

Coherence masking protection for speech in children and adults

Susan Nittrouer · Eric Tarr

© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract In three experiments, we tested the hypothesis that children are more obliged than adults to fuse components of speech signals and asked whether the principle of harmonicity could explain the effect or whether it is, instead, due to children's implementing speech-based mechanisms. Coherence masking protection (CMP) was used, which involves labeling a phonetically relevant formant (the target) presented in noise, either alone or in combination with a stable spectral band (the cosignal) that provides no additional information about phonetic identity and is well outside the critical band of the target. Adults and children (8 and 5 years old) heard stimuli that were either synthetic speech or hybrids consisting of sine wave targets and synthetic cosignals. The target and cosignal either shared a common harmonic structure or did not. An adaptive procedure located listeners' thresholds for accurate labeling. Lower thresholds when the cosignal is present indicate CMP. Younger children demonstrated CMP effects that were both larger in magnitude and less susceptible to disruptions in harmonicity than those observed for adults. The conclusion was that children are obliged to integrate spectral components of speech signals, a perceptual strategy based on their recognition of when all components come from the same generator.

Keywords Perceptual learning · Psycholinguistics · Speech perception

The traditional approach to the study of human speech perception has primarily focused on asking how specific

cues support the recovery of phonetic structure. Methods involve the generation of synthetic syllables in which all acoustic elements (known in aggregate as the *base*) are held constant across a series of stimuli at settings providing ambiguous information about phonetic identity, with one exception. That one signal component (known as the *cue*) is manipulated in a linear fashion across the series, spanning a range from a setting that disambiguates labeling in favor of one phonetic category to a setting that disambiguates labeling in favor of another phonetic category. These experiments typically focus on manipulating resonant properties of vocal tract cavities and/or acoustic consequences of narrow and brief vocal tract constrictions, because these properties have reliably been shown to underlie phonetic structure. All stimuli are presented to listeners for phonetic labeling multiple times, and a labeling function is derived from listeners' responses. This line of investigation has been tremendously useful in helping to uncover how systematic variation in small, well-defined bits of the signal, known as *acoustic cues*, specifies phonetic categories. Nonetheless, decades of research with this approach have failed to explain all aspects of human speech perception.

One reason for this shortfall in our attempts to construct a comprehensive account of human speech perception is surely that insufficient attention has been paid to examining how the various components of the complex signal are integrated into coherent streams. Why is it that disparate spectral components and temporally dispersed cues combine at all? Other branches of experimental psychology have been investigating such questions for years, but investigators in both psycholinguistic and psychoacoustic science have been slow to follow suit. As far back as the early 20th century, Gestalt psychologists described "laws of organization" (Wertheimer, 1923/1955) to explain how

S. Nittrouer (✉) · E. Tarr
Department of Otolaryngology, The Ohio State University,
915 Olentangy River Rd., Suite 4000,
Columbus, OH 43212, USA
e-mail: nittrouer.1@osu.edu

separate sensations at the periphery are fused into unitary percepts. Those principles included phenomena such as proximity, similarity, and common fate of visual elements, as well as closure, symmetry, continuity, objective set, and past experience with the signal. Although those principles were originally developed to explain visual perception, the attempts that have been made to explain how separate acoustic components of complex soundscapes are integrated have appropriated them (e.g., Bregman, 1990; Darwin & Carlyon, 1995). In general, however, efforts to understand how and why separate acoustic elements cohere in speech perception have been scarce, likely due to the collective focus of the field on explaining how human listeners recover phonetic units.

The relatively few experiments that have examined the question of how discrete components of the speech signal are fused in perception have reliably revealed that they coalesce such that any one is no longer available for individual inspection, except under very special circumstances. The duplex perception paradigm has been especially useful in informing us on this point. In this paradigm, the base of a synthetic syllable is presented to one ear, and the cue is presented to the other ear in appropriate temporal alignment. This configuration evokes two distinct but concurrent percepts: a fused percept of base+cue at the center of the perceptual space and the cue by itself off to one side. Results demonstrate that listeners can make phonetic judgments about the fused percept and can make auditory judgments about acoustic qualities of the isolated cue, such as whether it is rising or falling in frequency. What listeners cannot do is to make auditory judgments about the fused percept; the auditory qualities of the cue are lost to metacognitive inspection once it is fused with the base (Liberman, Isenberg, & Rakerd, 1981; Mann & Liberman, 1983; Whalen & Liberman, 1987).

Another experimental paradigm that provides corroborating evidence that separate acoustic elements in speech signals cohere so strongly that they cannot be individually inspected involves manipulating two cues in the same signal so that they either cooperate or conflict in how they bias listeners phonetically. The approach was developed to study speech perception by Fitch, Halwes, Erickson, and Liberman (1980) and involves designing stimuli so that one cue varies in equal-sized acoustic steps across a continuum. In one experimental condition, listeners are asked to discriminate pairs of adjacent stimuli on the basis of that one cue alone. In two additional conditions, another acoustic cue is set differently for each member of the pair such that it favors one or the other phonetic label represented by the continuum endpoints. In the *cooperating-cues* condition, each member of any given pair has both cues set to favor the same phonetic endpoint. In the *conflicting-cues* condition, each member of the pair has

those cues set to favor a different phonetic endpoint. The question addressed by this paradigm is whether human listeners organize speech signals exclusively to recover phonetic percepts, which might involve different mechanisms from those used in other instances of auditory perception. According to a general auditory account, discrimination should be better anytime two cues, rather than just one, differentiate the members of the pair, because those members are more different acoustically. According to a phonetic account, discrimination accuracy should vary depending on how well signals facilitate phonetic perception. Figure 1 shows the results from Fitch et al. and indicates that the phonetic account was supported. Discrimination was best in the cooperating-cues condition. In the conflicting-cues condition, adults actually showed poorer discrimination than they did for the *one-cue* condition. The two cues coalesced, such that neither could be inspected separately, and canceled each other out in terms of how they signaled a phonetic category. That finding sparked the conclusion that separate acoustic properties in the speech signal cohere in the course of perception and do so in a manner unique to signals arising from human speech production. Thus, it appeared that speech signals evoke perceptual mechanisms not evoked by other acoustic signals, an idea that has been strongly

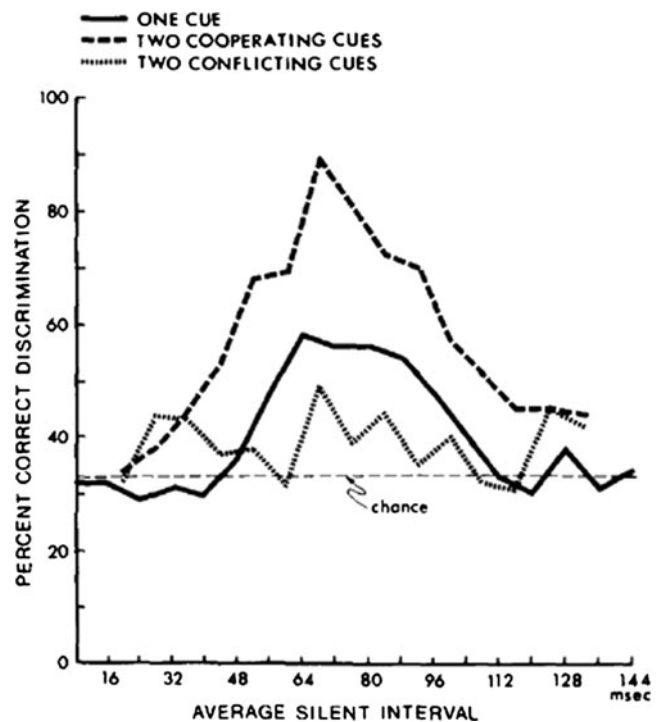


Fig. 1 Discrimination function from Fitch, Halwes, Erickson, and Liberman (1980) showing that performance was best when two cues cooperated in supporting phonetic judgments and poorest when two cues conflicted, with performance for the condition in which just one cue signaled phonetic identity intermediate between those two outcomes

criticized over the years (e.g., Diehl, Lotto, & Holt, 2004; Kluender, 1994; Lane, 1965; Miller, Wier, Pastore, Kelly, & Dooling, 1976) in a debate that continues (e.g., Galantucci, Fowler, & Goldstein, 2009; Lotto, Hickok, & Holt, 2009).

Another aspect of perceptual coherence for speech signals that remains unresolved involves the question of whether listeners automatically fuse signal components in this way, which would suggest that the principles governing that coherence are innate. Alternatively, this perceptual strategy might be acquired through years of experience listening to spoken language—years spent discovering how cues covary in natural speech and learning how to fuse them into unitary percepts. In 1984, Morrungiello, Robson, Best, and Clifton addressed that question, using the cooperating-/conflicting-cues paradigm. They found that 5-year-old children did not demonstrate the signal coherence that was a hallmark of adults' perception. Instead, these children discriminated stimuli in the conflicting-cues condition as well as they did in the one-cue condition, sparking the conclusion that children must need to learn how to fuse separate signal properties as adults do. Using slightly different stimuli, Nittrouer and Crowther (2001) attempted to replicate the result with adults and children but, instead, found that 5-year-old children were the only listeners to demonstrate the trend reported by Fitch et al. (1980). Adults actually demonstrated better discrimination in the conflicting-cues condition than in the one-cue condition, although their discrimination remained poorer in the conflicting-cues than in the cooperating-cues condition. That finding suggested that adults can recover the separate acoustic elements of speech signals under some conditions, an idea that has received and continues to receive support from others (e.g., Carney, Widin, & Viemeister, 1977; McMurray, Tanenhaus, & Aslin, 2002). The novel hypothesis to emerge from that study was that children may actually be more strongly obliged than adults to perceptually fuse speech signals, and that idea formed the basis of the present study. Here, it was again asked whether children show stronger perceptual coherence for speech signals than do adults, but a different paradigm was used.

