Language Learning and Development, 5: 94–114, 2009

ISSN: 1547-5441 print / 1547-3341 online
DOI: 10.1080/15475440802698524

Psychology Press
Taylor & Francis Group

Learning Prosodic Focus from Continuous Speech Input: A Neural Network Exploration

Bruno Gauthier and Rushen Shi
Department of Psychology, University of Québec, Montréal

Yi Xu
Department of Phonetics and Linguistics, University College London

This study uses connectionist modeling to explore whether and how infants might learn prosodic focus directly from continuous speech input. Focus is a communicative function that serves to put emphasis on a particular part of an utterance, and it is mainly encoded by pitch variations. The acquisition of focus entails two major difficulties. The first is that focus-related pitch patterns are confounded by other linguistic functions that also use pitch for their encoding, such as lexical tone in a tone language. Second, speakers have different pitch ranges, which further confounds the focus related pitch patterns. In three simulations using self-organizing neural networks, we explored how focus may be learned from continuous acoustic signals in Mandarin that were produced with co-occurring lexical tones and by multiple speakers. We used sentence-sized F_0 contours as well as their velocity profiles (D_1) as training input. Results show that both F_0 and D_1 contours provide information for focus learning, but only the D_1-trained network adequately handled the variability introduced by cross-gender differences. The recognition rate was analogous to human performance. Implications of these findings for theories of language acquisition and adult speech perception are discussed.

INTRODUCTION

During language acquisition, infants face the task of learning multiple linguistic functions, all of which are carried by a single speech signal. One important type of linguistic function to be learned is known as *focus*, which is a prosodic means to highlight (emphasize) a particular part (e.g., a word or phrase) of an utterance against the rest of the components (Cooper, Eady, & Mueller, 1985; Ladd, 1996; Xu, 2005). Detecting which element in an utterance is focused is particularly useful for language learners, since focused elements bring relevant information to the child's attention (Lœvenbruck et al., 2007). Focus can also be used as some form of verbal pointing for teaching a child a new word, consistent with the fact that mothers emphasize target

Correspondence should be addressed to Bruno Gauthier, Département de psychologie, Université du Québec à Montréal, Montréal, Québec, H3C 3P8, Canada. E-mail: gauthier.bruno@courrier.uqam.ca

words with exaggerated pitch peaks (Fernald & Mazzie, 1991). The acquisition of focus has so far received little attention in the field of language development, however. In this study we examine one particular issue about this developmental process, namely, whether adult speech contains sufficiently informative phonetic structure corresponding to word-level focus categories such that it is possible for infants to acquire them via unsupervised learning.

Focus can be expressed with the use of word order, e.g., topicalization, or the addition of a word particle specifically used for indicating focus. In Mandarin for example, the focus particle *shi* can be placed before a word that is intended to be emphasized (Li & Thompson, 1981). Another common linguistic device used in many languages for contrasting important parts of a message is prosody. Prosody highlights the word "cat" in "The CAT ate the mouse" in response to the question "Who ate the mouse?", mainly by expanding the pitch range of the focused component, compressing and lowering the pitch range of the postfocus components, but leaving the pitch range of the prefocus components largely intact (Cooper et al., 1985; Rump & Collier, 1996; Xu, 1999; Xu & Xu, 2005). Thus, the relative pitch differences of the whole utterance must be considered for the decoding of word-level focus. Focus also involves changes in duration, intensity, and vowel quality (de Jong, 1995; Turk & White, 1999; Xu, 1999; Xu & Xu, 2005), although F_0 is generally considered to be the primary cue for perceiving focus (Dahan & Bernard, 1996).

Before infants develop enough syntactic knowledge to learn focus through word order, they could use prosody to learn about word-level focus. A few studies exist in the literature that investigate children's processing of prosodic focus. These studies have concentrated on older children's speech production. For example, Ménard and colleagues (Ménard, Lœvenbruck, & Savariaux, 2006) recently explored the acoustic and articulatory correlates of focus in French-speaking children and adults. They found that 4- to 8-year-old children produce the neutral/focus contrast at the level of F_0. This finding is consistent with previous production work showing that prosodic focus is acquired by 3- to 4.5-year-old children (Allen & Hawkins, 1980; Hornby & Hass, 1970).

However, given infants' acute sensitivity to speech prosody, it is reasonable to believe that children may profit from perceptual cues to focus before they can produce those cues in their own speech. Abundant evidence from developmental speech perception research indicates that early in life, infants process intonational and rhythmic properties of input speech (Jusczyk, 1997). For example, newborns can distinguish between utterances produced in their native language versus those produced in a foreign language (Moon, Cooper, & Fifer, 1993), and, crucially, this ability is maintained when the signal is low-pass filtered, which eliminates phonetic and phonotactic cues (Mehler et al., 1988). Prosodic information thus seems to guide newborns' preference for their native language, an ability that most likely results from prenatal exposure to suprasegmental features. Infants' early sensitivity to prosody is not surprising, as it is known that the auditory system is functional from about 22 to 24 weeks of gestational age (Slater, 1998). Sensitivity to the sound structure of speech has indeed been observed in 33- to 38-week-old fetuses (DeCasper, Lecanuet, Busnel, Granier-Deferre, & Maugeais, 1994; Lecanuet & Granier-Deferre, 1993; Lecanuet et al., 1987). Further research indicates that newborns' capacity to distinguish between broad language classes is specifically based on rhythmic information (Nazzi, Bertoncini, & Mehler, 1998). In fact, as early as 1 month of age, infants distinguish between different stress locations based on durational differences alone and on the other acoustic correlates of stress (F_0 and intensity) (Spring & Dale, 1977).

Other studies also suggest that infants rapidly develop sharp sensitivities to the intonation structure of their native language (e.g., Kaplan, 1969; Morse, 1972). Jusczyk and colleagues (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993) showed that 6-month-old English-learning infants listen longer to lists of English than Norwegian (low-pass filtered) words, the former differing from the latter with respect to the pitch contour of final syllables (Haugen & Joos, 1972) and the pitch height of stressed syllables (Peters, 1997). English-learning infants are sensitive to the predominant stress patterns of English words at between 6 and 9 months old (Jusczyk, Cutler, & Redanz, 1993). During this period, infants also start using stress to segment words in continuous speech (Jusczyk, Houston, & Newsome, 1999), as adult English speakers do (Cutler & Norris, 1988). The role of prosody in language acquisition is further exemplified in studies showing that infants as young as 6 months of age use their sensitivity to prosody for marking syntactic units (Hirsh-Pasek et al., 1987; Nazzi, Nelson, Jusczyk, & Jusczyk, 2000; Soderstrom, Seidl, Nelson, & Jusczyk, 2003).

