Published in final edited form as:
J Exp Psychol Hum Percept Perform. 2009 August ; 35(4): 1245–1253. doi:10.1037/a0015020.

Children discover the spectral skeletons in their native language before the amplitude envelopes

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

Robert Packer
Washington State University

Abstract

Much of speech perception research has focused on brief spectro-temporal properties in the signal, but some studies have shown that adults can recover linguistic form when those properties are absent. In this experiment, seven-year-old English-speaking children demonstrated adult-like abilities to understand speech when only sine waves replicating the three lowest resonances of the vocal tract were presented, but failed to demonstrate comparable abilities when noise bands amplitude-modulated with envelopes derived from the same signals were presented. In contrast, adults who were not native English speakers, but were competent second language learners, were worse at understanding both kinds of stimuli than native English-speaking adults. Results showed that children learn to extract linguistic form from signals that preserve some spectral structure, even if degraded, before they learn to do so for signals that preserve only amplitude structure. We hypothesize that children's early sensitivity to global spectral structure reflects the role that it may play in language learning.

One of the first ideas that students of human communication encounter is the Acoustic Theory of Speech Production. Largely attributed to Fant (1960), this theory states that the speech signal primarily arises from a glottal source, which provides broad-band periodic sound to the vocal tract, and the vocal tract, which serves to filter that sound. Historically, the study of human speech perception has focused on the way that the vocal-tract filter shapes the speech signal to create linguistic form (primarily phonetic form). The acoustic properties arising from the actions of the vocal-tract filter that have been of most interest to investigators are those that are spectrally specific and temporally brief, properties that have come to be called “acoustic cues” (Repp, 1982). Scores of studies have been published documenting the relations between these kinds of properties and phonetic form, but the study reported here dealt with structure in the speech signal that is broader spectrally and longer temporally, structure that can be termed “global” in nature.

When we consider acoustic properties affiliated with long stretches of the speech signal (i.e., those that are the length of several segments or even words), properties arising from the glottal source generally come to mind. These properties fit under the heading of “prosody.” For example, fundamental frequency of the source spectrum provides information useful in speech perception, as does the degree of breathiness, and the temporal pattern of when the

nittrouer.1@osu.edu

source is active and when it is not (e.g., Bricker & Prusansky, 1976; Carrell, 1981; Jassem, 1971). These source properties inform mature listeners primarily about characteristics of the speech signal that are ***not*** phonetic in nature, such as speaker identity, but they also play an important role in language learning. "Prosodic bootstrapping" describes a process in which infants use prosodic cues to gain access to elementary syntactic structure, before lexical knowledge is well established (e.g., Gleitman & Wanner, 1982; Jusczyk, 1997; Morgan & Demuth, 1996; Soderstrom, Seidl, Kemler Nelson, & Jusczyk, 2003).

But the vocal-tract filter also imparts to the speech signal several kinds of structure that are best resolved from long signal stretches. These kinds of structure arise from general postural settings of the vocal tract, such as whether a language or speaker tends to use pharyngeal constrictions or heavy nasalization, from the continuously changing sizes and shapes of vocal tract cavities which determine formant frequencies, and even from modulations in overall amplitude due to the degree of vocal tract constriction. Likely because of the emphasis on acoustic cues, and their relation to explicitly phonetic form, the potential role of these more global kinds of structure went largely unexamined in early speech perception research. It was not until the early 1980s that a study breaking with traditional methods reported that adult listeners can both recognize a signal as speech and recover linguistic units when only the skeleton of spectral structure is presented (Remez, Rubin, Pisoni, & Carrell, 1981). In that study, the authors replaced the normally rich harmonic spectrum of speech with three time-varying sinusoids that traced the dynamic changes in the three lowest formants. The finding that adult listeners could perceptually integrate these impoverished signals into coherent linguistic forms could not be explained within the framework of speech perception theories existing at that time, theories that generally held that listeners gather acoustic cues from the signal and use them to recreate the string of phonetic segments intended by the speaker. So while the finding that listeners could make sense of signals consisting only of these spectral skeletons was interesting, the significance of the finding was not clear.

Later, another group of investigators used a different signal processing scheme to represent a different kind of global structure imparted to the signal by vocal tract actions: Shannon, Zeng, Kamath, Wygonski, and Ekelid (1995) described a study in which they divided the speech spectrum into two, three, or four bands, recovered the amplitude envelopes of each of those bands, and modulated noise bands limited by the same frequencies used to divide the original speech spectrum using those envelopes. In contrast to sine wave speech, which preserves only the global spectral structure, these amplitude envelopes largely eliminated the spectral structure of the signal. Nonetheless, Shannon et al. showed that adults were able both to recognize those signals as speech and to recover linguistic form. This finding had clinical significance because of its relation to the development of the cochlear implant, which had recently made its appearance. Cochlear implants provide little in the way of spectral structure, so if listeners do indeed understand speech by recovering brief, spectrally specific bits of the signal then these devices should not work. By 1995, however, results were clearly showing that hearing-impaired listeners were able to understand the signals provided by their cochlear implants. The findings of Shannon et al. served both to corroborate what we already knew hearing-impaired listeners could do, as well as to perturb our collective attention away from our earlier held focus on specific acoustic cues, and their relation to specific phonemes. It became apparent that our theories of human speech perception needed some tweaking.

