Available online at www.sciencedirect.com
SCIENCE DIRECT®
Infant Behavior & Development 27 (2004) 477–498
Infant Behavior & Development

Phonetic variability and prosodic structure in mothers

Melissa A. Redford*, Barbara L. Davis, Risto Miikkulainen

The University of Texas at Austin, TX, USA

Received 20 November 2003; received in revised form 10 May 2004; accepted 14 May 2004

Abstract

Spontaneous speech acoustics are highly variable. Such variability may be problematic for infants relying on phonological form to solve the segmentation problem. In the present study, acoustic measures of vowel duration and a computer model of speech segmentation were used to evaluate the problem of phonetic variability for a rhythm-based speech segmentation strategy. The specific questions under study were (1) whether or not mothers realized disyllabic vowel duration patterns consistently in spontaneous infant-directed speech, and (2) whether or not these patterns were distinctive enough in the context of an utterance to provide a useful cue for speech segmentation. Data from four English-speaking mothers indicated that the trochaic-like duration pattern may interact with phrase-position and with grammatical category, but when the resulting patterns are consistent, they provide useful segmentation cues for spontaneous infant-directed speech.

Keywords: Infant-directed speech; Speech segmentation; Phonetic variability; Acoustic measures; Computational model

When we listen to a language we have never heard before, we experience something like what the infant must experience when first confronting language—mystification. We know that there must be meaningful elements in the continuous stream of speech, but they are impossible to isolate. This is the segmentation problem, one of the major problems that infants must solve before they can understand and use language. A number of signal-based solutions have been proposed to solve this problem, but even the most successful proposals have not been evaluated against the inherent variability of natural speech acoustics. The present study provides the first such evaluation for a rhythm-based segmentation strategy. Specifically,

* Corresponding author. Present address: Linguistics Department, 1290 University of Oregon, Eugene, OR 97403-1290, USA. Tel.: +1 541 346-3789.
E-mail address: redford@darkwing.uoregon.edu (M.A. Redford).

0163-6383/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.infbeh.2004.05.001

we examined the constancy and distinctiveness of one cue to lexical stress—vowel duration—as a function of phrase position and grammatical category in the disyllabic words of spontaneous infant-directed speech.

1. Signal-based solutions to the segmentation problem

Signal-based solutions to the word segmentation problem have focused primarily on sound sequence regularities and prosodic structure. Computational studies suggest that syllable boundaries, which may often correspond to word boundaries, can be identified from language phonotactics, i.e., from the language-specific set of sound sequences at word edges (Martens, Daelemans, Gillis, & Taelman, 2002; Vroomen, van der Bosch, & de Gelder, 1998). Saffran, Aslin, and Newport (1996) have shown that 8-month-olds can recognize syllables in disyllabic nonsense words based solely on the transitional probabilities between adjacent syllables, i.e., the likelihood that certain sound sequences go together in previously heard sequences. Brent and Cartwright (1996) evaluated whether these ordering regularities could be used to segment words in natural infant-directed speech. Using an optimization algorithm and broadly-transcribed infant-directed speech input, they found that language phonotactics and cross-syllabic sound sequence regularities provided independent sources of information that could be used by the infant to segment natural language.

Although the evidence suggests that sound sequence regularities can be used to locate word boundaries in continuous speech, infants may attend to different aspects of the signal early in the segmentation process. Morgan and Saffran (1995) tested 6- and 9-month-old infants on their ability to group syllables into word-like units based on rhythm—encoded as a vowel duration pattern—and on the transitional probabilities between syllables. They found that 9-month-olds used both cues to identify word-like units, but that 6-month-olds were only sensitive to the rhythmic cue. Johnson and Jusczyk (2001) evaluated prosodic patterns of rhythm and coarticulation relative to sound sequence regularities in 8-month-old infants, and found that the prosodic suprasegmental patterns outweighed sound sequence regularities as cues to word boundaries for these infants. These studies suggest that infants may begin to solve the word segmentation problem by attending to ambient prosodic patterns before incorporating information about sound sequence regularities into the segmentation task.

Stress has received the most attention as a cue to word segmentation. It has been repeatedly suggested that infants attend to higher-level rhythm patterns to aid in word boundary identification (Gleitman & Wanner, 1982; Gleitman, Gleitman, Landau, & Wanner, 1988; Echols, 1993; Jusczyk, Cutler, & Redanz, 1993). This hypothesis was first proposed to explain the production biases observed in infants' first words, but it is also consistent with hypotheses about adult speech segmentation (e.g., Cutler, 1994), and has received support from infant speech perception experiments as well (e.g., Johnson & Jusczyk, 2001; Morgan & Saffran, 1995).

The production data indicate that English-learning infants mainly produce the stressed and word-final syllables of multisyllabic content words (Echols & Newport, 1992). Word-medial, unstressed syllables are very often omitted (Ingram, 1978; Klein, 1981). Echols (1996) has suggested that this bias arises from an attentional bias for stressed and word-final syllables in the continuous speech stream. These syllables attract attention because they are longer in duration than other syllables, especially in infant-directed speech (Albin & Echols, 1996), and so are proposed to be more salient. According to Echols (1996), infants' attention to stressed and word-final syllables allows them to extract a rhythmic word template

because together the syllables describe a strong-weak or trochaic stress pattern, which characterizes most familiar English words, e.g., "baby," "bottle," "diaper" (Cutler & Carter, 1987). Once infants extract the trochaic word template they can apply it to continuous speech to identify word boundaries (Cutler, 1994, 1996; Echols, 1996). In this way, infants solve the segmentation problem using a rhythm-based segmentation strategy.

Echols' (1996) developmental account of how infants come to extract a rhythmic word template and apply it to speech segmentation is valuable for tracing the way in which infants might use the signal to identify word-level stress patterns. One problem with the account, though, is that it relies on a phonetic factor that is not uniformly distributed across the signal.

Specifically, the claim that word-final syllables are more salient because they are longer may only be true for words that occur utterance-finally. In phrase-final position, final syllables undergo phrase-final lengthening, a phonetic effect that appears to be emphasized in infant-directed speech (Bernstein Ratner, 1986; Morgan, 1996). Albin and Echols (1996) acknowledged this issue—most of their multisyllabic word tokens were in utterance-final position—and countered by noting that some word-final lengthening was present in utterance-medial position. They also noted, however, that the effect was much more limited utterance-medially. This non-uniform distribution of word-final lengthening raises the problem of how a rhythmic word template could be extracted in attending to a cue that varies with phrase position.

A second and related problem emerges from the first. If the rhythmic form that aids in speech segmentation is a trochaic pattern, then how can similarities between stressed and word-final syllables help the infant extract the appropriate pattern? Albin and Echols (1996) anticipated this problem, and suggested that since stressed and word-final syllables are different in other respects, infants should be able to distinguish between the two, assuming that they are attending to multiple acoustic cues. However, it is equally possible that any acoustic similarities between word-final and stressed syllables due to phonetic processes would disrupt and/or conceal the phonologically-specified form of the word.

The strength of Echols' (1996) account is that it considers the influence of natural speech acoustics on the development of a rhythmic word template, making it possible to evaluate her proposal for spontaneous speech. Other research that supports a rhythm-based strategy of speech segmentation demonstrates that English-learning infants can group syllables together based on a trochaic stress pattern. But the research does not ask how infants come to extract this pattern as a word template from a variable signal, or whether a rhythm-based segmentation strategy would operate in spontaneous speech where phonetic effects may disrupt or conceal the phonologically-specified stress pattern. In the present study, we begin to address these questions by evaluating whether there is enough regularity in the signal to extract a rhythmic template, and whether or not such a template could be used to segment running speech.

