Children hear the forest (L)

Susan Nittrouer[a)]
Ohio State University

(Received 21 March 2006; revised 22 June 2006; accepted 8 July 2006)

How do children develop the ability to recognize phonetic structure in their native language with the accuracy and efficiency of adults? In particular, how do children learn what information in speech signals is relevant to linguistic structure in their native language, and what information is not? These questions are the focus of considerable investigation, including several studies by Catherine Mayo and Alice Turk. In a proposed Letter by Mayo and Turk, the comparative role of the isolated consonant-vowel formant transition in children's and adults' speech perception was questioned. Although Mayo and Turk ultimately decided to withdraw their letter, this note, originally written as a reply to their letter, was retained. It highlights the fact that the isolated formant transition must be viewed as part of a more global aspect of structure in the acoustic speech stream, one that arises from the rather slowly changing adjustments made in vocal-tract geometry. Only by maintaining this perspective of acoustic speech structure can we ensure that we design stimuli that provide valid tests of our hypotheses and interpret results in a meaningful way. *© 2006 Acoustical Society of America.*
[DOI: 10.1121/1.2335273]

PACS number(s): 43.71.Ft [MSS] Pages: 1799–1802

Once upon a time, the belief was common that phonetic segments line up in the acoustic speech stream like so many trees planted neatly in a row. Accordingly, great effort was put forth to describe these discrete elements in linguistic and acoustic terms. Several notions that continue to be cornerstones of our collective worldview were developed, such as the idea that phonetic segments can be described using linguistic features. So, we all agreed, a segment can be distinctively plus or minus voiced, tense or lax, rounded or unrounded, and so on (e.g., Chomsky and Halle, 1968). Many believed that the phonetic element and/or linguistic feature are appropriately described as collections of acoustic properties. Therefore, investigators searched for the acoustic correlates of segments or features, as indicated in this passage from Blumstein and Stevens (1981):

> "The major claim of a theory of acoustic invariance is that invariant acoustic properties can be derived directly from the acoustic signal, and these properties correspond to the phonetic dimensions which ultimately form the inventory of speech sounds used in natural language. Such a view provides an explicit characterization of the universal set of features, and, in particular, of the phonetic dimensions delineating natural classes. If it is the case that invariant properties structure phonetic dimensions, then such invariance provides an instantiation of particular feature systems." (pp. 27–28)

This view of the acoustic structure of speech and its relation to linguistic structure has enjoyed popular support for a long time.

Relatedly, the notion of categorical perception (e.g., Studdert-Kennedy, Liberman, Harris, and Cooper, 1970) fit the earlier worldview of what happens when a listener hears a sequence of these presumably discrete elements. Categorical perception was based on the idea that a phonemic category can be defined as a circumscribed range of settings for any one unique acoustic property.[1] According to this view, listeners reliably hear sounds with one of these settings as instances of that phonemic category, and similarly reject as instances of that category sounds with different settings. The effect is so strong, so it was hypothesized, that listeners fail to notice acoustic variability across the range of settings on the property that correspond to any one category. Thus, phonemes could effectively be defined as ranges of settings on specific acoustic dimensions. Even the concept of coarticulation rests upon the concept of linear representation of phonetic segments: The acoustic structure of any one segment can be influenced by surrounding phonetic elements, according to traditional accounts of coarticulation. This idea is based on the premise that there must be canonical settings for each phonetic segment. Overall, these notions of categoricalness and linear representation in the acoustic speech signal give us solace and form the framework of how speech is generally thought to be produced and perceived. These ideas have served as the basis of experimental design and theory building for a large body of research, including everything from speech recognition to dyslexia. Attempts to identify the acoustic properties that define phonemic categories (frequently referred to as "cues") have played central roles in the fields of psychology and linguistics for decades, and continue to do so.

