Published in final edited form as:
Lang Speech. 2014 December ; 57(0 4): 487–512.

Dynamic Spectral Structure Specifies Vowels for Adults and Children

Susan Nittrouer and **Joanna H. Lowenstein**
The Ohio State University, USA

Abstract

The dynamic specification account of vowel recognition suggests that formant movement between vowel targets and consonant margins is used by listeners to recognize vowels. This study tested that account by measuring contributions to vowel recognition of dynamic (i.e., time-varying) spectral structure and coarticulatory effects on stationary structure. Adults and children (four-and seven-year-olds) were tested with three kinds of consonant-vowel-consonant syllables: (1) unprocessed; (2) sine waves that preserved both stationary coarticulated and dynamic spectral structure; and (3) vocoded signals that primarily preserved that stationary, but not dynamic structure. Sections of two lengths were removed from syllable middles: (1) half the vocalic portion; and (2) all but the first and last three pitch periods. Adults performed accurately with unprocessed and sine-wave signals, as long as half the syllable remained; their recognition was poorer for vocoded signals, but above chance. Seven-year-olds performed more poorly than adults with both sorts of processed signals, but disproportionately worse with vocoded than sine-wave signals. Most four-year-olds were unable to recognize vowels at all with vocoded signals. Conclusions were that both dynamic and stationary coarticulated structures support vowel recognition for adults, but children attend to dynamic spectral structure more strongly because early phonological organization favors whole words.

Keywords

Vowels; children; vocoded speech; sine-wave speech; dynamic specification

1 Introduction

A central problem in speech perception research has always concerned how the continuous and variable signal of speech gets recognized as strings of discrete and invariant phonemes. Historically, one influential account suggested there are brief regions of invariant and broad spectral structure that are recovered and used for recognition, while regions of spectral change between those invariant sections are discarded perceptually (e.g., Blumstein & Stevens, 1979; Stevens, 1975; Stevens & Blumstein, 1978). A different perspective proposed that the time-varying, or dynamic, structure itself is the source of information

Corresponding author: Susan Nittrouer, Department of Otolaryngology, The Ohio State University, 915 Olentangy River Rd., Suite 4000, Columbus, OH 43212, USA, Susan.Nittrouer@osumc.edu.

regarding phonemic identity (e.g., Kewley-Port, Pisoni, & Studdert-Kennedy, 1983). A related idea is that information about adjacent segments is transmitted in parallel (Fowler, 1980).

Vowels provide a conspicuous focus for debate regarding the bases of speech perception because casual impression is that vowels are well represented by stable syllable sections. In fact, the earliest instrumental analyses of speech readily yielded descriptions of vowels based on formant frequencies at stable regions within syllables (e.g., Joos, 1948; Potter & Steinberg, 1950). As a consequence, perceptual accounts have tended to attribute vowel recognition to the frequencies of the first two or three formants at steady-state regions (e.g., Clark, 2003; Ferrand, 2007; Ladefoged, 1982; Tye-Murray, 2009), and research into the nature of vowel recognition for static spectral signals continues (e.g., Fox, Jacewicz, & Chang, 2011). Of course, these accounts must be accurate for vowels spoken in isolation, but it is the rare utterance occurring in natural environments that consists of an isolated vowel. A complete and veridical account of how human listeners recognize speech signals must explain vowel recognition in syllables with consonant margins.

1.1 Dynamic specification of vowels

In 1976, a study by Strange, Verbrugge, Shankweiler, and Edman showed that vowels produced in syllables were identified more accurately than vowels produced in isolation. That result led to the suggestion that listeners rely more on the movement of formants through vocalic syllable portions than on just the target formant frequencies at relatively steady-state regions of syllables for judging vowel identity. In order to test that account, Jenkins, Strange, and Edman (1983) studied vowel recognition of “vowelless” syllables, a procedure in which naturally produced consonant-vowel-consonant (CVC) syllables were edited so that the middle 50–65% of the syllable was attenuated to silence. The authors found that listeners recognized the syllables with silent centers more accurately than they recognized the excised syllable centers themselves. The explanation given by the authors was that formant movement arising from the vocal tract switching between closure at syllable margins and opening at syllable center provides more information about vowel identity than the steady-state formants at syllable center. The term *dynamic specification* was coined by Strange and colleagues (e.g., Strange, Jenkins, & Johnson, 1983) to capture the idea that time-varying formant structure within the syllable specifies vowel identity.

Later studies by Jenkins, Strange and colleagues provided still more support for the dynamic specification account. Vowel quality for silent-center syllables was judged accurately in sentence context (Strange, 1989), even when talkers or syllable context were switched in the middle of the sentence (Jenkins, Strange, & Miranda, 1994; Jenkins, Strange, & Trent, 1999). These studies provided robust support for the idea that dynamic spectral structure across the syllable plays a significant role in vowel perception. The current study is a continuation of that line of work, reflecting the sentiment of Jenkins and colleagues that “a final test of the adequacy of our descriptions [regarding the dynamic specification account] will come from studies using synthetically generated speech in which the dynamic sources of information are manipulated.” (Jenkins et al., 1983, p. 449). The objective of the current study was to capture and preserve the dynamic structure of CVC syllables as separately as

possible from other speech-like qualities in order to test the hypothesis that it is explicitly this kind of time-varying spectral structure that explains the effectiveness of syllable margins in vowel recognition.

1.2 Children's speech perception

The original unit of organization in both production and perception of speech is generally recognized as being something more akin to the whole word, rather than the phonemic segment (Allen & Hawkins, 1978; Ferguson & Farwell, 1975; Macken, 1979; Menn, 1978; Nittrouer, 2006; Vihman, 1996). Consequently, it is not surprising that empirical studies of children's speech perception have shown that children attend more than adults to formant transitions, rather than to brief sections of relatively steady-state spectral structure associated with some consonants, such as fricatives (e.g., Mayo, Scobbie, Hewlett, & Waters, 2003; Nittrouer, 1992; Nittrouer & Lowenstein, 2009; Nittrouer & Miller, 1997a, 1997b; Nittrouer and Studdert-Kennedy, 1987; Parnell and Amerman, 1978). Formant transitions better represent whole syllable structure, the very kind of structure upon which children's perceptual attention is focused. Those outcomes were used to make the prediction that children would rely strongly on formant movement for vowel perception, as well: in other words, dynamic specification should be apparent in children's labeling of vowels. That prediction was tested by Nittrouer (2007) in a study modeled after those of Jenkins, Strange and colleagues (Jenkins et al., 1983; Strange et al., 1983): vowel recognition for vowelless syllables was compared to recognition for the excised middles. However, the set of vowel choices used with children was more limited because children cannot tolerate as much stimulus uncertainty (Wightman & Kistler, 2005). Furthermore, the excised syllable centers were replaced with natural coughs because children have difficulty integrating signal sections across long intervals when only silence occupies those intervals (Murphy, Shea, & Aslin, 1989). Listeners fail to notice that sections are missing when they are replaced with coughs (Warren, 1970). Instead what is heard are two acoustic streams, one consisting of the cough and the other of a seemingly uninterrupted syllable.

Results of the Nittrouer (2007) study revealed that children performed fairly well with the vowelless stimuli – certainly above chance – and better than they performed with the excised middles. These outcomes were taken to indicate that children can use the dynamic spectral structure found at syllable margins for vowel judgments. Nonetheless, children did not perform as well as adults with the vowelless stimuli. Thus, some ambiguity in interpretations existed, sparking the idea that there might be additional information in the syllable margins that adults use in vowel judgments, but children do not. The idea specifically emerged that the additional information might be stationary effects of vowel coarticulation at syllable edges.

1.3 Stationary effects of coarticulation

In addition to formant transitions, the spectral structure at every location across a syllable is influenced by both vowel and consonant production, a phenomenon traditionally termed *coarticulation* (e.g., Daniloff & Hammarberg, 1973). Information about adjacent segments is transmitted in parallel across the syllable, and perceptual studies have established that listeners are sensitive to the way that broad spectral structure is shaped by adjacent segments

(e.g., Fowler & Brown, 2000; Manuel & Stevens, 1995). Because the acoustic consequences of coarticulation are seen in signal stretches that are brief and relatively stable, it was thought this phenomenon of parallel transmission might explain outcomes of experiments with vowelless syllables, without appealing to dynamic specification. According to this account, the explanation would be that listeners – at least adults – recover and use the consequences of vowel coarticulation to be found on stable spectral regions at syllable edges to make vowel decisions. Children would not be expected to attend to this local spectral structure as astutely, precisely because their focus is more on whole syllable structure. However, it would be impossible to separate potential effects of this kind of process from that of attending to dynamic spectral structure per se with the methods used in earlier experiments. Instead, some kind (or kinds) of signal processing would need to be implemented that differentially preserved these aspects of acoustic structure, which is what the current study sought to do.

1.4 Signal processing

In this study, two kinds of signal processing were applied to natural CVC syllables. One signal-processing technique involved creating sine-wave speech (SWS) replicas of the natural syllables. In this process, sine waves replicate the center frequencies of the first three formants, thereby retaining the dynamic formant structure of the syllable. However, the signals representing those formants are extremely narrow in frequency, with none of the temporal fine structure of natural speech. Nonetheless, at any one location the frequencies of those separate sine waves reflect the influences of both consonant and vowel gestures. Thus, these SWS signals preserve stationary coarticulated structure, as well as dynamic structure, but without the qualities that typically mark acoustic signals as speech. These signals are commonly viewed as the most distorted type of synthetic replicas of speech (Fox et al., 2011), but more than any other type of signal they can be characterized as capturing and preserving the essence of dynamic structure.

The other kind of signal created for this study involved noise vocoding (VOC), a process in which the speech spectrum is divided into some number of frequency channels, and the time-varying amplitude structure of each channel computed. Derived values are used to modulate noise, band-limited to the same channels. Each separate channel preserves amplitude structure across time, so the signal integrated across those channels is appropriately viewed as preserving what may be termed the *gross spectral envelope*. This is a broad, poorly defined spectral shape, somewhat akin to what would be obtained with a very wide analyzing filter in spectrographic analysis or from a talker with a high fundamental frequency (Assmann & Katz, 2000; Ryalls & Lieberman, 1982). These VOC signals lack fine temporal or spectral structure. Although they preserve gross spectral shape – such as changes in spectral tilt affiliated with vowel height – these signals provide only rudimentary dynamic formant structure. Listeners in the current experiment would need to recover vowel identity primarily from the broadly shaped spectra in this listening condition. It might be predicted that listeners would be able to achieve this feat, at least for signal sections from syllable middles. That prediction follows from models of recognition for static vowels, suggesting that it is really the broad shape of the spectrum that explains vowel recognition, rather than separate formant frequencies (e.g., Bladon & Lindblom, 1981;

Chistovich & Lublinskaya, 1979; Ito, Tsuchida, & Yano, 2001). However, the situation was more complicated in the current experiment because recognition was tested for signal sections at syllable edges, where spectral structure is heavily influenced by consonant production.