A problem with using the cooperating-/conflicting-cues paradigm to investigate questions related to perceptual coherence for speech signals is that the phonetic *informativeness* of the signal varies across conditions because the number of cues is manipulated. Children and adults assign different perceptual weights to the acoustic cues defining phonetic categories (e.g., Nittrouer, 1992, 2004; Nittrouer, Manning, & Meyer, 1993), so that manipulation may influence perception across conditions differently for listeners of different ages. It would be useful to have a way of decoupling perceptual coherence from the informativeness of the signal, and just such a paradigm is provided

by coherence masking protection. In this paradigm, the intensity level required to identify a low-frequency signal in noise is measured both when that signal is presented alone and when it is presented with a higher frequency signal that provides no additional information about identity and is well outside the critical band of the low-frequency signal. Similar procedures have been used by others to study auditory grouping for nonspeech signals (Hall & Grose, 1990) and even speech signals (Grose & Hall, 1992), but Gordon (1997) gets credit for developing the procedures used in this present work. In his studies, Gordon (1997) measured the signal intensity listeners needed to provide correct labels 79.4% of the time for voiced speech stimuli modeled after the vowels /i/ and /ε/. In the *F1*-only condition, only the first-formant target was presented, and that was done in a background of low-pass-filtered white noise with a cutoff of 1000 Hz. In the full-formant condition, a stationary *F2/F3* cosignal was presented with synchronous onset and offset to *F1* at a level 12 dB down from that of *F1*. Results revealed that adults' thresholds were 3.2 dB lower when the *F2/F3* cosignal was presented with *F1*, even though that cosignal provided no additional information regarding vowel identity and was outside the critical band of *F1*. Thus, the procedure permits the study of perceptual coherence without variation in how phonetically informative stimuli are. It also demonstrates one benefit of perceptual coherence of acoustic components: protection from masking.

In the present study, this procedure was used to examine perceptual coherence of vowel-related formants by adults and children who were 5 or 8 years of age. Including children provided an explicit test of the hypothesis that children are more obliged than adults to perceptually fuse components of a speech signal. This hypothesis derived from the findings of Nittrouer and Crowther (2001). It proposes that children are less easily deterred from this pattern of fusing speech components than are adults. Alternatively, children might demonstrate weaker perceptual coherence than do adults. As with experiments using cooperating and conflicting cues, the feat of integrating the *F1* target with the *F2/F3* cosignal might be viewed as perceptually sophisticated. The cosignal is spectrally distant from the target and provides no useful phonetic information. Experience with how these spectral components covary might be required before they can be fused perceptually. Consequently, children might not demonstrate the effect as strongly as adults.

Finally, the experiments described here were also designed to examine what principle might account for any patterns of integration or segregation observed. In this case, it was specifically asked whether perceptual coherence across vowel formants seems to be based on all formants sharing a common harmonic structure. This question was

addressed by a second experiment in which the $F1$ target did not share harmonic structure with the $F2/F3$ cosignal. Where adults are concerned, Gordon (1997) has already examined the potential contribution of harmonicity (having the low-frequency target and the high-frequency cosignal share a common harmonic structure) to coherence masking protection for speech signals. In order to manipulate harmonicity, Gordon (1997) replaced the low-frequency target with a narrow noise band. CMP effects were obtained in spite of this lack of harmonicity. In the second experiment reported here, different methods were used for disrupting harmonicity across target and cosignal, because of what is already known about children's processing of noise signals. Earlier experiments with whispered speech (Nittrouer & Lowenstein, 2009) and with noise vocoded speech (Nittrouer & Lowenstein, 2010; Nittrouer, Lowenstein, & Packer, 2009) have shown that children experience greater deficits in speech perception with noise-excited signals than do adults. Consequently, if reduced CMP effects were observed for children, but not for adults, with noise targets, interpretation would be difficult. For that reason all stimuli retained tonal qualities. Finally, a third experiment asked whether simply having a signal that possesses speechlike attributes is sufficient to evoke the effect, for children and/or adults. Alternatively, it could be that qualities of the signal cannot explain the effect. Instead, it might be attributable to listeners' using strategies in which signal components are fused when they are recognized as having been emitted by a common generator—in this case, a single vocal tract.

Experiment 1: replicating the original experiment, but with children

This experiment was designed to test the hypothesis that children are more obliged than adults to fuse separate spectral components of speech signals. Gordon's (1997) procedures were used, and support for the hypothesis would be obtained if children showed greater masking protection than did adults in the full-formant condition, as compared with the $F1$ -only condition.

Method

Listeners

Ninety-five listeners were tested in this experiment: 25 adults between the ages of 18 and 25 years; 32 children between 8 years, 0 months and 8 years, 11 months; and 37 children between 5 years, 2 months and 5 years, 11 months. All participants (or in the case of children, their parents on their behalf) reported having normal hearing, speech, and

language. None of the children had had more than five episodes of otitis media before the age of 3 years. At the time of testing, all participants passed hearing screenings of the frequencies of 0.5, 1.0, 2.0, 4.0, and 6.0 kHz presented at 25 dB HL to each ear separately.

Equipment and materials

Testing took place in a soundproof booth, with the computer that controlled stimulus presentation and recorded responses in an adjacent room. Hearing screenings were done with a Welch Allen TM-262 audiometer and TDH-39 headphones. Stimuli were presented using a Soundblaster digital-to-analog converter, a Samson Q5 headphone amplifier, and AKG-K141 headphones.

Two pictures on cardboard (6×6 in.) were used so that listeners could point to the picture representing their response choice after each stimulus presentation. One picture was of a dog biting a woman's leg (*bit*), and the other was of a man with playing cards in his hands and stacks of poker chips in front of him (*bet*).

Stimuli

Synthetic speech stimuli were created with the Sensimetrics "SenSyn" software, a version of the Klatt synthesizer. All stimuli were made at a 10-kHz sampling rate, with low-pass filtering below 5 kHz and 16-bit digitization. All stimuli were 60 ms long, which included 5-ms on and off ramps. Stimuli were modeled on the vowels /i/ and /ε/, with three steady-state formants. $F2$ and $F3$ were 2200 and 2900 Hz, respectively, for all stimuli. $F1$ was 375 Hz for /i/ and 625 Hz for /ε/. Formant bandwidths (at 3 dB below peak amplitude) were 50 Hz for $F1$, 110 Hz for $F2$, and 170 Hz for $F3$. Fundamental frequency (f_0) was stable at 125 Hz.

To create the $F1$ -only stimuli and the low-frequency portion of the full-formant stimuli, the two stimuli described above were low-pass filtered using a digital filter with attenuation starting at 1000 Hz, a transition band to 1250 Hz, and 50-dB attenuation above that. The /ε/ stimuli were used to create the high-pass portion of the full-formant stimuli. This was done by starting attenuation at 1250 Hz, with a transition band down to 1000 Hz and 50-dB attenuation below that. For the full-formant stimuli, this high-pass portion was combined with the low-pass $F1$ -only portions, using synchronous onsets and offsets to create stimuli that had identical high-pass characteristics. Making these stimuli in this way allowed precise control over the amplitude relations of $F1$ and the higher formants. In the full-formant stimuli, the $F2/F3$ cosignal was 12 dB lower than the $F1$ target, which matched the relative levels in Gordon (1997). Pilot work by us with 24 adults showed no differences in outcomes from those reported here when the

amplitude of the $F2/F3$ cosignal was varied between 16 and 24 dB below $F1$ in 2-dB steps. Consequently, maintaining relative amplitude across formants precisely as Gordon (1997) had done permitted the cleanest comparison of outcomes between his study and this one, and there was no compelling reason to deviate from his stimulus settings. Figure 2 shows smoothed spectra of the full-formant stimuli. It shows that only the frequency of $F1$ differs across /i/ and /ε/ conditions.

For use in training, synthetic versions of the words *bit* and *bet* were created from the full-formant stimuli by appending formant trajectories at the start and end of those stimuli. At the start, 40-ms transitions were appended with starting frequencies of 200, 1800, and 2300 Hz for $F1$, $F2$, and $F3$, respectively. Steady-state syllable portions were 100 ms for these word stimuli. At the end, 40-ms transitions were appended, with ending frequencies of 200, 1800, and 2900 Hz for $F1$, $F2$, and $F3$, respectively.

Flat-spectrum white noise was generated for masking purposes with a random-number generator in MATLAB. The noise was 600 ms long and was low-pass filtered below 1000 Hz in the same manner as the $F1$ -only stimuli: with a transition band to 1250 Hz and 50-dB attenuation in the stop band.