2. The problem of phonetic variability

Laboratory stimuli must be controlled. As a result, the problem of phonetic variation is neutralized. However, phonetic variability may result in two distinct cue-related problems, which would interfere with infants' ability to use a rhythm-based strategy to segment actual speech. We refer to these as problems of (1) cue constancy and (2) cue distinctiveness.

2.1. Cue constancy

In a critique of the prosodic "boot-strapping" hypotheses, Fernald and McRoberts (1996) argued that a phonetic cue must be consistently associated with the same linguistic structure in order for a rhythm-based segmentation strategy to work. Although they were concerned with the validity of such a strategy for syntactic parsing, their critique is also relevant for word segmentation. In order for a rhythm-based segmentation strategy to be valid for word segmentation, speakers must consistently realize phonological stress on individual words.

In English, stress is cued by vowel duration, pitch, and amplitude (Lehiste, 1970). All of these cues vary with factors other than lexical stress. For instance, a trochaically-stressed word should be realized with a long-short vowel duration pattern; however, this pattern interacts with other phonetic factors that influence vowel duration, such as intrinsic duration, consonantal context, or phrase position (e.g., Peterson & Lehiste, 1960). Thus, the lexically-specified long-short vowel duration pattern may in fact be realized as a short-short or a long-long pattern, depending on the phonetic factors interacting with stress. If the lexically-specified pattern is not consistently realized on disyllables, then it does not serve as a constant cue to speech segmentation. In this way, phonetic variability may result in a problem of cue constancy for prosodic theories of word-level segmentation.

2.2. Cue distinctiveness

Even if lexically-specified stress patterns achieve cue constancy they still may not be distinctive in the context of an utterance, and so would not provide a useful cue for speech segmentation. Take again the example of the lexically-specified long-short vowel duration pattern. If we consider this pattern in the context of additional vowel duration patterns—conditioned by morphological (Swanson, Leonard, & Gandour, 1992) and pragmatic factors (Oller, 1973) in addition to the previously mentioned phonetic factors—it is not immediately obvious how the lexically-specified pattern could stand out in an utterance. A trochaic-like, long-short vowel duration pattern could be lost in the context of connected speech, which is replete with long and short vowel alternations that have little to do with lexical stress. To illustrate, note that the likely vowel duration pattern for the utterance "look at these in daylight" is long-short-long-short-long-short, with only the last long-short alternation signaling the presence of a disyllable. How can an infant know which long-short alternation is relevant for distinguishing word boundaries, and which long-short alternation is due simply to phonetic variability?

Up to this point, we have argued against the possibility that phonological form can be identified in the signal, given the inherent variability of speech acoustics. The argument is based on what is typically true of adult-directed speech in most contexts. However, normal phonetic variability might be overcome in some contexts, when speakers are intent on realizing phonological form. For instance, Lindblom (1990, 1996) has demonstrated that speakers change the acoustic structure of the same word or phrase depending on their perception of the listeners' needs. When speakers perceive that the listeners need help in understanding their speech, they hyper-articulate, more fully realizing phonological forms. When speakers perceive no such need in the listeners, they hypo-articulate, reducing and obscuring phonological forms. If infants are perceived as weak listeners, speakers may be more prone to hyper-articulate, and so phonological forms may be better realized in infant-directed speech than in adult-directed speech (Snow, 1972).

It is also highly probable that certain words will be better realized than others in infant-directed speech. For example, Swanson et al. (1992) found that the vowel durations of content words were greater in infant-directed than in adult-directed speech, but the vowel durations of function words did not differ. They concluded that mothers emphasize meaningful words even more when speaking to infants than when speaking to adults. Function words, which are semantically light, are de-emphasized in both registers.

To summarize, phonetic variability may disrupt and/or conceal lexical stress in infant-directed speech if it is not countered by the speaker. If phonological forms are not realized consistently and distinctively in the signal, it becomes difficult to imagine how infants could use rhythm or any other aspect of the signal to solve the word segmentation problem.

3. The current study

This study evaluated the problem of phonetic variability for rhythm-based speech segmentation by studying the vowel duration patterns of disyllabic words in spontaneous, infant-directed speech. We focused on vowel duration because laboratory work has shown that infants can segment disyllabic words from a continuous stream of syllables using just this cue (e.g., Morgan & Saffran, 1995; Morgan, 1996). It is therefore relevant to ask whether infants could also use this cue to segment natural speech or whether phonetic variability in spontaneous speech input obviates its usefulness.

It is worth noting that even if phonetic variability conceals rhythm as encoded by vowel duration patterns, we cannot safely reject a rhythm-based solution to the segmentation problem since other phonetic cues to lexical stress would also need to be evaluated. However, if vowel duration patterns are found to encode lexical stress in spite of phonetic variation, then we can conclude that mothers provide direct information about phonological form in the signal. This information could be used by infants to identify words in spontaneous infant-directed speech.

The study was designed specifically to evaluate the constancy and distinctiveness of disyllabic vowel duration patterns. Cue constancy was defined as whether or not the same vowel duration pattern consistently reflected the phonologically-specified trochaic stress pattern. Cue distinctiveness was defined as whether or not a consistent vowel duration pattern stood out from other vowel duration patterns at the phrase level.

To evaluate the question of cue constancy we measured vowel durations of disyllabic words in spontaneous infant-directed speech in four English-speaking mothers. Mothers were recorded while interacting with their 6-month-old infants. At this age, infants should be able to segment words in a fluent speech stream based solely on a long-short vowel duration pattern (e.g., Morgan & Saffran, 1995). Our question was whether this trochaic-like pattern was consistently realized in these mothers' speech, i.e., whether or not vowel duration in the first syllable of disyllabic words was significantly greater than vowel duration in the second.

To evaluate the question of cue distinctiveness we measured segment durations across phrases containing disyllabic words. These measurements formed the input to a computational model, which extracts statistical regularities across time-varying input. The model, called a Simple Recurrent Network (SRN—Elman, 1990), was trained to identify syllable, word, and phrase boundaries given segment durations as input. The rationale was that if the disyllabic vowel duration patterns were

distinctive in the context of the other duration patterns across a phrase, then the SRN would be able to learn the word boundaries separately from syllable and phrase boundaries. If the disyllabic vowel duration patterns were not distinctive in context, then the SRN would not be able to distinguish between different boundary types. The SRN was used in preference to analytic, statistical methods because of the extreme high-dimensionality of the distinctiveness problem. Further, since the SRN was originally proposed by Elman (1990), it has become the computational method of choice to examine questions of language learnability (e.g., Aslin, Woodward, LaMendola, & Bever, 1996; Christiansen, Allen, & Seidenberg, 1998; Martens et al., 2002; Vroomen et al., 1998).

In both analyses, vowel duration patterns were analyzed as a function of phrase position and grammatical category. The focus on phrase position was chosen because it provides a well-known source of phonetic variability in vowel duration (Klatt, 1976; Oller, 1973). Such a focus is also interesting theoretically because a number of researchers have identified phrase-final position as a privileged position for speech segmentation. Echols (1996) considered the effect that phrase-final lengthening might have in making word-final syllables salient. In addition, a number of other researchers have argued that words in phrase-final position may be easier to segment because they are bounded on one side by a natural boundary (the pause at the end of an utterance) (Aslin, 1993; Aslin et al., 1996; Brent & Cartwright, 1996; Christiansen et al., 1998). Aslin (1993) and Aslin et al. (1996) have also argued that final position may make the recurring sound pattern of the word easier to remember (the recency effect). In this study, we evaluated the potential advantage of phrase-final position against the potential disadvantage of a disrupted trochaic-like long-short pattern.