These findings suggest that infants pay attention to prosody at a young age. The question that we address is whether the information in the sound stream is sufficiently rich to allow the child to abstract word-level focus. Different sources of variability complicate the surface realization of various linguistic functions (Perkell & Klatt, 1986), and focus cannot be an exception. Among the recognized sources of variability, cross-speaker differences have received the most attention (e.g., Johnson & Mullennix, 1997). Cross-speaker differences have also been shown to affect the acoustic manifestations of emphatic accent in French (Dahan & Bernard, 1996). In Dahan and Bernard, although all speakers' pitch increased on the target word and the increase spread from the peak-bearing syllable to the whole word, speakers' productions differed with respect to the location of the F_0 peak within the target word. One pervasive characteristic of cross-speaker variability concerns the impact of anatomical differences on the speech output, most notably between children and adults or between men and women (e.g., Ménard, Schwartz, & Boë, 2004). The different sizes and shapes of the vocal tract affect the vowel acoustic space, for example, merging the formants of different speech sounds together while creating separate formant clusters for the same vowel category (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952). With respect to intonation, female pitch is about three quarters of an octave higher than male pitch (Peterson & Barney, 1952). Figure 1a shows the F_0 contours (in hertz) of three-word declarative sentences (five syllables) produced by four male and four female native speakers of Mandarin, with either no focus or focus on word 1, word 2, or word 3 (data from Xu, 1999). While some within-gender variability can be observed, a similar pattern stands out for each focal condition. Between-gender variability is much larger, however. Not only did females produce higher pitch registers than males, but they also produced much wider pitch excursions on the focused components. Given the impact of between-gender variability on the surface realization of focus, it is unknown whether focus can be learned solely on the basis of F_0 when produced by multiple speakers.

A further complication to learning word-level focus from F_0 is that, in tone languages, F_0 is also crucial for perceiving lexical tones. The use of one acoustic cue for two or more linguistic functions may in principle cause particular challenges for the perceptual system. For example, because both word-level focus and question intonation raise F_0 at the sentence-final position, recognition of focus by adults is harder when it occurs at the end of a sentence than when it occurs at other sentence locations (Liu & Xu, 2005). On the other hand, adults can simultaneously perceive both focus and tones from the speech input (Liu & Xu, 2005). This is because

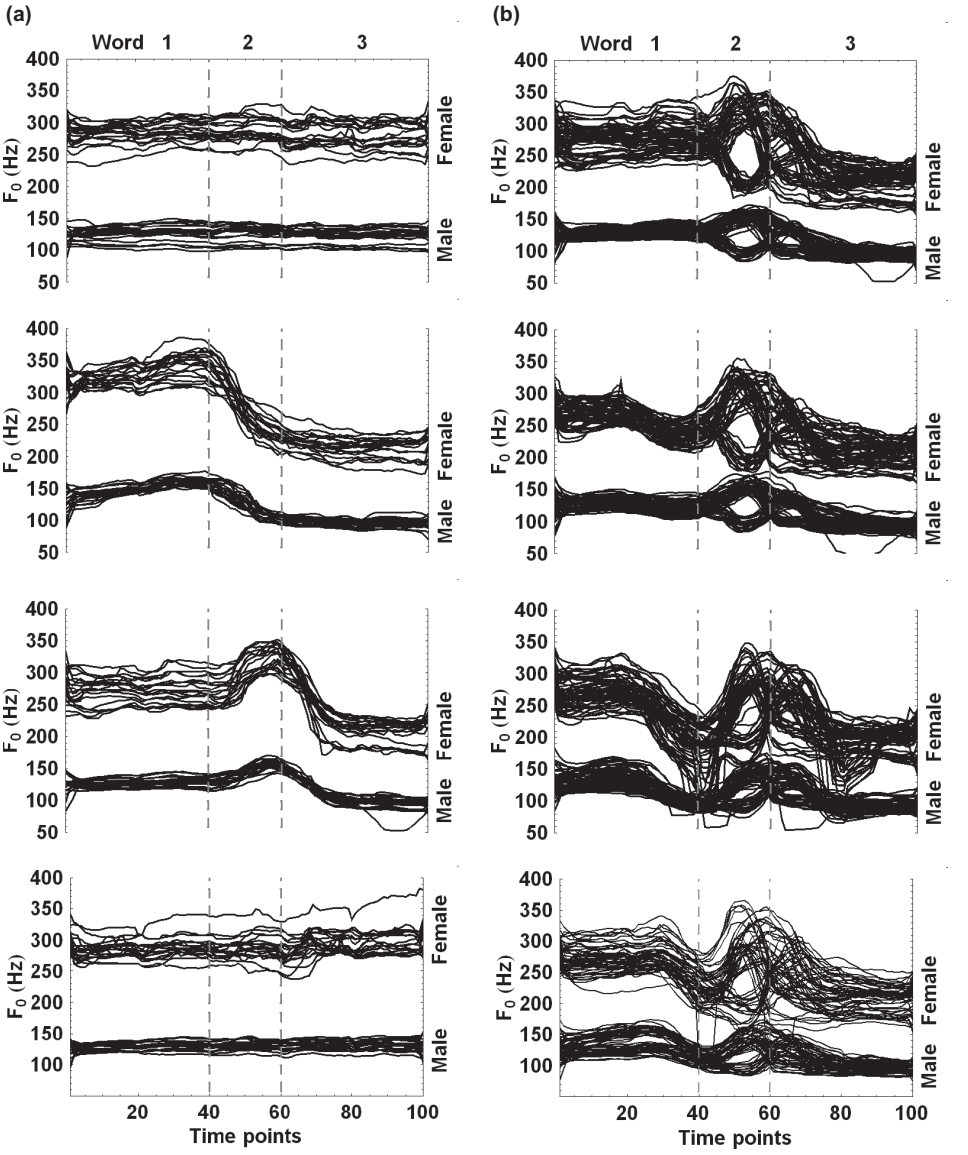

FIGURE 1 Variability in prosodic focus. (a) Variations in F0 contours of declarative sentences induced by cross-speaker and cross-gender variability (four male and four female speakers) in sentences produced with neutral focus (top panel), focus on word 1 (second from top), focus on word 2 (third from top), and focus on word 3 (bottom), each panel containing 40 exemplars. (b) Variability induced by the simultaneous encoding of focus and lexical tones (High, Rise, and Fall) on the second word preceded by High, Rise, Low and Fall (from top to bottom, respectively), each panel containing 120 exemplars.

in adult speech production in Mandarin, local F_0 contours of syllables in an utterance are mostly determined by lexical tones, while focus globally modulates the pitch range of the entire utterance (Xu, 1999).[1] The larger-scale pitch range modulations that signal focus operate in parallel with local pitch contours that encode lexical tones (Xu, 2005). It remains to be tested whether and how a naive learning system could learn focus when the same acoustic cue also signals varying lexical tones within the same utterances.