Procedure

Listeners visited the laboratory for a single session and were paid \$12 for their participation. As much as possible, procedures replicated those in Gordon (1997), but adjustments needed to be made because children were included. In particular, children do not tolerate long periods of testing near threshold, so not as many threshold estimates could be obtained. Partly to compensate for that fact, but also to ensure that children could label stimuli reliably, extensive training with clear exemplars was provided. These modifications (of obtaining fewer threshold estimates, but

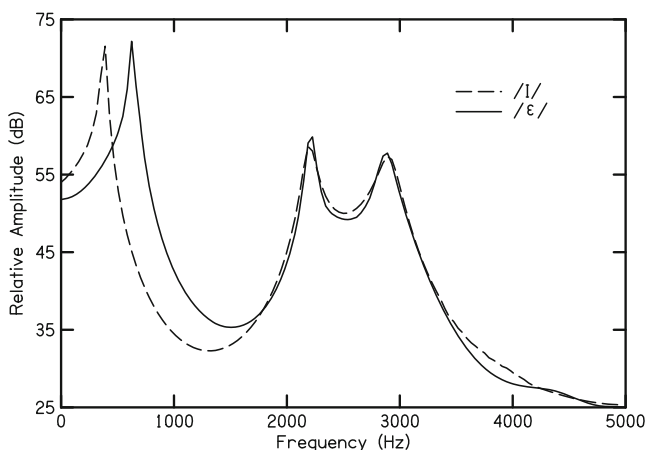


Fig. 2 Spectra of full-formant stimuli used in Experiment 1

of providing extensive training before testing) were first suggested by Aslin and Pisoni (1980) as ways to accommodate the special circumstances of working with children. Finally, feedback was not provided during testing itself. Gordon (1997) had provided feedback throughout testing. Previous work has shown that adults can modify their perceptual strategies by shifting the focus of their selective attention from one signal property to another, but that children cannot (e.g., Nittrouer, Miller, Crowther, & Manhart, 2000). Providing feedback during testing might have caused adults to modify their strategies over the course of data collection itself. That would likely not happen with children, so age-related outcomes would have been influenced due to this factor. Feedback was provided, however, during training to ensure that listeners reliably labeled stimuli and could perform the task before testing started.

General training The first general training involved the word stimuli. The experimenter introduced each picture separately and told the listener the name of the word associated with that picture. Listeners practiced pointing to the correct word and saying it after it had been spoken by the experimenter 10 times (5 times for each word). Having listeners both point to the picture and say the word ensured that they were correctly associating the word and the picture. Next, the synthetic words were presented over headphones at 74 dB SPL in random order without noise. The listener had to point to the correct picture and say the correct word. Feedback was provided. Fifty of these words (25 of each) were presented.

Next, the 60-ms full-formant stimuli were introduced, without noise. Listeners were instructed that they would be hearing “a little bit” of the word. They were told to continue pointing to the correct picture and saying the word that the little bit came from. Listeners heard 50 tokens of these samples (25 of each) at 74 dB SPL in random order, with feedback.

Condition-specific training and pre-test Training for the $F1$ -only condition followed because this condition was always presented first, which is what Gordon (1997) did. This training consisted of presenting 50 of the $F1$ -only stimuli at 74 dB SPL without noise and having listeners point to and say the word associated with that formant pattern. Feedback was provided.

Finally, up to 50 of these stimuli were presented without noise or feedback in the pre-test. As soon as the listener responded correctly to nine out of ten consecutive presentations, the training stopped. If 50 stimuli were presented without the listener ever responding correctly to nine out of ten consecutive presentations, that listener was not tested in that particular condition.

The last two training steps (the condition-specific training and the pre-test) were repeated before testing with the full-formant stimuli, using full-formant stimuli.

Adaptive testing An adaptive procedure (Levitt, 1971) was used to find the signal-to-noise ratio at which each listener could provide the correct vowel label 79.4% of the time. The noise was held constant throughout testing at 62 dB SPL, and the level of the signal varied. The initial signal level was 74 dB SPL. After three consecutive correct responses, the level of the signal decreased by 8 dB. That progression, or *run*, of decreasing signal level by 8 dB after three correct responses continued until the listener made one labeling error, at which time the level of the signal increased by 8 dB. That shift in direction of amplitude change is termed a *reversal*. Signal amplitude continued to increase until the listener responded with three correct responses, when another reversal occurred. During the first 2 runs (1 with decreasing amplitude and 1 with increasing), signal level changed by 8 dB on each step. During the next 2 runs, signal level changed by 4 dB. Across the next and final 12 runs, level changed by 2 dB on each step. The mean signal level at the last eight reversals was used as the threshold. No feedback was provided, and the stimuli were presented in an order randomized by the software.

Post-test After testing in each condition was completed, listeners heard ten stimuli at 74 dB SPL without noise and without feedback. They needed to respond correctly to nine of them. If they did not do so, their data were not included in the analysis.

Listeners had to meet the pre- and post-test inclusionary criteria for both conditions in order for their data to be included. This restriction ensured that the adaptive tracking procedure was not affected by listeners' not reliably knowing the vowel labels.

Results

One adult (4%), six 8-year-olds (19%), and fourteen 5-year-olds (38%) failed to meet either the pre- or post-test criterion described above. In all cases, these listeners failed to meet criterion for the *F1*-only condition. Failing to meet the criterion in the pre-test trials were the one adult, three 8-year-olds, and eleven 5-year-olds. The other three 8-year-olds and three 5-year-olds labeled *F1*-only stimuli adequately in the pretest but then failed to meet the criterion for the post-test trials. One of the 5-year-olds additionally failed to meet criterion for the full-formant post-test. That left 24 adults, twenty-six 8-year-olds, and twenty-three 5-year-olds with data to be included in the analyses.

Comparison of present results with Gordon (1997)

Methods for the present experiment differed slightly from those in Gordon (1997) because children participated. Therefore, the first step in analyzing these data was to see whether the magnitude of the CMP effect was similar for adults across the two studies. Table 1 shows labeling thresholds for all groups and both kinds of stimuli used in this experiment. Mean thresholds (and *SDs*) in Gordon's (1997) experiment were 58.5 dB (2.3 dB) for the *F1*-only condition and 55.3 dB (2.1 dB) for the full-formant condition. That means that adults in that earlier experiment showed 3.2 dB of masking protection. Adults in the present experiment showed 3.3 dB of masking protection. Thus, although thresholds were slightly higher in the present experiment, masking protection for the full-formant condition, as compared with the *F1*-only condition, was equivalent.

Age effects

A two-way analysis of variance (ANOVA) was performed on the thresholds shown in Table 1, with age as a between-subjects factor and number of formants (*F1* only or full formant) as within-subjects factors. Both main effects were found to be significant: age, $F(2, 70) = 39.38, p < .001$; formants, $F(1, 70) = 208.18, p < .001$. Those findings reflect the trends seen in Table 1: Thresholds were generally higher for younger than for older listeners and for the *F1*-only than for the full-formant stimuli. In addition, the age \times formants interaction was significant, $F(2, 70) = 15.05, p < .001$. This last outcome indicates that the magnitude of the formant effect increased with decreasing age. Means (and *SDs*) of differences (in decibels) between the *F1*-only and full-formant stimuli for adults, 8-year-olds, and 5-year-olds were 3.3 (3.5), 6.2 (3.8), and 9.2 (3.7), respectively. Those differences represent the CMP effect for each age group.

Matched *t*-tests were performed next on differences in thresholds for the *F1*-only and full-formant stimulus conditions for each group separately, to see whether CMP

Table 1 Means (and standard deviations) of labeling thresholds for Experiment 1

Age		Condition			
		<i>F1</i> only		Full Formant	
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Adults	24	61.2	3.4	57.9	1.4
8-year-olds	26	65.0	4.1	58.8	1.1
5-year-olds	23	70.2	3.8	61.1	2.9

effects were significant. In all cases, they were: adults, $t(23) = 4.57, p < .001$; 8-year-olds, $t(25) = 8.28, p < .001$; and 5-year-olds, $t(22) = 11.96, p < .001$.

Finally, the magnitude of the age-related difference in thresholds for each kind of stimulus (*F1* only or full formant) was indexed as a way of considering how close to adultlike children's responses were. Table 2 shows Cohen's *ds* (Cohen, 1988) for each possible combination of age groups. All comparisons show effects that are typically considered large ($d > 0.60$), but they are consistently smaller for full-formant than for *F1*-only stimuli. In particular, 5-year-olds showed thresholds closest to those of older children and adults for the full-formant speech stimuli. This finding lends support to the hypothesis that children are obliged to perceptually fuse the spectral components of speech signals. Without spectrally broad signals, which children fuse into unitary phonetic objects, children were greatly hampered in their perception.

Discussion

This experiment was conducted to test the hypothesis that young children are more obliged than adults to perceptually fuse the spectral components of speech. On the basis of earlier findings, it was hypothesized that young children are not inclined to perceptually segregate separate acoustic components in the speech signal. Although children have been found to weight formant transitions particularly strongly in their phonetic decisions (e.g., Nittrouer, 1993, 2005; Nittrouer & Studdert-Kennedy, 1987), this apparently does not happen via a process in which separate formant transitions are perceptually segregated from the spectral array and independently examined, with a postperception summation. Rather, children rely on broad spectral forms, perceived as unitary objects, for phonetic recognition. Such a perceptual strategy predicts that children will accrue greater benefit than adults from having broad spectral information available in the speech signal. A paradigm for measuring CMP developed by Gordon (1997) was used to test this hypothesis.

