 year olds participated in the testing.[1]

A. Natural-formant buck/bug

Nine 4 year olds, 11 6 year olds, and one 8 year old were unable to reach the criteria for having their data included because they failed either to label 90% of the best exemplars correctly during the pretest or to label 80% correctly during actual testing. Consequently, data were included for eight 4 year olds, 21 6 year olds, 23 8 year olds, and 21 adults.

Figure 2 shows mean labeling functions for each age group for the natural-formant *buck/bug* stimuli. Labeling functions were similar across age groups, suggesting that all listeners were responding similarly. Furthermore, it is clear that when the patterns of spectral change in formants across the utterance were appropriate for *bug*, all listeners responded with close to 100% *bug* responses; when these formant patterns were appropriate for *buck*, listeners responded with close to 0% *bug* responses. These functions are all extremely flat, as well, suggesting that children and adults did not weight vocalic duration very strongly at all in these stimuli. These patterns of responding replicate results for all contrasts created by editing natural stimuli in Nittrouer (2004).

TABLE I. Mean partial correlation coefficients for each age group, for each kind of acoustic property (vocalic duration and spectral) in each condition.

	Natural Formant		High/Low F1	
	Vocalic Duration	Spectral	Vocalic Duration	Spectral
4 year olds	0.21 (0.16)	0.90 (0.12)	0.46 (0.24)	0.65 (0.25)
6 year olds	0.19 (0.24)	0.87 (0.25)	0.39 (0.28)	0.72 (0.28)
8 year olds	0.27 (0.22)	0.84 (0.25)	0.48 (0.24)	0.67 (0.31)
Adults	0.17 (0.25)	0.88 (0.28)	0.45 (0.30)	0.73 (0.28)
Mean	0.21 (0.23)	0.87 (0.24)	0.44 (0.26)	0.70 (0.28)

In viewing labeling functions such as those in Fig. 2 it is difficult to get a sense of the variability in how listeners within a group labeled the stimuli, and so a sense of the variability in labeling across groups. For this reason we computed the means of the within-group standard deviations (of percent VOICED responses) across each of the 18 stimuli in this experiment. These means were 17 percent for adults, 17.5 percent for 8 year olds, 16 percent for 6 year olds, and 13 percent for 4 year olds. From these numbers it is concluded that variability in labeling was similar across groups.

The left side of Table I shows mean partial correlation coefficients (i.e., Pearson r) for each age group for the natural-formant stimuli. It can be seen that listeners in all age groups generally weighted spectral structure more than vocalic duration. A two-way analysis of varience (ANOVA) done on the coefficients, with property (vocalic duration vs spectral structure) and age as the main effects, showed a significant effect of property only, $F(1,76)=163.93$, $p < .001$. The Property X Age interaction was not significant. This result reveals that more of the variance in response patterns was explained by spectral structure than by vocalic duration, and this pattern did not vary with listener age.

B. High/low F1 stimuli

Seven 4 year olds, 13 6 year olds, one 8 year old, and one adult were unable to reach the criteria for having their data included during either the pretest or test. Consequently, data were included for ten 4 year olds, 19 6 year olds, 23 8 year olds, and 20 adults.

Figure 3 shows mean labeling functions for each age group for the high/low F1 stimuli. Again, children and adults appear to be weighting transitions heavily in their voicing decisions, and vocalic duration less so. This pattern appears similar across groups. In order to compare variability in labeling across groups we again computed means of the within-group standard deviations across each of the 18 stimuli in this experiment. Here we found means of 20 percent for adults, 17.5 percent for 8 year olds, 21 percent for 6 year olds, and 22.5 percent for 4 year olds. Again, these numbers were taken as an indication of similar variability across groups.

The right side of Table I displays mean correlation coefficients for each age group for these stimuli. A two-way ANOVA done on the coefficients with property and age as the main effects showed a significant effect of property only, $F(1,76)=17.10$, $p<0.001$, again revealing that correlation coefficients were greater for spectral structure than for vocalic duration. Again, the Property x Age interaction was not significant.

C. Comparison across stimulus types

Simple effects analysis was performed on correlation coefficients for each kind of acoustic property (vocalic duration and spectral structure) across the two stimulus types (natural-formant and high/low F1) separately to see if listeners weighted the acoustic properties differently depending on how much information was available in the overall spectral structure: Spectral structure more strongly signaled voicing in the natural-formant stimuli than in the high/low F1 stimuli. Both kinds of acoustic properties showed significant effects of stimulus type: vocalic duration, $F(1,76)=109.70$, $p<0.001$; spectral structure, $F(1,76)=59.97$, $p<0.001$. These results show that listeners weighted spectral structure less and vocalic duration more for the high/low stimuli than for the natural-formant stimuli. This difference in weighting strategy across the stimuli sets indicates that listeners pay more attention to spectral structure when that structure is more informative, as it was for the natural-formant stimuli compared to the high/low F1 stimuli in this experiment. When spectral structure is less informative, then listeners apparently turn their attention to other aspects of the acoustic structure that reliably signal phonetic identity, which in this case was vocalic duration. Neither the main effect of age nor the Stimulus Type x Age interaction were found to be significant in this analysis, indicating that adults and children showed similar shifts in perceptual attention across stimulus types. So adults and children based their voicing decisions on the same aspects of acoustic structure in both sets of stimuli, and the structure that they primarily used was the dynamic spectral structure of the formants across the syllables.

