Acoustic-based syllabic representation and articulatory gesture detection: Prerequisites for early childhood phonetic and articulatory development

Kevin L. Markey
Department of Computer Science and Institute of Cognitive Science
University of Colorado
Boulder, Colorado 80309-0430
markey@cs.colorado.edu

Abstract

We describe the perceptual foundations of a sensorimotor model of early childhood phonetic and articulatory development. The model's auditory perception is sensitive to prosodic and syllabic structure and simulates the categorical phonetic perception of late infancy. Importantly, the model relies on exclusively acoustic cues and their statistical distribution in the linguistic environment, avoiding prior assumptions of articulatory-acoustic correlations or linguistic contrasts which are inappropriate for a model of perceptual development. The model detects and categorizes speech segments, which, despite their acoustic basis, correlate with linguistic events and articulatory gestures. The resulting representation supports not only word recognition but also the unique demands of articulatory motor control and its development. In simulations examining the distinctiveness and faithfulness of the representation, we find that it preserves and makes explicit information about the phonetic properties of the acoustic signal.

Motivation

Human speech and human listening evolved together. It is therefore plausible that speech perception is specialized not only for word and sentence recognition but also for the unique demands of articulatory motor control and that it plays an important role in articulatory and phonological development. Perception's role in motor control is seldom acknowledged except as the source of target sounds to be imitated or learned. However, milestones of perceptual development always precede corresponding milestones of motor development. We hypothesize that the categorical character of auditory perception which underlies robust word recognition also acts as a grammar which defines the well-formedness of children's speech, shaping the distribution of sounds in their productive repertories. The connection is deeper: we conjecture that without the ability to parse speech into discrete perceptual events, learning to speak would be difficult, and it would not be possible to account for the compositional structure of speech which emerges in childhood.

In this paper we describe the perceptual foundations of a sensorimotor model of early childhood articulatory and phonetic development (Markey, 1993). The model features an auditory system that categorizes and recognizes speech sounds, an articulatory system that includes a realistic vocal tract model, and a central cognitive architecture that bridges the two. The environment in which the model resides is also simulated, including an adult speaker and objects to be referenced. Like an infant, the model's auditory perception is sensitive to prosodic and syllabic structure, organizes speech sounds syllabically (Jusczyk et al., 1993), and simulates the categorical phonetic perception of late infancy (e.g., Werker et al., 1981; Kuhl et al., 1992), although it does not attempt to explain the shift from acoustic to categorical phonetic discriminations (see, though, Jusczyk, 1993).

Importantly, the perceptual model relies on exclusively acoustic cues and their statistical distribution in the child's linguistic environment; it avoids prior assumptions of articulatory-acoustic correlations or linguistic contrasts. It is inappropriate to model perceptual development with features which assume prior knowledge of articulatory-acoustic correlations or semantic contrasts, knowledge the prelinguistic infant does not possess. Segments and categories the model detects are not the mature phonemes of adulthood; nor do they correspond to distinctive features (Chomsky & Halle, 1965), traditional articulatory features (Ladefoged, 1972), or any of myriad alternatives (e.g., Shillcock et al., 1992). The model's acoustic segments are longer in duration, at least long enough to capture coarticulation between contiguous consonants and vowels and to detect features of syllable structure. They correspond to coarse-grained changes in voicing, friction, and spectra. They identify the most salient spectral features of an utterance.

Although such segments and categories are based on exclusively acoustic measures, they correlate with linguistic events and delineate articulatory gestures. The model's auditory perception is specialized to segment and categorize acoustic feedback into discrete phonetic events which closely correspond to discrete gestures learned by the vocal tract's articulatory apparatus. To imitate words, the model need not solve the hard problem of relating continuous speech sound and continuous vocal tract motion. It learns the correspondence between one discrete sequence of events and another. The model's babble conforms to the linguistic environment by learning to match the simplest categories of sounds. Syllabic representations are stored in a lexicon and shape whole word pronunciation.