Outcomes clearly supported the hypothesis posed by this study: Children showed significantly stronger CMP effects than did adults, and the younger the children, the stronger the effects. These effects are quantified by the

Table 2 Cohen's *ds* for age-related differences in thresholds for Experiment 1

	<i>F1</i> only	Full Formant
Adults/8-year-olds	1.02	0.73
Adults/5-year-olds	2.49	1.40
8-/5-year-olds	1.32	1.04

difference in thresholds measured when listeners are presented with *F1*-only versus full-formant stimuli. Children had elevated thresholds, as compared with adults, for both kinds of stimuli, but they were disproportionately more elevated for *F1*-only stimuli. That pattern of results suggests that children benefit greatly from having complete spectral information about the speech signal, which they fuse into unitary percepts. They could have shown raised thresholds for the *F1*-only condition and CMP effects similar to those of adults, leading to equally elevated thresholds in both conditions. But they did not. They showed enhanced CMP effects. This finding is important because the perceptual feat of integrating spectral components to recover a unitary perceptual object, which provides protection from masking, seems sophisticated. Yet the enhanced performance of children, as compared with that of adults, means that children can perform these perceptual tasks at least as well as adults. The question left unanswered by this first experiment was what explains this enhanced spectral integration for children? For that matter, it is not clear from this one experiment what perceptual principle explains CMP for adults. It could be that the effect is based on different principles for listeners of different ages. If true, that might help explain why the effect was stronger for children than for adults.

In particular, it seemed possible at the conclusion of this first experiment that adults might rely strongly on what Bregman (1990) terms a *schema-based* principle of auditory grouping but that children might depend strongly on a *primitive* principle. Primitive principles of auditory grouping are those arising from the structure of the sound source itself, such as the harmonic relationship among spectral components. According to this account, listeners automatically—without learning—group spectral components together if they have the same harmonic structure. Schema-based principles are those that involve knowing which components of a complex auditory scene should be grouped together, perhaps because they are all part of a familiar pattern. According to this account, listeners use this stored knowledge of familiar patterns to integrate related signal components. In the case of the speech signals in this first experiment, the components of the full-formant stimuli might be integrated because they are all recognized as arising from a common generator, a single vocal tract. Because schemas generally require that perceivers know which parts of a complex scene should be grouped together, they are often learned, but they do not need to be. There are innate schemas.

In order to test the hypothesis that children's outcomes might reflect primitive grouping principles while adults' results demonstrate learned schemas, it was necessary to design stimuli that explicitly disrupt one of the primitive

principles known to facilitate perceptual grouping of sounds, but without diminishing adults' demonstration of CMP. If this was done and children showed drastic reductions in their CMP effects, the hypothesis would be supported that children's demonstration of CMP in this first experiment depended on a primitive principle, but adults' CMP did not. Fortunately, earlier work by Gordon (1997) suggested just the right stimulus manipulation.

Experiment 2: testing the principle of harmonicity

In the first experiment, it was discovered that children showed evidence of more strongly fusing disparate spectral components of the signal than adults did. This finding seems to refute the idea that part of perceptual learning involves discovering how to fuse related signal components to form unitary objects. Children already did so, suggesting that their enhanced perceptual integration might be explained by primitive principles of auditory grouping. This second experiment was undertaken to test that hypothesis.

Gordon (1997) tested two primitive principles as possible explanations for CMP in adults' perception of speechlike stimuli. One principle tested by Gordon (1997) was that of harmonicity, the idea that formants sharing a common harmonic structure should cohere. The other was that formants with synchronous onsets and offsets should cohere. In that work, the principle of harmonicity was tested by replacing the $F1$ target with a narrow band of noise. The principle of synchrony was tested by perturbing the start and/or end of higher formants, relative to $F1$. Gordon (1997) found that adults continued to demonstrate CMP when harmonicity was disrupted, but not when formant synchrony was disrupted. Following those outcomes, harmonicity was manipulated in this second experiment as a way of examining whether primitive principles underlie children's strong tendency to integrate spectral components of the speech signal. Because Gordon's (1997) results showed that adults' perceptual integration was not interrupted by disruptions in harmonicity, it was not expected to be here. On the other hand, if primitive principles of auditory grouping need to be maintained in order for children to fuse components of speech signals, children's responses should be more strongly (and negatively) perturbed by a disruption in harmonicity across formants. This second experiment tested that prediction. However, stimulus design in this experiment necessarily differed from that in Gordon (1997), because children's speech perception is more disturbed than that of adults by the use of any speech stimuli made up of noise. Therefore, all spectral components were kept tonal in nature.

Method

Listeners

Twenty-five adults, twenty-seven 8-year-olds, and twenty 5-year-olds participated. New listeners were recruited for this study, but all met the same criteria as in [Experiment 1](#).

Equipment and materials

The same equipment and materials as those used in [Experiment 1](#) were used in this experiment.

Stimuli

The $F1$ -only stimuli were the same as those used in [Experiment 1](#). They had an f_0 of 125 Hz, and $F1$ was either 375 Hz (for /i/) or 625 Hz (for /ε/). The full-formant stimuli were created in the same way as in the first experiment, except that the $F2/F3$ signal component was derived from a stimulus generated with a 175-Hz f_0 .

Procedure

Training was the same as in [Experiment 1](#). The general training still consisted of words and full-formant stimuli that had a common harmonic structure across target and cosignal. Condition-specific training and pretesting with the stimuli to be used in each condition followed. Adaptive testing was also the same as in [Experiment 1](#), except that the order of presentation of conditions varied across listeners. Half of the listeners first heard $F1$ -only stimuli, and half first heard full-formant stimuli.

Results

Data could not be included from five adults, seven 8-year-olds, and seven 5-year-olds. In all cases involving adults and 8-year-olds, data had to be excluded because the listener failed the post-test with $F1$ -only stimuli. Two 5-year-olds similarly failed to meet the post-test criterion for the $F1$ -only condition; the other five 5-year-olds failed to meet the pre-test criterion for the $F1$ -only condition. Five of those seven 5-year-olds additionally failed to meet one of the criteria with the full-formant stimuli, whereas only one 8-year-old and no adults failed to meet criterion performance with the full-formant stimuli. More adults failed to meet criteria in this experiment than in [Experiment 1](#): 20% in this experiment and 4% in [Experiment 1](#). Similarly, more 8-year-olds failed to meet one of the criteria for having their data included in the statistical analyses in this experiment, as compared with the first experiment: 26% and 19%, respectively. However, the percentages of 5-year-olds

failing to reach criterion were similar across the experiments: 38% and 35% in [Experiment 1](#) and [2](#), respectively. Of course, these percentages reflect the numbers of listeners who failed to reach criterion with *F1*-only stimuli, because no listener, in either experiment, met criterion with *F1*-only and failed with full-formant stimuli. Nonetheless, only 1 listener in [Experiment 1](#) (a 5-year-old) failed a pre- or post-test with full-formant stimuli. In this second experiment, five 5-year-olds failed to meet criteria on a pre- or post-test with full-formant stimuli. Data were included from 20 adults, twenty 8-year-olds, and thirteen 5-year-olds.

Table 3 shows mean thresholds (and *SDs*) for each age group for the *F1*-only and full-formant stimuli. As was found in [Experiment 1](#), it appears that thresholds were slightly higher for children than for adults overall, but children showed larger CMP effects. In fact, adults did not demonstrate CMP effects at all for these disharmonic stimuli. Statistical analyses support these impressions. A two-way ANOVA performed on thresholds revealed significant main effects of age, $F(2, 50) = 23.25, p < .001$, and number of formants, $F(1, 50) = 80.83, p < .001$, as well as a significant age \times formants interaction, $F(2, 50) = 25.69, p < .001$. This significant interaction reflects the finding that the magnitude of the CMP effect decreased with increasing listener age, as was observed in the first experiment. Matched *t*-tests performed on differences in thresholds for the *F1*-only and full-formant stimulus conditions showed significant CMP effects for 5-year-olds, $t(12) = 8.29, p < .001$, and 8-year-olds, $t(19) = 6.42, p < .001$, but not for adults ($p > .10$). In the present experiment, the mean CMP effect was 7.1 (3.1) for 5-year-olds and 5.1 (3.5) for 8-year-olds. These values represent a reduction in magnitude of effect of 2 dB (for 5-year-olds) and 1 dB (for 8-year-olds) from the first experiment, which used stimuli with a consistent harmonic structure across formants. Nonetheless, these effects for these disharmonic stimuli are larger than what was observed for adults, even when formants shared the same harmonic structure ([Experiment 1](#)). And again, adults showed no CMP in this second experiment.