The focus on grammatical category was chosen under the assumption that speakers control the extent to which phonological form is phonetically realized, as in the acoustic differences between clear and casual speech registers (Lindblom, 1990, 1996). It is possible that overall infant-directed speech is less variable than adult-directed speech because infants are perceived as weak listeners. It is also possible that infant-directed speech emphasizes lexical stress more on words deemed more important within the context than on those that are not in focus (e.g., Swanson et al., 1992). In particular, it might be expected that English-speaking mothers will mark lexical stress more prominently on nouns than verbs or other parts of speech, since English-speaking mothers focus nouns in infant-directed speech (Fernald & Morikawa, 1993; Choi & Gopnik, 1995).

Because cue constancy and cue distinctiveness were evaluated using different methodologies, we present methods and results pertaining to each of these in separate sections. The results from the evaluation of cue constancy are presented first.

4. Part 1: cue constancy

Disyllabic vowel duration patterns were evaluated to determine whether mothers consistently marked disyllables with lexically correct trochaic stress when speaking to their infants. Interactions between vowel duration patterns and phrase position, as well as grammatical category were analyzed. We predicted that phrase-final lengthening would disrupt the trochaic long-short pattern. We further predicted that the typical focus on objects rather than action would lead mothers to realize the citation stress pattern on nouns more consistently than on other grammatical categories.

Table 1
Monosyllabic, disyllabic, and polysyllabic words are shown as a percentage of the overall number of words in each speech sample

	Syllables per word		
	1	2	>2
Mother 1	86.27%	13.10%	0.63%
Mother 2	85.34%	13.68%	0.98%
Mother 3	87.57%	11.42%	1.01%
Mother 4	81.56%	17.46%	0.98%

5. Method

5.1. Speakers

Four American-English speaking mothers were recorded while interacting with their 6-month-old infant sons and daughters.[1] The recordings took place in an acoustically-treated experimental room filled with many play-things during two 30-minute sessions occurring one week apart. Both the mother and infant wore vests with small microphones connected to FM transmitters. Signals were recorded onto separate HiFi audio channels (for details, see Kehoe, Stoel-Gammon, Buder, 1995). Since the recordings were collected for studying infant vocal development, the mothers' spontaneous, infant-directed speech was as natural and as unguarded as is possible in an experimental situation. However, the mothers' speech was goal-directed in that it was meant to elicit spontaneous vocalization from their infants. If an infant was particularly quiet, the mother was encouraged to try to elicit imitative vocalizations.

5.2. Speech sample

Each mother's infant-directed speech on the hour-long tapes was transcribed by two listeners. Utterance boundaries were determined according to perceived starting and stopping points indicated by silent pauses, not according to grammatical criteria. Individual words in the sample were coded as monosyllables, disyllables or polysyllables and by initial, medial, or final utterance position, in isolation, or as an iterative series. Overall, 85.15% of the words were monosyllabic, 13.95% were disyllabic, and very few words were polysyllabic (0.09%). The absolute percentages differed somewhat from speaker to speaker, but each sample consisted primarily of monosyllabic words, with some disyllabic words, and very few polysyllabic words (Table 1). The majority of disyllabic words were produced in utterance-medial (42.32%) and utterance-final (39.32%) position.

Intertranscriber reliability on these data was extremely high. Transcribers disagreed on the perceived number of syllables on less than 1% of the total number of words.

[1] The mother–infant dyad recordings used in the current study were a subset of those collected by Carol Stoel-Gammon and colleagues at the University of Washington for the purpose of investigating the role of maternal input in children's acquisition of intrinsic and extrinsic vowel length.

Table 2
The different disyllabic types and tokens are presented by grammatical category and number of medial/final pairs

Nouns	Pairs	Verbs	Pairs	Others	Pairs
Baby	8	Doing	3	Better	2
Bucket	1	Going	3	Diff'rent	1
Doggy	1	Happen	3	Funny	1
Honey	1	Happen'd	1	Fussy	1
Mama	1	Needed	1	Over	4
Minute	1	Saying	1	Pretty	1
Mister	1	Talking	3	Sleepy	1
Morning	1	Tickle	1		
Paper	1				
People	1				
Puppet	2				
Puppy	3				

All matching phrase-medial and phrase-final disyllables were selected for acoustic measurement. These were matched by both speaker and word type. For example, the disyllable *puppet* was selected for measurement when one mother produced the following phrases:

(1) You want your *puppet* friend?
(2) What happened to the *puppet*?

Sixty-one matching pairs were found. Nine were discarded either because the recording was too noisy, or because a word perceived to occur in utterance-final position did not precede a sufficiently long silent pause (only pauses 260 ms and greater in duration were taken to signal an utterance boundary (Fisher & Tokura, 1996)). Three other pairs (two adverbial and one interjective) were eliminated because, according to dictionary pronunciations, the medial/final disyllable stress pattern was iambic (weak-strong), and the set was meant to include only trochaically-stressed disyllables. Of the remaining 49 pairs (98 disyllables) there were 22 nouns, 16 verbs, and 11 "other", including adverbs, adjectives, and prepositions. The higher frequency of nouns, compared to verbs or other grammatical categories, was consistent with the reported noun-bias in English infant-directed speech (Fernald & Morikawa, 1993; Choi & Gopnik, 1995). Table 2 lists disyllabic types spoken by the same speaker that occurred in phrase-medial and phrase-final position.

5.3. Duration measurements

Vowel durations were measured using Kay Elemetric Computerized Speech Laboratory (CSL) software. The entire utterance was entered in analog form from a Marantz Portable Cassette Recorder (PMD201) and digitized by the Kay external module (Model 4300) at a rate of 10,000 samples per second. The utterance was displayed concurrently in two views: as an oscillogram, and as a broadband spectrogram. Once the disyllable was located, vowel durations were measured. One cursor was positioned at the onset of the vowel, the other at the offset. The time of onset was automatically subtracted from the time of offset to yield duration in seconds. For the most part, vowel boundaries were established at the edges of obstruent consonant boundaries, and so were identified by an abrupt onset/offset of periodicity

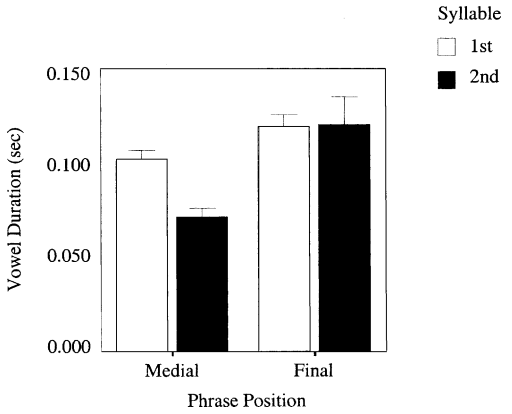

Fig. 1. Vowel duration in medial and final phrase position. The long-short pattern is trochaic-like. The long-long pattern is not.

and/or amplitude. If a vowel was bounded on one or both sides by a sonorant, then the boundary was identified by a change in the frequency characteristics and amplitude of the waveform. If a vowel was bounded on one or both sides by another vowel and the two vowels did not form a diphthong, boundaries were marked either by a very short pause (e.g., 24 ms) or by an abrupt change in formant structure. Such situations occurred when the final vowel of a syllable preceded a syllable with no consonantal onset and both vowels were stressed. For example, in the phrase "Do we turn you over?" the /iu/ of *you* and /oa/ of *over* are both stressed and the vowels are the nuclei of two different syllables. Decisions on boundary location were confirmed with auditory playback. On a separate occasion, 15% of the data were randomly selected and remeasured to assess reliability. A two-tailed *t*-test indicated that there were no systematic differences between the old and new measures for this subset of data. The average vowel durations in the two measurement sets were very similar (106.2 ms versus 109.2 ms) and the spreads were similarly large (SD = 68.9 and 72.7, respectively). On average the old and new duration measures differed by plus or minus 9.26 ms (SD = 10.7).