As children master the production of new subsyllabic sounds, they quickly generalize them to new lexical contexts, revealing a primitive compositional structure. Poorly mastered and erroneously pronounced sounds reveal the

same structure. Elemental articulatory skills seem to be acquired in the context of larger more abstract phonological patterns, but phonological competence may not be acquired or demonstrated without articulatory skills.

In order to resolve this dilemma, the model views speech as a hierarchical control problem. An abstract phonological level of control composes each utterance out of one or more elemental sounds, choosing which among lower level articulatory controllers is most likely to generate each sound. The discrete perceptual events of segmented auditory feedback thus regulate the timing and decisions of the abstract level of phonological control. With continuous proprioceptive feedback as a guide, the articulatory control level choreographs the exquisite timing of vocal tract and pulmonary motions necessary to produce each elemental sound. A non-hierarchical motor control strategy is possible, but demands more structure from the perceptual machinery. The hierarchical control strategy seems to offer a more parsimonious distribution of structure between articulatory and perceptual components and synergistic developmental strategies.

An acoustic-based phonetic representation

Auditory perception's *input* is an unsegmented acoustic representation of parental speech or feedback from the model's own speech. Its *output* is a phonetic representation of the sequence of acoustic segments and spectral categories detected in the utterance. We use "duck" as an example to introduce how the model segments and categorizes speech parcels. Figure 1a illustrates the model's continuous acoustic input as a schematic spectrogram of "duck". Its vertical axis is frequency (using the "Bark" scale), the horizontal axis is time, and dark patches represent high energy sound. The thick dark bands represent formants, the changing resonant frequencies of the vocal tract; the variously-shaded vertical area near 450 msec is a burst of noise. The horizontal bar near zero frequency represents the underlying vibration of the vocal cords.

A reasonable first step in parsing the unsegmented acoustic signal is to divide it into broadly classified periods of sound — continuous periods of silence, voicing, and friction (aspiration or frication). This approach yields a period of voicing between 40 and 330 msec, friction between 370 and 410 msec, and three periods of silence. This is plausible, given the broad segmentation of speech patterns by the auditory nerve's adaptation properties (Seneff, 1988), but it is clearly not sufficient to identify linguistically relevant spectral features.

To further divide the utterance, we consider locating temporally stable spectral patterns associated with the steady-state portion of vowels and other sonorants, fricatives, and aspiration. Consonants and diphthongs are the result of the vocal tract in motion, and hence are not associated with a single static acoustic pattern. Even the spectral pattern which accompanies some vocal tract closures, such as the nasal undertone which accompanies /m/, /n/, and /N/,[1] does not

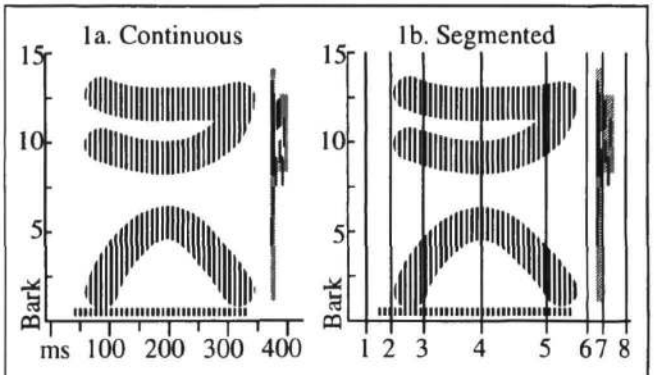

Figure 1: Schematic spectrogram of "duck" without (1a) and with (1b) segmentation.

Segments are numbered as follows: (1, 6, 8) Silence, (2) Prevoicing, (3, 5) Transitions [dA] and [Ak], (4) Static segment [A], and (7) Burst [k^h].

reveal the consonant's place-of-articulation.

Consequently, consonant identity can be determined only indirectly from *dynamic* spectral patterns. Points of maximum spectral change are the most salient portion of an utterance for consonant and syllable perception in adults (Furui, 1986; see also Lindblom & Studdert-Kennedy 1967) and are especially salient for young children (Nittrouer & Studdert-Kennedy, 1987; Nittrouer, 1992). Thus, a reasonable second step in parsing this signal is to identify periods of maximal and minimal spectral change.