At the same time that investigators were discovering that adults could recover linguistic form from processed signals preserving only the global structure created by the supra-glottal vocal tract, scientists examining the development of speech production and perception were finding that this kind of structure may play an important role in language acquisition. For

example, Charles-Luce and Luce (1990; 1995) showed that children, unlike adults, do not perform detailed acoustic-phonetic analyses when recognizing spoken words; instead children use strategies based on more global acoustic properties. Boysson-Bardies, Sagart, Halle and Durand (1986) showed that the long-term spectra of 10-month-old infants resembled those of adults in their native language community. So before they had uttered their first real word, these infants had discovered the postural vocal tract settings that help to define their native language.

Another group of investigators asked the question of whether or not children show the same trading relations between the spectro-temporal properties used in phonetic judgments that adults show. Work with adults had already demonstrated that the settings of some relevant acoustic cues that evoked specific phonetic judgments from listeners could be influenced by the settings of other properties. For example, Mann and Repp (1981) showed that adults would label noises with lower spectral means as [s] when formant transitions in trailing vocalic portions were appropriate for an alveolar constriction; but if those formant transitions were appropriate for a palatal constriction, then the mean frequency of the noise had to be higher for listeners to label the syllable as starting with [s]. Many psycholinguists of the time considered the spectral shape of obstruent noises to be a primary, context-independent feature supporting correct classification by human listeners; vocalic formant transitions were considered secondary (e.g., Stevens & Blumstein, 1978). According to this view, human listeners learn how the secondary properties vary with the primary properties through extensive experience, and so the child must surely use spectral shape initially to label obstruent noises (Stevens, 1975).

In 1987, Nittrouer and Studdert-Kennedy reported a study in which they asked if children showed the same trading of acoustic cues as adults. In other words, did children appreciate how the setting of a “primary” acoustic property could be affected by a “secondary” property? The expectation was that children would not show trading relations of the same magnitude as adults did. Contrary to that expectation, however, Nittrouer and Studdert-Kennedy observed that children's labeling responses actually indicated ***stronger*** effects of the following formant transitions on the spectral composition of noises associated with specific sibilant decisions. Although that finding seemed at odds with theoretical notions of the day, it was nonetheless largely supported by subsequent research (e.g., Mayo, Scobbie, Hewlett, & Waters, 2003; Nittrouer, 1992; Nittrouer & Miller, 1997a; 1997b; Watson, 1997). Within the context of findings showing that adults can comprehend speech with only global spectral structure (e.g., Remez et al., 1981) and that infants replicate in their own babbled productions the global properties of speech produced by adults in their language community, a new hypothesis evolved (e.g., Nittrouer, 2006). That hypothesis suggested that children use global spectral structure to gain initial access to linguistic form. Then, as they acquire experience with their native language, they discover the spectro-temporal details specific to phonetic form in that language.

The study reported here sought to compare the abilities of native English-speaking children and adults to recognize signals processed to preserve only the spectral skeletons or the amplitude envelopes (two forms of global structure created by the vocal-tract filter). This was considered an appropriate and timely focus because results could extend our understanding of the development of speech perception, just as our perspectives concerning initial perceptual capacities and subsequent learning are evolving. Traditionally, theories of language acquisition have assumed that phonetic units are available to infants very early in life (Kuhl, 1987, provides a substantive review), but now the idea is emerging that infants must discover these units through protracted perceptual learning (e.g., Beckman & Edwards, 2000; Nittrouer, 2001; Werker & Yeung, 2005). So the question arises of how children ever gain a toehold into a signal as highly continuous and variable as the speech signal. This

study explored the possibility that such a toehold may be provided by some form of global acoustic structure, most likely the spectral skeleton.

One study has presented children with amplitude envelope stimuli created from isolated words and sentences with high syntactic and semantic context (Eisenberg, Shannon, Schaefer Martinez, Wygonski, & Boothroyd, 2000). Results showed that 5- to 7-year-old children were poorer at recognizing both isolated words and sentences than adults and older children when presented with the same number of channels, spurring the conclusion that "… speech pattern recognition is still maturing in children 5 to 7 years of age" (p. 2709). It is not clear whether the authors are suggesting that some aspect of auditory development is not completed by 5 to 7 years of age, or that children at these ages have not learned to recover perceptually meaningful form from those impoverished signals. Trying to determine which of these alternative explanations accounts for their results was one goal of the present experiment.

It has been a long-held perspective that any measured difference in children's and adults' speech perception may be attributable to immature sensory pathways in children (e.g., Sussman, 1993), and so it might be that the 5- to 7-year-old listeners in the Eisenberg et al. (2000) study did not process the amplitude envelope signals in the same way as adults. Consequently, the maturity of the auditory systems in listeners in the current study was an issue, and needed to be controlled. To do so we included a group of listeners with mature auditory systems, but with different linguistic experiences from those of our English-speaking adults: namely, second language learners of English. We could have used listeners with any native language who had learned English as a second language, but chose to use native speakers of Mandarin Chinese because it has been demonstrated that adult Chinese listeners recognize Chinese sentences processed to preserve only amplitude envelopes as well as adult English listeners recognize English sentences processed in the same way (Fu, Zeng, Shannon, & Soli, 1998).