These two kinds of processed speech signals were used in this study to examine the veracity of the dynamic specification account of vowel recognition. Ideal processing strategies for this purpose would permit an absolute disassociation of dynamic spectral structure from stationary coarticulated structure. Of course, that degree of disassociation is impossible because the dynamic structure of formants cannot be retained without those formants having some absolute frequency. Similarly, broad spectral shape cannot be meaningfully retained without some change across time, albeit slow. Within these constraints, the two signal-processing algorithms selected for use provided a reasonable test of the hypothesis: The VOC signals preserved stationary coarticulated structure, but only a poor depiction of dynamic structure. The SWS signals provided stationary coarticulated structure, with the added character of finely represented dynamic structure. Thus, any benefit to response accuracy for the SWS signals over the VOC signals could be attributed to that dynamic structure, especially since other speech-like qualities were missing in the SWS signals. It is critical to emphasize that this study was not undertaken as a test of which of these processing methods – noise vocoding or sine-wave synthesis – would evoke better vowel recognition. Rather, the experiment was conducted to evaluate as well as possible the contribution of dynamic structure to vowel recognition when all other kinds of speech structure were either eliminated or separately quantified.

Based on the kind(s) of information primarily preserved by each processing strategy, specific predictions could be made regarding adults' performance in the current experiment: (1) if the dynamic specification account of vowel recognition is accurate and primarily explains recognition, adults should perform well with SWS stimuli, and poorly with VOC stimuli; (2) if instead the earlier experiments with vowelless syllables suffered from a confound of dynamic and stationary coarticulated structure and listeners actually derive vowel identity primarily from stationary coarticulated structure, adults should show no benefit in performance for the SWS stimuli over VOC stimuli; (3) if both kinds of information (dynamic and stationary coarticulated) account for outcomes to a more or less similar extent, adults should perform better with SWS stimuli, but still be able to recognize vowels reasonably well with the VOC stimuli.

The primary prediction regarding children's performance derives from the theoretical account of early speech organization as being focused on whole words, or syllables, with perceptual attention directed to dynamic spectral structure. According to this position, children should show a greater difference in performance between SWS and VOC signals than adults, with a strong benefit observed for SWS stimuli. Dynamic spectral structure is the one aspect of speech well preserved in SWS signals, in spite of their degraded quality. Outcomes of experiments using sentence-length materials show that children perform comparably to adults with SWS signals, but more poorly with VOC signals, an outcome explained precisely by the hypothesis that children rely on dynamic spectral structure (Nittrouer & Lowenstein, 2010; Nittrouer, Lowenstein, & Packer, 2009).

1.5 Choice of syllable sets

This study used some stimuli from Nittrouer (2007), selected specifically to vary the extent of coarticulation at syllable edges. Two vowel contrasts were employed: a (mostly) high–low contrast with the vowels /ɪ/ and /æ/ and a (mostly) front–back contrast with the vowels /æ/ and /ʌ/. Although it is difficult to design vowel contrasts that involve differences purely in height or frontedness, these contrasts come as close to doing so as possible, and they allow the use of real or quasi-real words. In production terms, the high–low contrast can largely be implemented with a difference in jaw height. In this case, both vowels are produced towards the front of the mouth, so tongue posture is similar for both. The front–back contrast reflects primarily a difference in tongue posture: fronted or backed.

Two consonant contexts were used: /b/ and /d/. The labial /b/ is produced with the lips, not the tongue. Syllable production for /bVb/ sequences can be modeled largely by having the vocal tract assume a static shape across the syllable, and varying only the cross-sectional area of the lips (Manuel & Stevens, 1995). The alveolar /d/ involves the tongue blade, which must touch the alveolar ridge. Because of the anatomical constraints of the tongue, this means the tongue body must be raised and fronted for stop closure. These differences in production lead to the expectation that /d/ production would limit coarticulation with vowel production to a greater extent. This limitation has been dubbed *coarticulatory resistance*. Its existence is supported by findings of greater variability in formant frequencies (especially F2) across vowels at the start of the transition for labial consonants compared to alveolars (e.g., Recasens, 1985; Sussman, McCaffrey, & Matthews, 1991).

Given the syllables to be used in the current experiment, predictions could be made about vowel labeling with whole syllables, before any parts of the middles were removed. Firstly, the high–low contrast in the /b/ context (/b/, high–low) should result in the most accurate vowel labeling across processing methods because production largely consisted of only changes in jaw height and lip aperture. Jaw height for each vowel can be set throughout syllable production and closure achieved with only lip movement, so there is information about the vowel throughout. Listeners should do well with these syllables, even with VOC stimuli, where dynamic spectral structure is lacking. The front–back contrast in the /d/ context (/d/, front–back) should show the poorest performance in vowel labeling when dynamic spectral structure is missing, as it is in VOC stimuli. In these syllables, the tongue body is constrained by the requirements of stop production at either edge of the syllable. As a result, it is difficult to coarticulate vowels at those edges because vowels are also dependent on tongue posture. That means there is little stable coarticulatory structure at syllable edges. On the other hand, these syllables have strong dynamic specification of vowel identity, arising from the fact that the tongue needs to move from that alveolar constriction to the vowel target. If the dynamic specification account is correct, listeners should have difficulty labeling these stimuli even in their whole form with the VOC stimuli, but should perform well with SWS stimuli.

Accuracy of vowel labeling with VOC stimuli for the two remaining syllable sets should fall intermediate between the two syllable sets described above (i.e., the /b/, high–low syllable set and the /d/, front–back syllable set). For production of syllables in the /b/, front–back set,

the tongue body is integral to vowel production, but not to stop production. This means that the tongue can be positioned for the vowel largely throughout syllable production, leading to stationary coarticulatory effects at syllable edges. However, lip closure at syllable margins has some effect on the frequency of F2, which is more important to this vowel contrast than to the high–low contrast. For production of syllables in the /d/, high–low set, jaw height largely accounts for the vowel distinction. The tongue is involved in stop production, but unlike bilabial constrictions, alveolar constrictions affect jaw height to some extent. Thus, these last two syllable types allowed more coarticulation than the front–back contrast in the /d/ context, but not as much as the high–low contrast in the /b/ context.

Finally, signal sections of two sizes were removed from the middles of these syllables, and replaced with coughs. (1) A section half the size of the vocalic syllable portion was removed from the temporal middle of the syllables. Given the duration of the syllables and the fundamental frequency (F0) of the talker, this left an average of 64 ms on either side of the cough, which should be long enough to preserve sufficient dynamic spectral structure to support vowel recognition, if that structure specifies vowels. (2) In another manipulation, all but the first three and last three pitch periods were removed from the middle of the syllables, and replaced with coughs. Less dynamic structure can be preserved in sections this short, so vowel recognition should be negatively affected, according to the dynamic specification account.

1.6 Summary

In summary, the current experiment was undertaken to explore the acoustic and perceptual basis of vowel recognition. The act of producing a CVC syllable gives rise to a dynamic spectral structure in which the spectrum is shaped at every instance by both consonant and vowel. The goal of the current work was to test the extent to which adults and children base their vowel decisions on the dynamic structure itself, rather than on the stationary consequences of coarticulation. Natural tokens of CVC sequences were used, and processed in each of two ways. VOC stimuli preserved the relatively stable spectral structure arising from coarticulation, but had severely restricted dynamic spectral structure. SWS stimuli preserved both kinds of signal structure, although other qualities of speech were lost. Children participated, as well as adults, because it has been found that children attend more than adults to the kind of dynamic spectral structure hypothesized to underlie the accurate recognition of vowels when sections of the syllable's middle are missing. Sections of two sizes were removed from the middle of syllables, with different predictions about how each manipulation would affect outcomes.

2 Method

2.1 Listeners

Twenty adults between the ages of 18 and 32 years, 40 seven-year-olds ranging in age from seven years; one month to seven years; five months, and 12 four-year-olds ranging in age from four years; three months to four years; 11 months participated. Twice the number of seven-year-olds participated as adults because each seven-year-old was tested with only one

vowel contrast (high– low or front–back). Each adult was tested with both vowel contrasts. Four-year-olds were tested with the high–low vowel contrast only.

None of the listeners (or their parents, in the case of children) reported any history of hearing or speech disorder. All listeners passed hearing screenings consisting of the pure tones of .5, 1, 2, 4, and 6 kHz presented at 25 dB HL to each ear separately. Parents of the four- and seven-year-olds reported that their children were free from significant histories of otitis media, defined as six or more episodes during the first three years of life.

Even though all listeners (or their parents) reported normal speech and language abilities, they were nonetheless screened. Adults were given the reading subtest of the Wide Range Achievement Test 4 (Wilkinson & Robertson, 2006) and all demonstrated better than a 12th grade reading level. Children were given the Goldman Fristoe 2 Test of Articulation (Goldman & Fristoe, 2000) to screen for speech disorders, and were required to score at or better than the 30th percentile for their age in order to participate. For seven-year-olds, that meant they could produce no more than four errors. Thirty-seven of the 40 seven-year-olds had no errors, one had two errors, and two had three errors. The four-year-olds showed a broader range of performance (from the 31st to the 98th percentile), with a mean score at the 56th percentile (*SD* = 21). Children were also given the Peabody Picture Vocabulary Test–4th Edition (PPVT-IV) (Dunn & Dunn, 2007). This is a test of receptive vocabulary. Children were required to achieve a standard score of at least 92 (30th percentile) on this task. The mean PPVT-IV standard score for seven-year-olds was 118 (*SD* = 11), which corresponds to the 88th percentile. The mean PPVT-IV standard score for the four-year-olds was 120 (*SD* = 11), which corresponds to the 91st percentile. These scores indicate that these children had receptive vocabularies roughly 1 *SD* above the mean of the published normative sample.

2.2 Equipment and materials

All speech samples were recorded in a soundproof booth, directly onto the computer hard drive, using an AKG C535 EB microphone, a Shure M268 amplifier, and a Creative Labs Soundblaster 16-bit analog-to-digital converter. A waveform editor (WAVED) (Neely & Peters, 2005) was used for recording and editing. An acoustic analysis program (Praat) (Boersma & Weenink, 2009) was used for spectral analysis.

All testing took place in a soundproof booth, with the computer that controlled stimulus presentation in an adjacent room. Hearing was screened with a Welch Allyn TM262 audiometer using TDH-39 headphones. Stimuli were stored on a computer and presented through a Creative Labs Soundblaster card, a Samson headphone amplifier, and AKG-K141 headphones. This system has a flat frequency response and low noise. Custom-written software controlled the audio presentation of the stimuli. The experimenter recorded responses with a keyboard connected to the computer. Two pictures (8 in. × 8 in.) were used with each syllable set to represent response labels. These included a baby's bib for *bib*, a baby babbling for *bab*, bubbles for *bub*, a girl showing off a picture she drew for *did*, a man for *dad*, and an unexploded firecracker for *dud*.