Table 3 Means (and standard deviations) of labeling thresholds for [Experiment 2](#)

Age	<i>n</i>	Conditions			
		<i>F1</i> only		Full Formant	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Adults	20	60.4	1.9	60.7	1.9
8-year-olds	20	65.8	4.4	60.7	2.7
5-year-olds	13	70.4	3.6	63.3	3.0

When the mean thresholds shown in Table 3 are compared with those obtained in [Experiment 1](#) (shown in Table 1), it seems that thresholds are similar across the two experiments for each listener group for the *F1*-only stimuli. However, thresholds for full-formant stimuli appear slightly higher in this second experiment. To examine this apparent result, a series of two-group *t*-tests was conducted comparing thresholds for each condition, for each age group. No age group showed a significant difference in thresholds across experiments for the *F1*-only stimuli, but all groups showed significantly higher thresholds for the full-formant condition in this second experiment, as compared with the first: adults, $t(42) = 5.70, p < .001$; 8-year-olds, $t(44) = 3.23, p = .002$; and 5-year-olds, $t(34) = 2.2, p = .035$.

Discussion

This experiment was conducted to examine the magnitude of the CMP effect for each age group when stimuli lack a common harmonic structure between the low-frequency target and the high-frequency cosignal. Before conducting this experiment, it was predicted that adults would maintain CMP effects of the same magnitude as those observed for stimuli with a common harmonic structure across formants, as were used in the first experiment. That prediction followed from the findings of Gordon (1997) showing that adults maintained CMP effects in spite of the *F1* target being presented as a narrow noise band. Clearly, adults fused the noise-excited *F1* with a harmonic *F2/F3* cosignal in that experiment. The more unpredictable issue going into this experiment was thought to be whether or not children's results would show a reduction in the magnitude of CMP effects. If they did, it would indicate that children likely use harmonicity as the basis for assigning formants to a common source.

Contrary to expectations, it was found that disruption in harmonicity due to change in the *f0* of the *F2/F3* cosignal was sufficient to eliminate the CMP effect for adults. It remained intact for children, which means that the principle of harmonicity could not be posited as explaining children's especially strong coherence of signal components in speech perception. Something else must account for children's strong and seemingly obligate tendency to fuse spectral components in the speech signal such that they form a unitary percept.

Of course, the difference in outcomes for adults in this experiment, as compared with what Gordon (1997) reported, needs to be considered. In Gordon's (1997) experiment, the *F1* target was presented as a narrow noise band, not as a speechlike signal with a different harmonic structure than the cosignal. Apparently, noise components and tonal components can be fused into unitary percepts, at least for speech signals. Ecological support for that idea is

provided by everyday speech perception: Listeners consistently integrate periodic and noise components of the speech signal into percepts of voiced fricatives produced by a single speaker. However, that sort of perceptual integration apparently does not occur when it comes to two voiced components with different harmonic structures. Likely, the reason is that listeners (at least adults) rely on harmonic structure in speech signals to separate the acoustic scene according to different speakers. In fact, adults can use even small differences in harmonic structure to segregate competing speech signals into different streams, which allows them to track a single speaker in a background of noise from other speakers (e.g., Assmann & Summerfield, 1990; Brox & Nootboom, 1982). In spite of those outcomes, however, it has been demonstrated that adults are able to sum across formants differing in harmonic structure in order to make a phonetic judgment, if necessary. For example, Darwin (1981) synthesized three-formant vowels (in hVd or hVt contexts) with either a single f_0 across formants or a different f_0 for each formant. Adults were asked to label both vowel quality and “number of sounds” making up each vowel. Outcomes revealed that adults judged there to be more sounds (or sources) when f_0 differed across formants, but that did not interfere with their labeling of vowel quality. Those adults were able to perform a sort of summation across formants to derive vowel categories, even when the spectral components did not fuse into unitary percepts. With the present paradigm, this kind of summing was not productive for the task at hand, which was protecting against noise masking. For that, signal components must be fused. In these experiments, in fact, listeners were able to make judgments about vowel category on the basis of F_1 alone. The finding that all adults and 8-year-olds who failed to meet criteria for having their data included in statistical analyses did so for the F_1 -only condition, rather than for the full-formant condition, is complementary to Darwin’s results. These listeners were able to assign vowel labels to the full-formant stimuli, even though (in the case of adults) formants were not perceptually fused into unitary objects. On the other hand, a few 5-year-olds, an age group that generally seemed unaffected by the disharmonic nature of the stimuli, had difficulty assigning labels to full-formant stimuli. These results highlight the disconnection that appears to exist between mechanisms underlying phonetic labeling and perceptual integration.

In summary, the first two experiments revealed that children demonstrate CMP effects greater in magnitude than those of adults and are less easily perturbed from invoking this perceptual strategy, even when the target and cosignal lack a common harmonic structure. Thus, at least one primitive principle of auditory grouping appears an unlikely candidate to explain this strong perceptual inte-

gration in children. However, the mechanism that does underlie children’s strong tendency to fuse components of speech signals is not discernible from these two experiments. The third experiment was designed to explore one more possibility, that children fuse signal components on the basis of a strategy in which all elements apparently arising from a common generator should be integrated to form a unitary object.

Experiment 3: when sine waves are heard as speech

The purpose of this third experiment was to further examine conditions under which CMP might be observed for adults and children. The primary hypothesis addressed was that children’s strong tendency to fuse spectral components of speech signals might be based on a strategy of perceptual organization in which components are fused when they are recognized as emanating from a single speaker. To test this hypothesis, a procedure developed by Gordon (2000) was again used. In this procedure, a target signal that explicitly lacks the qualities of speech is presented in combination with a speechlike cosignal to see whether that target will be recruited into the speech percept, which will produce CMP effects. Because the target lacks the typical qualities of speech, integration of signal components, if observed, cannot be attributed to properties of the signal itself. If that integration is observed, it provides evidence that those components are integrated because they are recognized as belonging together. The alternative possibility is that the nonspeech target will be segregated perceptually from the speech cosignal and used by itself to assign labels, a strategy that will not result in CMP.

Following Gordon’s (2000) procedures, stimulus design in this third experiment used low-frequency sine waves as the F_1 targets, combined with the synthetic speech F_2/F_3 cosignal of Experiment 1. Sine waves lack speechlike qualities themselves, so they meet the criterion for this experiment. Replicating Gordon’s (2000) procedures, those sine wave targets were the same frequencies as the F_1 targets in Experiment 1 and 2: 375 and 625 Hz. These frequencies are harmonics of the 125-Hz f_0 used to generate the F_2/F_3 cosignal. Extending procedures of Gordon (2000), low-frequency sine waves other than the 375- and 625-Hz tones in the previous experiments were also used. These other tones were deliberately selected to be out of alignment with the harmonic structure of the F_2/F_3 synthetic speech cosignal. This manipulation was used to further test the hypothesis that harmonicity might explain, at least to some extent, the perceptual integration of target and cosignal that leads to CMP: If CMP effects were present (or greater) when the sine wave targets were harmonics of the f_0 of the cosignal, but not otherwise,

harmonicity could be invoked to explain the phenomenon, at least to some extent. If CMP effects were similar in magnitude regardless of whether the target was or was not a harmonic of the cosignal, harmonicity could not explain any of the effect.

Another way in which procedures in this experiment differed from those in Gordon (2000) had to do with the labels that listeners were asked to apply to the sine wave targets. Gordon (2000) had the adults in that study label these targets as *high* or *low*, but they applied the vowel labels (/i/ and /ε/) to the hybrid, full-formant stimuli. Even though these labels for nonspeech tones seem natural and obvious to adults, they are actually abstract. Children learn them through (even rudimentary) musical training, which all 5-year-olds in this study may not have had. Partly for that reason, but also to keep procedures consistent across experiments, listeners used the same vowel labels for the sine wave targets as for the full-formant stimuli in this experiment. Although it might have been a bit unnatural for older listeners to assign a phonetic label to a nonspeech tone, it was considered preferable for the youngest children. In any case, the pre- and post-tests provided the opportunity to identify and dismiss listeners who had difficulty assigning phonetic labels to these nonspeech signals.

Method

Listeners

Sixty-nine new listeners participated in this experiment: 20 adults, twenty-seven 8-year-olds, and twenty-one 5-year-olds. All participants met the criteria for participation described in Experiment 1.

Equipment and materials

The same equipment and materials were used in this experiment as in the first two experiments.

Stimuli

Six sets of stimuli were developed: three with target $F1$ -only signals consisting of single sine waves and three with those target sine waves + $F2/F3$ cosignals. All three conditions with $F2/F3$ cosignals used the cosignal from Experiment 1, so all had a harmonic structure based on an f_0 of 125 Hz.

The three $F1$ -only conditions consisted of two sine waves each. One of those conditions used sine waves that corresponded to the center frequency of $F1$ in /i/ and /ε/ from Experiment 1 and 2 (375 and 625 Hz), and that condition will be referred to as the mid- $F1$ condition. That condition replicates procedures in Gordon (2000). The goal

in designing the other two conditions was to perturb those sine waves away from values harmonically related to the 125-Hz f_0 of the cosignal, while maintaining $F1$ values that would reasonably be expected to lead to /i/ and /ε/ percepts. The range of formant frequencies measured by Peterson and Barney (1952) for individual vowels provided initial estimates of which values could be used. Pilot testing confirmed that listeners readily recognized /i/ and /ε/ with those formant frequencies. Changing sine wave $F1$ values to higher frequencies while maintaining /i/ and /ε/ percepts was easily accomplished: Both were perturbed by half the value of the 125-Hz fundamental (i.e., 63 Hz), moving them to 438 Hz (/i/) and 688 Hz (/ε/). The stimulus sets with those values will be described as the high- $F1$ condition. Moving the $F1$ values to lower frequencies was slightly problematic. They could be perturbed only by one third of the value of the 125 Hz f_0 before they ceased being clear exemplars of the intended vowels. Thus, the low- $F1$ condition had sine waves of 337 and 587 Hz for /i/ and /ε/, respectively.