By this method, there are eight acoustic segments in "duck". Their location in the utterance is portrayed in Figure 1b, which divides the time axis by segment rather than equal units of time. After an initial *silence* (segment 1), a period of *prevoicing* (segment 2: 50 msec) is detected during which vocal-cord vibration is audible, but the spectrum of the vocal tract's resonant sound structure is obscured by the /d/'s closure. Once a spectrum becomes apparent, it is scanned for the relative degree of spectral change. *Transitions* (3, 5 at 100 and 300 msec) are segments corresponding to maximal spectral change; formant slopes are greatest. The *static segment* (4: 200 msec) corresponds to a period of minimal spectral change; formants are relatively flat. After a period of silence (6: 350 msec) during the unvoiced /k/'s closure, a *burst* (7: 370 msec) of intense but rapidly decaying friction is detected when the contact of tongue body and velum is released. This is followed by a final period of silence (8).

Once segmented, the next step is to draw a sample static spectrum from each static segment and match it with prototype categories of steady-state sounds. Likewise, the model samples dynamic spectral properties during transition or burst segments and matches the sample against prototype

1. UNIBET (MacWhinney, 1991) instead of IPA phonetic symbols are used. Some UNIBET symbols which differ from IPA's include: N as in pi*ng*, T e*th*er, D ei*th*er, S *sh*oe, Z a*z*ure, I b*i*t, E b*e*t, & b*a*t, A b*u*t, U f*oo*t, O l*aw*, 6 *a*bove.

categories of dynamic sounds. The model learns an inventory of prototype categories for static and transition spectra from its linguistic environment. In the example above, let us assume that the model has already learned an inventory of prototype spectral categories. Then static spectral properties of segment 4 match the prototype static spectrum corresponding to the vowel [A]. The dynamic spectral properties of segments 3, 5, and 7 match transition spectrum prototypes corresponding to demisyllables [dA] and [Ak], and aspirated stop consonant release [k^h] respectively.

Form of the phonetic representation

This process generates information about the type, absolute order, and spectral properties of each acoustic segment. Other requirements of the complete sensorimotor model such as motor control, lexical access, and short term memory place additional constraints on the optimal form of the model's phonetic representation. They demand a representation whose size does not vary with the length of the utterance, which encodes relative order of acoustic segments, and which nonetheless faithfully captures important acoustic properties of the entire utterance and admits an accurate metric and efficient algorithm for determining the phonetic distance among utterances.

The model employs a phonetic feature vector with one unit for each feature. A unit's activation is increased each time the property it encodes is detected in an utterance, by a degree inversely related to the acoustic distance between speech token and feature prototype. *Spectral features* encode spectral properties and their relative order in the utterance. *Segmental features* encode acoustic segment properties and their relative order in the utterance.

Spectral features

Prototype categories of static and transition spectra provide the raw material for spectral features. Activations corresponding to prototypes of [dA], [A], [Ak], and [k^h] are increased as each segment is detected in "duck".

Transition and static segment categories implicitly encode spectral properties and their relative order. For example, in "duck" the [dA] segment encodes the spectral change which occurs as the tongue moves from an alveolar closure for /d/ or /t/ to the opening for the intended vowel /A/. Static segment [A] must follow transition segment [dA]. Segment [Ak] encodes the spectral change which occurs as the tongue makes contact with the velum for the /k/. It cannot precede [dA] except in a multi-syllabic utterance.

Segmental features

Acoustic segment classifications and a contextual encoding of segmental order are the ingredients of segmental features. Segment type is distinguished by five acoustic cues — voicing, friction, spectral change, spectral stability, or moments of relatively intense, quickly changing sounds. Eight segment types observed in English are described in Table 1. As "duck" is perceived, segment type activations are increased as each segment is encountered.