2.3 Stimuli

Natural productions of /bɪb/, /bæb/, /bʌb/, /dɪd/, /dæd/, and /dʌd/ from a male talker were used. Although not all of these syllables are real words on their own, they are all phonotactically legal sequences in English. The six syllables were grouped into four sets, depending on how the vowel was being contrasted (high–low or front–back) and consonant context (/b/ or /d/). For the high– low contrast, the vowels /ɪ/ and /æ/ were used; for the front–back contrast, the vowels /æ/ and /ʌ/ were used. The talker had a F0 of roughly 122 Hz, with no extreme breathiness or fry. The talker was recorded producing five samples of each syllable in random order. Samples were digitized at a 22.05-kHz sampling rate with 16-bit digitization. One production of each syllable that the authors agreed was representative of that syllable was used as a stimulus in the listening experiment. Table 1 lists the F1, F2, F3, and F0 at onset, midpoint, and offset for each syllable. To illustrate what these syllables look like, Figure 1 shows spectrograms of /dɪd/, /dʌd/, and /bæb/. All stimuli created from these natural productions (without further processing) will be described as unprocessed (UP) stimuli.

A separate talker, a woman, was recorded producing several coughs. Various sections of these coughs were used to replace stimulus sections removed from the middle of the syllables. Using coughs from a talker of a different gender helped ensure that the coughs would be perceived as a separate stream from the speech.

2.3.1 Processed stimuli—These natural stimuli were processed to create both SWS and VOC stimuli. To make the SWS stimuli, each UP syllable was analyzed in Praat (Boersma & Weenink, 2009) and the tracks for the first three formants were extracted. The formant tracks were examined for anomalies and adjusted if necessary, then transferred to TONE software (Tice & Carrell, 1997) for sinewave synthesis. The resulting stimuli retained the center frequencies of the first three formants. TONE was used to synthesize syllables at a 22.05-kHz sampling rate.

A MATLAB routine was used to create the VOC stimuli from the UP syllables. All signals were first low-pass filtered below 8000 Hz. The spectrum was divided into four channels, a number shown to support recognition of changes in vowel quality for adults and five- to seven-year-olds in consonant-vowel (CV) syllables (Eisenberg, Shannon, Schaefer Martinez, Wygonski, & Boothroyd, 2000). Cutoff frequencies between channels were .8, 1.6, and 3.2 kHz. This meant that three channels covered the frequency range affiliated with the first three formants (i.e., below 3 kHz), so there was some correspondence between the SWS and VOC stimuli in this respect. Each channel was half-wave rectified, and low-pass filtered using a 160-Hz high-frequency cutoff. Results were used to modulate white noise limited by the same band-pass filters as those used to divide the speech signal into channels. Overall rms amplitude of SWS and VOC stimuli was equalized to match the UP stimuli.

Figure 2 shows spectrograms for VOC *bib* and *bub*, and illustrates that dynamic structure is highly restricted for these kinds of signals, but the structure arising from CV coarticulation is present throughout the vocalic sections, including the edges.

2.3.2 Missing middles—Next, portions were removed from the middle of each UP stimulus in each of two sizes, and replaced with sections of the coughs that were the same in duration as the removed syllable portions. All waveform editing was done with cuts made at positive-going zero crossings at the start of a pitch period. For the "50% cough" stimuli, this meant that as close as possible (preserving whole pitch periods), syllables had the middle 50% of the vowel replaced with a cough section. This was done by measuring the duration of the vocalic (unclosed) portion of the syllable. The temporal middle of that portion was found, and 25% of the vocalic duration was removed from either side of that, leaving an average of 64 ms on each side of the cough. The "three pitch period" stimuli had all but the first three and last three pitch periods of the vowel replaced with a cough section, preserving on average 26 ms on either side. Precise locations of starts and ends of these sections were recorded to the millisecond. Those times were then used to replace exactly the same portions of the SWS and VOC stimuli with exactly the same cough sections. This resulted in three kinds of stimuli for each processing type: whole, 50% cough, and three pitch period.

In summary, there were 12 stimulus conditions: 2 vowel contrasts × 2 consonant contexts × 3 processing types. Each condition was presented separately, but with stimuli of all three kinds of missing middles included (whole, 50% cough, and three pitch period). Each stimulus was presented 20 times. This meant that testing in each condition included a total of 120 stimuli: 2 vowels × 3 kinds of missing middles × 20 tokens of each.

2.4 Procedures

All procedures were approved by the local Institutional Review Board. Informed consent was obtained prior to testing.

2.4.1 General procedures—Listeners came into the laboratory for two sessions, typically on consecutive days. Screenings were administered first, then testing was conducted. Adults heard one vowel contrast (high–low or front–back) in the first session and the other in the second session, which meant that they were presented with six conditions each day: 2 consonant contexts × 3 processing types. Seven-year-olds and four-year-olds heard only one contrast, presented over the two sessions. For four-year-olds, this was always the high–low contrast.

In testing with adults, the stop context always alternated. Half the adults were presented with the /b/ context first and half heard the /d/ context first. Testing started with the other context in the second session. All SWS and VOC stimulus conditions were presented before UP stimuli in each session so that listeners would not be biased by hearing the UP stimuli first. Half the listeners started with SWS stimuli and half started with VOC stimuli. Thus, a possible presentation order for an adult might be as follows. In session #1, for the high–low contrast:

d-SWS, b-VOC, d-VOC, b-SWS, d-UP, b-UP. Then in session #2, for the front–back contrast:

b-VOC, d-SWS, b-SWS, d-VOC, b-UP, d-UP. In all, there were 16 possible orders of presentation for adults.

For four- and seven-year-olds the conditions presented to adults for a single contrast were presented with the same assortment in order, but divided over the two test sessions. The PPVT-IV was administered at the end of the second session.

2.4.2 Training—Training was provided separately before testing within each stimulus condition and consisted of whole syllables from that stimulus condition. The pictures were introduced one at a time by the experimenter, and the appropriate label told to the listener. There was a brief discussion about how the picture related to the word (e.g., "This girl is showing you what she *did*"). Next the experimenter said each of the two syllables five times in random order, for a total of ten syllables. Listeners were asked to say the word they heard and point to the appropriate picture for their response. Having both pointing and labeling responses helped to ensure that the listener was paying attention to the task and could correctly match the word to the picture.

For the SWS and VOC stimulus conditions, listeners were told they would hear a robot saying the words: for SWS, a squeaky-voiced robot, and for VOC, a scratchy-voiced robot. Listeners heard ten whole syllables via the computer (five of each vowel) in random order and were asked to match them to the pictures. Feedback was given, if necessary. Then listeners heard the whole syllables again and had to respond correctly to nine out of ten of the stimuli without feedback to proceed to testing. If an error was made, feedback was provided for one more set. Listeners were given three chances to meet the criterion of responding correctly to nine of ten stimuli, without feedback. If they did not reach that criterion by the third set of training, they did not proceed to testing for that stimulus type.

2.4.3 Testing—Testing immediately followed training in each stimulus condition. Listeners were told that the experimenter had a cough when she was recording the robot (or person, in the case of the UP stimuli), so they might hear some coughing while the robot or person was talking. Listeners were instructed to ignore the coughing and just continue reporting which word the robot or person said. No feedback was provided.

Game boards with ten steps were used with children, who moved a marker to the next number on the board after each block. Cartoon pictures were presented on a computer monitor and a bell sounded as additional reinforcement indicating the end of each block.

2.5 Analyses

The dependent measure in this experiment was the percentage of stimuli in each condition recognized as containing the originally produced vowel. Patterns across stimulus conditions were examined using statistical analyses on percent correct scores. Arcsine transforms were used because percentages were close to 100% correct in some cases, particularly for the UP stimuli. Precise p values are reported when $p < 0.10$; otherwise results are reported as *not significant* (NS). Where multiple comparisons were involved, Bonferroni adjustments were used. Partial eta squares were used to index the size of specific effects.

Statistical analyses were performed to examine predictions related to outcomes in four areas. These predictions were based on what would likely be found, if the dynamic specification account accurately explains vowel recognition.

2.5.1 Syllable sets—Predictions were made regarding the relative ease of the four syllable sets. It was predicted that /b/, high–low stimuli would evoke the most accurate recognition across processing methods and missing middles because articulatory gestures used for the consonant and for the vowel are the most independent in these syllables. The /d/, front–back syllable set was predicted to result in the poorest performance across processing and missing middles because of the strong interdependence of articulators for consonant and vowel production. The other two syllable sets were predicted to show responses intermediate to those two sets in terms of accuracy.

2.5.2 Processing effects—The next class of predictions had to do with processing effects. If the dynamic specification account is correct, then recognition should generally be better for the SWS stimuli, compared to the VOC stimuli, because SWS stimuli preserve dynamic spectral structure, as well as stationary coarticulated structure. The VOC stimuli largely preserve only the latter.

2.5.3 Missing middles—Predictions were made regarding the effects of removing either half of the vocalic middle or all but three pitch periods at either edge. These predictions were expected to interact with the type of signal processing. For the SWS signals, it was predicted that vowel recognition would remain high when half of the syllable was available because that manipulation preserves sufficiently long stretches of dynamic spectral structure. Accuracy would likely drop precipitously when only three pitch periods were available at either edge for these SWS signals, because little dynamic structure can be preserved in sections so short. For VOC signals, decline in recognition accuracy was predicted to be gradual across missing middles. Recognition for these vowelless stimuli would depend on the stationary coarticulated structure, which would be present in each slice of the signal, although in diminishing strength as distance from the vowel target increased. Thus, accuracy should decline gradually as signal sections are removed.

2.5.4 Age effects—Finally, predictions were made about how children's performance might differ from that of adults. Although children might perform generally more poorly than adults, the age-related difference should be disproportionately stronger for the VOC stimuli. With those stimuli, listeners who rely more strongly on dynamic spectral structure (as children do, compared to adults) would likely be more seriously disadvantaged.

3 Results

3.1 Adults

3.1.1 Syllable sets—Predictions were made regarding the relative difficulty of individual syllable sets. The first outcome examined had to do with those predictions. Table 2 shows percent correct responses (and *SD*s) for each syllable set, across processing types and missing middles. As predicted, recognition was best for the /b/, high–low syllable set and poorest for the /d/, front–back syllable set, shown in the first and last columns of the table, respectively. Recognition was intermediate to those two outcomes for the other two syllable sets: /b/, front–back and /d/, high–low. These scores are shown in the middle columns of Table 2.

A two-way analysis of variance (ANOVA) with vowel contrast and stop context as within-subjects factors was performed on adults' recognition scores shown in Table 2. Both factors were significant: vowel contrast, $F(1, 19) = 10.68$, $p = 0.004$, $\eta^2 = 0.15$, and stop context, $F(1, 19) = 74.78$, $p < 0.001$, $\eta^2 = 0.42$. The effect sizes indicate that stop context accounted for more variability in responding than did vowel contrast. The interaction of vowel contrast × stop context was also significant, $F(1, 19) = 12.79$, $p = 0.002$, $\eta^2 = 0.02$, but the effect size was small.

3.1.2 Processing effects—Responses for whole syllables only were analyzed as an initial step in separating the effects of dynamic spectral and stationary coarticulated structure on vowel recognition. Table 3 shows percent correct responses for whole syllables, for each vowel contrast × stop context and processing type. Adults were most accurate for UP, nearly as accurate for SWS, and least accurate for VOC. These patterns suggest that the reduction in dynamic spectral structure for the VOC stimuli had negative effects on vowel labeling because that was specifically the structure that was lacking in these stimuli. Recognition of VOC stimuli was especially poor for /d/, front–back stimuli, much lower than for any other kind of syllable. That outcome had been predicted because recognition of VOC stimuli would be highly dependent on the stationary consequences of coarticulation, and in this syllable, coarticulation was highly constrained.