In summary, there were low-, mid-, and high- $F1$ stimulus conditions, and within each of those conditions, there were $F1$ -only and full-formant stimuli. The full-formant stimuli paired a sine-wave $F1$ with an $F2/F3$ cosignal consisting of synthetic speech.

Procedure

The same procedures as those used in Experiment 1 were used in this experiment, but the order of presentation was randomized. One of the six conditions was selected to be the first for a participant, somewhat randomly but partly based on what other listeners of the same age had heard as their first condition: Presentation order of the six stimulus sets was randomized across listeners within each age group. Condition and number of formants were then alternated across presentations, with care given to not immediately repeat either one. For example, if the first stimulus set happened to be the low $F1$ -only condition, the second set had to be a full-formant condition, and it had to be one of the other $F1$ conditions (mid or high). The third stimulus set had to return to the number of formants presented in the first set, and the remaining $F1$ condition was used. The fourth through sixth stimulus sets alternated through $F1$ conditions in the same order as the first three had, presenting the stimuli with one or three formants that had not been presented for that condition in the first round.

The two kinds of general training provided in Experiment 1 were presented: words and synthetic, full-formant stimuli having the same harmonic structure across target and cosignal. Before testing with each of the six stimulus sets, practice using 50 presentations in no background noise was provided, with feedback (i.e., condition-specific training).

The pre-test followed, and listeners had to respond correctly to nine out of ten consecutive presentations, with no feedback, to proceed to testing. Finally, a post-test with no noise and no feedback ensured that listeners could label nine out of ten presentations reliably.

After testing in each condition, adults and 8-year-olds were asked what the stimuli sounded like. Five-year-olds were not queried, because that sort of metaperceptual task is too abstract for children that young.

Results

Eight 8-year-olds and six 5-year-olds were unable to label nine out of ten items correctly in either the pre- or post-test for one of the conditions, and so their data were eliminated from the analysis. Failures were evenly distributed across the three conditions, and in this case, evenly distributed across the *F1*-only and full-formant stimuli. The percentage of children who could not reliably label the stimuli was 29% for both children’s groups, which is within the range of percentages found for the first two experiments (19%–38%). All adults were able to label items correctly in all pre- and post-tests. Consequently, no evidence was found that listeners encountered particular difficulty using phonetic labels with the nonspeech targets. Data were included for 20 adults, nineteen 8-year-olds, and fifteen 5-year-olds.

Although differences in stimulus construction across conditions were not explained to listeners, it was apparent from their descriptions of the stimuli that the sine wave targets were not heard as speech by adults or 8-year-olds. The full-formant stimuli, on the other hand, were described by listeners as unambiguously sounding like speech.

Table 4 shows means (and *SDs*) for each age group for each stimulus type. A three-way ANOVA was performed on these thresholds, with age as a between-subjects factor and condition (low, mid, or high *F1*) and number of formants (one or three) as within-subjects factors. Results of that analysis are shown in Table 5. The main effects of age and formants were significant, indicating that thresholds generally decreased with increasing age and that thresholds were

generally lower for full-formant than for *F1*-only stimuli. Although close ($p = .067$), the main effect of condition was not significant, suggesting that thresholds were similar across conditions of low, mid, and high *F1*. The age \times formants interaction was significant, reflecting the apparent finding that the magnitude of the CMP effect decreased with increasing age. The age \times condition interaction was not significant, indicating that age-related differences in thresholds were comparable across conditions of low, mid, and high *F1*. Finally, the three-way interaction was significant, which meant that more analyses needed to be done before these effects could be thoroughly understood.

Simple effects analyses were performed on the data in Table 4 for each age group separately, with condition and number of formants as within-subjects factors. These results are shown in Table 6. Results for adults revealed no significant effects. Thresholds were similar regardless of condition or the number of formants.

Looking next at the results for 8-year-olds, a significant condition effect is observed. This result is probably attributable to the fact that 8-year-olds had thresholds for the *F1*-only stimuli that were roughly 3 dB lower in the high-*F1* condition than in the other two conditions. That outcome would lower the mean threshold for the high-*F1* condition, as compared with the other two conditions. In any event, neither adults nor 5-year-olds showed a condition effect. Therefore, the “almost” significant condition effect found in the three-way ANOVA must be attributable to this trend of 8-year-olds. A significant effect of number of formants was also obtained for 8-year-olds, indicating that thresholds were lower for the full-formant than for the *F1*-only stimuli. There was a significant condition \times formants interaction as well, which meant that the magnitude of the CMP effect differed across conditions.

Five-year-olds did not demonstrate a significant condition effect, but a significant formants effect was obtained. That outcome reflected the finding that 5-year-olds had lower thresholds for the full-formant than for the *F1*-only stimuli. There was also a significant condition \times formants interaction. Inspection of Table 4 reveals that thresholds for

Table 4 Means (and standard deviations) of labeling thresholds for Experiment 3, each condition shown separately

Age	Condition												
	Low <i>F1</i>				Mid <i>F1</i>				High <i>F1</i>				
	<i>n</i>	<i>F1</i> only		Full Formant		<i>F1</i> only		Full Formant		<i>F1</i> only		Full Formant	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Adults	24	57.6	2.3	56.7	1.0	57.5	1.1	55.6	1.4	57.0	1.4	56.3	0.8
8-year-olds	19	63.0	5.7	57.6	2.0	63.8	6.4	57.9	4.2	60.6	3.9	57.4	0.9
5-year-olds	15	68.7	5.9	61.7	4.4	69.8	5.7	60.7	5.0	70.2	5.8	59.5	5.1

Table 5 Results of three-way ANOVA performed on thresholds from Experiment 3

Effect	<i>df</i>	<i>F</i>	<i>p</i>
Age	2,51	36.80	<.001
Condition (low, mid, or high)	2,102	2.78	.067
Formants (<i>F1</i> or full)	1,51	95.67	<.001
Age×condition	4,102	1.60	n.s.
Age×formants	2,51	18.93	<.001
Condition×formants	2,102	2.24	n.s.
Age×condition×formants	4,102	4.66	.002

Precise *p* values are shown if they are less than .10; n.s. (not significant) is shown for values greater than .10.

F1-only stimuli rose slightly as the frequencies of the sine waves increased across conditions, but thresholds of full-formant stimuli decreased slightly across these conditions. These effects appear minor.

CMP effects (in decibels) were computed for individual listeners, and means were calculated across groups for each condition. Table 7 displays mean values for each group and each condition. For 8-year-olds, two of the three conditions showed CMP effects between the 6.2 and 5.1 dB obtained in Experiment 1 and 2, respectively. Five-year-olds showed effects across all conditions that were similar to the 9.2 and 7.1 dB obtained in Experiment 1 and 2. Thus, for these youngest listeners, combining a single sine wave with a stable synthetic speech cosignal was sufficient to reliably elicit CMP, regardless of whether that sine wave had a harmonic relation to the synthetic speech component or not.

For adults, there appears to be small, positive CMP effects, especially for the mid-*F1* condition, even though

Table 6 Results of simple effects analyses performed on thresholds from Experiment 3, for each age group separately

Effect	<i>df</i>	<i>F</i>	<i>p</i>
Adults			
Condition	2,102	0.65	n.s.
Formants	1,51	1.93	n.s.
Condition×formants	2,102	1.07	n.s.
8-year-olds			
Condition	2,102	5.33	.006
Formants	1,51	32.30	<.001
Condition×formants	2,102	4.78	.010
5-year-olds			
Condition	2,102	0.19	n.s.
Formants	1,51	87.10	<.001
Condition×formants	2,102	5.81	.004

Precise *p* values are shown if they are less than .10; n.s. (not significant) is shown for values greater than .10.

Table 7 Mean coherence masking protection effects (and standard deviations) for each age group, in each condition for Experiment 3

Age	<i>n</i>	Low <i>F1</i>		Mid <i>F1</i>		High <i>F1</i>	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Adults	20	0.9	2.2	1.9	1.4	0.6	1.2
8-year-olds	19	5.4	5.7	5.9	5.4	3.2	4.0
5-year-olds	15	7.0	4.9	9.1	5.6	10.7	6.5

These values are given in decibels and represent the differences in thresholds between the *F1*-only and full-formant conditions. Positive values indicate masking protection.

the overall formant effect was not significant in the simple effects analysis. For that reason, suspicion was raised that at least the effect for the mid-*F1* stimuli might be significant; perhaps it was attenuated in the simple effects analysis by smaller effects in the other conditions. To examine that possibility, matched *t*-tests were performed on all comparisons between the *F1*-only and the full-formant conditions. For both 5- and 8-year-olds, all three tests were highly significant (*p* < .001). For adults, only the comparison for the mid-*F1* condition was highly significant, *t*(19) = 6.04, *p* < .001. The comparison for the high-*F1* condition was significant, although not as strongly, *t*(19) = 2.32, *p* = .032, and the comparison for the low-*F1* condition was not significant, *t*(19) = 1.81, *p* = .086. Thus, even though the simple effects analysis did not reveal an overall difference in thresholds for the *F1*-only and full-formant stimuli, these *t*-tests showed that two of the three conditions did demonstrate some effect for adults. In particular, the mid-*F1* condition, which preserved a harmonic relationship across formants, showed a small effect in the predicted direction. That effect was smaller than what was demonstrated by adults for synthetic stimuli in Experiment 1 and smaller than the effect Gordon (2000) showed for these exact stimuli. However, the effect was present.