Table 1: Segment Types and Defining Acoustic Cues

Segment type	Voice	Friction	Transient	Stable	Unstable
Silence (0)	0	0	0	0	0
Prevoicing (P)	1	0	0	0	0
Static (S)	1	0	0	1	0
Transition (T)	1	0	1	0	0
Voiced friction (Z)	1	1	0	1	0
Voiceless friction (F)	0	1	0	1	0
Transition friction (H)	0	1	1	0	0
Burst (B)	0	1	1	0	1

The model encodes the relative, not absolute, order of segment types. It does so by forming a cluster of the three most recently detected acoustic segment types. This coding scheme is similar to Wickelgren's (1969) context-dependent allophone sequence encoding. Many segment type clusters encode important linguistic cues. Contiguous prevoicing and transition segments (PT or TP) indicate a voiced stop consonant in many contexts (e.g., the initial consonant in "duck"); a transition-silence-burst subsequence (T0B) signals a word-final unvoiced stop (e.g., the final consonant in "duck"); a transition-static-transition (TST) cluster usually signals a consonant-vowel-consonant syllable; and a static-transition-static (STS) cluster signals a diphthong.

Because of the role of the three-segment clusters in encoding the temporal patterns of voicing, friction, silence, and syllable structure, we call them "prosodic triads", where "prosodic" is used here in the Firthian sense as suprasegmental but subsyllabic and syllabic structural features.

Superimposed activations and phonetic distance

To summarize, there are four classes of phonetic features — acoustic segment types, prosodic triads, static spectrum categories, and transition spectrum categories. Static and transition categories are learned from the linguistic environment. The phonetic representation is an activity pattern over a vector of units, one for each feature. Activations accumulate until the end of an utterance. This superposition of contextual and global phonetic features is a faithful representation (Smolensky, 1990) of any one-syllable utterance.

Phonetic categorization and category learning

Phonetic features are learned. The model builds an inventory of prototype acoustic segments that correspond to phonetic feature categories. They represent those transition and static spectra recognized as relatively distinct according to an acoustic distance measure and statistically important according to their relative frequency in parental speech.

Each time an audible segment is detected, auditory perception measures the acoustic distance between the token's

spectral properties and the corresponding inventory of spectral prototypes, activating each by an amount inversely related to acoustic distance. New prototype categories are learned by a simple competitive process (Grossberg, 1976; Carpenter & Grossberg, 1987). If a token segment is close to existing prototypes, winning prototypes are updated. Otherwise, a new prototype is added probabilistically, using the current token as the new prototype. Such probability is proportional to the token's distance from its nearest neighbor, but is low enough to discourage a proliferation of gratuitous categories.

Relationship between acoustic segments, gestures, and linguistic units

Transitions usually signify demisyllables, diphthongs, or glides. Static segments signify the stable portion of vowels, nasals, approximants, and fricatives. Grounding the sensorimotor model is a one-to-one correspondence between phonetic segments and articulatory gestures — its fundamental perceptual and articulatory building blocks. Several investigators have proposed that humans *produce* speech as a sequence of articulatory gestures as from a musical score (Browman & Goldstein, 1989; Saltzman & Munhall, 1989; Kelso et al., 1986). Moreover, Fowler (1991) argues that humans *perceive* speech as gestures instead of as phonemes.

The model's mechanism for gesture perception is its detection of spectral transitions. This is based on the observation that transitions correspond approximately to the zenith of articulatory gestures and that static segments correspond to the ends of gestures. It is also motivated by research showing that spectral transitions are essential for consonant and syllable perception in adults (Furui, 1986) and young children (Nittrouer, 1992), as well as by the performance of automatic speech recognition systems which use dynamic spectral data (e.g., Lee, 1990). Moreover, distinctive nonlinear acoustic changes usually occur between a gesture's terminal configurations (Stevens, 1989), enhancing spectral transitions and ensuring their reliability as linguistic codes.

Perceptual Model Implementation and Simulations

We have implemented the perceptual model and are currently integrating it into the full sensorimotor model. Here we report results showing the properties of the transition and static spectra, the distinctiveness and faithfulness of its demisyllabic encoding. The results indirectly indicate the reliability of the segmentation algorithm.