The other possible explanation for the results of Eisenberg et al. (2000) is that the sensory information in the amplitude envelope stimuli is processed in the auditory system similarly by the young and older listeners, but that the young listeners are not able to organize this information in such a way as to recover perceptual form. Visual analogies of what must be accomplished in listening to these acoustically impoverished signals abound, and include examples such as the classic Rubin's vase, which can be recognized as a vase or two portraits: The sensory information is sparse, and the image that is recovered depends on how the perceiver organizes that information. In the work described here, the question was asked if different groups of listeners, and so listeners with different amounts and types of experience with the linguistic form to be recovered, were affected differently by different kinds of signal processing. The answer to this question could inform us as to what kinds of signal structure aid children in language learning.

In sum, the current experiment compared speech recognition for signals processed to preserve either spectral skeletons or amplitude envelopes in order to explore the roles of these forms of structure in speech perception and its development. Sentences were processed to preserve only spectral skeletons by creating sine wave analogs of those sentences. Global amplitude structure for sentences was obtained using the methods of Shannon and colleagues (1995). The focus of this study was on the question of whether children recover linguistic form from spectral skeletons or amplitude envelopes as well as adults do in speech perception. The answer to this question could extend our understanding of how children acquire language. Non-native speakers of English were incorporated into the study to serve as a "control" group with mature auditory pathways, but without the native language experience of our native adult listeners.

Method

Participants

Data are reported for 120 participants: 40 adult native English speakers, 40 seven-year-old native English speakers, and 40 adult native Mandarin speakers who were competent second-language speakers of English. All Mandarin speakers had started learning English between the ages of 12 and 14, had spoken it an average of 14 years, and were in the United States as either graduate students, post-doctoral fellows, or junior faculty. All adults were between 18 and 38 years of age, with mean ages of 25 years for the English-speaking adults, and 27 years for the Mandarin-speaking adults. To participate, all participants had to pass a hearing screening and demonstrate appropriate abilities to comprehend and produce spoken English. Half of the participants in each group listened to the sine wave stimuli and half listened to the amplitude envelope stimuli, making a total of six groups of listeners. In all but two of these groups the proportions of male and female listeners were exactly the same (ten in each group): For the sine wave stimuli, twelve English-speaking adult listeners were female; for the amplitude envelope stimuli, twelve Mandarin-speaking listeners were female.

Equipment

All speech samples were recorded in a sound-proof booth, directly onto the computer hard drive, via an AKG C535 EB microphone, a Shure M268 amplifier, and a Creative Labs Soundblaster analog-to-digital converter. Perceptual testing took place in a sound-proof booth, with the computer that controlled the experiment in an adjacent room. The hearing screening was done with a Welch Allen TM262 audiometer and TDH-39 earphones. Stimuli were stored on a computer and presented through a Samson headphone amplifier, and AKG-K141 headphones.

Stimuli

Thirty-six four-word sentences were either adopted from Boothroyd and Nittrouer (1988) or developed according to the same rules used for developing those sentences: All sentences consisted entirely of monosyllabic words, were syntactically appropriate for English, but semantically anomalous. Most of these sentences had some version of a subject-verb structure (e.g., *Dumb shoes will sing. Knees talk with mice*.), although five had a command structure (e.g., *Paint your belt warm*.) These sentences can be found in Appendix A. Six sentences were used for practice, and thirty were used for testing. Sentences were recorded by an adult male speaker of American English at a 44.1-kHz sampling rate with 16-bit digitization. Figure 1 displays a natural speech sample of *Late forks hit low* (top panel), the sine wave replica (center panel), and the four-channel amplitude envelope replica (bottom panel). All sentences were equalized for mean RMS amplitude across sentences before any processing was done.

For creating the sine wave (SW) stimuli, a Praat routine written by C. Darwin and available at his website was used to extract the center frequency of formants. Stimuli were generated from these formant frequencies, and spectrograms were compared to spectrograms of the original stimuli to ensure that the trajectories of the sine waves matched those of the formant frequencies.

To create the amplitude envelope (AE) stimuli, a MATLAB routine was written. Both four- and eight-channel AE stimuli were created and played for listeners because we wanted to have AE stimuli that would produce recognition scores in the same ballpark as those we expected to obtain for the SW stimuli, and that was difficult to predict before the experiment. All signals were first low-pass filtered with an upper cut-off frequency of 8,000

Hz. For the four-channel stimuli, cut-off frequencies between bands were 800, 1,600, and 3,200 Hz; for the eight-channel stimuli, cut-off frequencies were 400, 800, 1,200, 1,800, 2,400, 3,000, and 4,500 Hz. Each channel was half-wave rectified, and results used to modulate white noise limited by the same band-pass filters as those used to divide the speech signal into channels. Resulting bands of modulated noise were low-pass filtered using a 160-Hz high-frequency cut-off, and combined.