Discussion

The last experiment was conducted to test the hypothesis that children’s strong tendency to fuse disparate spectral speech components might be based on a strategy that has them grouping together sensory inputs that seem to emanate from a common generator. This was accomplished by using a target signal that lacked the acoustic qualities of speech. By presenting it synchronously with a signal that possessed those speechlike qualities, the opportunity was afforded listeners to recruit that signal into the speechlike percept. If they did, CMP effects would be observed. If listeners perceptually segregated that target signal from the cosignal, no CMP would be seen.

Results of this experiment showed that in all three conditions, children demonstrated substantial CMP effects.

In most cases, these effects were similar in magnitude to what was observed for speech stimuli in [Experiment 1](#) and [2](#). That was true even when the harmonicity of the stimuli was disrupted. Consequently, it may be concluded that children group these disparate spectral components together on the basis of the expectation that the signal is speech; harmonicity is not necessary. The one exception to this conclusion was the high-*F1* condition, where 8-year-olds showed a diminished CMP effect. However, that result was not due to these children's having especially high thresholds for full-formant stimuli; thresholds were similar for these stimuli across the three conditions. Rather, that result was obtained because 8-year-olds' thresholds were lower for the *F1*-only stimuli in that high-*F1* condition than in the other two conditions. It is not clear why that would be, but that particular outcome does not negate the general finding that children showed evidence of CMP with these hybrid stimuli similar in magnitude to what they showed for synthetic speech. As long as the full-formant stimuli could be recognized as speechlike, children showed the effect.

For adults, findings were quite different. These mature listeners were found to have greatly reduced CMP for all conditions in this third experiment, as compared to findings for synthetic speech stimuli in [Experiment 1](#). Unlike children, adults did not strongly incorporate that nonspeech target into a unitary percept, even though they reported hearing these stimuli unambiguously as speechlike. Of the three conditions, the CMP effect was observed only for the mid-*F1* condition, where signals preserved a harmonic relationship across target and cosignal. However, even there, it was reduced from [Experiment 1](#). There is no obvious reason why the results of this experiment differ from those of [Gordon \(2000\)](#), who did not find any diminishment in effect for these stimuli from what was observed with synthetic speech. Nonetheless, the trends are clear: Adults can much more easily than children be deterred from perceptually integrating signal components so strongly that any one component cannot be segregated and independently examined. Put another way, children exhibit stronger perceptual coherence for speech signals than do adults, a trend that has been previously reported ([Nittrouer & Crowther, 2001](#)). Furthermore, the primitive principle of harmonicity appears to explain CMP in adults' responding, at least to a small extent. Children, on the other hand, seemed to group formants together into unitary percepts if they were recognized as originating from a common generator.

General discussion

This study was undertaken primarily to test the hypothesis that young children are more strongly obliged than adults to

perceptually integrate components of the speech signal. The notion of being obliged to perceptually integrate signal components means that children's perceptual strategies are resistant to being perturbed from fusing components in that manner. A second goal of the study was to examine the extent to which the tendency to fuse signal components is likely based on primitive principles of auditory grouping, such as harmonicity, or on a strategy in which signal components are fused if they likely arose from a single generator—or speaker, in this case. In general, any improvement in our understanding of the mechanisms underlying the organization of speech signals should extend our understanding of speech perception. More practicably, this information could help in the design of more effective treatments for individuals who face problems related to the processing of linguistically significant signals, a group that likely includes individuals with dyslexia and cochlear implant users, to name a few. For example, finding differences in the perceptual strategies of adults and children can highlight how auditory prostheses might be fit differently for listeners of different ages.

The major outcome of this study was that young children showed greater coherence masking protection than did adults for the speech signals used here and were less readily perturbed from integrating spectral components in that way than were adults. The effect for children was not restricted to conditions in which all components shared a harmonic relationship, or even to conditions in which all components had harmonic structure. It was more related to listener characteristics than to signal characteristics. These outcomes support the hypothesis that young children are more strongly obliged than adults to fuse spectral components when those components are recognized as being part of a speech signal.

An implication of the present study concerns notions of how acoustic components come to be integrated perceptually. These are the principles that form the basis of auditory scene analysis, or ASA ([Bregman, 1990](#)). Principles associated with ASA are generally described as being either primitive or learned phenomena. Primitive principles of auditory grouping are those arising from the structure of the sound itself, and [Bregman](#) suggests that these principles can be invoked by listeners in the absence of experience, or learning. Learned principles are generally described as including all those that involve schemas or some sort of pattern recognition. They require that the listener recognize which components of the auditory scene naturally fit together. Schemas require that the user analyze the sensory data to determine whether a particular schema should be applied. Two examples of schema-based auditory grouping cited by [Bregman](#) are music and speech. When one hears a busker playing a musical instrument on a busy street corner, for example, the components of the sound reaching our ears

that arise from the musical instrument are readily separated from the street sounds and fused into the music stream. According to ASA, that happens because we have had experience listening to various instruments and know to expect that a particular set of sounds belongs together. Similarly, according to this view, the various spectral components of the speech signal are grouped together because of our experience listening to speech. Davis and Johnsruide (2007) put a finer point on this illustration with their discussion of speech containing clicks. When clicks are combined with an ongoing speech signal, native speakers of most languages hear those clicks as belonging to a separate spectral stream from the speech (e.g., Fodor & Bever, 1965; Garrett, Bever, & Fodor, 1966). However, listeners of languages that make use of clicks perceptually integrate those clicks into the speech stream. This variation in perceptual organization occurs because of differences in listeners' experiences with language. If an individual has grown up in a language community that has clicks as part of its phonetic inventory, that individual expects clicks to have been generated by the speaker. If an individual grew up learning a language without clicks, that individual does not expect speakers to be producing clicks when they talk. Thus, this is an example of a schema-based grouping principle for speech that needs to be learned.

The results for the present set of experiments serve as a reminder that not all schema-based principles of auditory grouping are learned. They also suggest that some principles commonly considered to be primitive may not be innate. In this study, it was found that the younger the listener, the more strongly spectral components were fused, even when those components did not share a common harmonic structure. As long as signal components could have been generated by a single speaker, children fused those components. The listeners with more experience were the ones who were more influenced by the harmonic structure of the signals. These outcomes mean that notions of which principles are innate and which are learned need some adjustment. It may be that for some sorts of sensory inputs, humans innately group together elements that arise from a common generator. Experience then provides opportunity to explore those signals and eventually discover that those elements share certain attributes, such as a common harmonic structure in the case of speech, a discovery that becomes critical to mature and optimal patterns of speech perception.

Of course, there is one obvious potential contradiction to this explanation, which is that the children in this study all had at least 5 years of experience listening to speech. Consequently, it is possible that they could have learned to group together the spectral components that generally arise from a moving vocal tract, even when it means overlooking the fact that those components do not share a harmonic

relationship. However, attributing the outcomes for young children to such an account would require the construction of a cumbersome developmental model, because adults did not so readily fuse spectral components when the principle of harmonicity was violated. The hypothesized course of development would need to propose that infants initially group spectral components if they share a common harmonic structure. At some time over their first 5 years, they learn to group similar components on the basis of the expectation that they were generated by a common speech source. At that precise moment in ontogenesis, that expectation comes to trump the need for common harmonic structure, but only temporarily. By adulthood, the perceptual organization of speech is again dependent on spectral components sharing the same harmonic structure—or so this model would suggest. That developmental course seems needlessly awkward and unsubstantiated.

Another possibility that might reconcile these findings is that speech may be one sort of auditory event that holds a special status as a sensory signal because of the benefits afforded by being able to communicate, especially if an organism is young. For that reason, innate schemas for processing speech may have been selected through evolution. Although speculative and in need of more explicit testing, there is some evidence to support that suggestion. For example, infants as young as 9 months of age have been found to integrate acoustic cues in speech perception at least as well as adults (Eilers, Oller, Urbano, & Moroff, 1989). Also, the long-term spectra of babbling productions from 10-month-olds have been found to vary across languages to match the long-term spectra of adults' speech in those communities, suggesting that infants discover broad spectral characteristics of their native language before they recognize the spectro-temporal details associated with phonetic inventories (de Boysson-Bardies, Sagart, Halle, & Durand, 1986).