In Gauthier, Shi, and Xu (2007a, 2007b), unsupervised artificial neural networks were presented with syllable-sized F_0 contours and their velocity profiles as input. The results showed that, despite extensive variability due to cross-speaker differences, contextual variations, and simultaneous focus encoding, the system could still learn lexical tones and that velocity profiles provide better perceptual cues than F_0 contours. A velocity profile consists of a continuous instantaneous rate of change of the corresponding F_0 contour. It is shown in Gauthier et al. (2007a) that the mathematical process of deriving velocity profiles automatically removes a large portion of speaker differences in terms of individual pitch heights, which makes velocity profile a more direct indicator of phonetic categories. In fact, velocity profile by definition should be consistent across speakers. However, it remains unknown as to how an unsupervised learning system may simulate word-level focus acquisition under comparable circumstances of variability as in the previous tone-learning studies. Figure 1b shows the variability that lexical tones can add to F_0 contours, thus making it more difficult to abstract word-level focus.

The goal of this study is to investigate whether unsupervised neural networks can learn word-level focus from continuous dynamic speech signal produced by multiple speakers in various lexical tone conditions. We tested the hypothesis that dynamic F_0 information (i.e., velocity profile) is useful not only for learning lexical tones, as shown in Gauthier et al. (2007a, 2007b), but also for learning word-level prosodic focus. Three simulations using unsupervised artificial neural networks were performed. In each simulation, natural speech data were used for training and testing self-organizing maps (SOMs; Kohonen, 1982, 1995). If dynamic F_0 information indeed provides robust cues for learning word-level focus, the networks should develop clusters corresponding to the focal categories produced by the speakers. If the networks cannot extract any meaningful structure from F_0 patterns or their velocity profiles, this would indicate that other sources of information may be necessary for learning word-level prosodic focus.

METHODS

The Self-Organizing Map

The SOM is widely applied as a method of statistical analysis in different research areas. It was first proposed for modeling topographic mapping of sensory input in the brain, for example the tonotopic representation of sound frequencies in the primary auditory cortex (Kohonen, 1982). The advantage of using the SOM for studying learning processes lies not only in its self-organizing principles but also in its visualization and abstraction properties: by mapping a high-dimensional

[1]More specifically, the pitch range of the focused component is expanded, the pitch range of the postfocus components is lowered and compressed, and the pitch range of prefocus components remains largely intact (see Figure 1a).

input space onto a lower-dimensional grid, the SOM compresses information while preserving important geometric relationships in the data (Kohonen, 1998). The Kohonen networks can be designed to contain a large number of units (e.g., several hundred). This makes it possible to examine whether the units on the maps, which initially have no relationship with one another, gradually form neighborhood structures during training. In other words, one can observe whether clusters of units sharing similar characteristics can be developed based on the properties of training input (e.g., acoustic information), whether the clusters correspond to categories being learned (e.g., tones), and whether clusters show any further internal structure. This process is comparable to category formation in early language development. Since the number of map units does not a priori correspond to the number of categories to be learned, this type of design reflects closely the initial state of naïve learning. Recently, unsupervised methods similar to the SOM have been adopted for testing various hypotheses in language acquisition, for example, to evaluate the role of statistical learning in language acquisition and to examine specific mechanisms underlying early speech/language development (e.g., Behnke, 1998; Guenther & Gjaja, 1996; Shi, 1996; Shi, Morgan, & Allopenna, 1998; Shi, 2006). In the present study, the SOM was used to explore what cues are essential for deriving word-level focus categories and whether focus may be learned directly from the speech input without any feedback. The detailed algorithm of the SOM is presented in the Appendix.

In our previous SOM modeling of the acquisition of lexical tones, we used syllable-sized input (Gauthier et al., 2007a, 2007b) because the temporal domain of tone production is known to be the syllable (Xu, 1999, 2005). The temporal domain of focus, in contrast, is the entire utterance rather than just the focused word, as described earlier. The acquisition of focus thus requires sentence-sized rather than syllable-sized input. To evaluate the impact of cross-speaker and cross-gender variability on word-level focus, we conducted three simulations that varied in the amount of variability in the input. In the first simulation, the networks were trained with sentences produced by a single adult female speaker. Simulation 2 presented the network with input produced by four male speakers. Finally, Simulation 3 involved the most natural input corpus, produced by four male and four female speakers.

Input corpus and representation. The global input corpus contains 3,840 declarative sentences produced by four adult male and four adult female native speakers of Mandarin (speaking to adults) at a normal speech rate (about 5.6 syllables per second on average) (data from Xu, 1999). Each sentence is composed of three words (subject, verb, object). Each word contains one or two CV syllable(s) where C is a sonorant (/m,n/), except when the Low (L) tone occurs on the fourth syllable, which starts with /d/. The subject and object are disyllabic and the verb is monosyllabic [subject: 'maomi', with the High tone (H) on the first syllable and four different tones on the second syllable: 'kitty' — H-H, 'cat-fan' — H-Rising (R), 'cat-rice' — H-Low (L), or 'cat-honey' — H-Falling (F); verb: 'mo' — H, 'na' — R, or 'mai' — F ('stroke', 'take', 'sell'); object: 'maomi' — H-H or 'madao' — L-H ('kitty', 'saber')]. The sentences were produced in various focal and tonal conditions: (a) neutral focus (focus0), (b) focus on word 1 (focus1), (c) focus on word 2 (focus2), and (d) focus on word 3 (focus3), with an equal number of occurrences in each focal condition. These four different conditions were elicited with focus-inducing leading questions (e.g., "Who is stroking Kitty?" to elicit focus on the first word of the sentence "THE SUBJECT strokes Kitty"; see Xu, 1999, for details). Inspection of the input data prior to the simulations revealed that for the initial focus (i.e., the subject word) and the final

focus (the object word), both syllables in each word carried focus information, and since the first and last syllables always carried the High tone, they were excluded from the training and testing tasks. The second, third, and fourth syllables were produced with varying tones (H, R, L, F on the second syllable; H, R, F on the third syllable; and H, L on the fourth syllable, as just described), for a total of 24 different tonal patterns, each produced an equal number of times. To summarize, eight speakers each produced five repetitions of 4 focal and 24 tonal patterns, for a total of $8 \times 5 \times 4 \times 24 = 3{,}840$ sentences.