Procedures

All stimuli were presented under headphones in a sound-proof booth at 68 dB SPL. Practice was the same for listeners hearing both the AE and SW stimuli, but listeners in the SW condition were given one additional task prior to practice because it can be difficult for some listeners to perceptually integrate SW stimuli into forms that are speech-like. A similar problem has not been reported for AE stimuli. To facilitate listeners' abilities to hear the SW stimuli as speech-like, we played a story before they heard the practice sentences. The story was presented first as natural speech, and then as SW replicas derived from that speech. A picture book without words accompanied the story.

Each of the six practice sentences was presented, either as an AE or SW signal. For listeners in the AE condition, three each were presented as four- and eight-channel AE stimuli. For each practice sentence, the natural version was played first, and the listener was instructed to listen and repeat it. Next the listener was told “Now you will hear a robot say the sentence. Listen carefully so you can repeat it.” The notion of a robot was invoked to help personify the stimuli. For both the AE and SW conditions, listeners were asked at the end of the six-sentence practice if they were able to recognize the “robot's” productions as speech. It was made clear that the question was whether they could hear it as if it were speech, rather than could they understand every word. If a listener either was unable to repeat any of the unprocessed sentences or failed to report hearing the processed signals as speech, they were dismissed. Thirteen individuals in addition to the 120 listeners whose data are reported here heard the practice sentences and were dismissed because they were unable to hear the sine wave stimuli as speech: six each from the children's and Mandarin-speakers' groups; one from the adult English-speakers' group. None of our listeners had difficulty hearing AE stimuli as speech. Listeners were paid for their participation, regardless of whether they could hear the stimuli as speech, so there was no reason for any listener to falsely state that the stimuli were heard as speech if they were not.

During testing, the order of presentation of the sentences was randomized independently for each listener. Each sentence was presented three times in a row. It is a common impression when listening to speech degraded in some way that recognition might improve if the sentence could just be heard again. Because of that impression, it was thought that recognition might improve with repeated trials, and the amount of improvement might vary across listener groups; that is, perhaps some listeners would benefit more than others from hearing each sentence again. If true, then optimal performance would not be obtained for one or more groups with just one trial.

The listeners presented with the AE stimuli heard half of them as four-channel and half as eight-channel stimuli. The selection of sentences to be presented as four- or eight-channel stimuli was randomly made for each listener by the software, and presentation of these two kinds of stimuli was intermingled during testing with the stipulation that no more than two four- or eight-channel stimuli could be presented in a row.

After hearing the sentences in their processed forms, these sentences were played to listeners in their unprocessed forms. None of the listeners had difficulty comprehending those unprocessed sentences.

Results

The percent of words recognized correctly across all sentences within any one condition served as the dependent measure. Table 1 shows mean percent correct word recognition for each listener group on each trial. Mean recognition did not improve by more than 10% from the first to the third trial for any group, and all improvement was within one standard deviation of individual trial means. Consequently, mean results across the three trials were used in subsequent analyses.

Figure 2 shows mean correct word recognition for each listener group for each kind of stimulus. (For the Mandarin-speaking adults, there was no correlation between length of English experience and recognition scores.) Table 2 shows the results of one-way analyses of variance (ANOVAs), and post hoc comparisons performed on percent correct word recognition scores. All statistics were computed using arc sine transforms because some listeners scored below ten percent correct in some conditions.

Significant main effects of group were obtained for all three conditions; therefore post hoc comparisons between groups were examined. For the eight-channel AE stimuli, mean correct recognition for each group differed from that of every other group ($\eta^2 = .75$): English-speaking adults performed best, followed by English-speaking children, and lastly by Mandarin-speaking adults. For the four-channel AE stimuli, the children and Mandarin-speaking adults performed similarly, but both groups performed more poorly than the English-speaking adults ($\eta^2 = .41$). For the SW stimuli, the children and English-speaking adults performed the same, and both of those groups performed better than the Mandarin-speaking adults ($\eta^2 = .56$). So these between-group results show that children and Mandarin-speaking adults performed more poorly than English-speaking adults with the AE stimuli, regardless of how many channels were used. From this result support was garnered for the conclusion that amplitude structure differs across languages, and children must discover that structure in their native language. From results with the SW stimuli, support was garnered for the conclusion that global spectral structure also differs across languages, which suggests that children must discover that kind of structure, as well. However, apparently they have already done so by seven years of age.

Comparisons across stimulus types for each group were also examined, and recognition scores are replotted in Figure 3 to illustrate within-group comparisons for the three conditions. Table 3 shows results of *t* tests comparing performance across conditions for each listener group: For comparisons of four- and eight-channel AE stimuli, these were within-groups *t* tests; for comparisons of either the four- or eight-channel AE stimuli with SW stimuli, these were between-groups *t* tests. Again, these analyses were performed using arc sine transforms. For listeners hearing the AE stimuli, performance was better with the eight-channel than with the four-channel stimuli, for all groups. When we look across AE and SW stimuli, we find for all three groups that listeners performed better with the eight-channel AE stimuli than with the SW stimuli. But for the comparison of four-channel AE and SW stimuli, children were the only group to show a difference in recognition scores: Children hearing the SW stimuli out-performed children hearing the four-channel AE stimuli.