Table 4 shows results of *t*-tests comparing responses for each processing type for each syllable set separately. For the /b/, high–low syllable set, there were no differences in accuracy of responding across processing types: all *t*-tests produced non-significant results. This had been predicted to be the syllable set that would produce the most accurate recognition scores because coarticulation is at its maximum. Looking down the table it is seen that there was no difference between the UP and SWS processing conditions for the other three syllable sets, but the comparison of UP and VOC was significant for all three of those syllable sets. These results support the suggestion that the diminishment of dynamic spectral structure in VOC signals resulted in poorer performance: when coarticulation was not especially strong, listeners needed that dynamic spectral structure to make vowel judgments. That dynamic spectral structure was present in the SWS stimuli, but was lacking in the VOC stimuli. For a similar reason, the SWS versus VOC comparison was significant for both front–back contrasts, where having dynamic structure available in the signal was predicted to be most useful. Particularly in the case of the /d/, front–back syllable set, where coarticulation is most constrained, dynamic spectral structure should play the largest role in recognition.

3.1.3 Missing middles—Next, recognition scores across stimuli based on the amount of syllable center removed were examined. Figure 3 shows percent correct responses for adults for each processing type and amount of missing middle, for each syllable set separately. As expected, performance remained high for UP stimuli, regardless of how much of the syllable was removed. For the two kinds of processed signals, specific predictions were made about how recognition would be affected across missing middles, according to the availability of dynamic spectral and stationary coarticulated structure. If dynamic spectral structure is important to vowel recognition, there should be good recognition for the 50% cough stimuli

for SWS processing, where dynamic spectral structure is well preserved. A decrement in performance would be predicted for the three pitch period stimuli, where dynamic structure is restricted. Indeed, that pattern was found for SWS stimuli for all four syllable sets, but especially for the /d/ context. For VOC stimuli, on the other hand, there appears to be a linear decline in recognition accuracy as the amount of syllable replaced by the cough increases, for all but the /d/, front–back stimuli. For those stimuli, recognition is poor, even for whole syllables. Of course, it is tricky trying to interpret these patterns in the absence of more fine-grained adjustments to the amount of syllable sections remaining, but the patterns do suggest that recognition for VOC stimuli declines linearly as the amount of syllable replaced by cough increases, while recognition for SWS stimuli remains high until some critical point, presumably when dynamic structure ceases to be sufficient.

An index was obtained of how much information about vowel identity was preserved by each kind of missing-middle stimulus by calculating differences in labeling scores for (1) whole syllables versus 50% cough stimuli and (2) 50% cough versus three pitch period stimuli. These two differences were termed *change score 1* and *change score 2*, respectively. This computed measure was used instead of repeated-measures ANOVAs on the recognition scores themselves because it provides a better test of the magnitude of change. Repeated-measures ANOVAs, even if post hoc contrasts are done, can only assess if the measures differed, not the magnitude of difference, or change. Table 5 shows these scores.

The change scores shown in Table 5 reinforce impressions obtained from Figure 3. Looking first at UP stimuli, little change in performance across missing middles is found for stimuli in the /b/ context. The *t*-tests performed on the change scores in the /b/ context are not significant for UP stimuli. However, for stimuli in the /d/ context, change score 2 was significantly larger than change score 1, meaning that performance degraded significantly more from the 50% cough to the three pitch period stimuli than from the whole syllable to the 50% cough stimuli. That finding must be due at least partly to the fact that the /d/ context constrains vowel coarticulation, but it might also be explained by the fact that there is less dynamic spectral structure present with only about 26 ms of signal in the three pitch period stimuli than with the roughly 64 ms of signal in the 50% cough stimuli. With little coarticulatory information or dynamic spectral structure upon which to make decisions, adults performed more poorly.

Results for SWS stimuli help to clarify the story. For all syllable sets, there is little change in performance from whole syllables to 50% cough stimuli, but a rather steep drop between 50% cough and three pitch period stimuli. In fact, *t*-tests done on the change scores were significant for all four syllable sets, indicating that the change between 50% cough and three pitch periods is greater than that observed for whole syllables and 50% cough. These outcomes support the conclusion that listeners can use dynamic spectral structure for making vowel decisions, and that structure is adequately preserved when the middle half of the syllable is missing. When only three pitch periods are preserved, however, recognition deteriorates.

Results for VOC stimuli are different still. Other than the /d/, front–back contrast where performance had already deteriorated for whole syllables, performance degrades in a linear fashion as missing middles become larger. The change scores are not significantly different in these other conditions. These results indicate that the property being lost as progressively larger syllable portions are removed has a linear effect on recognition. That trend fits the specific suggestion about what would be found if it is stationary coarticulated structure that explains listeners' abilities to recognize vowelless syllables. That kind of structure is distributed across the syllable, and is gradually lost as progressively longer sections are removed.

In summary, results from adults support the contention that dynamic spectral structure supports vowel labeling decisions. The SWS stimuli, which preserve that sort of structure especially well with at least the 50% missing-middle stimuli, supported more accurate labeling responses than did the VOC stimuli in which that structure is not well preserved. However, performance never reached chance levels for the VOC stimuli, suggesting that adults likely use information from coarticulation, as well, in their vowel labeling decisions. Outcomes from children should help shed light on these conclusions because children would be predicted to use dynamic spectral structure, but not necessarily stationary coarticulatory effects, in making vowel decisions.

3.2 Children

Even though four-year-olds only listened to the vowel contrast predicted to be easiest (high–low), the task was difficult for them with the VOC stimuli. All 12 four-year-olds met criterion for participation with the UP stimuli, and nine out of 12 reached criterion with the SWS stimuli. However, only five four-year-olds (less than half) reached criterion to participate with the VOC /b/ stimuli, and no four-year-old reached criterion with VOC /d/ stimuli. Because of this difficulty with VOC stimuli, subject recruitment was discontinued with four-year-olds and data from the four-year-olds who were tested were not included in analyses for any condition. Nonetheless it seemed worthwhile to report this attempt to test four-year-olds because of the implication that young children depend heavily on dynamic spectral structure for vowel recognition: when that structure was missing, as it largely was in the VOC stimuli, four-year-olds were greatly disadvantaged in their efforts to label vowels.

All 40 of the seven-year-olds reached criterion for participation in testing with the UP stimuli. Almost all reached criterion with the SWS stimuli: specifically, for /b/, high–low stimuli, 20 out of 20 children reached criterion; for /d/, high–low stimuli, 19 out of 20; for /b/, front–back stimuli, 19 out of 20; and for /d/, front–back stimuli, 18 out of 20. However, the VOC stimuli presented problems for these seven-year-olds: 16 out of 20 reached criterion for participation with the /b/, high– low stimuli, but for the other three sets, just 11 out of 20 seven-year-olds met criterion.

The fact that more four- and seven-year-olds were able to reach criterion with the VOC stimuli for the /b/, high–low syllable set than for any of the other syllable sets can be seen as evidence that children are sensitive to stationary coarticulatory effects, but only in a context where that structure is very strong.

3.2.1 Syllable sets—Table 6 shows percent correct responses (and *SD*s) for seven-year-olds for each syllable set, across processing types and missing middles. Overall, scores were lower than those of adults (Table 2). Possible reasons for that age-related difference are examined in the Section 3.3. When it comes to relative accuracy across syllable sets, however, the pattern is the same for children as that observed for adults. In particular, it appears seven-year-olds were best at recognizing the /b/, high–low stimuli and poorest at recognizing the /d/, front–back stimuli. As was done with adults' data, a two-way ANOVA was performed on these scores. In this case, vowel contrast was a between-subjects factor and stop context was a within-subjects factor. The effect of vowel contrast did not reach statistical significance, $F(1, 38) = 3.03$, $p = 0.09$, but stop context was significant, $F(1, 38) = 14.20$, $p < 0.001$, $\eta^2 = 0.12$. The failure to find a significant effect of vowel contrast differs from what was obtained with adults' data. Nonetheless, the general pattern replicates what was found with adults in that stop context had a greater effect on vowel recognition than did vowel contrast. For these seven-year-olds, the interaction of vowel contrast × stop context was not significant. It was for adults, but with a very small effect size.

3.2.2 Processing effects—Table 7 displays percent correct recognition for whole syllables only, for each syllable set and processing type. Seven-year-olds clearly did as well as adults with UP stimuli, but it appears their performance was poorer for both types of processed signals, even though they could hear the entire syllable. That trend suggests that children may simply be deleteriously influenced by signal processing of any kind. At the same time, children's performance appears better for SWS than for VOC stimuli, supporting the suggestion that children are sensitive to dynamic spectral structure, but not so much to stationary coarticulated structure. Table 8 presents results from *t*-tests comparing recognition scores for each processing condition, for each syllable set. All differences in scores for UP and both sorts of processed stimuli were significant. When it comes to SWS versus VOC stimuli, all differences were significant before Bonferroni adjustments were applied, and with one exception, after the adjustment, as well. Thus it may be concluded that children performed better with SWS than with VOC stimuli, providing support for the prediction that children would be able to make use of dynamic spectral structure, but not stationary coarticulated structure, in their vowel decisions.

3.2.3 Missing middles—Figure 4 shows percent correct responses for seven-year-olds for each processing type and missing middle, for each syllable set separately. This figure shows that the findings of Nittrouer (2007) were replicated: seven-year-olds maintained high recognition scores for UP stimuli, except in the case of the three pitch period stimuli with /d/ context. Thus, children are capable of recognizing vowels when much of the middle portions of the syllables are missing, as long as natural signals are presented.

Comparing Figure 3 (for adults) and Figure 4, it can be seen that seven-year-olds performed more poorly than adults with both kinds of processed signals, but disproportionately more so with the VOC signals. In fact, with the exception of the /b/, high–low contrast, seven-year-olds' performance with VOC stimuli fell below chance as soon as 50% of the middle was replaced with coughs. For SWS stimuli, however, children's performance never fell below

chance. This last result suggests that children made use of dynamic spectral structure to a great extent.

Change scores for correct responding for whole syllables versus 50% cough and 50% cough versus three pitch periods were also calculated for the seven-year-olds, as had been done for adults. Those scores are shown in Table 9. For adults, these scores generally indicated that response accuracy remained high for UP stimuli across all three kinds of missing middles. For SWS stimuli, adults maintained performance when 50% of the syllable was replaced with cough sections, but then performance dropped when only three pitch periods remained. For VOC stimuli, performance fell in a linear fashion across the three kinds of stimuli.

For seven-year-olds, trends were different. For UP stimuli, these children maintained performance across all three kinds of missing middles only for stimuli in the /b/ context. In the /d/ context, seven-year-olds showed degraded performance for UP stimuli when only three pitch periods at either syllable edge remained: *t*-tests done on change scores were significant, suggesting that whatever information adults culled from those three pitch periods was not useful to seven-year-olds.