Stimuli are synthetic consonant-vowel (CV) syllables generated by an adaptation of Haskins Laboratories' ASY articulatory synthesizer (Rubin et al., 1981). The synthesizer employs 6 vocal tract articulators representing a total of 10 degrees of freedom to specify the vocal tract's configuration. Voicing and frication is modeled by a simple mechanical-acoustical model of respiration. Low level dynamics of the articulators are simulated as simple gestures which resemble the motion of a critically damped spring of a given stiffness from some displacement to its equilibrium point. A script language specifies a queue of gestures and proprioceptive triggering conditions (an intrinsic timing device) necessary to produce an utterance. Each stimulus employs one of three stop consonants (b, d, g) and one of ten target vowels. Optimal vocal tract configurations necessary to accurately render each target vowel are determined, and gesture script templates are designed and fine-tuned by the experimenter for bV, dV and gV frames. Selecting target vowel and consonant

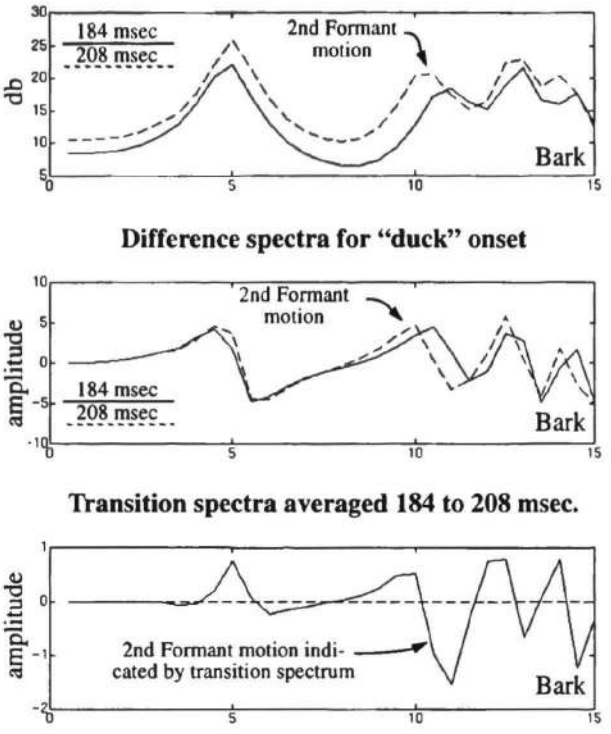

Figure 2: Static and transition spectra for onset demisyllable of "duck"

Steps in spectral analysis are shown above. The first step is conversion from Hertz to Bark frequency and Gaussian smoothing. The peak near the 10 bark detector is the 2nd formant, dropping in frequency between the two sample times of 184 and 208 msec. The difference spectrum effectively normalizes the spectrum for changes in amplitude. The 2nd formant frequency shift is evident

The 2nd formant's change is captured in the transition spectrum during the period of maximum spectral change. The positive value at 10 bark reflects the formant's drop in frequency and the increased amplitude of the difference spectrum at that frequency. The negative value at 11 bark occurs behind the formant's motion, where the difference spectrum has fallen in amplitude.

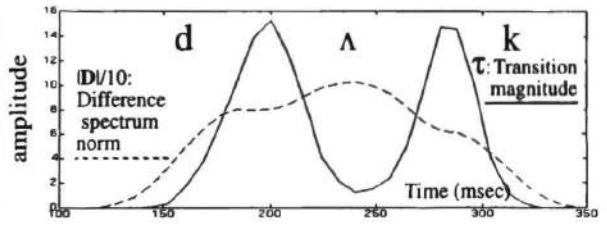

Figure 3: Transition magnitude for duration of "duck"

The transition magnitude τ (solid line) and the norm of the difference spectra (dashed line) are shown for the entire utterance in the graph above. Also shown are the approximate locations of linguistic segments. Spectra in figure 2 correspond to the first transition peak between /d/ and /A/.

at random, we generate 1,350 syllable tokens — about 45 tokens per syllable type.