To simulate the learning of word-level focus, we used as input sentence-sized F_0 contours that contained no syllable boundary information. Each input token corresponded to the global F_0 contour of a sentence, represented by a 30-data point vector, i.e., equally distanced discrete values taken from the time-normalized syllable-sized F_0 curves of the three middle syllables. The F_0 extraction procedure consisted of manually segmenting the sentences at syllable boundaries, converting vocal pulses into F_0 values, smoothing the resulting F_0 contours, and taking a predefined number of discrete F_0 values from each syllable (see Xu, 1999). These contours were quasi-continuous in the sense that, at the sampling frequency of 10 points per syllable, for a male voice of 120 Hz, there was one representative point every two vocal periods in a syllable of 180 msec in the dataset (Xu, 1999). The F_0 contours were minimally if at all interrupted by vocal tract closures during production (at least in the middle three syllables) since only sonorant consonants were used (e.g., see Lehiste & Peterson, 1961). There were no pauses within any of the sentences. F_0 input vectors were in hertz. The velocity profiles of F_0 were generated according to:

$$D_{1i} = 0.5(F_{0_{i+1}} - F_{0_{i-1}}) \tag{1}$$

which yields the discrete first derivatives of F_0. The computation of velocity by every three points is known as central differentiation and is commonly used in data analysis because of its speed, simplicity, and accuracy (Bahill, Kallman, & Lieberman, 1982).[2] To illustrate the effect of the F_0 to D_1 transformation, Figure 2 shows a minimal pair of three-word sentences (five syllables, all with High tones) with focus on word 1 and focus on word 2. The sentences were produced by a male and a female speaker saying 'maomi mo maomi' ('kitty strokes kitty'[3]) in answering the questions 'Who is stroking Kitty?' (prompting focus on word 1) and 'What is Kitty doing to Kitty?' (prompting focus on word 2). Note that the large differences between the male and female speakers clearly visible in the F_0 patterns (Figure 2a) virtually disappear in the velocity profiles (Figure 2b). This is due to the differentiation process, which eliminates the constant term of a function (Gauthier et al., 2007a).

Training and testing phases. To recapitulate, the simulations involved six distinct networks corresponding to three training conditions (Simulations 1 to 3 with increased variability) and two types of training/testing data (F_0 and D_1). The training corpora of Simulations 1, 2, and 3, respectively, contained 240, 960, and 1,920 stimuli (half of the input corpus), which were

[2]Note that here D_1 is computed based on time-normalized F_0 contours for which duration information is partially lost. Although this reduction in duration accuracy may in turn affect the accuracy of D_1 profiles, it is unlikely that this reduction in accuracy would affect the simulation outcome, judging from the results of Gauthier et al. (2007a, 2007b).

[3]The use of the same word in the sentence initial and final positions in Xu (1999) was based on multiple considerations, including ease of segmentation, tonal composition, and, most crucially, lack of alternative real words. The sentence in Mandarin clearly distinguishes the subject and the object as different entities.

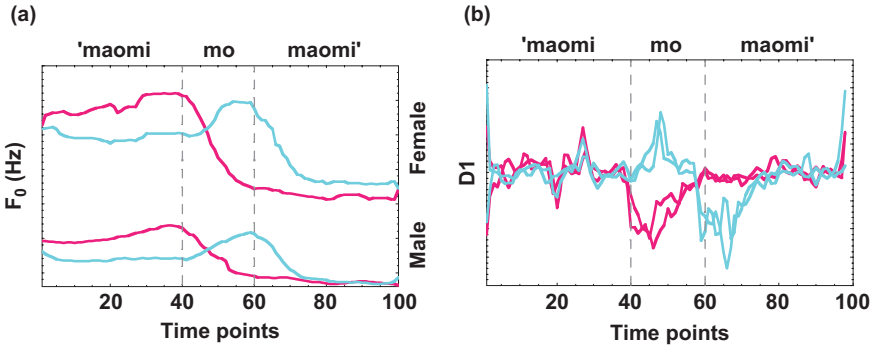

FIGURE 2 Examples of velocity profiles transformed from F_0 contours. **(a)** F_0 contours of a sentence consisting of only the High tone in Mandarin, spoken by a male (bottom curves) and a female (top curves), with focus on word 1 'maomi' (magenta) or word 2 'mo' (cyan). **(b)** D_1 profiles obtained from the F_0 contours in (a) using equation (1). (This figure appears in color online.)

randomly presented to the networks 30 times each. During training, the SOM implemented a recursive regression algorithm that mapped the continuous input distribution $P(x)$, $x_i \varepsilon X$, onto a discrete output space composed of multiple processing units. A basic SOM is illustrated in the Appendix (Figure 7) as a linear array of four units (filled dots) connected to a one-dimensional input space. Each unit is pointing to a specific location in the input space with a scalar-valued connection weight vector, or receptive field center (empty dots). Each unit also defines a receptive field (bold horizontal lines) that encompasses all data points closer to its center than to any other unit's center. The neural maps in our simulations were much larger. They were each squared arrays of 900 units (30×30), and thus did not entail a predefined number of (four) categories. Each unit was connected to the input space with a 30-dimensional (i.e., 30-point) weight vector, so as to correspond to the 30 F_0 or D_1 points from each input utterance. Each unit was initiated with a connection weight vector that represents a linear (F_0 or D_1) trajectory in the form $ax + b$, whose minimum and maximum values fell within the range of the input space. Note that the initiated vector values on the units were totally random, bearing no direct relation to the values of any input utterances. Training began after the initialization step, and the connection weight vectors were gradually transformed into representative contours of the input in the following way. Each time an input token was presented to the network, it was compared to each weight vector to determine the closest unit in terms of the Euclidean distance. The winning unit and its neighboring units on the map were shifted to better fit the data based on the learning algorithm (see details in the Appendix). This process proceeded according to the learning step size, which decreased exponentially from 0.7 to 0.01 during the learning task. The neighborhood function, which implemented lateral activation between units, initially included all map units, went through exponential decay, and activated only a single unit at the end of training. As mentioned earlier, the trained SOM can be used for revealing the neighborhood structure of multidimensional input spaces.[4]

[4]Note again that each unit on the 2-dimensional map contained a 30-dimensional vector analogous to a whole F_0 or D_1 contour (see Kohonen, 1989, for details on the compression from a high-dimensional input space to a low-dimensional map).