6. Results and discussion

The disyllabic vowel durations were analyzed in a 4-way univariate, analysis of variances (ANOVA). There were three fixed factors and one random factor. The three fixed factors were: grammatical category (noun, verb, other); disyllabic word position within the phrase (medial or final); and, vowel position within the disyllable (first syllable, second syllable). Individual disyllabic tokens were nested within speakers, and treated as random factors. Degrees of freedom were calculated using the Satterthwaite Method, appropriate to a mixed model. Results indicated a significant main effect for Phrase Position [$F(1,1.37) = 73.55$, $p < 0.05$], and significant two-way interactions between Syllable and Phrase Position [$F(1,8.08) = 5.81$, $p < 0.05$] and Syllable Position and Grammatical Category [$F(2,4.2) = 6.95$, $p < 0.05$]. Importantly, the main effect for Speaker and the higher order interactions with Speaker were not statistically significant, indicating that the pattern of results was similar across the four speakers.

The interaction between syllable and phrase position conformed to our expectation (Fig. 1). The vowel duration pattern within a disyllable was trochaic-like when it occurred in phrase-medial position, but was

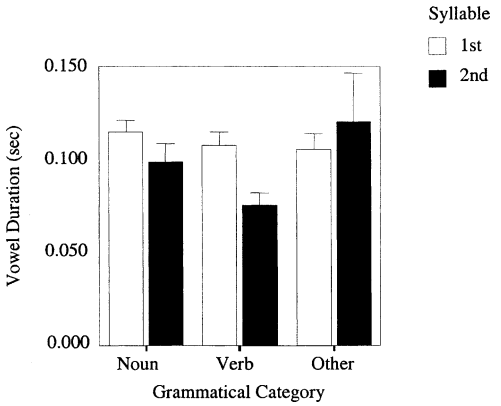

Fig. 2. The vowel duration pattern within noun, verb and other grammatical categories. Disyllabic verbs exhibit the most trochaic-like pattern of vowel duration.

transformed by phrase-final lengthening at the end of a phrase. In phrase-final position, the vowel of the second syllable (V2) was lengthened and the vowel duration pattern became long-long instead of long-short. The difference between V2 duration in phrase-medial and phrase-final position was substantial—on the order of 50 ms (phrase-medially $M = 71.86$ ms, phrase-finally $M = 121$ ms). The variance in durations was also greater in phrase-final position (phrase-medially $\sigma^2 = 1$, phrase-finally $\sigma^2 = 10$), suggesting that the phrase-final long-long pattern was less consistent than the trochaic-like phrase-medial pattern.

The duration of the first vowel (V1) also differed as a function of Phrase Position. Phrase-medial V1 durations were shorter than phrase-final V1 durations, but variance was low in both positions (medial $M = 102.39$ ms, $\sigma^2 = 1$; final $M = 119.92$ ms, $\sigma^2 = 2$). This suggests that phrase-final lengthening may affect both syllables of an utterance-final disyllabic word, although the effect is smaller on V1 than it is on V2 (17 ms difference versus 50 ms difference).

In contrast to the interaction between Syllable Position and Phrase Position, the interaction between Syllable Position and Grammatical Category did not conform to our prediction (see Fig. 2). The vowel duration pattern was most trochaic-like for disyllabic verbs. The difference between V1 and V2 duration was a little over 30 ms (V1 $M = 108.47$ ms; V2 $M = 75.91$) and variance was low (V1 $\sigma^2 = 2$; V2 $\sigma^2 = 1$), suggesting a highly consistent duration pattern. The vowel durations in disyllabic nouns also tended towards a long-short pattern, but this tendency was reduced. The difference between V1 and V2 was less than 16 ms (V1 $M = 115.55$ ms; V2 $M = 99.18$) and variance was higher for V2 (V1 $\sigma^2 = 2$; V2 $\sigma^2 = 5$), suggesting some inconsistency in the pattern. The disyllables for other grammatical categories exhibited no reliable vowel duration pattern. Although V1 was generally shorter than V2 (V1 $M = 111.15$ ms; V2 $M = 120.77$), V2 variance was high (V1 $\sigma^2 = 2$; V2 $\sigma^2 = 15$), suggesting a lack of consistency in the pattern.

In summary, the analysis of vowel durations in disyllables of spontaneous infant-directed speech shows that a trochaic-like stress pattern was realized consistently in phrase-medial position and on verbs. Thus, the trochaic-like long-short pattern achieves cue constancy in certain phrase positions and for certain grammatical categories. It therefore satisfies the first condition for being a useful cue for speech segmentation, albeit for a subset of the disyllables. The aim of the following section is to evaluate the distinctiveness of this cue and of the other more variable patterns when they are embedded in a connected speech context.

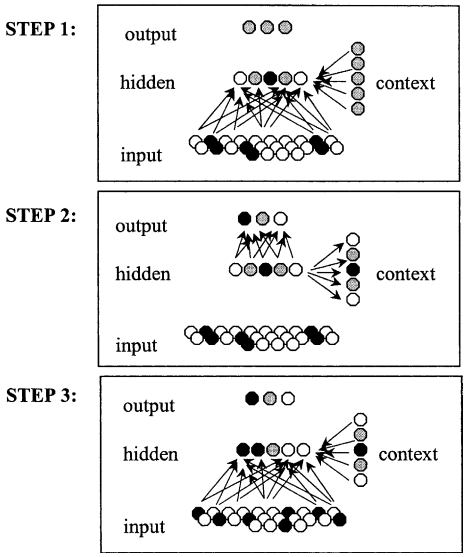

Fig. 3. A schematic representation of the architecture and algorithm of a Simple Recurrent Network (SRN) is shown.

7. Part 2: cue distinctiveness

Distinctive patterns can be defined as regular patterns that differ either from other regular patterns or stand out in a background of noise. To test whether the disyllabic vowel duration patterns were distinctive according to this definition, we used an SRN to determine whether or not the disyllabic patterns stood out among the duration patterns of adjacent monosyllabic and polysyllabic words in the context of a phrase. The SRN was used because, unlike most computational models, it is able to extract regular patterns from time-varying input by accessing a kind of dynamic memory (Elman, 1990). The memory is created by adding an extra bank of processing units, called *context units* that contain the network's internal representation of past input. As shown in Fig. 3, each new input to the network's processing units, called *hidden units*, is supplemented with input from the context units, providing the network with a temporal context for processing the current information. After processing the input, the hidden units pass their information to the output units, which make some task decision, as well as to the context units. At the next time step, the context units pass the same information back to the hidden units, which also receive the next new input, and the process repeats itself. Over time, the information stored in the context units represents information about the serially-ordered inputs over multiple time steps.