To generate each token, Gaussian noise is added to each vowel's articulatory parameters. There is no way to determine if each resulting sound actually corresponds to the intended syllable type or even whether it is a legal sound of English without actually listening to it. The experimenter transcribes each token, rejecting any non-English sound and rejecting any sound whose phonetic transcription does not agree with the intended type. The 892 remaining tokens are divided into training, validation, and test sets for supervised classification tasks.

Sounds are sampled once every 8 msec for periodic sound (voicing) amplitude, aperiodic sound (frication or aspiration) amplitude, and 256-frequency power spectrum. Spectral analysis converts the power spectrum from Hertz to Bark, performs Gaussian smoothing over time, and normalizes for total amplitude by computing a difference spectrum. The momentary transition spectrum is the first derivative of the difference spectrum with respect to time. The transition magnitude is the L_1 norm of the transition spectrum minus the L_1 norm of the difference spectrum. The peak transition spectrum is sampled as an average of the momentary transition spectra for a 50 msec period centered at the point when the transition magnitude is at a local maximum. An example is presented in Figures 2 and 3, above. Further details are available in Markey (1993).

We then measure phonetic distance $\mathcal{D}$ among all syllable stimuli. This is determined as a function of the linear correlation ρ between each pair of token transition spectra $\mathcal{D}(\mathbf{T}^i, \mathbf{T}^j) = 1.0 - \rho(\mathbf{T}^i, \mathbf{T}^j)$ (e.g., Pomerleau, 1993). We use this distance measure here and as the basis for unsupervised classification of parental speech by the full sensorimotor model because it factors out irrelevant differences in scale between two spectra.

As a further test of the encoding's faithfulness, a multilayer back propagation network is trained to classify syllables by eight features for vowel quality and consonantal place-of-articulation using transition spectra as input. The same test is performed with both transition and difference spectra as input, since acoustic information is lost by the transition spectrum. As a control, we train a time-delay neural network (TDNN; Waibel et al., 1989) on the same task, using as input smoothed bark spectra over the entire duration of the consonant-vowel transition (averaging 168 msec). Each network is trained until a validation dataset indicates overtraining and then is tested using separate test dataset.

Results

The average phonetic distance between pairs of syllables of the same type is 0.211. The average distance between pairs of syllables of different types is 0.955. Thus, tokens of the same type appear to be relatively clustered in the representational space. A tabulation of average cross-distances for all bV syllable types appears in Table 2.

Table 2: Mean phonetic distance measured between bV syllable tokens

	bA	ba	bo	bO	bu	bU	b&	bE	bi	bI
bA	0.18	0.89	0.92	0.95	1.06	0.53	1.19	0.80	1.25	1.09
ba		0.33	0.88	0.72	0.94	0.92	0.81	0.96	1.15	0.93
bo			0.39	0.85	1.05	0.93	0.79	0.75	1.02	0.86
bO				0.13	0.96	1.03	0.86	0.91	1.29	1.12
bu					0.62	1.18	0.99	1.01	0.81	0.77
bU						0.33	1.23	0.79	0.21	1.21
b&							0.08	0.65	0.98	0.85
bE								0.13	1.06	0.63
bi									0.12	0.80
bI										0.09

Syllable types with the same vowel type but with different consonants (gE, dE, bE) are less distant with respect to each other (see Table 3) than syllables with the same consonant but different vowels (bE, bi, bo, b&, etc., below). Several tokens in the simulation corpus (especially Co and CO syllables) are not phonetically distinct. These tokens are also difficult for the experimenter to transcribe phonetically.

Table 3: Additional Phonetic Distance Comparisons

Same Vowel				Hard Discrimination			
	bE	dE	gE		bo	do	go
bE	0.13	0.55	0.50	bo	0.39	0.52	0.39
dE		0.08	0.57	do		0.43	0.39
gE			0.10	go			0.27

Supervised classification tests also suggest that the transition spectrum is a robust demisyllable representation. Syllable classification errors by the network which uses only the transition spectral coefficients as inputs is 5.29%. Error rates are substantially lower when both transition and difference spectra coefficients are used as inputs, dropping to 0.85% when both transition and difference spectra coefficients are rescaled to range between 0 and 1.