For SWS stimuli, performance degraded linearly and slightly in the /b/ context across the three types of missing-middle stimuli, such that *t*-tests done on the change scores were not significant. That was true even though both sets of change scores showed small decrements in performance, from whole syllables to 50% cough (change score 1), and from 50% cough to three pitch periods (change score 2). For stimuli in the /d/ context, performance fell by a similar amount to what was observed for the /b/ context between the whole syllable and 50% cough stimuli, but then dropped precipitously for the three pitch period stimuli. This pattern matches more closely what was found for adults, and suggests that seven-year-olds were using dynamic spectral structure for making decisions about vowel identity. When that structure was missing because not enough of the syllable remained, seven-year-olds were unable to recognize the vowels.

Change scores for VOC stimuli confirm what is observable from Figure 3: performance dropped precipitously from whole syllables to 50% cough stimuli. Because performance was at chance with 50% of the middle missing, it could not deteriorate any further for the three pitch period stimuli. Seven-year-olds were largely unable to make vowel decisions with these stimuli even when 50% of the syllables remained intact. Without dynamic spectral structure, these children were greatly hampered.

3.3 Age effects

Finally, performance of adults and seven-year-olds was compared. Two-group *t*-tests were performed on recognition scores from adults and seven-year-olds for the 50% cough stimuli, for SWS and VOC stimuli separately. These tests were not performed on UP stimuli because mean performance for both groups was very close to 100% correct. Only the 50% cough stimuli were used because those were the ones where some dynamic spectral structure would be retained, even in the absence of the vowel targets at syllable center. Thus, outcomes for these stimuli should provide the most sensitive test of the dynamic specification account of vowel recognition. Results of these *t*-tests are shown in Table 10, and reveal that all

contrasts were significant. Children performed more poorly than adults with all sets of these stimuli. That runs counter to what would be expected, if only dynamic spectral structure explains vowel recognition.

Although seven-year-olds performed more poorly than adults with both kinds of processed stimuli, the magnitude of those age effects could have differed. To look at the magnitude of differences between age groups for each processing type, Cohen's *d*s (Cohen, 1988) were computed using results for the 50% cough stimuli, and these are shown in Table 11. For all four syllable sets, children's performance was closer to that of adults with the SWS stimuli than with the VOC stimuli, as indicated by smaller Cohen's *d*s for the former. Because the SWS stimuli preserved dynamic spectral structure better than the VOC stimuli, this trend supports the suggestion that it was precisely that structure that children were using in their vowel recognition.

4 Discussion

A central goal of speech perception research has been to reconcile the continuous and variable nature of speech signals with the discrete and invariant nature of the percepts they evoke. Vowels have served as an especially interesting example of that paradox. Approaches positing that listeners base their vowel judgments on the frequencies of formants at relatively stable portions of syllable centers only work for vowels spoken in isolation, which are not commonly found in many languages. Jenkins, Strange and colleagues have suggested that recognition does not depend on formant frequencies at any discrete time in syllable production, but rather on dynamic structure across the syllable. Specifically, time-varying spectral structure associated with movement between the vocalic center of the syllable and the margins of that syllable specifies vowel identity, according to these authors. That perspective, termed dynamic specification, was supported by results of experiments in which adults listened to syllables with the middle sections replaced by silence, and recognized the vowels with great accuracy (e.g., Jenkins et al., 1983; Strange et al., 1983).

In 2007, Nittrouer reported on adults' and children's vowel recognition for syllables with two lengths of the middle replaced by coughs. The hypothesis going into that study was that children should perform similarly to adults with these sorts of stimuli because children depend especially strongly on just the sort of dynamic spectral structure thought to underlie vowel recognition in silent-center experiments. While children did quite well, they did not perform exactly as adults did with those stimuli. One possible explanation could have been that children's performance is simply more disrupted by stimuli rendered incomplete when coughs are substituted for sections of speech. On the other hand, the possibility existed that some kind of acoustic structure other than dynamic spectral structure might account for some part of adults' responding. Perhaps children are not as sensitive to that other type of structure in the signal. The current experiment was undertaken to explore the basis of vowel recognition by adults and children.

4.1 Results from adults

Results of this current experiment both support and extend the idea that vowel quality is specified by dynamic spectral structure across syllables. Looking first at results only from

adults, it was found that adults recognized vowels with close to perfect accuracy, as long as robust dynamic spectral structure was available to them. Recognition with the SWS stimuli was highly accurate, even when the middle half of the syllable was missing. As evidence of the important role of dynamic spectral structure, recognition was poorer for VOC than SWS stimuli for those syllables missing their middle halves. At the same time, recognition with VOC stimuli did not drop to chance levels for adults when the middle half of the syllable was missing. That may have been due to VOC signals preserving some level of dynamic spectral structure, albeit degraded, but it might instead be that another kind of structure contributes to vowel recognition, as well. It is proposed that an additional basis of adults' abilities to recognize vowels from brief sections of syllable margins comes from the effects of CV coarticulation on spectral structure at every slice of the syllable. Adults seem able to use that structure, even at the very edges of CVC syllables, to uncover vowel identity. A source of support for that conclusion comes from the findings that recognition was best in those syllables where coarticulatory effects on stable structure across the syllable are greatest (/b/ contexts) and poorest in precisely those conditions where coarticulation of consonant and vowel is most constrained, but where dynamic spectral structure is strong (/d/ contexts).

The specific suggestion being made is that both dynamic and stationary coarticulated structure is used by adults to make decisions regarding vowel identity. The SWS stimuli preserved both sorts of structure fairly well, but the VOC stimuli robustly preserved only structure associated with coarticulation; dynamic spectral structure was less salient. When the middle half of the syllable was removed, adults' performance remained at ceiling for SWS stimuli, but dropped for VOC stimuli. If this experiment had only involved removing the middle 50% of the syllable, these results might be equivocal regarding the basis of these findings. When even more of the syllable was removed, however, the basis for these outcomes becomes clearer. Dynamic spectral structure was diminished greatly when only the first three and last three pitch periods were left intact: there can only be so much spectral change across such brief sections. In addition, although performance dropped for both sorts of processed stimuli in this condition, it was more precipitous for SWS than VOC stimuli in most cases. This combination of results from adults across the two conditions of missing middles supports the suggestion that the contributions of both dynamic spectral and stationary coarticulated structure are used in vowel recognition. That conclusion is bolstered by the finding that performance with VOC stimuli is most hindered when the stop being produced most strongly constrains vowel coarticulation at the margins, as happens with /d/. This last finding argues against the conclusion that overall differences in performance for SWS and VOC stimuli were simply due to VOC stimuli being more difficult to comprehend.

4.2 Results from children

Linguistic organization for young children tends to be more syllable focused and less segmental than that of adults. Therefore it was not surprising to find that children in the current study were effective in their use of dynamic spectral structure – which reflects the syllable as a whole – but were poorer at using broad spectral structure in brief signal sections – which arises from coarticulation of separate segments. For example, four-year-olds were completely unable to recognize vowels, even for whole syllables, when VOC /d/

stimuli were presented. Performance of seven-year-olds with most VOC stimuli dropped to chance levels as soon as half of the syllable was removed. These results support the conclusion that children as old as seven years are poor at using the stationary consequences of coarticulation for vowel recognition. Nonetheless, when CV coarticulation was relatively strong, as in the VOC /b/, high–low stimuli, seven-year-olds were more likely to reach test criterion. These outcomes mean that seven-year-olds were not completely unable to use stationary coarticulated structure, but the amount of that coarticulation had to be great for them to do so.

When it came to SWS stimuli, seven-year-olds performed better than with VOC stimuli, but not as well as adults. This result further confirms that these seven-year-olds could effectively use dynamic spectral structure to uncover vowel quality, but could not use stationary coarticulated structure very well. The SWS stimuli preserved both sorts of structure. The decline in performance, compared to that of adults, may be explained by their inability to recover vowel quality from the effects of coarticulation at syllable edges; the advantage in their own performance for SWS compared to VOC stimuli is explained by the well-preserved dynamic spectral structure of the SWS stimuli.

An alternative suggestion to the conclusions reached above might be that children are negatively impacted to a greater extent than adults by any kind of signal processing. However, that conclusion is contradicted by earlier studies of children's speech perception: when presented with complete sentences composed of SWS replicas, seven-year-olds have been found to perform nearly as well as adults (Nittrouer & Lowenstein, 2010; Nittrouer et al., 2009). Thus, at least at the sentence level, children recognize speech signals as well as adults when those signals are processed by replacing formant tracks with sine waves. The finding in this experiment that children heavily weighted dynamic spectral structure in a specifically phonetic decision (i.e., of vowel identity) matches results from a variety of earlier experiments. Studies using traditional labeling tasks have shown that young children (i.e., under seven years of age) weight formant transitions more than adults in consonant decisions (e.g., Greenlee, 1980; Nittrouer, 1992, 2004; Wardrip-Fruin & Peach, 1984). Taken together, results from children across studies emphasize the important role that dynamic spectral structure has on speech perception by young children. The interpretation of these findings has typically been that they reflect the fact that children's earliest unit of linguistic organization is not the phonetic segment, but rather something larger, such as the syllable or word (Nittrouer & Studdert-Kennedy, 1987; Nittrouer et al., 2009). Attention to the kind of dynamic structure preserved by sine-wave replicas of speech should help young children parse the speech signal into syllable-sized units because these transitions are fairly stable across disyllables. Eventually they discover the details of those units, a process that coincides with a growing refinement of phonemic organization. Findings from the current study fit with these assertions.

4.3 Summary

The current experiment provided support for the dynamic specification account of vowel recognition. Adults and children performed well on this recognition task with SWS signals when the middle portions of syllables were replaced with coughs, especially when the

portion removed was 50% of the syllable, rather than all but three pitch periods. Those results likely can be attributed in large part to listeners' abilities to use dynamic spectral structure. However, it appears that the structure arising from coarticulation at distinct locations within the syllable also aids vowel recognition, especially for adults who use that form of structure quite extensively and efficiently. Thus it is concluded that both dynamic spectral structure and structure arising from coarticulation of consonant and vowel spread across the syllable are used by mature listeners to recognize vowels, and children are able to use dynamic spectral structure developmentally sooner than they can use the stationary consequences of coarticulation.

Acknowledgments

Funding

This work was supported by the National Institute on Deafness and Other Communication Disorders, the National Institutes of Health (grant no. R01 DC000633).

References

Allen, G.; Hawkins, S. The development of phonological rhythm. In: Bell, A.; Hooper, JB., editors. Syllables and segments. Amsterdam, The Netherlands: North Holland Publishing Company; 1978. p. 173-185.

Assmann PF, Katz WF. Time-varying spectral change in the vowels of children and adults. Journal of the Acoustical Society of America. 2000; 108:1856–1866. [PubMed: 11051512]

Bladon RA, Lindblom B. Modeling the judgment of vowel quality differences. Journal of the Acoustical Society of America. 1981; 69:1414–1422. [PubMed: 7240572]

Blumstein SE, Stevens KN. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America. 1979; 66:1001–1017. [PubMed: 512211]

Boersma, P.; Weenink, D. Praat: Doing phonetics by computer (Version 5.1.1). 2009. Retrieved from www.praat.org

Chistovich LA, Lublinskaya VV. The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hearing Research. 1979; 1:185–195.