To test whether or not a trochaic-like long-short vowel duration pattern may provide infants with a distinctive cue for speech segmentation, the SRN was trained to identify syllable, word, and phrase boundaries from sequential segment duration inputs, representing the complete utterances in which the disyllables occurred. If the disyllabic vowel duration pattern is distinct from the other vowel duration patterns in the phrase, the SRN will learn to distinguish between syllable and word boundaries. If the disyllabic pattern is not distinctive, the SRN will not be able to differentiate between monosyllabic and disyllabic words. The prediction was that the trochaic-like pattern, which occurred mainly in phrase-medial position, would

be obscured by the alternation of content and function words, which are characterized by long and short vowels, respectively (Swanson et al., 1992). Another prediction was that the long-long vowel duration pattern would be distinctive because it would be uniquely associated with trochaically-stressed disyllabic words occurring in phrase-final position.

8. Method

8.1. Speech sample

The speech sample consisted of all the utterances containing the 49 phrase-medial/-final disyllabic pairs, analyzed in the previous section. The average length of the 98 phrases was 6.5 syllables with a range of 3–17 syllables. Segment durations were measured across the entire utterance, in the manner described below.

8.2. Duration measurements

Vowel and consonant durations across the phrase were measured in the same way as the disyllabic vowel durations. The Kay Elemetrics CSL displayed the utterance in concurrent views as an oscillogram and as a spectrogram. The exact criteria guiding boundary decisions for vowels were described in the previous section. Criteria guiding boundary decisions for two consonant types—sonorants and obstruents—were as follows:

(1) If a sonorant was bounded by a vowel or another sonorant, boundaries were marked on the spectrogram by a change in formant structure. If bounded by a vowel, the change in formant structure was often accompanied by a moderate, but abrupt decrease or increase in energy. If bounded by an obstruent, a boundary was established at the edge of the obstruent.
(2) If an obstruent was bounded by a vowel or sonorant, boundaries were marked on the spectrogram by an abrupt onset or offset of formant structure, which corresponded to an abrupt increase or decrease in the amplitude of the periodic waveform on the oscillogram. If an obstruent consonant was bounded by another obstruent, the boundaries could only be determined if at least one of the consonants was characterized by noisy energy, and the spectral center of that energy changed from one consonant to the next. In the case of adjacent stop consonants with no medial release burst, each consonant was assigned half of the overall closure duration.

As before, 15% of the data were randomly selected and remeasured on a separate occasion to assess reliability. Two-tailed *t*-tests indicated that there were no systematic differences between the old and new measures for obstruents, sonorants, and vowels in the subset of data. The average obstruent durations in the two measurement sets were identical (84 ms, SD = 46.4 and 44.7 for old and new measures respectively), and the two sonorant durations and vowel durations were very similar (sonorant, 64 ms versus 67 ms, SD = 37.4 and 37.7; vowel, 104 ms versus 108 ms, SD = 82.3 and 83.7). On average the old and new duration measures differed by plus or minus 10.1, 8.7, and 12.1 ms (SD = 11.6, 9.5, and 17.2) for obstruents, sonorants, and vowels, respectively.

8.3. The Simple Recurrent Network

The SRN was developed and run in the Light, Efficient, Network Simulator (LENS, Rohde, 1999). LENS is a fast, reliable, non-commercial simulator that has the additional advantage of providing the end user with detailed control over all aspects of the network's architecture as well as its training and learning algorithms. Sequences of segments, representing the selected phrases, were presented as input along with their durations. Three input units registered segment identity—obstruent, sonorant, vowel—and 21 units registered duration in 10 ms bins (durations of over 200 ms were represented by a single unit). Activation of the duration units was normally distributed around the target duration bin, so that similarity in duration values would be encoded. Several different tests were run using different input representations, and it was established that (a) segment identity was necessary, but not sufficient for learning different types of boundaries; and (b) the SRN could not distinguish between disyllabic word-initial and word-internal boundaries on the basis of consonant duration patterns.

The activation patterns of the input and context layer were fed forward to the output layer via a hidden layer of five units. The resultant activation pattern of the hidden layer was then copied over to the five context units at the same time as it was passed forward to the output layer. The output units determined boundary type based on the hidden layer activation pattern. A syllable boundary corresponded to the activation of one output unit, a word boundary to two units, and a phrase boundary to all three units. The relative activation levels of the different output units indicated what the SRN learned on the basis of the input, since each unit uniquely corresponded to a particular boundary type.

The SRN was trained to identify whether or not a previous segment was associated with a syllable, word, or phrase boundary based on information about the current segment and the series of preceding segments. This lag task best instantiates the idea that disyllabic word boundaries can be extracted on the basis of vowel-vowel duration patterns, since this type of pattern requires a comparison of vowel durations.

To maximize our data, we used the one-out method for training and testing. The SRN was trained on 96 phrases and tested on the remaining two, which were a matched pair containing the same disyllabic word in medial and final positions. Thus, 49 sets of simulations were run so that each of the disyllabic pairs could be tested. Each set consisted of eight simulations, which differed in their initial conditions (i.e., the initial random weight values assigned to the hidden, context, and output units).

During each simulation, the SRN received 100,000 presentations of the phrases drawn randomly from the 96 in the training set. Testing on the novel pair of phrases occurred at intervals, which increased as training continued. Initially testing occurred after every 20 presentations for the first 100 cycles, then every 100 for 10,000 cycles, and finally every 1,000 for the remaining cycles.

9. Results and discussion

The extent to which the SRN could distinguish between word and syllable boundaries for nouns, verbs, and "other" disyllables in differing phrase positions provided a measure of vowel pattern distinctiveness as well as an opportunity to evaluate distinctiveness as a function of Phrase Position and Grammatical Category. If a pattern was distinctive, the SRN would use it to learn that the two syllables of a disyllable formed a single word. If a pattern was not distinctive, the SRN would equate the disyllable's syllable

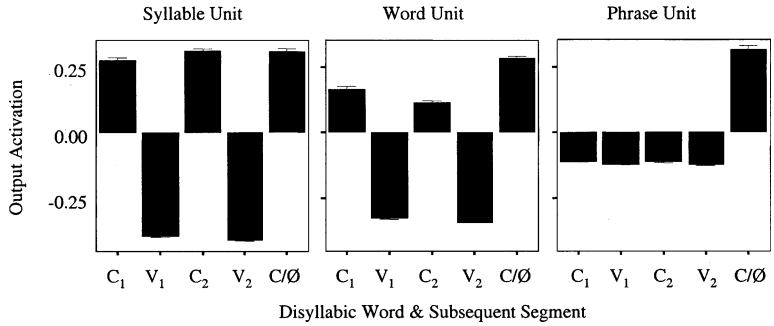

Fig. 4. The activation levels for the different output units are shown for five segments: the disyllable's word-initial (C_1) and word-internal (C_2) boundaries; the disyllable's syllabic nuclei (V_1 and V_2); and the first post-word segment (C/Ø), which was a word-initial boundary in half of the test phrases and a phrase boundary in the other half. Positive activation levels indicate above-average activation; negative levels indicate below-average activation.

boundary with a word boundary, since most of the input to the model consisted of monosyllabic words (see Table 1).

Since testing occurred throughout training, output error could be plotted over time. Such plots showed that most learning took place during the initial training cycles, and very little learning occurred later. There were also no apparent differences in the rate at which the SRN learned to identify the different types of disyllables in different phrase positions. We therefore focus on differences in the activation of the output units to different segments after learning occurred—a measure of how well the SRN learned to identify boundaries.