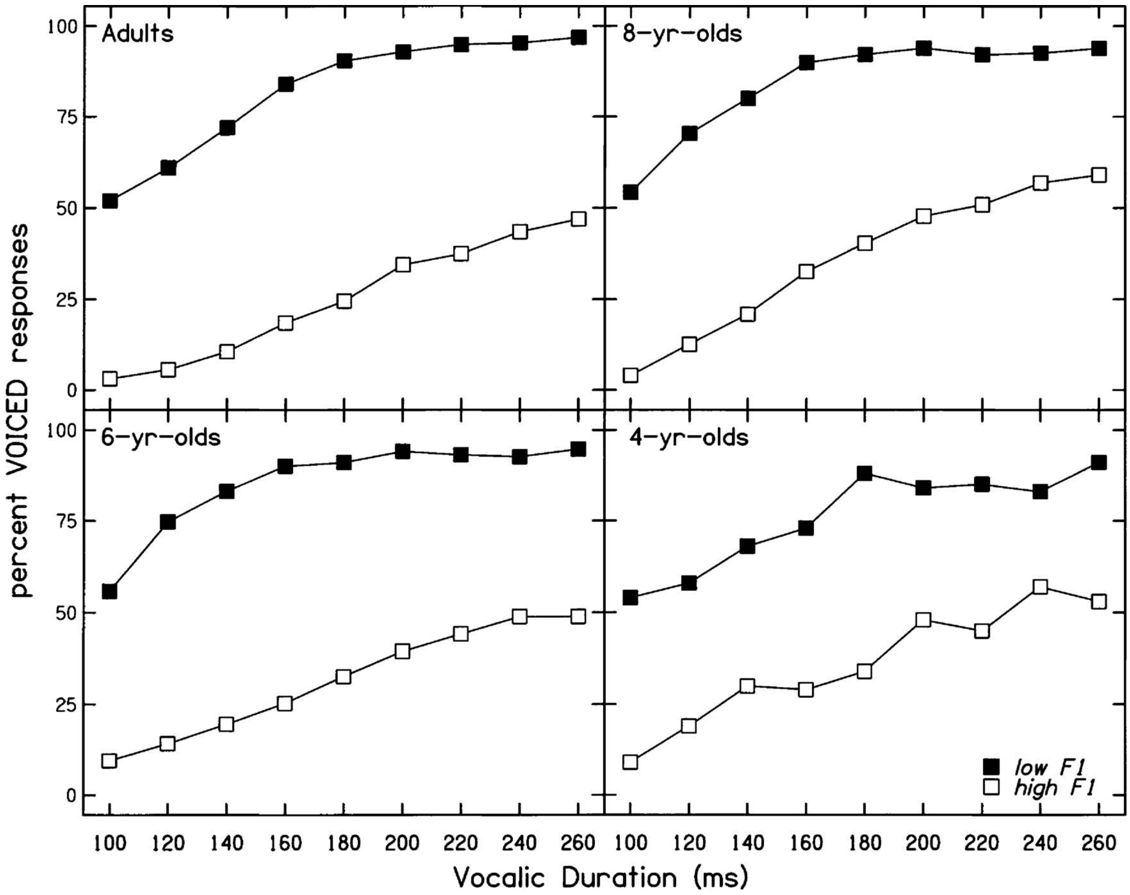

FIG. 3. Labeling functions for the high/low F1 stimuli.