Discussion

This experiment was designed to measure speech recognition for English-speaking adults, English-speaking children, and Mandarin-speaking adults, using sentences that were syntactically correct but semantically anomalous. Seven-year-old children learning English were found to recognize stimuli that preserved only the spectral skeleton of the speech

signal as well as native English-speaking adults. Both the child and adult native speakers of English recognized these SW stimuli better than native Mandarin-speaking adults who had learned English as a second language. But when presented with signals that preserved only amplitude envelopes, the children performed similarly to the Mandarin-speaking adults, and both of these groups performed more poorly than the English-speaking adults. The finding that native Mandarin speakers who are fluent in English performed so poorly with both sorts of global signals suggests that variability in performance may be largely explained by differences in language experiences that would influence a listener's ability to recover linguistic form from these signals. These results for native Mandarin speakers diminish the possibility that age-related differences observed by Eisenberg et al. (2000) and by us are explained by immature auditory systems in children. Of course, it is possible that Mandarin-speaking adults and English-speaking children would have demonstrated similar results with the AE stimuli for different reasons (differences in language experience on the part of Mandarin-speaking adults and immature auditory systems on the part of children), but that seems unlikely. The principle of parsimony encourages us to adopt one common explanation for the results of both groups, and the finding that children performed similarly to adults for the SW stimuli indicates that children's auditory systems processed at least one kind of impoverished signal maturely.

The most significant implications of these results involve language development. By seven years of age, these children demonstrated native skill in using spectral skeletons, but not amplitude envelopes. Evidence is accumulating that children are not adult-like in their use of the spectrally specific, temporally brief properties of the acoustic speech signal (i.e., acoustic cues) until much later, around the age of puberty (Hazan & Barrett, 2000). Taken together these results support the broader suggestion that global spectral structure plays a pivotal role in language acquisition. It may be that children are able to perceptually organize global spectral structure in the speech signal in order to start recovering linguistic form early in life, and the ability to do so facilitates the discovery of other sorts of structure by helping children divide the speech stream into separate linguistic units, such as words. Based on results of this study, we can only say that children acquired mature sensitivity to global spectral patterns in their native language by seven years of age, but evidence from others suggests that children may learn about this spectral structure much earlier, possibly within the first year of life (Boysson-Bardies, et al., 1986).

The focus of much speech perception research has traditionally been on discrete spectro-temporal properties, and methods of investigation have involved manipulating these properties in syllable-length signals to evaluate how those manipulations affect phonetic decisions. In accordance with that work, research on infant speech perception manipulated the same properties to see what infants could discriminate. (Labeling tasks can not be used with infants.) This collective body of research has led to the perspective that infants are sensitive to the acoustic cues relevant in signaling phonetic segments, effectively from birth (e.g., Aslin, Pisoni, Hennessy, & Perey, 1981; Eimas, Siqueland, Jusczyk, & Vigorito; 1971; Kuhl, 1979). But infants rarely hear speech signals the length of syllables; instead, they hear speech in phrase or sentence length units. Because infants lack lexical knowledge, one early chore facing the child must surely be learning how to chisel out linguistically meaningful stretches of the signal. The question therefore arises of how children can identify stretches of the speech signal that form separate linguistic units, such as phrases within sentences or words within phrases. Regarding the first of these, several studies have shown that prosodic markers, such as pauses, are readily used by infants to parse the signal into clause-length units (e.g., Hirsh-Pasek et al., 1987; Kemler Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989). However, speakers do not mark word boundaries with pauses, so the mystery remains of how infants learn to separate individual words from the ongoing speech signal. The suggestion made here is that infants could be using recurring patterns of global spectral

structure to perceptually isolate and recognize the repeating strings. And although that suggestion is speculative at present, there are now two kinds of findings to motivate it: Young children rely on formant transitions more than adults for many phonetic decisions (e.g., Greenlee, 1980; Krause, 1982; Nittrouer, 1992; Nittrouer & Studdert-Kennedy, 1987; Wardrip-Fruin & Peach, 1984), and children recognize stimuli preserving only global spectral structure in an adult-like fashion sooner than they recognize stimuli preserving only global amplitude structure (this study) or discrete sequences of spectro-temporal structure (this study compared with Hazan & Barrett, 2000). When combined these findings suggest that global spectral structure may play an important role in early language learning. That structure could help infants begin to separate stretches of the ongoing speech signal that form linguistic units. There is already some evidence that infants use an acoustic property not related to individual phonetic segments to begin recovering separate words from the ongoing speech signal: Specifically, Mattys, Jusczyk, Luce, and Morgan (1999) found that 9-month-olds were more sensitive to a prosodic cue (i.e., stress) than to phonotactic constraints in recognizing word boundaries. Although prosody is a property arising largely from the laryngeal source of speech (while sine wave analogs preserve structure imposed by the filter), this finding nonetheless is one more piece of evidence that children can use global structure to begin recognizing word-length units. Once these units are extricated, children are better able to discover the details of those units (i.e., discrete spectro-temporal properties), forming more precise representations of each lexical item.