The testing corpora used in Simulations 1, 2, and 3, respectively, involved 480, 1,920, and 3,840 stimuli. Half of these stimuli were used during the training and the other half were novel so as to verify the network's generalization capacity. During the recall task of the test phase, each input token was presented once to the network and was assigned to the closest unit. The testing set contained as much variability as the training set and was presented in an orderly fashion (all exemplars of neutral focus, focus on word 1, focus on word 2, and focus on word 3). The number of input tokens projected onto each unit was indexed into a global firing frequency matrix. Units that responded at least once during the recall task were treated as operable units, while those that did not respond to any input vector were nonoperable units and were not further considered. Categorized units were defined as operable units that responded to a single focus category at least 68% of the time during recall. This criterion was based on a chi-square test, according to which the null hypothesis stipulates that all units respond equally to more than one category, thus to the four focus categories with an equal probability of 25%. Any different outcome would in principle allow us to reject this chance-level hypothesis. However, our goal was to determine the observed firing frequency for a category that would offer enough evidence to reject the null hypothesis to a much more strict level, namely, a level that determined those units sensitive to one focus category more than any combination of the three others. The critical value at a 0.001 confidence level was found to be $\chi^2(3) = 12.84$, above which the firing probability of a unit to a single focal category is >0.68. Units responding to multiple focal categories, none of which was dominant, corresponded to ambiguous units.

Measures

The performance of the trained networks was first evaluated in terms of rate of success for categorized units, which corresponded to the percentage of input tokens from one focus category landing on the same-class categorized units. This performance is independent of the one for tone categorization, which has previously shown to be effective for the same corpus using syllable-sized F_0 and D_1 profiles as input (Gauthier et al., 2007b) as discussed earlier. To assess whether categorized units were grouped into clusters corresponding to focus categories, the neural maps were quantitatively colored by associating distinct categories with distinct colors. The coloring of the maps reveals the distribution of the testing input corpus to the outside observer. If focus categories were well separated in the data, the color map should be divided into regions by classes that group similar units together, and we could thus infer that within-cluster variability is trivial. Finally, the similarity of within-cluster units on which the network could rely to infer focus information was shown in the internal maps. An internal map after training is a graphic display of a unit's connection weight vector showing F_0 or D_1 trajectory as a function of time, analogous to the input tokens shown in Figure 2. By plotting on the same graph the trajectories of multiple units responding to the same category, we were able to examine the internal representation of each externally identified focus cluster developed by the networks.

RESULTS

The results of Simulation 1 (Figure 3a) show high performances for each focus condition in both F_0 and D_1, suggesting that sentences with distinct focus locations can be distinguished solely on

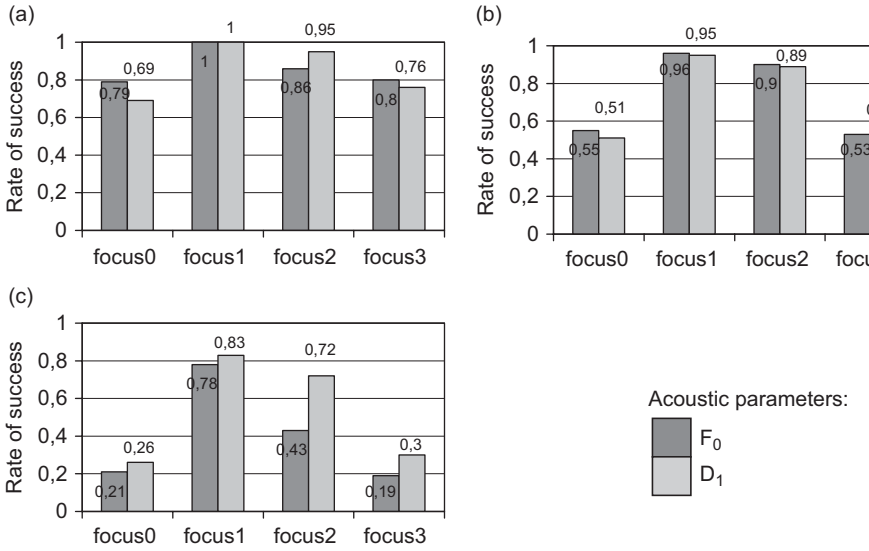

FIGURE 3 Rate of success for Simulations 1, 2 and 3. The vertical axis of each panel shows the rate of success for each learning condition as a function of focus location in the utterance (focus0 = neutral focus; focus1 = focus on word 1; focus2 = focus on word 2, and focus3 = focus on word 3). (**a)** Simulation 1: Training and testing with data produced by one female speaker. (**b)** Simulation 2: Training and testing with data produced by four male speakers. (**c)** Simulation 3: Training and testing with data produced by four female and four male speakers.

the basis of global F_0 contour produced by a single speaker. In contrast, the results of Simulation 2 with four male speakers (Figure 3b) show that both F_0 and D_1 networks responded positively to only focus1 and focus2 conditions, while confusing focus0 and focus3. In this regard, the results of Simulation 2 are comparable to the perceptual results obtained with adult human subjects, who cannot perceive final focus as easily as nonfinal focus categories (Liu & Xu, 2005). (A simulation involving input data from four female speakers yielded comparable results to those reported here for the four male speakers.) Finally, the D_1 network in Simulation 3 also achieved a performance comparable to that of human subjects, especially considering the fact that the human perceptual performance reported in Liu and Xu (2005) was based on only two speakers, one male and one female. However, the F_0 network performance was lower, especially for focus2, which did not correspond to any previously observed human response patterns (Figure 3c).

Color maps were used to better understand the outcome of the adaptation process that took place during the simulations. Quantitative coloring of the neural maps was achieved by associating distinct categories with distinct colors. Focus color maps were obtained for F_0 and D_1 networks of Simulation 3 by associating the four focus categories with four colors produced with the RGB color system. Neutral focus, represented by yellow, was specified by a mix of red and green in the vector [1,1,0] (focus1 = magenta [1,0,1]; focus2 = cyan [0,1,1]; focus3 = blue [0,0,1]).

Each map unit was then associated to a three-dimensional vector, the values of which were specified according to the unit's firing probabilities for each focus class. Units responding to only a particular focus condition thus appeared in a saturated color, while units sensitive to multiple classes yielded mixed "impure" colors. Therefore, if the input data contain categorical information, the color maps will reveal each category as a distinct region. Black color represents nonoperable units that remained inactive during the testing phase.