The activation levels of the three output units at each segment in the test phrases were averaged across the final 20 tests (i.e., those tests beginning after training cycle 81,000). These were then calibrated by subtracting the overall average activation of a particular unit from the unit's specific activation to a particular segment. A two-way (Segment × Output Unit) ANOVA performed on the calibrated activation levels established that the SRN *had* learned to distinguish between syllable, word, and phrase boundaries [$F(8,11571) = 480.49$, $p < 0.01$]. Fig. 4 shows that the output unit sensitive to syllable boundaries was appropriately activated (above zero) when a segment corresponded to a word-initial boundary, a word-internal boundary, or a phrase boundary. This is appropriate activation since syllable boundaries coincide with word and phrase boundaries. The same unit was appropriately inactive (below zero) when a segment did not correspond to a boundary. Fig. 4 also shows the output units sensitive to word and phrase boundaries. Phrase units were appropriately activated at phrase boundaries, and inactive to all other inputs.

Unlike the syllable and phrase units, the output unit sensitive to word boundaries was active where it should not have been—at the word-internal (syllable) boundary in disyllabic words (e.g., C_2 in Fig. 4). This is because most of the words in the speech sample were monosyllabic even though they all contained at least one disyllable, and so there was a strong correlation between syllable and word boundaries. Importantly, though, the word unit activation appeared to be greater at the word-initial boundary than at the word-internal boundary, indicating that the SRN was able to identify disyllabic word boundaries, albeit less well than syllable or utterance boundaries.

A subsequent analysis focused on the word unit's activation level at the disyllabic word-initial and word-internal boundaries to evaluate differences in distinctiveness for the different grammatical types of

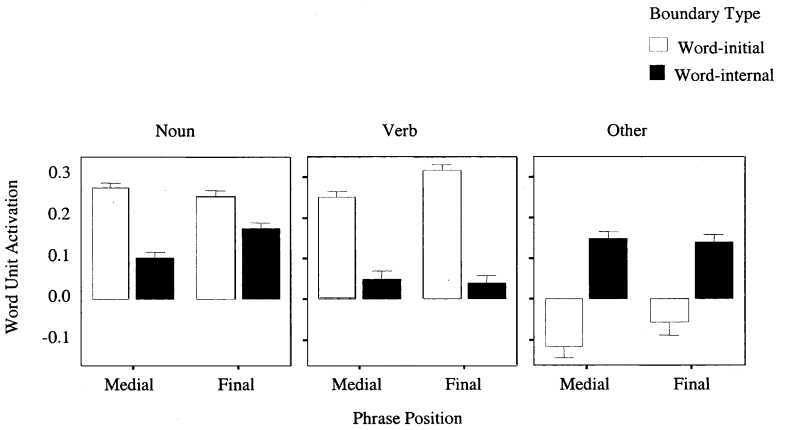

Fig. 5. Activation levels for the output unit sensitive to word boundaries for the word-initial and word-internal boundaries of disyllables in phrase-medial and phrase-final position.

disyllables in different phrase positions. Greater activation at the word boundary with lower activation at the internal boundary would indicate that a word was more easily identified as a single unit and so was distinctive. Equal activation at the two boundaries would indicate that the word was being identified as two monosyllabic words.

As shown in Fig. 5, the word unit's activation level at the disyllable's initial and internal boundaries varied with Phrase Position and Grammatical Category [$F(2,4872) = 4.242$, $p < 0.05$]. The word unit was highly (and appropriately) active at the word-initial boundaries of disyllabic nouns and verbs, and relatively suppressed at the word-internal boundary—especially for disyllabic verbs. Phrase position affected word unit activation for disyllabic nouns. The SRN distinguished between initial and internal boundaries less well when these words occurred in phrase-final position than when they occurred in phrase-medial position [$F(1,2012) = 4.03$, $p < 0.05$]. Fig. 5 indicates, however, that the phrase-final disyllabic nouns and verbs were learned.

Fig. 5 also shows that the SRN identified the word-internal boundary of other types of disyllables as a word-initial boundary, and failed altogether to identify the actual word-initial boundary. To determine whether this failure resulted from the vowel-initial structure of a few of the tokens in this category (e.g., "over"), such words were eliminated and the ANOVA was repeated. The results were that the word unit was activated at the word-initial boundary in addition to the word-internal boundary, but there were no significant differences in the activation levels, indicating that the SRN still failed to identify these disyllabic words based on their vowel duration pattern.

An SRN, like any neural network, learns by extracting regularities from the input data. The regularities are often not captured by simple statistics, since they may be high-order and nonlinear. Nonetheless, the learning differences reflected in the significant interaction between Boundary Type, Phrase Position, and Grammatical Category (present even when vowel-initial words are eliminated from the analysis [$F(2,4104) = 4.342$, $p < 0.05$]) suggest that such an interaction also exists in the vowel duration data, even though, it was not statistically significant. A re-analysis of the vowel duration data confirms the trend (Fig. 6), which is important for understanding the relative distinctiveness of the different disyllabic vowel duration patterns.

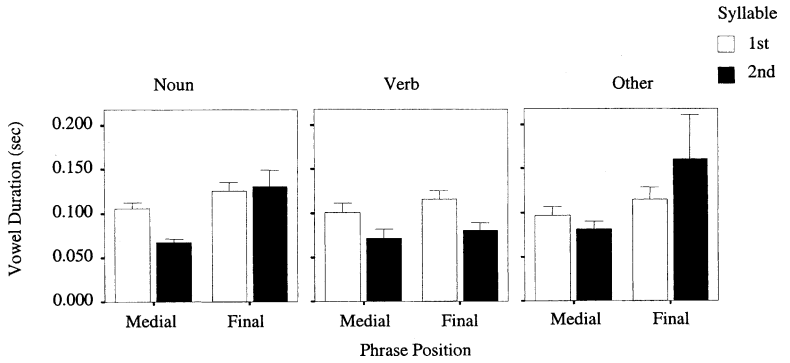

Fig. 6. The vowel duration pattern is shown to vary with grammatical category and phrase position.

Fig. 6 shows that a trochaic-like, long-short pattern characterized disyllabic verbs in both phrase positions, and disyllabic nouns in phrase-medial position. A long-long pattern characterized the vowels of disyllabic nouns in phrase-final position, a short-short and long-variable pattern characterized those of other disyllables in phrase-medial and phrase-final position respectively. Given these patterns and the SRN results on the disyllabic word-initial and word-internal boundaries, we conclude that the trochaic-like pattern is highly distinctive in the context of a phrase, and that a long-long pattern is also distinctive in phrase-final position. The short-short pattern and the variable pattern are not distinctive, and therefore, not possible for the SRN to learn.

10. General discussion

The hypothesis that infants can use a strong-weak stress pattern to extract disyllabic words from a continuous speech stream is only viable if stress is systematically realized on disyllables and is distinctive from the other sound patterns occurring in spontaneous infant-directed speech. The present study focused on vowel durations to evaluate whether this correlate of linguistic stress is consistently realized in a long-short pattern on trochaically-stressed disyllables and whether the pattern is distinct from other vowel duration patterns in a phrase. The long-short pattern was found to be realized more consistently in phrase-medial position than in phrase-final position, and more consistently for some grammatical categories than others. This pattern was also found to be distinctive, as was a hybrid pattern resulting from the interaction between stress and phrase-final lengthening. Thus, the study shows that even a single phonetic cue may consistently and distinctively code phonological form in natural infant-directed speech. A rhythmic word template can be extracted from the signal in spite of phonetic variability, and used to segment spontaneous speech.