IV. DISCUSSION

In this experiment, listeners were presented with synthetic stimuli that varied in two ways: vocalic duration varied along a continuum from short to long, and stimulus-internal spectral structure differed across all or most of the stimulus depending on the voicing of the final stop. All listeners were strongly influenced in their phonetic judgments by the spectral structure of those stimuli. Vocalic duration did not heavily influence responses for any age group. These findings contrast sharply with previous results from adults and children for synthetic stimuli that carefully controlled the spectral structure across most of the syllable, except for the final portion, and there varied it for only one formant, F1 (e.g., Greenlee, 1980; Krause, 1982; Nittrouer, 2004; Wardrip-Fruin and Peach, 1984). For those stimuli it was found that children, but not adults, used that extremely circumscribed spectral structure in their voicing decisions. Adults instead turned their perceptual attention to vocalic duration for making voicing decisions in those experiments, as they have been found to do in similar experiments dating back 50 years (e.g., Denes, 1955; Raphael, 1972; Wardrip-Fruin, 1982). Thus it may be concluded from results across those earlier experiments and this current experiment that younger listeners seek out dynamic spectral structure, and use whatever information of that nature they can find in the signal. Adults, on the other hand, do not use dynamic information that is impoverished. Perhaps it is the case that adults are more facile at shifting their perceptual attention as the situation demands: In the case of synthetic stimuli in earlier final-voicing experiments, adults may have shifted their attention completely to the temporal information when the spectral information was impoverished. Be that as it may, when available, adults weight dynamic spectral information more than the temporal information for final voicing decisions, and much as children do.

Of course, it is always tempting to attribute any age-related differences in speech perception to possible age-related differences in auditory sensitivities for the properties being manipulated. So, for example, the earlier results with synthetic speech stimuli for syllable-final stop voicing might indicate that adults are more sensitive than children to temporal properties. By the same reasoning, an auditory hypothesis would have to predict that children are more sensitive than adults to spectral glides because children have been found to weight that information more than adults. Nittrouer and Lowenstein (2007) examined the auditory hypothesis by manipulating temporal and spectral properties in nonspeech signals to replicate the structure of stimuli in earlier speech perception experiments with syllable-final voicing. Discrimination thresholds were obtained for nonspeech signals that varied in either duration or spectral structure. Similar sensitivity was found for adults and children for stimulus duration and spectral glides across three sine waves corresponding to the first three formants. When only the glide of the lowest frequency sine wave was manipulated (the condition corresponding most closely to those earlier synthetic speech stimuli), adults were actually slightly more sensitive than children to those glides. So, no evidence was found from experiments with nonspeech stimuli to support the contention that perceptual weighting strategies for speech signals are based on auditory sensitivity.

The findings reported here replicate the finding of Summers (1988) showing that adults make use of differences in F1 frequency across the syllable. The current study extends that work by demonstrating that children show similar effects. Of course, because the high/low F1 stimuli manipulated the onset and middle portions of the stimuli together, this study was unable to distinguish whether listeners were specifically weighting the onset transition or the steady-state information in their decisions. Whatever the case, it is clear that listeners of all ages use spectral structure that is not temporally constrained to the acoustic region traditionally associated with the segment they are being asked to label in those experiments: In the high/low F1 condition, the spectral structure at the ends of the syllables was ambiguous between the voiced and voiceless replicas.

A broader implication of the results reported here is that perhaps the traditional view of human speech perception must be revised. Perhaps our theory building regarding this important process has been ill served by the view that listeners extract discrete acoustic cues and then perform a cognitive translation of these properties into phonetic units. The results of the current study are consistent with a model of speech perception in which listeners attend to overall spectral structure, the kind of structure that arises from the relatively slow articulatory movements within the vocal tract. When the fine spectral structure within the syllable portion generally considered to be associated with the specific phonetic segment listeners were being asked to label was set to be ambiguous with regard to stop voicing, the weight that listeners assigned to vocalic duration increased slightly, but still decisions were largely based on the dynamic spectral pattern associated with the production of the whole syllable. This trend was as apparent for children as for adults. So, perhaps listeners attend primarily to overall patterns of spectral change during speech perception. Perhaps it is primarily in our psychophysical experiments where we restrict the sensory information available for decision making that we find that some listeners are able to turn their attention to other properties. And it should not surprise us that the listeners who are most capable of doing so are mature listeners who are native speakers of the language being manipulated, and so who have had the most experience hearing how other properties covary with overall spectral structure. For words ending in voiced and voiceless stops, this suggestion derives from the numerous studies showing that English-speaking adults are better able to use vocalic duration (when spectral structure is constrained) than English-speaking children (Greenlee, 1980; Krause, 1982; Nittrouer, 2004; Wardrip-Fruin and Peach, 1984) or adults who are native speakers of a language without a vowel-length distinction associated with final voicing (Crowther and Mann, 1994; Flege and Wang, 1989).

In summary, perhaps the early emphasis on the horizontal view, as exemplified by the work of Cooper and Liberman (Liberman, 1996) took us down a blind alley. To be sure, these investigators eventually turned their attention to alternative models of speech perception, as Liberman explains in his 1996 book. Nonetheless the view of the speech signal as a collection of discrete cues persisted in theory and experimental procedure. Have we spent decades dissecting speech stimuli and manipulating individual "cues," all the while ignoring the importance of integrated spectral structure to human speech perception? Only further investigation using paradigms that do not rely exclusively on manipulation of separate acoustic cues can answer this question.