Figure 4 shows the color maps obtained when testing the networks on the general distinction between focused and neutral-focus sentences in Simulation 3. The F_0 color map shows no internal organization. In contrast, on the D_1 color map, focus information is embedded in a circular

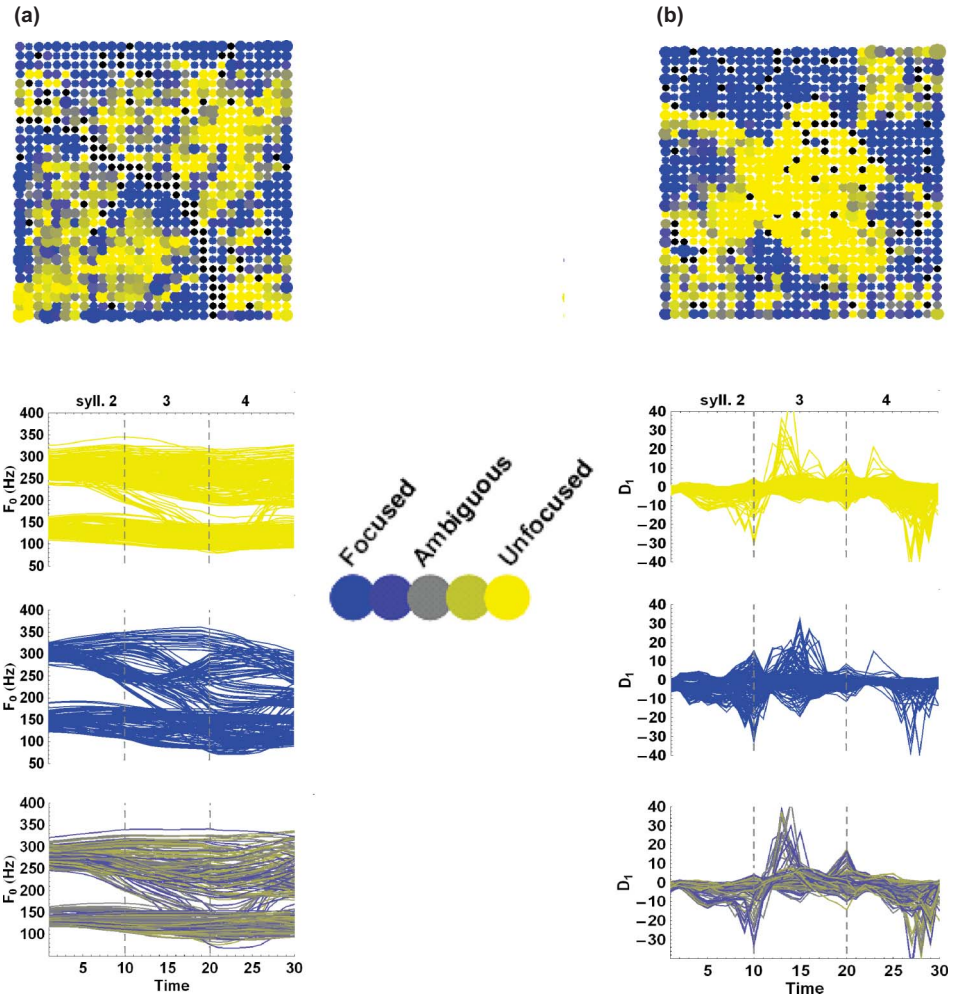

FIGURE 4 Color and internal maps of Simulation 3 showing the focus versus unfocused distinction by **(a)** F_0 and **(b)** D_1. (This figure appears in color online.)

fashion, with focused sentences situated in the periphery and the neutral-focus sentences (as well as sentences with final focus) in the more central areas. The D_1 internal maps (lower graphs) reveal that the center units represent relatively flat pitch patterns in comparison to the peripheral units, which display wider pitch excursions.

In the detailed color F_0 map (Figure 5a), the focus1 category takes the form of multiple randomly distributed unit clusters, while the focus2 category appears as a network of river-like channels running through the whole map. In the D_1 map (Figure 5b), focus0 and focus3 categories overlap in the central region, reminiscent of human performance (Liu & Xu, 2005), while focus1 and focus2 form clusters distributed in the periphery.

Further analyses of units' sensitivities reveal that the bottom right focus1 cluster mainly responds to sentences for which the focused component is a Low tone (77%), while the rightmost zone of the top focus1 cluster responds to Rise (91%), the middle zone to High (88%), and the leftmost zone to Fall (77%) (Figure 6). Similar organization can be observed for focus2 clusters. The leftmost unit cluster responds to focused High tone (75%), the bottom cluster to focused Rise (90%), and the rightmost cluster to focused Fall (83%).

GENERAL DISCUSSION

The results of this study show that unsupervised neural networks can develop focus-specific clusters from continuous dynamic speech signals produced by multiple speakers in various lexical tone conditions, which may eventually lead to the acquisition of focus. The performance of focus recognition is not uniform across all focus locations. In particular, final focus is not highly distinguished from a no-focus condition. Interestingly, this is similar to human focus perception, which also shows high confusion rates between final focus and no focus (Liu & Xu, 2005). Like human adults, the network performance at other focus locations was successful. The overall results thus suggest that despite variability due to lexical tones and multiple speakers, it is possible for a naïve system to develop phonetic categories that allow the recognition of word-level focus from continuous F_0 input at a level that approaches the performance of normal adult listeners. This finding indicates the effectiveness of sentence-sized F_0 input for focus recognition. This is interesting from the learning point of view, as there is evidence that sentences and clauses are among the initial units perceived by preverbal infants (Hirsh-Pasek et al., 1987; Nazzi et al., 2000; Soderstrom, Nelson, & Jusczyk, 2005). Hence, it is possible that infants may begin the learning of focus at an early age.

Furthermore, the results of the trained networks in the three simulations indicate that different focus categories can be contrasted according to the portion of the sentence that the emphasis is placed on. Previous production work has shown that nonfinal focused components are characterized by widened pitch span, and postfocused components by suppressed and lowered pitch span (Xu, 1999). The internal maps of Simulation 3 (Figures 4, 5, and 6) corroborate such findings. The internal representation developed by the D_1 network shows that focus-sensitive units are characterized by increased (positive and negative, depending on the tone) velocity compared to the neutral units, and that most postfocus syllables show more negative velocity (see Figures 2 and 6), indicating an active suppression of F_0 after focus. Previous work has also shown that final focus is only marginally different from neutral focus (Xu, 1999). This resemblance of final focus to neutral focus is seen as the mixed cluster in the center of the color map of Simulation 3