11. Phonetic instantiations of prosodic structure

The finding that a trochaic-like vowel duration pattern is systematically realized in infant-directed speech on at least some types of words is perhaps less surprising than the finding that this pattern is highly distinctive in the context of a phrase. We had expected that phonetic variability in vowel durations

would obscure any pattern due to stress, particularly in phrase-medial position where a long-short duration pattern could arise from the alternation of content and function words (e.g., Swanson et al., 1992). The present finding of distinctiveness suggests that sound patterns due to phonetic processes are sufficiently random to stand in contrast to sound patterns due to phonological processes, which are more regular. In this way, our findings validate laboratory research on infant speech segmentation, which ignores the problem of phonetic variability when investigating infants' ability to extract and use sound patterns to locate word boundaries (e.g., Jusczyk et al., 1993; Morgan & Saffran, 1995; Morgan, 1996).

On the other hand, laboratory studies may idealize speech stimuli in another way that could potentially misrepresent the usefulness of single cues to speech segmentation. For example, in Morgan and Saffran (1995), six-month old infants used only vowel duration to segment continuous speech into disyllabic word-like units. Although we found that a long-short vowel duration pattern was consistently and distinctively realized on a subset of the disyllables in these spontaneous infant-directed speech samples, it may be premature to conclude that vowel duration alone provides sufficient information for speech segmentation. In these data, the largest average difference between the long and short vowels of the trochaic-like pattern was three times less than the difference Morgan and Saffran used to encode rhythm (50 ms versus 150 ms). This difference between spontaneous speech and laboratory stimuli may translate into a difference in the perceptual salience of the long-short pattern. When this possibility is coupled with the finding that the long-short pattern is consistently realized only on a subset of all disyllabic words, it becomes especially likely that if infants use trochaic stress to segment speech, then they must abstract the pattern from multiple sources of information available in the signal. Such a suggestion is not new—others have emphasized the importance of multiple cues in speech segmentation (e.g., Albin & Echols, 1996; Johnson & Jusczyk, 2001; Morgan, 1995), but the present study supports this point by evaluating acoustic data from continuous, spontaneous infant-directed speech.

12. Potential effects of cue interactions on segmentation

The usually consistent and highly-distinctive nature of the long-short pattern in the present sample of spontaneous infant-directed speech means that at least one cue to prosodic structure is available to the infant to be used for word boundary identification. The presence of this cue does not, however, imply that the segmentation problem is solved by linguistic rhythm alone. Alternative solutions have been proposed that capitalize on the natural boundary occurring at the end of an utterance (e.g., Aslin, 1993; Aslin et al., 1996), or on infants' remarkable sensitivity to segment co-occurrence statistics (Saffran et al., 1996). Such solutions have been found to be especially powerful when combined with information about rhythmicity. For instance, Christiansen et al. (1998) used an SRN to show that trochaic stress, co-occurrence statistics, and utterance boundaries all provided useful and independent sources of information for speech segmentation.

Signal-based solutions to the segmentation problem, such as the one proposed by Christiansen et al. (1998), that incorporate multiple sources of information are probably preferable to those that rely on a single underlying pattern (e.g., trochaic stress). However, future research on the use of multiple cues might evaluate whether the interaction between these possible sources of information consistently supports speech segmentation. This type of investigation will require an evaluation of the natural speech signal as opposed to transcribed speech, which has usually been favored in computational studies designed to examine the usefulness of multiple cues (e.g., Brent & Cartwright, 1996; Christiansen et al., 1998). For

example, the present results indicate that the potential advantage of phrase-final position may be mediated to some extent by the variable realization of trochaic stress in this position. The long-long pattern was learned less well than the trochaic-like pattern, probably because the phonetic effect of phrase-final lengthening did not apply uniformly within a category or between categories.

13. A role for syllables

Mehler, Dommergues, Frauenfelder, & Segui, (1981) and Mehler, Dupoux, Nazzi, & Dehaene-Lambertz (1996) have also suggested that the syllable plays an important role in speech segmentation. The SRN results from the present study support this suggestion. The SRN easily learned to identify syllable boundaries, and did so with greater accuracy than word boundaries (see Fig. 4). Moreover, it appears to have learned the syllable boundaries from segment sequencing patterns alone, as evidenced by the difficulty the SRN had in recognizing boundaries for vowel-initial syllables. Prosodic information was unnecessary. In this respect, our model behaved like previous models, which have shown that syllable boundaries can be identified on the basis of phonotactics (Martens et al., 2002; Vroomen et al., 1998). Our results are especially compelling because we provided the SRN with only a bare minimum of segment information, namely, segment manner class, whereas previous studies have provided the models with more complete specifications of segment identity (e.g., place, manner, voicing).

Because the identification of syllable boundaries requires less information than the identification of (multi-syllabic) word boundaries, the segmentation problem may be best solved in steps. Infants may first isolate syllables, then bind them into words, when appropriate, based on prosodic information or syllable co-occurrences at the end of a phrase. This stepwise approach would also be the most efficient means for learning English, given the preponderance of monosyllabic words in the simplified speech used with children (see Table 1). Aslin (1993) makes this same point, having also noted a preponderance of monosyllabic words in the infant-directed speech sample he analyzed.

In spite of its efficiency, and the fact that very young infants are sensitive to syllables as units of speech (Bertoncini & Mehler, 1981; Bertoncini, Floccia, Nazzi, & Mehler, 1995), a syllable-first approach to solving the segmentation problem is controversial for English. For instance (Cutler, Mehler, Norris, & Segui, 1986) explicitly reject the idea that English-speaking adults use syllables to segment speech, and Mattys, Jusczyk, Luce, & Morgan (1999) find that when phonotactic and rhythmic cues are in conflict, English-learning infants segment speech based on rhythmicity.

It should be noted, however, that English stress is a property of the syllable. So, infants must first have some idea of these linguistic units before information about stress becomes relevant—even if this idea is very rudimentary. For example, it would be sufficient for infants to think of syllables as those chunks of sound defined by low amplitude or aperiodic moments in the sound wave. These proto-syllable units would consist of vowel nuclei with consonantal onsets and offsets in the sense that the low amplitude or aperiodic sounds (consonants) blend into the high amplitude, periodic ones (vowels) through formant transitions. In other words, preverbal infants could identify syllable-like units based on coarticulatory information, to which they are clearly sensitive (Johnson & Jusczyk, 2001), instead of on phoneme units, which may not be conceptualized until later in acquisition (Werker & Tees, 1999). Thus, a two-step model in which infants first isolate syllables, then bind them into words based on rhythmicity may be the most appropriate model to explain how infants solve the segmentation problem. Such a proposal is similar to Mehler et al.'s (1996) TIGRE proposal, which suggests that infants' extract the rhythmic property

of a language by first attending to the duration and amplitude characteristics of vowels. We argue that consonants also deserve a role in this process as they not only delimit the vowels, but are also melded with the vowel via coarticulation.

14. The speaker's role

It may turn out that the aspects of the signal that infants end up attending to the most are the ones that caregivers emphasize in their speech. Although some may argue that infant-directed speech does not help infants acquire language (e.g., Pinker, 1994), research has shown that this speech register has more in common with clear speech, where phonological forms are well realized, than with casual speech, where they are not (Lindblom, 1990, 1996). For instance, clause boundaries are more clearly delineated in infant-directed than adult-directed speech (Bernstein Ratner, 1986), content words are more emphasized (Swanson et al., 1992), and stressed and word-final syllables are more fully realized (Albin & Echols, 1996). The fact that infants acquire content words before function words (Gerken & McIntosh, 1993), and produce stressed and word-final syllables before unstressed word-medial syllables (Echols & Newport, 1992) suggests a causal connection between what the caregiver emphasizes and what the infant acquires.