But several events over the last 25 years have shaken the foundations of this worldview for some of us. In 1981, four scientists published a paper that would cause us to question our most basic beliefs. Remez, Rubin, Pisoni, and Carrell (1981) showed us that listeners can recover linguistic structure when all of the traditional cues to phonemic identity that we held so dear were eliminated from the signal. These investigators synthesized signals consisting only of the dy-

[a)]Electronic mail: nittrouer.1@osu.edu

namic spectral patterns of speech, and yet listeners were able to understand those signals. At the same time, engineers were designing devices that could be implanted in the cochleas of individuals with profound hearing loss to stimulate the auditory nerve directly. Of course, without the hair cells to provide frequency resolution the best these devices could do was to provide a single channel of information about the acoustic wave form of speech, and many of us predicted unmitigated failure for these devices. But in sharp contrast to our predictions, implant users somehow managed to use that horribly impoverished signal to recover linguistic structure. Since the days of those early implants, cochlear implants have improved, but still implant users have access to only a few channels of information. Even at that, the amplitude envelopes of those few channels is primarily what is received, without even the spectral skeleton provided by sine wave replicas of speech. Nonetheless, experiments with deaf and hearing listeners alike demonstrate that these envelopes serve speech recognition well (Smith, Delgutte, and Oxenham, 2002; Zeng *et al.*, 2004). Clearly both the dynamic changes in spectral resonances and the amplitude envelopes of speech can provide sufficient information to support recognition of linguistic units without traditional speech cues.

Simultaneously with the occurrence of these changes in our understanding of what speech perception entails, investigators interested in production were modifying their worldview. In particular, the idea that there are discrete motor commands for individual phonemes was questioned, and instead the notion was hatched that task-specific goals reduce the multiple degrees of freedom inherent in vocal-tract mechanics. According to this perspective, relations among movements of various articulators in time and space instantiate linguistic structures such as stress, syllable cohesion of intervocalic consonants, and even phonetic identity in the resulting signal (e.g., Kelso, Saltzman, and Tuller, 1986). It became clear that the acoustic speech signal was not shaped by discrete actions of individual articulators. Instead, relatively slow movements such as the rise and fall of the jaw are punctuated by the actions of more rapidly moving articulators, such as the lips, but all these actions are coordinated to impose linguistic structure across the length of an utterance.

It was against these theoretical backdrops that the idea of the developmental weighting shift (DWS) for speech perception emerged. Specifically this idea began with the thought to examine in children's perception what was then termed "trading relations." The premise of trading relations was that settings for one acoustic property could trade with settings for another acoustic property in the phonetic judgments of listeners. For example, Mann and Repp (1980) found that the frequency of a fricative noise that supported a [s] judgment could be lower if the vowel formant frequencies were appropriate for a more apical, rather than velar, place of constriction. The original question posed by Michael Studdert-Kennedy and this author (Nittrouer and Studdert-Kennedy, 1987) was "Would children show this same shift in acceptable settings for one property, based on settings of another property?" Fricative-vowel stimuli similar to those of Mann and Repp were used in this initial exploration of the question, and the results were dramatic. Not only did children show the trading relation described by Mann and Repp, but their responses seemed to be even more strongly influenced by formant transitions than those of adults.

At the time and under the rubric of "trading relations," fricative noise frequency and dynamic formant transitions were viewed as being equal in kind and similar in nature. Each property was one separate bit of the acoustic fiber, a discrete clue that could tell the listener what that temporally discrete element, the fricative, was. Over time, however, the significance of the changes in perspective of speech perception and production described above became apparent: The consonant-vowel (CV) formant transition is not equal in kind to the other acoustic properties that we, as a field, have generally termed "cues." Instead, the CV (or VC) formant transition is one short piece of the larger, continuously changing spectral pattern that arises from the relatively slow modifications made to overall vocal-tract posture. It is these slow changes that children first notice in the speech around them. Gradually, through experience with a native language, children discover the other acoustic properties that are relevant for phonetic identity in their native language. It is not that adults cannot and do not use the dynamic components of the speech signal to recover linguistic structure. Clearly they do. Results with sine wave speech (e.g., Remez *et al.*, 1981) illustrate that fact. Thus, adults are sensitive to both the global spectral structure that arises from the relatively slow modulations of the vocal tract, as well as to the acoustic details imposed by articulatory factors such as the precise shape of a fricative constriction or the exact timing between two gestures. Children, on the other hand, attend primarily to acoustic changes that arise from the slow modulations of the vocal tract, learning about the phonetic significance of details of the signal only as they gain experience with their native language.