In conclusion, this experiment was undertaken to examine children's abilities to recover linguistic form from acoustic speech signals processed in each of two ways that largely eliminate the spectro-temporal properties traditionally associated with phonetic structure. We know these processing algorithms preserve access to linguistic form for adult listeners who are native speakers of the language being examined. Results presented here indicate that children can recover linguistic form for processed signals that preserve global spectral structure as well as adult, native speakers of the same language, but can not do so when only amplitude envelopes are preserved. Therefore it is concluded that children discover the global spectral structure of their native language before the global amplitude structure. We presume that those spectral skeletons have heuristic value for language learning.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders grant R01 DC000633. The authors thank Vignesh Hariharan for help with programming. For further information regarding this article, contact the first author at Otolaryngology – Head & Neck Surgery, The Ohio State University, Cramblett Hall, Room 4126, 456 West 10th Avenue, Columbus, OH 43210.

Appendix A: Sentences

Practice sentences

1. Cooks run in brooms.
2. Ducks teach sore camps.
3. Find girls these clouds.
4. Gangs load near sweat.
5. Great shelf needs tape.
6. Throw his park hand.

Test sentences

Blue chairs speak well.

Cars jump from fish.

Cats get bad ground.

Cups kill fat leaves.

Drive my throat late.

Dumb shoes will sing.

Fan spells large toy.

Feet catch bright thieves.

Green hands don't sink.

Hard checks think tall.

Jobs get thick hay.

Knees talk with mice.

Late forks hit low.

Late fruit spins lakes.

Lead this coat home.

Lend them less sleep.

Let their flood hear.

Paint your belt warm.

Pink chalk bakes phones.

Sad cars want chills.

Slow dice buy long.

Small lunch wipes sand.

Soap takes on dogs.

Socks pack out ropes.

Soft rocks taste red.

Suits burn fair trail.

Teeth sleep on doors.

Thin books look soft.

Trucks drop sweet dust.

Wide pens swim high.

References

Aslin RN, Pisoni DB, Hennessy BL, Perey AJ. Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. Child Development. 1981; 52:1135–1145. [PubMed: 7318516]

Beckman ME, Edwards J. The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. Child Development. 2000; 71:240–249. [PubMed: 10836579]

Boothroyd A, Nittrouer S. Mathematical treatment of context effects in phoneme and word recognition. Journal of the Acoustical Society of America. 1988; 84:101–114. [PubMed: 3411038]

Boysson-Bardies, B.; Sagart, L.; Halle, P.; Durand, C. Acoustic investigations of cross-linguistic variability in babbling. In: Lindblom, B.; Zetterstrom, R., editors. Precursors of early speech. Stockton Press; New York: 1986. p. 113-126.

Bricker, PD.; Pruzansky, S. Speaker recognition. In: Lass, NJ., editor. Contemporary issues in experimental phonetics. Academic Press; New York: 1976. p. 295-326.

Carrell TD. Effects of glottal waveform on the perception of talker sex. Journal of the Acoustical Society of America. 1981; 70:S97.

Charles-Luce J, Luce PA. An examination of similarity neighbourhoods in young children's receptive vocabularies. Journal of Child Language. 1995; 22:727–735. [PubMed: 8789521]

Charles-Luce J, Luce PA. Similarity neighbourhoods of words in young children's lexicons. Journal of Child Language. 1990; 17:205–215. [PubMed: 2312642]

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J. Speech perception in infants. Science. 1971; 171:303–306. [PubMed: 5538846]

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

Fant, G. Acoustic theory of speech production. Mouton; The Hague, Netherlands: 1960.

Fu Q, Zeng FG, Shannon RV, Soli SD. Importance of tonal envelope cues in Chinese speech recognition. Journal of the Acoustical Society of America. 1998; 104:505–510. [PubMed: 9670541]

Gleitman, L.; Wanner, E. Language acquisition: The state of the art. In: Wanner, E.; Gleitman, L., editors. Language acquisition: The state of the art. Cambridge University Press; Cambridge, UK: 1982. p. 3-48.

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

Hazan V, Barrett S. The development of phonemic categorization in children aged 6–12. Journal of Phonetics. 2000; 28:377–396.

Hirsh-Pasek K, Kemler Nelson DG, Jusczyk PW, Cassidy KW, Druss B, Kennedy L. Clauses are perceptual units for young infants. Cognition. 1987; 26:269–286. [PubMed: 3677573]

Jassem W. Pitch and compass of the speaking voice. Journal of the International Phonetic Association. 1971; 1:59–68.

Jusczyk, PW. The discovery of spoken language. The MIT Press; Cambridge: 1997.