If a causal connection exists between what caregivers emphasize and the nature of infant acquisition, then the present findings may provide some additional insight into why nouns and verbs are acquired before other parts of speech (Gentner, 1982). Whereas the vowel duration pattern of disyllabic nouns and verbs exhibited cue constancy and distinctiveness, the patterns associated with other parts of speech (i.e., adjective, adverbs, and prepositions) were highly variable, and so not distinctive. The suggestion is that these common words "pretty," "fussy," "little" and so on, are not the focus of the mother's communicative intent to the infant, and so their phonological forms are realized less well than the words that are more important to that intent, e.g., "baby," "tickle." This suggestion might be worth pursuing in a study designed to investigate whether the emphasis on nouns and verbs versus adjective and adverbs is greater in infant-directed versus adult-directed speech.

Although nouns and verbs were realized more consistently and distinctively than other parts of speech, they were not realized equally consistently and distinctively. Contrary to our expectations, the lexically-specified long-short pattern was best realized on disyllabic verbs rather than nouns. This is because in our sample, which is representative of infant-directed speech in English, verbs with a disyllabic root were very uncommon. Most of the time, disyllabic verbs were comprised of a monosyllabic root plus an inflectional ending (Table 2). Since grammatical affixes are nearly always reduced, the long-short vowel duration pattern was almost certain to emerge in these verbs. Interestingly, though, our results show that the tendency to reduce inflectional endings is so strong that, unlike in nouns, the trochaic-like pattern is preserved even in phrase-final disyllabic verbs. The SRN results suggest that verbs are therefore highly distinctive, and in phrase-final position they are possibly more distinctive than disyllabic nouns. Thus, we find an interaction between a mother's tendency to phonetically reduce functional morphemes (Swanson et al., 1992) and her tendency to emphasize phonological stress in infant-directed speech (Echols & Albin, 1996). In this case, the interaction between these tendencies appears to be beneficial to word segmentation, presenting a counter-example to the problem mentioned earlier with reference to stress and phrase-final lengthening on the possible adverse effects of cue interactions.

15. Conclusion

This study of spontaneous infant-directed speech suggests that prosodic structure is discernable in the phonetics of the highly-variable speech signal directed to infants. Our results bolster the claim that infants are able to use regularities in the signal to solve the segmentation problem. Mothers consistently realize a trochaic-like vowel duration pattern on a subset of the disyllables of infant-directed speech. Such a pattern was also found to be distinctive in the context of the other vowel duration patterns characterizing a phrase.

Two other findings have relevance for continuing work on how infants solve the segmentation problem. First, the finding that syllable boundaries were more distinctive than word boundaries in spontaneous infant-directed speech suggests a two-step model of speech segmentation. English-learning infants might first chunk the speech stream into syllables, then bind the syllables into words (when appropriate) on the basis of rhythm. Second, the finding that cue interactions can sometimes help and sometimes hinder speech segmentation suggests that researchers should carefully consider interaction effects when proposing models of infant speech segmentation that incorporate multiple sources of information.

Acknowledgments

The research was supported by an NIH National Research Service Award F32-DC00459-02 granted to the first author. We would like to thank Carol Stoel-Gammon for generously providing us with the spontaneous infant-directed speech sample.

References

Albin, D., & Echols, C. (1996). Stressed and word-final syllables in infant-directed speech. *Infant Behavior and Development, 19*, 401–418.

Aslin, R., et al. (1993). Segmentation of fluent speech into words: Learning models and the role of maternal input. In B. de Boysson-Bardies (Ed.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 305–315). The Netherlands: Kluwer Academic Publishers.

Aslin, R., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Lawrence Erlbaum Associates.

Bernstein Ratner, N. (1986). Durational cues which mark clause boundaries in mother-child speech. *Journal of Phonetics, 14*, 303–309.

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Moraea and syllables: Rhythmical basis of speech representations in neonates. *Language and Speech, 38*, 311–329.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development, 4*, 247–260.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93–125.

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. *Journal of Child Language, 22*, 497–529.

Christiansen, M., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221–268.

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua, 92*, 81–104.

Cutler, A. (1996). Prosody and the word boundary problem. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 87–100). Mahwah, NJ: Lawrence Erlbaum Associates.

Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 113–121.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.

Echols, C. (1996). A role for stress in early speech segmentation. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 151–170). Mahwah, NJ: Lawrence Erlbaum Associates.

Echols, C. (1993). A perceptually-based model of children's earliest productions. *Cognition*, *46*, 245–296.

Echols, C., & Newport, E. (1992). The role of stress and position in determining first words. *Language Acquisition*, *2*, 189–220.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Fernald, A., & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 365–388). Mahwah, NJ: Lawrence Erlbaum Associates.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, *64*, 637–656.

Fisher, C., & Tokura, T. (1996). Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development*, *67*, 3192–3218.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity vs. natural partitioning. In S. Kuczay II (Ed.), *Language development: Syntax and semantics* (pp. 301–334). Hillsdale, NJ: Lawrence Erlbaum.

Gerken, L., & McIntosh, B. (1993). The interplay of function morphemes and prosody in early language. *Developmental Psychology*, *2*, 448–457.

Gleitman, L., Gleitman, H., Landau, B., & Wanner, E. (1988). Where learning beginnings: Initial representations for language learning. In F. Newmeyer (Ed.), *Language: Psychological and biological processes* (pp. 150–193). Cambridge: Cambridge University Press.

Gleitman, L., & Wanner, E. (1982). Language acquisition: The state of the art. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3–48).

Ingram, D. (1978). The role of the syllable in phonological development. In A. Bell & J. -B. Hooper (Eds.), *Syllables and segments* (pp. 143–155). Amsterdam: North Holland.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 548–567.

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Infants' preferences for the predominant stress patterns of English words. *Child Development*, *64*, 675–687.

Kehoe, M., Stoel-Gammon, C., & Buder, E. (1995). Acoustic correlates of stress in young children's speech. *Journal of Speech and Hearing Research*, *38*, 338–350.

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustic Society of America*, *59*, 1208–1221.

Klein, H. (1981). Early perceptual strategies for the replication of consonants from polysyllabic lexical models. *Journal of Speech and Hearing Research*, *24*, 535–551.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, *99*, 1683–1692.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht.

Martens, E., Daelemans, W., Gillis, S., & Taeleman, H. (2002). Where do syllables come from? In W. Gray (Ed.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 637–642). Fairfax, VA, George Mason University.

Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 101–115). Mahwah, NJ: Lawrence Erlbaum Associates.

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*, 298–305.

Morgan, J. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language*, *35*, 666–688.

Morgan, J., & Saffran, J. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66*, 911–936.

Oller, K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, *54*, 1235–1247.

Peterson, G., & Lehiste, I. (1960). Duration of syllabic nuclei in English. *Journal of the Acoustical Society of America*, *32*, 693–703.

Pinker, S. (1994). *The language instinct*. New York: William Morrow.

Rohde, D. (1999). *LENS: The light, efficient network simulator. Technical Report CMU-CS-99-164*, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Snow, C. (1972). Mothers' speech to children learning language. *Child Development*, *43*, 549–565.

Swanson, L., Leonard, L., & Gandour, J. (1992). Vowel duration in mothers' speech to young children. *Journal of Speech and Hearing Research*, *35*, 617–625.

Vroomen, J., van der Bosch, A., & de Gelder, B. (1998). A connectionist model for bootstrap learning of syllabic structure. *Language and Cognitive Processes*, *13*, 193–220.

Werker, J., & Tees, R. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology*, *50*, 509–535.