This suggestion regarding the development of speech perception skills is supported by findings for perceptual development in general: There is evidence that children glean the global structure of sensory input before discovering the details. For example, Kimchi, Hadad, Behrmann, and Palmer (2005) presented sets of visual patterns that either matched on global structure, but not on local structure, or vice versa to adults and children (ages 5 to 14 years). Participants were alternately asked to judge similarity based on global or local structure. Results showed that children performed as well as adults when asked to judge similarity based on global structure, but errors increased with diminishing age when the task was to judge local structure.

The suggestion that children initially focus their perceptual attention on the acoustic consequences of relatively slow vocal-tract movements finds support from studies of the development of speech production, as well. These global signal components most consistently provide information about places of constriction. In reviewing data from a report by Vihman (1996), Studdert-Kennedy (2000) shows that young children rarely make errors of place in their productions. One of children's earliest accomplishments involves learning to produce the more general movements of the vocal tract, and that includes being able to get from one constriction that is appropriate in their native language to another. Only later do they learn to refine shapes for consonant constrictions and to precisely coordinate timing among various gestures. Work by de Boysson-Bardies, Sagart, Halle and Durand (1986) provides another, elegant illustration of the fact that children initially attend to the overall changes in vocal-tract geometry of those around them. These investigators computed the long-term spectra of adults and 10-month-olds whose native languages were French, Cantonese, or Algerian. The resulting spectra clearly showed that the infants' spectra already resembled those of the adults they hear speaking everyday. What this means is that these children were learning about the large postural adjustments that are typical of the languages they would come to speak, adjustments having to do with the larynx, the pharynx, and the velum. Again, this perspective of speech development has correlates in another developmental literature: Thelen (e.g., 1985) has shown that children initially acquire a general leg motion for walking, without differentiation of the individual joints. Only later do they acquire independent control over the actions of the hips, knees, and ankles.

In summary, the theoretical perspective being offered here is that children initially attend strongly to the global changes in the acoustic speech signal arising from relatively slow modifications in vocal tract postures. This perspective has explanatory power for many earlier findings. For example, we would not expect, and have not found, that adults and children differ in their labeling of stimuli when dynamic signal components are pretty much all that are available. An example of this situation is provided by the English weak fricatives [*θ*] and [f]. These fricatives differ very little from each other in noise spectra; accordingly, Harris (1958) found that adults weight formant transitions strongly in decisions regarding their identity. Nittrouer (2002) hypothesized that under these conditions adults and children would weight similarly fricative noise spectra and formant transitions, and results supported that position: Both adults and children depended largely on formant transitions for their fricative decisions. Another example of a situation in which we would not expect weighting of formant transitions to differ for adults and children is when formant transitions do not differ across the stimuli listeners are being asked to label. Mayo and Turk (2004) constructed continua of stop-vowel stimuli with the same place of constriction across each continuum, but with different voicing characteristics for the stops. Thus, formant trajectories were the same across stimuli. As expected, the primary difference among stimuli was when those formants switched from being excited by an aperiodic to a periodic source. Given the constraint on information conveyed by the dynamic character of formant transitions in this contrast, adults and children did not differ in the extent to which formant transitions contributed to their voicing decisions.