ACKNOWLEDGMENT

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

[1] These numbers reveal that 37 percent of our 4 year olds were unable to label practice syllables if there was no voicing during closure or release burst. We wished to examine the hypothesis that this high failure rate may be at least partly explained by young children's inattention to the ends of syllables when there is not a salient marker to those syllable endings. To examine that possibility, we appended 12 ms of a natural /k/ burst taken from a *buck* sample to the ends of all natural-formant stimuli, 100 ms after stimulus offset. When we played these stimuli to a group of 15 4 year olds not included in the main study, using the same practice and test procedures, we found that only three children were unable to label the practice stimuli (20 percent). Testing was completed with the remaining children, and mean partial correlation coefficients computed on the data from their labeling results were 0.07 for vocalic duration and 0.97 for spectral structure. Although these values suggested a pattern of responding in which these 4 year olds weighted vocalic duration less and offset transitions more than any listener group in the main study, they were within a standard deviation of the mean correlation coefficients found for 4 year olds in the study (see Table I). Moreover, running the reported statistical analyses with these children included in the sample did not change any outcomes.

Crowther, C. S., and Mann, V. (**1994**). "Use of vocalic cues to consonant voicing and native language background: The influence of experimental design," Percept. Psychophys. **55**, 513–525.

Denes, P. (**1955**). "Effect of duration on the perception of voicing," J. Acoust. Soc. Am. **27**, 761–764.

Flege, J. E., and Wang, C. (**1989**). "Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast," J. Phonetics **17**, 299–315.

Goldman, R., and Fristoe, M. (**2000**). *Goldman Fristoe 2: Test of Articulation* (American Guidance Service, Inc., Circle Pines, MN).

Greenlee, M. (**1980**). "Learning the phonetic cues to the voiced-voiceless distinction: A comparison of child and adult speech perception," J. Child Lang **7**, 459–468.

Jastak, S., and Wilkinson, G. S. (**1984**). *The Wide Range Achievement Test-Revised* (Jastak Associates, Wilmington, DE).

Krause, S. E. (**1982**). "Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults," J. Acoust. Soc. Am. **71**, 990–995.

Liberman, A. M. (**1996**). *Speech: A special code* (MIT Press, Cambridge, MA).

Mirenda, P., and Beukelman, D. R. (**1987**). "A comparison of speech synthesis intelligibility with listeners from three age groups," Augmentative and Alternative Communication **3**, 120–128.

Morrongiello, B. A., Robson, R. C., Best, C. T., and Clifton, R. K. (**1984**). "Trading relations in the perception of speech by 5-year-old children," J. Exp. Child Psychol. **37**, 231–250.

Nittrouer, S. (**1996**). "The discriminability and perceptual weighting of some acoustic cues to speech perception by three year olds," J. Speech Hear. Res. **39**, 278–297.

Nittrouer, S. (**2004**). "The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults," J. Acoust. Soc. Am. **115**, 1777–1790.

Nittrouer, S., Estee, S., Lowenstein, J. H., and Smith, J. (**2005**). "The emergence of mature gestural patterns in the production of voiceless and voiced word-final stops," J. Acoust. Soc. Am. **117**, 351–364.

Nittrouer, S., and Lowenstein, J. H. (**2007**). "Children's weighting strategies for word-final stop voicing are not explained by auditory sensitivities," J. Speech Lang. Hear. Res. **50**, 58–73.

Nittrouer, S., and Studdert-Kennedy, M. (**1987**). "The role of coarticulatory effects in the perception of fricatives by children and adults," J. Speech Hear. Res. **30**, 319–329.

Parnell, M. M., and Amerman, J. D. (**1978**). "Maturational influences on perception of coarticulatory effects," J. Speech Hear. Res. **21**, 682–701.

Raphael, L. J. (**1972**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**, 1296–1303.

Raphael, L. J., Dorman, M. F., Freeman, F., and Tobin, C. (**1975**). "Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies," J. Speech Hear. Res. **18**, 389–400.

Raphael, L. J., Dorman, M. F., and Liberman, A. M. (**1980**). "On defining the vowel duration that cues voicing in final position," Lang Speech **23**, 297–307.

Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psychol. Bull. **92**, 81–110.

Reynolds, M. E., and Jefferson, L. (**1999**). "Natural and synthetic speech comprehension: Comparison of children from two age groups," Augmentative and Alternative Communication **15**, 174–182.

Summers, W. V. (**1987**). "Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses," J. Acoust. Soc. Am. **82**, 847–863.

Summers, W. V. (**1988**). "F1 structure provides information for final-consonant voicing," J. Acoust. Soc. Am. **84**, 485–492.

Wardrip-Fruin, C. (**1982**). "On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants," J. Acoust. Soc. Am. **71**, 187–195.

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.