Table 4: Supervised Training Error Rates

Method and Inputs	% Error
Backpropagation, 44 Transition spectrum coefficients as inputs.	5.29%
Backpropagation, 44 Transition and 44 Difference spectrum coefficients.	1.77%
Backpropagation, 44 Transition and 44 Difference spectra coefficients rescaled.	0.85%
TDNN, 44 Bark spectrum coefficients for each of 21-8 msec frames, 3-frame input window.	1.59%

The TDNN network's performance is similar. This is not surprising, as it must discover those static and dynamic features necessary to classify each syllable (Waibel et al., 1989).

Typical errors for all methods include single mistaken features, features just under threshold, or confusion in deciding the place-of-articulation.

Supervised classification results are encouraging for the use of transition data for segmentation and syllable classification without needing to learn segmentation strategies first. However, results point out that both dynamic and static spectral information is essential for accurate classification.

Although the sensorimotor model and these simulations employ synthetic speech, it may be possible and desirable to extend the method to automatic human speech recognition.

Acknowledgments

This research is supported by NSF Presidential Young Investigator award IRI-9058450 and grant 90-12 from the James S. McDonnell Foundation to Michael C. Mozer. We thank Haskins Laboratories for making available its ASY software, and Troy Sandblom for his rewriting it for C and Unix. I acknowledge the support, guidance and coaching by Mike Mozer and Lise Menn. I also thank Alan Bell for his assistance with acoustic phonetics, and anonymous reviewers for their helpful comments.

References

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6,* 102-151.

Carpenter, G.A. and Grossberg, S. (1987). ART2: self-organization of stable category recognition codes for analog input patterns. *Applied Optics, 26,* 4919-4930.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* New York: Harper & Row.

Fowler, C.A. (1991). The perception of phonetic gestures. In I. G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech.* Hillsdale, NJ: Lawrence Erlbaum

Furui, S. (1986). On the role of spectral transition for speech perception. *J. Acoustical Soc. America, 80 (4),* 1016-1025.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics, 23,* 187-202.

Jusczyk, P.W. (1993). From general to language-specific capacities: the WRAPSA Model of how speech perception develops. *J. Phonetics, 21,* 3-28.

Jusczyk, P.W., Jusczyk, A.M., Kennedy, L.J., Schomberg, T. and Koenig, N. (1993). Young infants' retention of information about bisyllabic utterances. Submitted.

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

Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606-608.

Ladefoged, P. (1982). *A Course in Phonetics.* 2nd Ed. Harcourt Brace Jovanovich, New York.

Lee, K.-F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 38,* 599-609.

Lindblom, B.E.F. and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *J. Acoustical Soc. America, 42,* 830-843.

MacWhinney, B. (1991). *The CHILDES Project.* Hillsdale, NJ: Erlbaum.

Markey, K.L. (1993). A sensorimotor model of early childhood phonological development. Thesis Proposal. Technical Report CU-CS-695-93. Boulder, CO: University of Colorado, Department of Computer Science.

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

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *J. Phonetics, 20,* 351-382.

Pomerleau, D.A. (1993). Input reconstruction reliability estimation. In Hanson, S.J. et al. (Eds.), *Neural Information Processing Systems, 5,* 279-286.

Rubin, P., Baer, T. and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *J. Acoustical Soc. America, 70 (2),* 321-328.

Saltzman, E.L. and Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1 (4),* 333-382.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *J. Phonetics, 16,* 55-76.

Shillcock, R., Lindsey, G., Levy J. and Chater, N. (1992). A phonologically motivated input representation for the modeling of auditory word perception in continuous speech. In *Proc. Fourteenth Annual Conference of the Cognitive Science Society* (pp. 408-413). Hillsdale, NJ: Erlbaum.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46,* 159-216.

Stevens, K.N. (1989). On the quantal nature of speech. *J. Phonetics, 17,* 3-45.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-37,* 328-393.

Werker, Janet F., J.H.V. Gilbert, K. Humphrey, and R.C. Tees. (1981). Developmental aspects of cross-language speech perception. *Child Development, 52,* 349-355.

Wickelgren, W. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76,* 1-15.