Clark, G. Cochlear implants: Fundamentals and applications. New York, NY: Springer; 2003.

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed.. Hillsdale, NJ: Erlbaum; 1988.

Daniloff RG, Hammarberg B. On defining coarticulation. Journal of Phonetics. 1973; 1:239–248.

Dunn, L.; Dunn, D. Peabody picture vocabulary test. 4th ed.. Bloomington, MN: Pearson Education Inc.; 2007.

Eisenberg LS, Shannon RV, Schaefer Martinez A, Wygonski J, Boothroyd A. Speech recognition with reduced spectral cues as a function of age. Journal of the Acoustical Society of America. 2000; 107:2704–2710. [PubMed: 10830392]

Ferguson CA, Farwell CB. Words and sounds in early language acquisition. Language. 1975; 51:419–439.

Ferrand, CT. Speech science: An integrated approach to theory and clinical practice. 2nd ed.. Boston, MA: Allyn and Bacon; 2007.

Fowler CA. Coarticulation and theories of extrinsic timing. Journal of Phonetics. 1980; 8:113–133.

Fowler CA, Brown JM. Perceptual parsing of acoustic consequences of velum lowering from information for vowels. Perception & Psychophysics. 2000; 62:21–32. [PubMed: 10703253]

Fox RA, Jacewicz E, Chang CY. Auditory spectral integration in the perception of static vowels. Journal of Speech, Language, and Hearing Research. 2011; 54:1667–1681.

Goldman, R.; Fristoe, M. Goldman Fristoe 2: Test of articulation. Circle Pines, MN: American Guidance Service, Inc.; 2000.

Greenlee M. Learning the phonetic cues to the voiced-voiceless distinction: A comparison of child and adult speech perception. Journal of Child Language. 1980; 7:459–468. [PubMed: 7440672]

Ito M, Tsuchida J, Yano M. On the effectiveness of whole spectral shape for vowel perception. Journal of the Acoustical Society of America. 2001; 110:1141–1149. [PubMed: 11519581]

Jenkins JJ, Strange W, Edman TR. Identification of vowels in “vowelless” syllables. Perception & Psychophysics. 1983; 34:441–450. [PubMed: 6657448]

Jenkins JJ, Strange W, Miranda S. Vowel identification in mixed-speaker silent-center syllables. Journal of the Acoustical Society of America. 1994; 95:1030–1043. [PubMed: 8132897]

Jenkins JJ, Strange W, Trent SA. Context-independent dynamic information for the perception of coarticulated vowels. Journal of the Acoustical Society of America. 1999; 106:438–448. [PubMed: 10420634]

Joos M. Acoustic phonetics. Language. 1948; 24:1–136.

Kewley-Port D, Pisoni DB, Studdert-Kennedy M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. Journal of the Acoustical Society of America. 1983; 73:1779–1793. [PubMed: 6223060]

Ladefoged, P. A course in phonetics. 2nd ed.. Orlando, FL: Harcourt Brace; 1982.

Manuel SY, Macken MA. Development reorganization of phonology: A hierarchy of basic units of acquisition. Lingua. 1979; 49:11–49.

Manuel SY, Stevens KN. Formant transitions: Teasing apart consonant and vowel contributions. Proceedings of the 13th International Congress of Phonetic Sciences. 1995:436–439.

Mayo C, Scobbie JM, Hewlett N, Waters D. The influence of phonemic awareness development on acoustic cue weighting strategies in children’s speech perception. Journal of Speech, Language, and Hearing Research. 2003; 46:1184–1196.

Menn, L. Phonological units in beginning speech. In: Bell, A.; Hooper, JB., editors. Syllables and segments. Amsterdam, The Netherlands: North-Holland Publishing Company; 1978. p. 157-172.

Murphy WD, Shea SL, Aslin RN. Identification of vowels in"vowel-less” syllables by 3-year-olds. Perception & Psychophysics. 1989; 46:375–383. [PubMed: 2798031]

Neely, ST.; Peters, JE. WAVED (Version 2.7). Omaha, NE. Boys Town National Research Hospital; 2005. Retrieved from http://audres.org/rc/index.html

Nittrouer S. Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. Journal of Phonetics. 1992; 20:351–382.

Nittrouer S. The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. Journal of the Acoustical Society of America. 2004; 115:1777–1790. [PubMed: 15101656]

Nittrouer S. Children hear the forest. Journal of the Acoustical Society of America. 2006; 120:1799–1802. [PubMed: 17069277]

Nittrouer S. Dynamic spectral structure specifies vowels for children and adults. Journal of the Acoustical Society of America. 2007; 122:2328–2339. [PubMed: 17902868]

Nittrouer S, Lowenstein JH. Does harmonicity explain children’s cue weighting of fricativevowel syllables? Journal of the Acoustical Society of America. 2009; 125:1679–1692. [PubMed: 19275325]

Nittrouer S, Lowenstein JH. Learning to perceptually organize speech signals in native fashion. Journal of the Acoustical Society of America. 2010; 127:1624–1635. [PubMed: 20329861]

Nittrouer S, Lowenstein JH, Packer R. Children discover the spectral skeletons in their native language before the amplitude envelopes. Journal of Experimental Psychology: Human Perception and Performance. 2009; 35:1245–1253. [PubMed: 19653762]

Nittrouer S, Miller ME. Developmental weighting shifts for noise components of fricativevowel syllables. Journal of the Acoustical Society of America. 1997a; 102:572–580. [PubMed: 9228818]

Nittrouer S, Miller ME. Predicting developmental shifts in perceptual weighting schemes. Journal of the Acoustical Society of America. 1997b; 101:2253–2266. [PubMed: 9104027]

Nittrouer S, Studdert-Kennedy M. The role of coarticulatory effects in the perception of fricatives by children and adults. Journal of Speech and Hearing Research. 1987; 30:319–329. [PubMed: 3669639]

Parnell MM, Amerman JD. Maturational influences on perception of coarticulatory effects. Journal of Speech and Hearing Research. 1978; 21:682–701. [PubMed: 745369]

Potter RK, Steinberg JC. Toward the specification of speech. Journal of the Acoustical Society of America. 1950; 22:807–820.

Recasens D. Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. Language and Speech. 1985; 28:97–114. [PubMed: 4087987]

Ryalls JH, Lieberman P. Fundamental frequency and vowel perception. Journal of the Acoustical Society of America. 1982; 72:1631–1634. [PubMed: 7175033]

Stevens, KN. The potential role of property detectors in the perception of consonants. In: Fant, G.; Tatham, MAA., editors. Auditory analysis and perception of speech. New York, NY: Academic Press; 1975. p. 303-330.

Stevens KN, Blumstein SE. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America. 1978; 64:1358–1368. [PubMed: 744836]

Strange W. Dynamic specification of coarticulated vowels spoken in sentence context. Journal of the Acoustical Society of America. 1989; 85:2135–2153. [PubMed: 2732388]

Strange W, Jenkins JJ, Johnson TL. Dynamic specification of coarticulated vowels. Journal of the Acoustical Society of America. 1983; 74:695–705. [PubMed: 6630725]

Strange W, Verbrugge RR, Shankweiler DP, Edman TR. Consonant environment specifies vowel identity. Journal of the Acoustical Society of America. 1976; 60:213–224. [PubMed: 956528]

Sussman HM, McCaffrey HA, Matthews SA. An investigation of locus equations as a source of relational invariance for stop place categorization. Journal of the Acoustical Society of America. 1991; 90:1309–1325.

Tice, B.; Carrell, T. TONE: Tone-analog waveform synthesizer. Lincoln, NE: University of Nebraska; 1997.

Tye-Murray, N. Foundations of aural rehabilitation: Children, adults, and their family members. Clifton Park, NY: Delmar; 2009.

Vihman, MM. Phonological development: The origins of language in the child. Cambridge, MA: Blackwell Publishers Ltd.; 1996.

Wardrip-Fruin C, Peach S. Developmental aspects of the perception of acoustic cues in determining the voicing feature of final stop consonants. Language and Speech. 1984; 27:367–379. [PubMed: 6536845]

Warren RM. Perceptual restoration of missing speech sounds. Science. 1970; 167:392–393. [PubMed: 5409744]

Wightman F, Kistler D. Informational masking of speech in children: Effects of ipsilateral and contralateral distracters. Journal of the Acoustical Society of America. 2005; 118:3164–3176. [PubMed: 16334898]

Wilkinson, GS.; Robertson, GJ. The Wide Range Achievement Test (WRAT). 4th ed.. Lutz, FL: Psychological Assessment Resources; 2006.

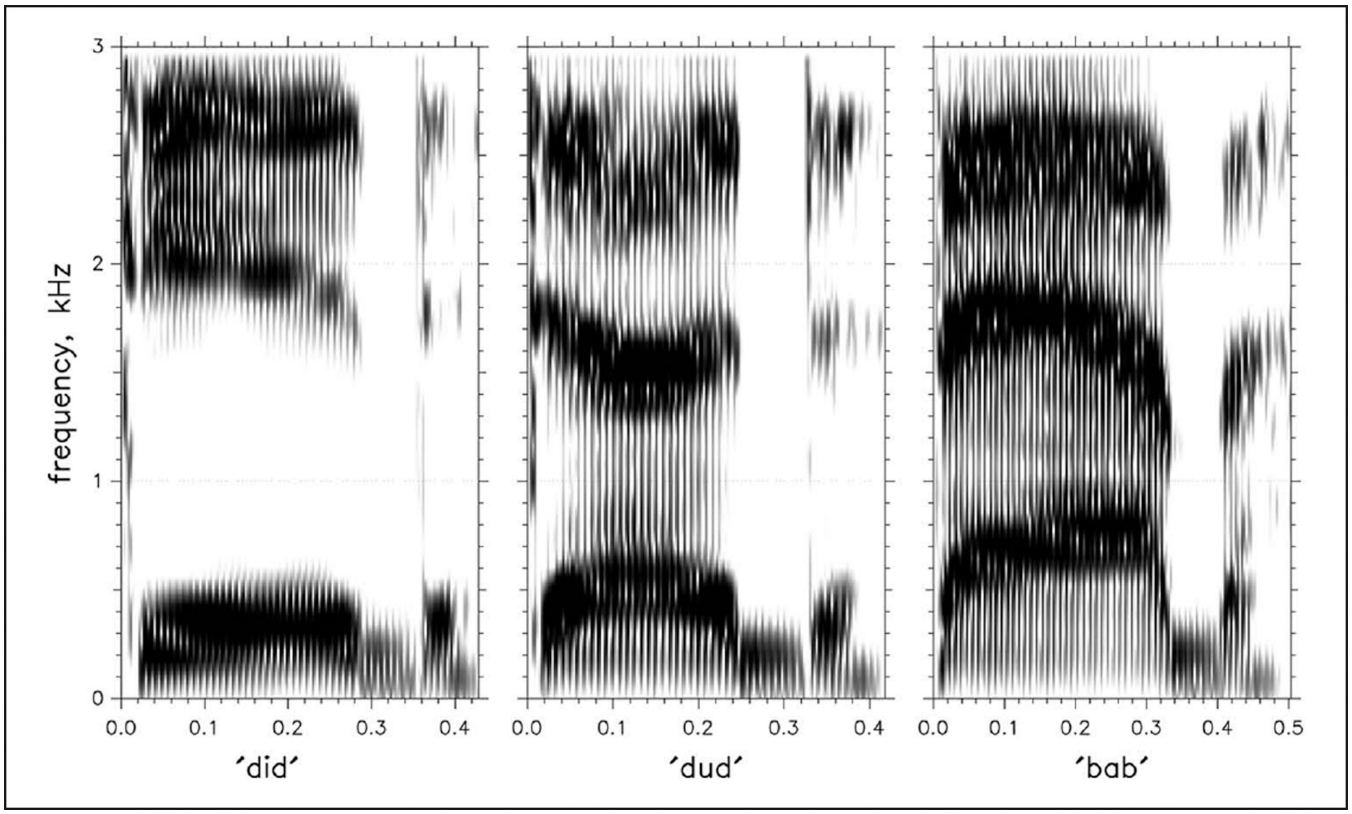

Figure 1.
Spectrograms of natural, unprocessed /dɪd/, /dʌd/, and /bæb/.