None of what has been written here is meant to support or challenge specific theories of phonology, such as whether phonology is word or phoneme based. Human speech is structured at many levels, and presumably each level has significance in communication. Furthermore, children ultimately need to know about structure at each of these levels for their native language in order to attain adult levels of language competency. Many intriguing questions are left to be explored, such as if children explicitly need to learn about each kind of structure or if some structure is automatically recoverable. The purpose of this Letter has merely been to illustrate that we have only recently recognized the significance of global structure in language processing. From that perspective we realize that the formant transition is a specific instance of general vocal-tract dynamics. In the conduct of speech perception experiments, particularly those involving children, we must be mindful of how stimulus design relates to both global and local signal structure. Such attention will help us avoid designing stimuli that violate natural constraints, as well as help us interpret results more appropriately. In general it is fair to say that while we have been busy arranging and rearranging the details of the speech signal in our experiments, it has been the global structure that children have been noticing. Children initially hear the forest, not the trees.

This Letter was originally composed as a Reply to a Comment written by Catherine Mayo and Alice Turk. Although they withdrew their Comment, the decision was made to let this Letter stand on its own. The author thanks Dr. Mayo and Dr. Turk for their stimulating exchange, and several reviewers, including Susan Rvachew, for their comments. This work was supported by research Grant No. R01 DC000633 from the National Institute on Deafness and Other Communication Disorders.

[1] The terms "phoneme" and "phonetic segment" are generally used to refer to the abstract unit and the physical segment as instantiated in the acoustic signal, respectively. According to the view of speech to be presented here, the distinction becomes rather blurry.

Blumstein, S. E., and Stevens, K. N. (**1981**). "Phonetic features and acoustic invariance in speech," Cognition **10**, 25–32.

Chomsky, N., and Halle, M. (**1968**). *The Sound Pattern of English* (MIT Press, Cambridge, MA).

de Boysson-Bardies, B., Sagart, L., Halle, P., and Durand, C. (**1986**). "Acoustic investigations of cross-linguistic variability in babbling," in *Precursors of Early Speech*, edited by B. Lindblom and R. Zetterström (Stockton Press, New York), pp. 113–126.

Harris, K. S. (**1958**). "Cues for the discrimination of American English fricatives in spoken syllables," Lang Speech **1**, 1–7.

Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (**1986**). "The dynamical perspective on speech production: Data and theory," J. Phonetics **14**, 29–59.

Kimchi, R., Hadad, B., Behrmann, M., and Palmer, S. E. (**2005**). "Microgenesis and ontogenesis of perceptual organization—Evidence from global and local processing of hierarchical patterns," Psychol. Sci. **16**, 282–290.

Mann, V. A., and Repp, B. H. (**1980**). "Influence of vocalic context on perception of the /ʃ/-/s/ distinction," Percept. Psychophys. **28**, 213–228.

Mayo, C., and Turk, A. (**2004**). "Adult-child differences in acoustic cue weighting are influenced by segmental context: children are not always perceptually biased toward transitions," J. Acoust. Soc. Am. **115**, 3184–3194.

Nittrouer, S. (**2002**). "Learning to perceive speech: How fricative perception changes, and how it stays the same," J. Acoust. Soc. Am. **112**, 711–719.

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

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**, 87–90.

Studdert-Kennedy, M. (**2000**). "Imitation and emergence of segments," Phonetica **57**, 275–283.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., and Cooper, F. S. (**1970**). "Theoretical notes. Motor theory of speech perception: A reply to Lane's critical review," Psychol. Rev. **77**, 234–249.

Thelen, E. (**1985**). Developmental origins of motor coordination: Leg movements in human infants," Dev. Psychobiol. **18**, 1–22.

Vihman, M. M. (**1996**). *Phonological Development* (Blackwell, Oxford, UK).

Zeng, F. G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y. Y. *et al.* (**2004**). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," J. Acoust. Soc. Am. **116**, 1351–1354.