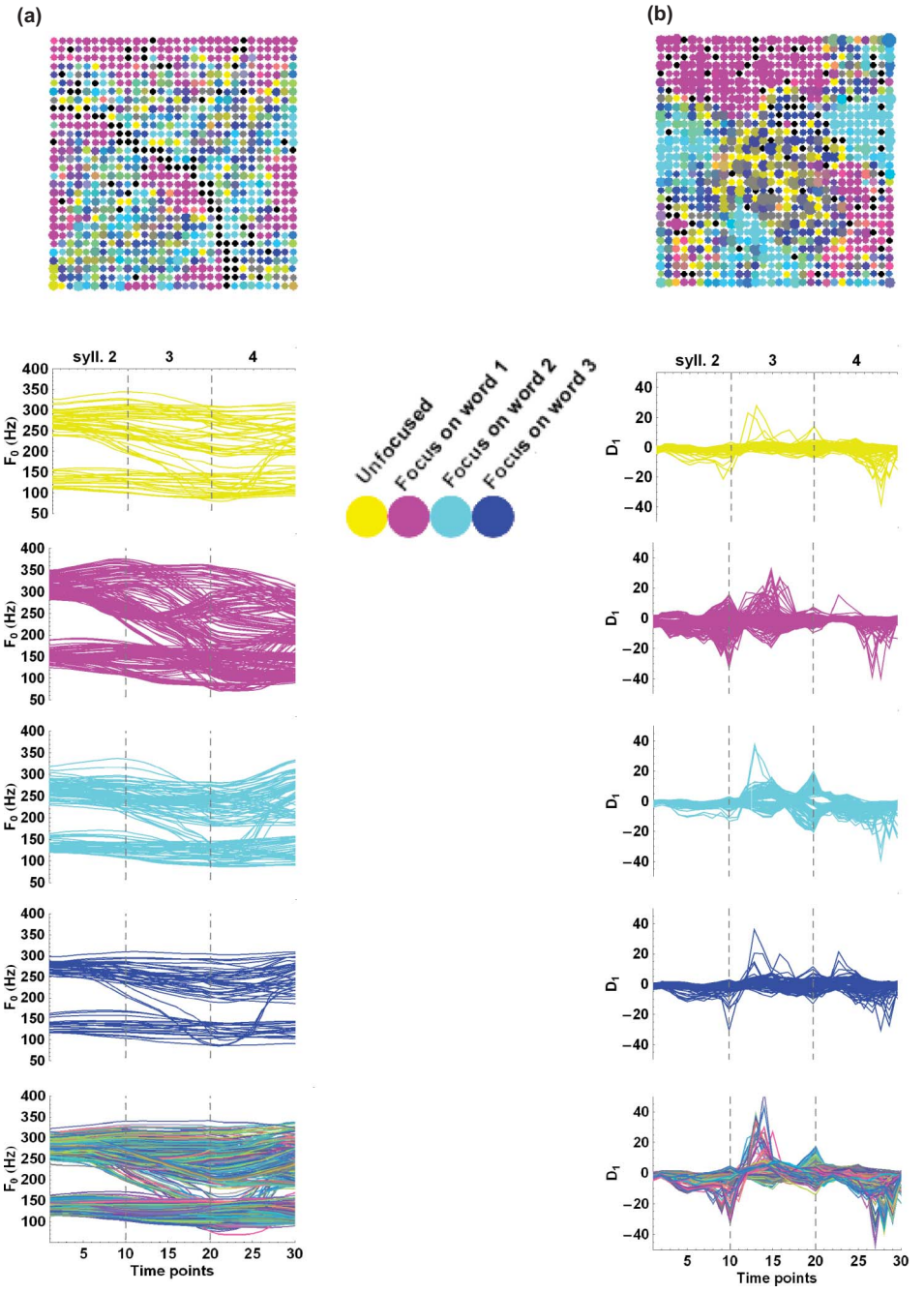

FIGURE 5 Detailed color and internal maps of Simulation 3 showing focus locations distinction by **(a)** F_0 and **(b)** D_1. (This figure appears in color online.)

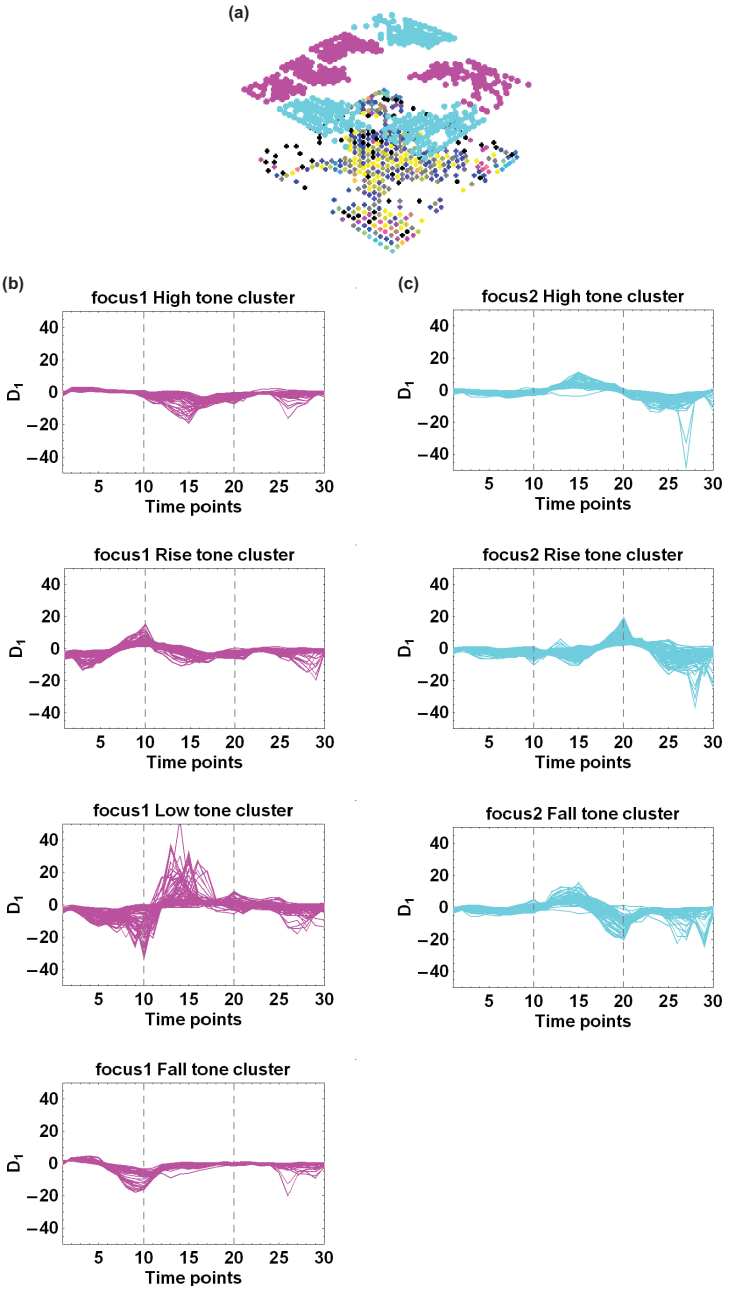

FIGURE 6 Color map (**a**) and internal maps of focus1 (**b**) and focus2 (**c**) clusters developed by the D_1 network in Simulation 3 as a function of tone. In (**a**), focus1 and focus2 clusters are lifted up for clarity. (This figure appears in color online.)

(Figures 4 and 5) and is shown in the similarity of the learned internal maps of these two focus categories.