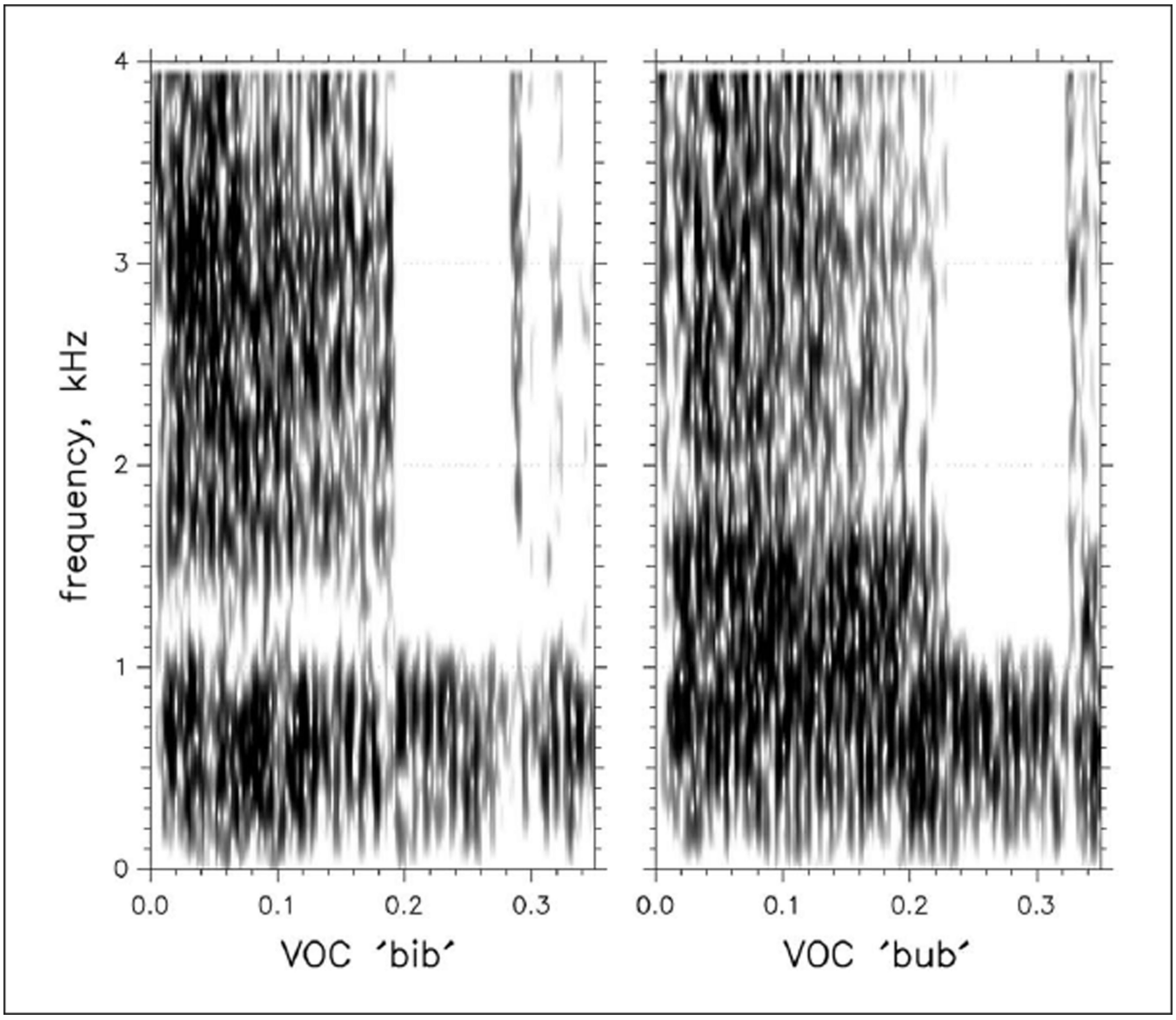

Figure 2.
Spectrograms of vocoded /bɪb/ and /bʌb/.

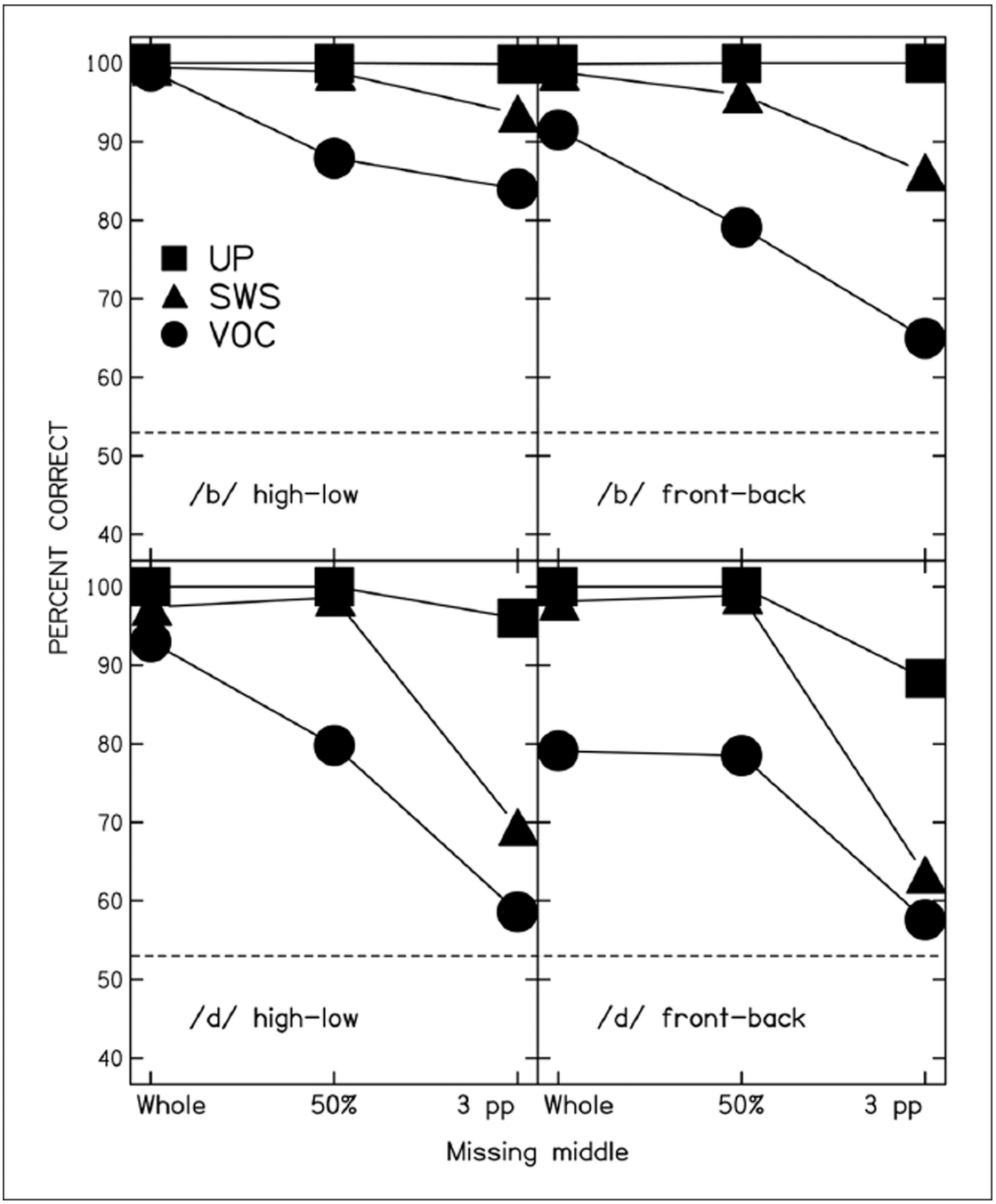

Figure 3.

Percent correct responses for adults for each processing type and kind of missing middle, for each syllable set separately. The dotted line represents the upper limit of random responding.