Kemler Nelson DG, Hirsh-Pasek K, Jusczyk PW, Cassidy KW. How the prosodic cues in motherese might assist language learning. Journal of Child Language. 1989; 16:55–68. [PubMed: 2925815]

Krause SE. Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults. Journal of the Acoustical Society of America. 1982; 71:990–995. [PubMed: 7085987]

Kuhl PK. Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. Journal of the Acoustical Society of America. 1979; 66:1668–1679. [PubMed: 521551]

Kuhl, PK. Perception of speech and sound in early infancy. In: Salapatek, P.; Cohen, L., editors. Handbook of infant perception, Vol. 2, From perception to cognition. Academic Press; New York: 1987. p. 275-382.

Mann VA, Repp BH. Influence of preceding fricative on stop consonant perception. Journal of the Acoustical Society of America. 1981; 69:548–558. [PubMed: 7462477]

Mattys SL, Jusczyk PW, Luce PA, Morgan JL. Phonotactic and prosodic effects on word segmentation in infants. Cognitive Psychology. 1999; 38:465–494. [PubMed: 10334878]

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

Morgan, JL.; Demuth, K. Signal to syntax. Erlbaum; Mahwah, NJ: 1996.

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

Nittrouer S. Challenging the notion of innate phonetic boundaries. Journal of the Acoustical Society of America. 2001; 110:1581–1597. [PubMed: 11572368]

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

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

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

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

Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science. 1981; 212:947–949. [PubMed: 7233191]

Repp BH. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin. 1982; 92:81–110. [PubMed: 7134330]

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science. 1995; 270:303–304. [PubMed: 7569981]

Soderstrom M, Seidl A, Kemler Nelson DG, Jusczyk PW. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. Journal of Memory and Language. 2003; 49:249–267.

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

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

Sussman JE. Auditory processing in children's speech perception: Results of selective adaptation and discrimination tasks. Journal of Speech and Hearing Research. 1993; 36:380–395. [PubMed: 8487530]

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

Watson, JMM. Unpublished doctoral dissertation. Queen Margaret College; Edinburgh: 1997. Sibilant-vowel coarticulation in the perception of speech by children with phonological disorder.

Werker JF, Yeung HH. Infant speech perception bootstraps word learning. Trends in Neurosciences. 2005; 9:519–527.

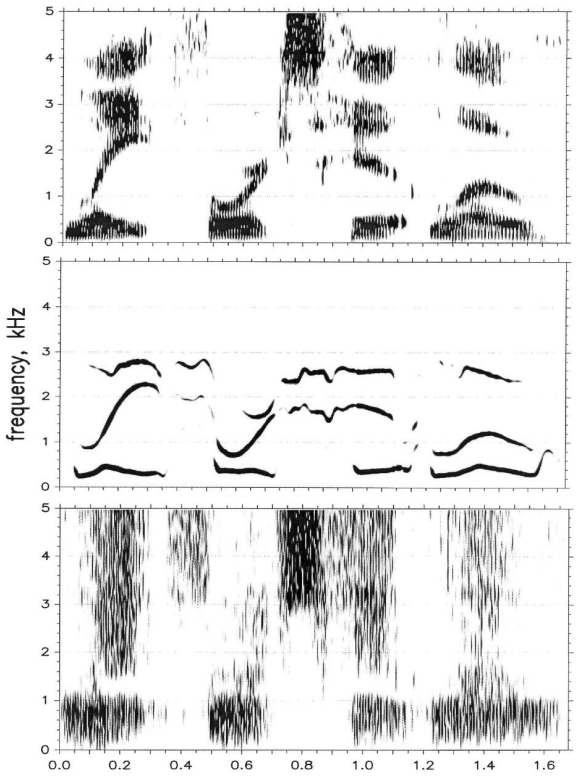

FIGURE 1.
Spectrograms of the sentence “Late forks hit low.” (Top) Natural speech sample from which SW and AE stimuli were derived. (Middle) SW stimulus. (Bottom) Four-channel AE stimulus.

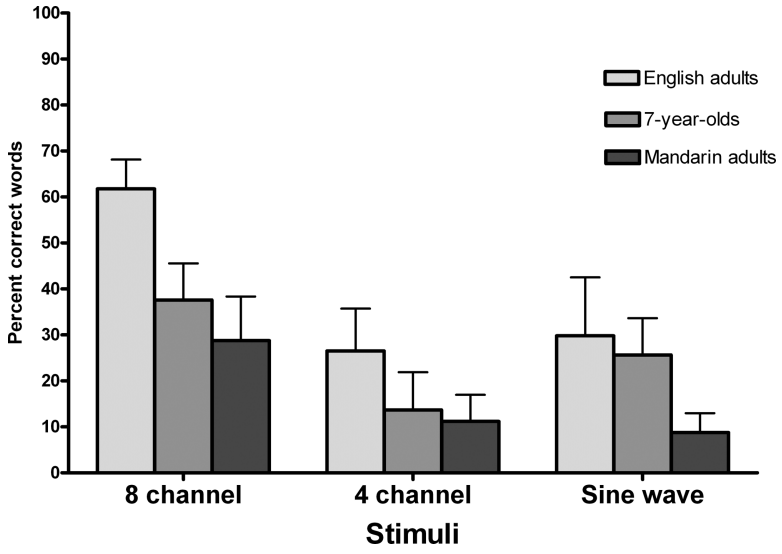

FIGURE 2.
Percent correct recognition for each of the three listener groups, by stimulus type.