Evidence of perceptual integration in other contexts

The paradigm used in this study, CMP, examined potential benefits of a perceptual strategy for speech that integrates spectral components across broad regions covering more than a critical band, the presumed integration window of the peripheral auditory system, regardless of whether all those components contribute phonetic information or not. In this case, protection against noise masking was accrued when listeners integrated across broad spectral slices of the signal. Benefits have similarly been reported when listeners integrate across temporal stretches of the signal longer than the presumed sliding window of 200 ms (Näätänen, 1990). In particular, experiments asking listeners to identify vowel quality for syllables that lack any steady-state or even target formant frequencies have shown that listeners can do so

with only very brief portions of formant transitions at each syllable margin (Jenkins, Strange, & Edman, 1983; Strange, Jenkins, & Johnson, 1983). More than 200 ms can be missing from the center of these syllables, and yet listeners readily perceive them as unitary events and recognize the vowel that should occupy the silent center. Results are typically described as demonstrations of the importance of dynamic syllable structure in speech perception, but they also illustrate how listeners use expectations about speech signals to perceptually organize signals. Furthermore, children as young as 3 years of age have been found to show these temporal integration effects, so the phenomenon cannot be easily attributed to listeners' learning how formant transitions are associated with steady-state formant frequencies in the signaling of vowel identity (Nittrouer, 2007).

If it ain't broke, why fix it?

A question that arises as a result of the findings reported here concerns why the strong perceptual integration found for children's speech perception diminishes with development. In the case of CMP especially, there seems to be a perceptual advantage to integrating across the speech spectrum as strongly and robustly as children were found to do: It protected against masking. Why, then, does that perceptual strategy shift so that older listeners more readily isolate the low-frequency target? The answer to that question likely is that there are different perceptual advantages to be gained from learning to attend to the acoustic details or individual components of the speech signal. Although the paradigm used here showed that listeners who were able to ignore characteristics of the signal originating at the source (i.e., the harmonic structure) showed greater masking protection, other paradigms have demonstrated that attention to details of the acoustic speech signal, such as harmonic structure, provides benefits to psycholinguistic processing.

For much of the history of speech perception research, the role of acoustic structure in psycholinguistic processing has been thought to end with phonetic recognition. Once the string of phonemes in the heard speech signal is extracted, acoustic details are no longer needed, this view has held. The conventional wisdom has been that phonemes are used exclusively in pretty much all subsequent processes. Acoustic details have been viewed as undesirable variability to be filtered out through normalization (e.g., Gerstman, 1968; Halle, 1985; Pisoni, 1985). However, that view has more recently received challenges from some speech scientists (e.g., Goldinger, 1998; Pisoni, 1997; Port, 2007). Acoustic details such as speaker-specific harmonic structure, dialectically specified vowel formants, and specific voice onset times have been found to support many psycholinguistic processes, such as serial recall of linguistic materials

(Goldinger, 1990), lexical access (McMurray et al., 2002), and speech recognition in noisy environments (Nygaard, Sommers, & Pisoni, 1994). Presumably, this is what allows listeners to follow the target speaker in acoustic backgrounds consisting of many voices—the classic cocktail party. Thus, children may initially show a predisposition to strongly integrate the acoustic components of the speech signal, even when it means ignoring signal detail. That strategy may provide some benefits in the early stages of language acquisition. However, learning to attend to separate acoustic attributes of the speech signal apparently provides benefits characteristic of mature psycholinguistic processing.

Summary

In summary, this study was undertaken (1) to test the prediction that young children are more likely to fuse the spectral components of speech signals than are adults and (2) to examine whether a primitive grouping principle (harmonicity) underlies the effect or whether, instead, it arises because children group together signal components that all seem to come from a single generator. The results of three experiments supported the prediction and revealed that the principle of harmonicity could not explain the effect. Instead, results seem to arise because children apply a schema (at least for speech) of fusing signal components if they are recognized as having come from a single generator. These outcomes suggest that commonly held views about the innateness of some principles of perceptual organization and the learnedness of others might require modification.

Acknowledgments This work was supported by Grant R01 DC000633 from the National Institutes of Health, National Institute on Deafness and Other Communication Disorders. The contributions of Christopher Chapman and Joanna H. Lowenstein to stimulus generation and programming are gratefully acknowledged. We also thank Amanda Caldwell and Don Sinex for reading earlier drafts of the manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Aslin, R. N., & Pisoni, D. B. (1980). Some developmental processes in speech perception. In G. H. Yeni-Komshian, J. H. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: 2. Perception* (pp. 67–96). New York: Academic Press.
- Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 88, 680–697.
- Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

- Brox, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*, 23–36.
- Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop constants differing in VOT. *Journal of the Acoustical Society of America*, *62*, 961–970.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, *33A*, 185–207.
- Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Hearing* (pp. 387–424). San Diego, CA: Academic Press.
- Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*, 132–147.
- de Boysson-Bardies, B., Sagart, L., Halle, P., & Durand, C. (1986). Acoustic investigations of cross-linguistic variability in babbling. In B. Lindblom & R. Zetterstrom (Eds.), *Precursors of early speech* (pp. 113–126). New York: Stockton.
- Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179.
- Eilers, R., Oller, D. K., Urbano, R., & Moroff, D. (1989). Conflicting and cooperating cues: Perception of cues to final consonant voicing by infants and adults. *Journal of Speech and Hearing Research*, *32*, 307–316.
- Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, *27*, 343–350.
- Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, *4*, 414–420.
- Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, *71*, 1138–1149.
- Garrett, M. F., Bever, T. G., & Fodor, J. A. (1966). The active use of grammar in speech perception. *Perception & Psychophysics*, *1*, 30–32.
- Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *AU-16*, 78–80.
- Goldinger, S. D. (1990). Effects of talker variability on self-paced serial recall. *Research on Speech Perception Progress Report 16* (pp. 131–326). Bloomington: Indiana University Press.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.
- Gordon, P. C. (1997). Coherence masking protection in speech sounds: The role of formant synchrony. *Perception & Psychophysics*, *59*, 232–242.
- Gordon, P. C. (2000). Masking protection in the perception of auditory objects. *Speech Communication*, *30*, 197–206.
- Grose, J. H., & Hall, J. W. (1992). Comodulation masking release for speech stimuli. *Journal of the Acoustical Society of America*, *91*, 1042–1050.
- Hall, J. W., & Grose, J. H. (1990). Comodulation masking release and auditory grouping. *Journal of the Acoustical Society of America*, *88*, 119–125.
- Halle, M. (1985). Speculations about the representations of words in memory. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 101–114). Orlando, FL: Academic Press.
- Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in “vowelless” syllables. *Perception & Psychophysics*, *34*, 441–450.
- Kluender, K. R. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 173–217). San Diego, CA: Academic Press.
- Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, *72*, 275–309.
- Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*, 467–477.
- Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, *30*, 133–143.
- Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*, 110–114.
- Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, *14*, 211–235.
- McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, B33–B42.
- Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, *60*, 410–417.
- Morrongiello, B. A., Robson, R. C., Best, C. T., & Clifton, R. K. (1984). Trading relations in the perception of speech by 5-year-old children. *Journal of Experimental Child Psychology*, *37*, 231–250.
- Näätänen, R. (1990). The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences*, *13*, 201–208.
- Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics*, *20*, 351–382.
- Nittrouer, S. (1993). The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. *Journal of Speech and Hearing Research*, *36*, 959–972.
- Nittrouer, S. (2004). 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*, *115*, 1777–1790.
- Nittrouer, S. (2005). Age-related differences in weighting and masking of two cues to word-final stop voicing in noise. *Journal of the Acoustical Society of America*, *118*, 1072–1088.
- Nittrouer, S. (2007). Dynamic spectral structure specifies vowels for children and adults. *Journal of the Acoustical Society of America*, *122*, 2328–2339.
- Nittrouer, S., & Crowther, C. S. (2001). Coherence in children’s speech perception. *Journal of the Acoustical Society of America*, *110*, 2129–2140.
- Nittrouer, S., & Lowenstein, J. H. (2009). Does harmonicity explain children’s cue weighting of fricative-vowel syllables? *Journal of the Acoustical Society of America*, *125*, 1679–1692.
- Nittrouer, S., & Lowenstein, J. H. (2010). Learning to perceptually organize speech signals in native fashion. *Journal of the Acoustical Society of America*, *127*, 1624–1635.
- Nittrouer, S., & Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research*, *30*, 319–329.
- Nittrouer, S., Lowenstein, J. H., & Packer, R. (2009). Children discover the spectral skeletons in their native language before the amplitude envelopes. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1245–1253.
- Nittrouer, S., Manning, C., & Meyer, G. (1993). The perceptual weighting of acoustic cues changes with linguistic experience. *Journal of the Acoustical Society of America*, *94*, S1865.
- Nittrouer, S., Miller, M. E., Crowther, C. S., & Manhart, M. J. (2000). The effect of segmental order on fricative labeling by children and adults. *Perception & Psychophysics*, *62*, 266–284.

- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42–46.
- Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175–184.
- Pisoni, D. B. (1985). Speech perception: Some new directions in research and theory. *Journal of the Acoustical Society of America*, 78, 381–388.
- Pisoni, D. B. (1997). Some thoughts on “normalization” in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego, CA: Academic Press.
- Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, 25, 143–170.
- Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74, 695–705.
- Wertheimer, M. (1923). Laws of organization in perceptual forms. *Psychologische Forschung*, 4, 301–350. [Also published in (1955) W. D. Ellis (Ed.), *A source book of Gestalt psychology* (4th ed., pp. 71–88). New York: Humanities Press].
- Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, 237, 169–171.