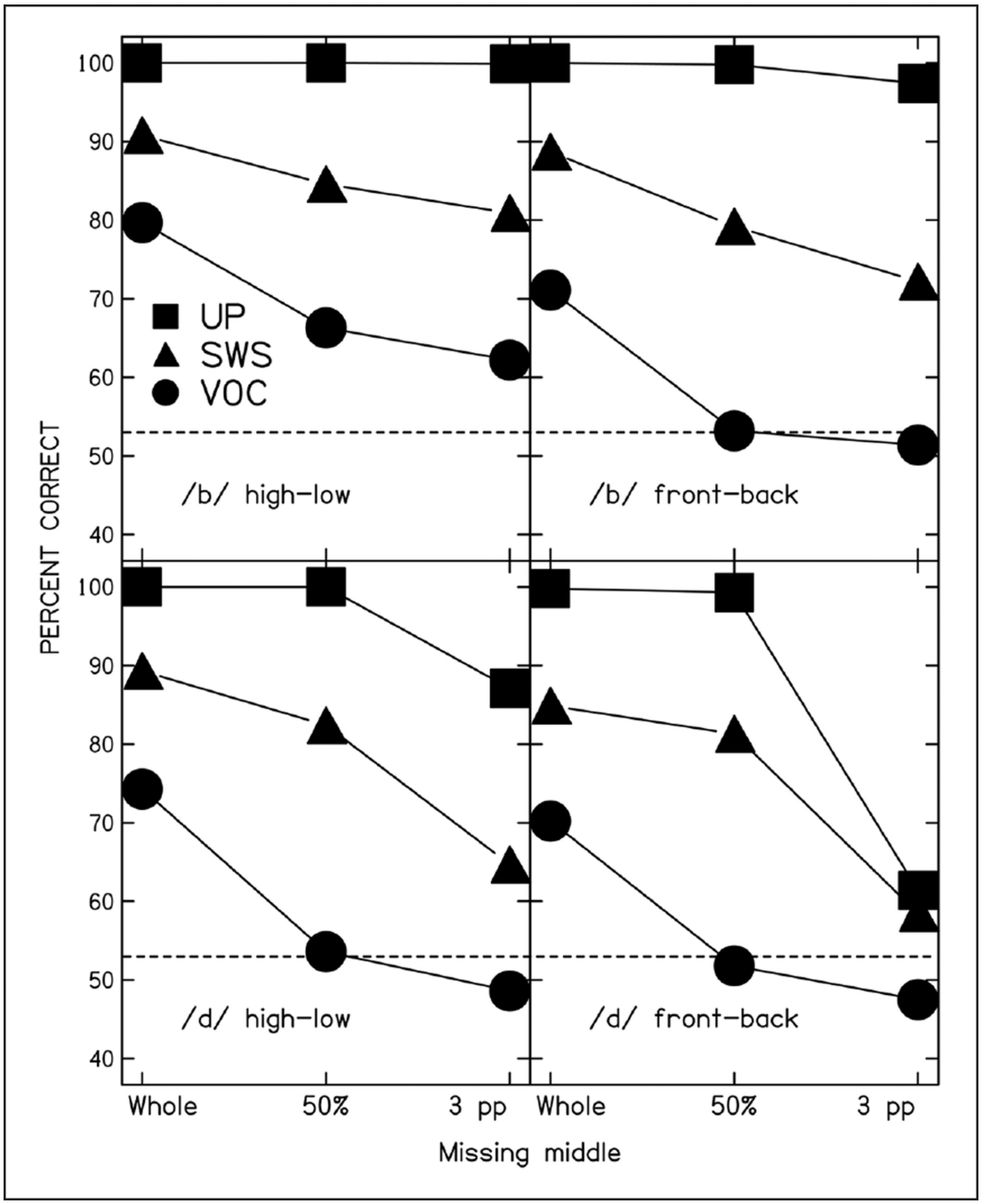

Figure 4.

Percent correct responses for seven-year-olds for each processing type and kind of missing middle, for each syllable set separately. The dotted line represents the upper limit of random responding.

Table 1

Formant frequencies (F1, F2, F3) and fundamental frequency (F0) for the original syllables.

	F1			F2			F3			F0		
	Onset	**Middle**	**Offset**	**Onset**	**Middle**	**Offset**	**Onset**	**Middle**	**Offset**	**Onset**	**Middle**	**Offset**
/bæb/	517	764	441	1690	1776	1378	2412	2530	2358	123	120	123
/bɪb/	366	441	334	1830	1852	1583	2466	2509	2422	123	113	110
/b^b/	484	463	409	1260	1303	1184	2444	2530	2466	126	124	122
/dæd/	388	603	398	1852	1830	1680	2627	2552	2530	132	124	122
/dɪd/	280	366	323	2067	2035	1776	2670	2681	2627	130	125	123
/d^d/	431	528	409	1755	1518	1669	2573	2412	2562	125	120	118

Table 2

Mean percent correct vowel recognition for each syllable set for adults, across processing types and missing middles. Standard deviations are in parentheses.

/b/, high–low	/b/, front–back	/d/, high–low	/d/, front–back
95.9	90.9	88.2	85.0
(4.4)	(5.4)	(5.0)	(5.5)

Table 3

Mean percent correct vowel recognition for whole syllables for adults, for each syllable set × processing type. Standard deviations are in parentheses. UP: unprocessed stimuli; SWS: sine-wave stimuli; VOC: vocoded stimuli.

	UP	SWS	VOC
/b/, high–low	100.0 (0.0)	99.5 (1.7)	99.0 (2.5)
/b/, front–back	99.9 (0.6)	98.9 (3.0)	91.5 (11.5)
/d/, high–low	100.0 (0.0)	97.4 (5.2)	93.0 (10.2)
/d/, front–back	100.0 (0.0)	98.1 (6.7)	79.1 (21.5)

Table 4

Results from *t*-tests for response scores from adults for whole syllables. Degrees of freedom are 19. Actual *p* values are listed for $p < 0.10$, but given as NS for $p > 0.10$. UP: unprocessed stimuli; SWS: sine-wave stimuli; VOC: vocoded stimuli.

Source	*t*	*p*	*Bonferroni*
/b/, high–low			
UP vs. SWS	−1.38	NS	NS
UP vs. VOC	−1.82	0.084	NS
SWS vs. VOC	−0.74	NS	NS
/b/, front–back			
UP vs. SWS	−1.53	NS	NS
UP vs. VOC	−3.95	<0.001	0.01
SWS vs. VOC	−3.15	0.005	0.05
/d/, high–low			
UP vs. SWS	−2.62	0.017	NS
UP vs. VOC	−3.87	0.001	0.01
SWS vs. VOC	−1.89	0.075	NS
/d/, front–back			
UP vs. SWS	−1.67	NS	NS
UP vs. VOC	−5.93	<0.001	0.001
SWS vs. VOC	−4.75	<0.001	0.001

Table 5

Change scores for adults of (1) whole syllables versus 50% cough and (2) 50% cough versus three pitch periods, and results for *t*-tests comparing those scores. Standard deviations are in parentheses. Degrees of freedom are 19.

Change score	(1) Whole – 50% Mean difference (SD)	(2) 50% – 3 pp Mean difference (SD)	*t*	*p*
UP				
/b/, high–low	0.0 (0.0)	0.1 (0.4)	–	NS
/b/, front–back	−0.1 (0.4)	0.0 (0.0)	–	NS
/d/, high–low	0.0 (0.0)	3.9 (7.8)	−2.21	0.040
/d/, front–back	0.0 (0.0)	11.6 (10.6)	−4.86	<0.001
SWS				
/b/, high–low	0.6 (2.2)	5.1 (7.9)	−2.58	0.019
/b/, front–back	2.8 (7.5)	9.8 (8.9)	−3.17	0.005
/d/, high–low	−1.4 (4.4)	29.1 (15.9)	−7.86	<0.001
/d/, front–back	−0.8 (5.4)	35.5 (12.9)	−11.09	<0.001
VOC				
/b/, high–low	11.0 (16.6)	3.8 (10.1)	1.33	NS
/b/, front–back	12.3 (14.6)	14.0 (15.2)	−0.30	NS
/d/, high–low	13.1 (11.7)	21.2 (16.6)	−1.36	NS
/d/, front–back	0.8 (18.2)	20.8 (14.3)	−3.52	0.002

Table 6

Mean percent correct vowel recognition for each syllable set for seven-year-olds, across processing types and missing middles. Standard deviations are in parentheses.

/b/, high–low	/b/, front–back	/d/, high–low	/d/, front–back
86.6	84.2	82.4	76.5
(9.4)	(7.9)	(8.1)	(7.0)

Table 7

Mean percent correct vowel recognition for whole syllables for seven-year-olds, for each syllable set × processing type. Standard deviations are in parentheses. UP: unprocessed stimuli; SWS: sine-wave stimuli; VOC: vocoded stimuli.

	UP	SWS	VOC
/b/, high–low	100.0 (0.0)	90.8 (16.1)	79.7 (15.5)
/b/, front–back	100.0 (0.0)	88.7 (14.8)	71.1 (15.4)
/d/, high–low	100.0 (0.0)	89.3 (16.2)	74.3 (17.1)
/d/, front–back	99.8 (1.1)	84.9 (17.7)	70.2 (13.1)

Table 8

Results from *t*-tests for seven-year-olds with whole syllables. UP: unprocessed stimuli; SWS: sine-wave stimuli; VOC: vocoded stimuli.

Source	*t*	*df*	*p*	Bonferroni
/b/, high–low				
UP vs. SWS	−3.29	19	0.004	0.05
UP vs. VOC	−6.28	15	<0.001	0.001
SWS vs. VOC	−2.87	15	0.012	0.05
/b/, front–back				
UP vs. SWS	−4.35	18	<0.001	0.01
UP vs. VOC	−9.90	10	<0.001	0.001
SWS vs. VOC	−3.36	10	0.007	0.05
/d/, high–low				
UP vs. SWS	−3.62	18	0.002	0.01
UP vs. VOC	−6.76	10	<0.001	0.001
SWS vs. VOC	−2.69	10	0.023	NS
/d/, front–back				
UP vs. SWS	−4.45	17	<0.001	0.01
UP vs. VOC	−10.79	10	<0.001	0.001
SWS vs. VOC	−3.77	10	0.004	0.05

Table 9

Change scores for seven-year-olds of (1) whole syllables versus 50% cough and (2) 50% cough versus three pitch periods, and results for *t*-tests comparing those scores. Standard deviations are in parentheses. UP: unprocessed stimuli; SWS: sine-wave stimuli; VOC: vocoded stimuli.

Change score	**(1)**	**(2)**	***t***	***df***	***p***
	Whole – 50%	**50% – 3 pp**			
	Mean difference (SD)	**Mean difference (SD)**			
UP					
/b/, high–low	0.0 (0.0)	0.1 (0.4)	−1.0	19	NS
/b/, front–back	0.2 (0.6)	2.3 (4.9)	−1.85	19	0.079
/d/, high–low	0.0 (0.0)	12.7 (18.1)	−3.13	19	0.006
/d/, front–back	0.5 (2.2)	37.8 (11.6)	−13.74	19	<0.001
SWS					
/b/, high–low	6.2 (14.0)	3.7 (8.5)	0.61	19	NS
/b/, front–back	9.4 (11.8)	6.8 (7.9)	0.69	18	NS
/d/, high–low	6.7 (12.3)	17.7 (16.2)	−2.13	18	0.047
/d/, front–back	3.6 (7.3)	22.7 (16.3)	−4.30	17	<0.001
VOC					
/b/, high–low	13.6 (11.2)	3.9 (7.0)	2.53	15	0.023
/b/, front–back	18.0 (20.7)	1.8 (7.3)	2.25	10	0.048
/d/, high–low	20.5 (14.9)	5.0 (4.9)	3.24	10	0.009
/d/, front–back	18.6 (14.5)	4.3 (9.5)	2.37	10	0.040

Table 10

Outcomes of two-group *t*-tests performed on recognition scores for adults and seven-year-olds for 50% cough stimuli. SWS: sine-wave stimuli; VOC: vocoded stimuli.

	t	*df*	*p*
SWS stimuli			
/b/, high–low	3.51	38	0.001
/b/, front–back	4.05	37	<0.001
/d/, high–low	5.66	37	<0.001
/d/, front–back	4.73	36	<0.001
VOC stimuli			
/b/, high–low	4.38	34	<0.001
/b/, front–back	3.84	29	<0.001
/d/, high–low	4.66	29	<0.001
/d/, front–back	4.33	29	<0.001

Table 11

Cohen's *d*s indexing the effect of differences in performance between adults and seven-year-olds for the 50% cough stimuli, for each syllable set and processing type. SWS: sine-wave stimuli; VOC: vocoded stimuli.

	SWS	VOC
/b/, high–low	1.03	1.32
/b/, front–back	1.37	1.62
/d/, high–low	1.37	2.08
/d/, front–back	1.38	1.93