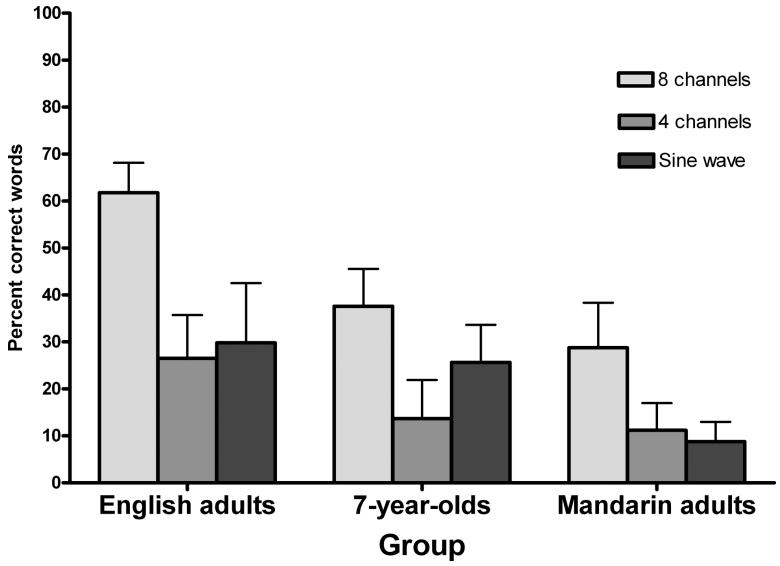

FIGURE 3.
Percent correct recognition for each of the three stimulus types, by listener group.

Table 1

Mean percent correct word recognition for each trial in each condition.

		English adults	7 year olds	Mandarin adults
8 channel AE	Trial 1	57.30 (8.55)	33.15 (9.18)	23.10 (9.41)
	Trial 2	63.10 (6.34)	38.50 (8.32)	29.90 (10.25)
	Trial 3	64.90 (5.64)	41.15 (8.39)	33.35 (10.42)
	Mean	***61.77 (6.39)***	***37.60 (7.94)***	***28.78 (9.58)***
4 channel AE	Trial 1	21.90 (8.66)	12.20 (8.18)	9.10 (4.46)
	Trial 2	27.15 (9.65)	13.95 (8.19)	11.90 (6.37)
	Trial 3	30.50 (10.25)	14.90 (8.81)	12.60 (7.21)
	Mean	***26.52 (9.20)***	***13.68 (8.20)***	***11.20 (5.77)***
SW	Trial 1	24.60 (12.90)	22.05 (7.23)	6.75 (4.03)
	Trial 2	31.45 (13.07)	26.80 (8.20)	9.25 (4.17)
	Trial 3	33.45 (12.53)	28.10 (9.25)	10.35 (4.64)
	Mean	***29.83 (12.67)***	***25.65 (7.98)***	***8.78 (4.19)***

Table 2

Results of the statistical analyses for group effects, using arc sine transforms.

	F ratio or t	P
8-channel AE	84.66	<.001
English-speaking adults vs. children	9.07	<.001
English- vs. Mandarin-speaking adults	12.61	<.001
Children vs. Mandarin-speaking adults	3.54	<.001
4-channel AE	20.07	<.001
English-speaking adults vs. children	4.95	<.001
English- vs. Mandarin-speaking adults	5.90	<.001
Children vs. Mandarin-speaking adults	0.94	*NS*
SW	35.90	<.001
English-speaking adults vs. children	1.22	*NS*
English- vs. Mandarin-speaking adults	7.87	<.001
Children vs. Mandarin-speaking adults	6.65	<.001

F ratios are for oneway ANOVAs, with group as the between-subjects factor; *t* tests are for post hoc comparisons. Degrees of freedom are 2,57 for the ANOVAs, 57 for the *t* tests. Computed *p* values are presented. Bonferroni corrections for three multiple comparisons require a *p* of less than .0033 for the comparison to be significant at the .01 level. All post hoc comparisons met that criterion.

Table 3

Results of the *t* tests done to examine stimulus effects for each group, using arc sine transforms.

	t	P
English-speaking adults		
8- *vs.* 4-channel AE	18.60	<.001
8-channel AE *vs.* SW	9.44	<.001
4-channel AE *vs.* SW	0.82	NS
Children		
8- *vs.* 4-channel AE	10.01	<.001
8-channel AE *vs.* SW	4.61	<.001
4-channel AE *vs.* SW	4.59	<.001
Mandarin-speaking adults		
8- *vs.* 4-channel AE	8.70	<.001
8-channel AE *vs.* SW	9.21	<.001
4-channel AE *vs.* SW	1.36	NS

The test of 8- vs. 4-channel AE stimuli involved within-group comparisons, and each had 19 degrees of freedom. The test of 8-channel AE vs. SW stimuli and of 4-channel AE vs. SW stimuli involved two-group comparisons, and each had 38 degrees of freedom.