The first two simulations show that F_0 is sufficient for carrying information about initial and medial focus. This suggests that F_0 probably carries important perceptual cues for focus, which agrees with the previous finding that focus is encoded mainly through pitch range adjustments (Xu, 1999). However, when speech with cross-gender variability was used to train and test the networks, as in Simulation 3, only the dynamic information represented by D_1 was useful for distinguishing focus satisfactorily. As described in the introduction, D_1 has the advantage of being largely free of individual speakers' pitch range differences, as can be seen in Figure 2. The pitch range differences become much more extensive when male and female data are both included in the input. As explained in Gauthier et al. (2007a, 2007b), taking the derivative of a curve to compute D_1 eliminates the constant term that specifies the curve's Y-intercept, which would automatically remove most of the cross-gender pitch register differences. D_1 is therefore more reflective than F_0 of what the speakers do during their articulation rather than who they are. In this sense, it is the part of the information in the acoustic signal that is articulatorily the most relevant and provides the best cues for the perception of both lexical tone (Gauthier et al., 2007a, 2007b) and prosodic focus (the present data). The results of Simulation 3 also show that when final focus was not substantially different from neutral focus in the production, its learned clustering was also not highly distinct from neutral focus, consistent with the observation that human listeners cannot perceive final focus as easily as nonfinal focus categories (Liu & Xu, 2005). Map units responding to focus1 and focus2 were clearly separated, however, and their respective internal maps (Figure 6b, c) revealed the similarity structure within each cluster.[5]

These results, together with the lexical tone results which we obtained previously (Gauthier et al., 2007a, 2007b), demonstrate that the D_1 profile of the fundamental frequency of the speech signal carries sufficient information to convey both focus and lexical tones simultaneously in a tone language. The fact that both initial and medial focus formed multiple clusters, as shown in the D_1 color map (Figure 5b), suggests that the simulated process may be considered as some type of initial input sorting that eventually would lead to proper focal categorization, which raises the question as to how the system can link these clusters to form such categories. With respect to whether tones can be learned independently from focus, the results obtained in our previous work show that lexical tones can be directly derived from syllable-sized D_1 profiles (Gauthier et al., 2007a, 2007b). In the present study, detailed analyses of the clusters revealed that each of them mainly responded to sentences for which the focused component was a single tone. Such patterning reflects the fact that the fundamental frequency is used for simultaneously encoding tones and focus. It also suggests that it is possible for infants to use sentence-sized patterns during lexical tone learning, by noticing the similarities between syllable-sized chunks of sentences, a strategy that is not much different from focusing directly on the syllable.

The breakdown of focus categories into multiple clusters related to tonal identity also suggests that focus may be learned with the help of prior knowledge about tones, by using some prototypical representation of tones from a separate "syllabic" layer of processing. Once the contribution of tones is removed, the "clause" or sentence layer may better represent focus categories in terms of the individual clusters that are grouped together. Thus, we could predict that young infants should be able to discriminate F_0 variations due to focus, but the ability to

[5]Quantitative clustering methods for the SOM are detailed in Vesanto & Alhoniemi (2000).

contrast different focus categories in the presence of varying tone should emerge only after tonal acquisition. Future studies using network simulations and infant perceptual experimentation need to specifically test this predicted order. This prediction is plausible given the tonal perceptual abilities shown in prelinguistic human infants (Mattock, 2004) and the near-perfect tonal categorization performance in our neural network simulations of tone learning (Gauthier et al., 2007a, 2007b). It is also consistent with theories of language acquisition such as the PRIMIR model (Werker & Curtin, 2005), according to which a basic perceptual level of representation first extracts regularities from the speech signal, achieves clustering based on input similarity, and then combines various learned structures for accomplishing higher-level acquisition tasks. The findings of the present study may have broader implications for understanding how infants exposed to a nontone language such as English may also benefit from D_1 in achieving phonetic encoding of prosodic focus, as complex interactions also exist in English between lexical stress, sentence type, and focus (Liu & Xu, 2007; Xu & Xu, 2005).

Admittedly, the speech material used in the present study does not resemble the kind of phonetic input infants typically receive, given that it is adult-directed speech produced in the laboratory, while infants are likely to receive a mixture of naturally produced adult- and infant-directed speech (ADS and IDS). On the other hand, as a first attempt, our study provides important insight into the likely mechanisms that the learning system with no built-in linguistic knowledge may rely on for discovering structures underlying specific phonetic categories. We have shown that with sentence-sized F_0 contours and velocity profiles as input, a naïve learning system can develop clusters that correspond well to focus categories intended by the speakers. Future research should directly test human infants using infant-directed speech. It is hard to predict whether IDS is better than ADS for learning focus, however. IDS is known to be characterized by higher pitch level, expanded pitch range, and wider frequency sweeps (Fernald & Mazzie, 1991; Fernald et al., 1989), which may mean that there is more variability in IDS that could make focus learning difficult. On the other hand, focus has been shown to have fixed pitch ranges that cannot be exceeded by further increase in the amount of emphasis (Chen, 2003), which means that focus could resist interference from other factors that also introduce pitch range variations.

Finally, the current results show that, although F_0 simultaneously carries information about tone, focus, and speaker gender, it is possible for a naïve system to develop focus-specific clusters if the right input is chosen by the learner. For example, the use of sentence-sized contours may lead to focus learning, whereas syllable-sized contours may lead to the development of tone-specific clusters. The reliance on F_0 contours may lead to the development of gender-specific clusters, while D_1 profiles may filter out the gender information and help develop the tonal or focus clusters. The learning system can potentially develop multiple structures (e.g., tone, focus, gender) by responding to one input characteristic at a time, as long as that characteristic shows clear patterns. If a particular characteristic supports more than one structure, it is still possible that the learner develops each structure, but some other constraints would be needed to organize the learned structures so that they can serve different functions. One question, which is not addressed in the present study, is how the learning system determines which generalizations are linguistically important. Future studies need to explore the factors that can constrain the learning system such that linguistically relevant principles are treated differently than nonlinguistic structures.

In summary, the present study shows that a simple unsupervised learning mechanism can develop focus-specific clusters from continuous dynamic speech signal produced by multiple

speakers in various lexical tone conditions. We found that the perceptual formation of focal clusters could be directly based on sentence-sized continuous F_0 contours and F_0 velocity profiles. However, in comparison with tone learning demonstrated in previous work (Gauthier et al., 2007a, 2007b), the present results suggest that learning focus may be more difficult, because focus exaggerates the pitch of the focused tones, making them either extra high or extra low, resulting in nonunimodal cluster distributions on a map. It awaits future research to find out how such complex distributions can be coherently represented to allow focus acquisition. Interestingly, as discussed earlier in this section, focal clusters in the present simulations show subdistribution of tones. In contrast, tonal clusters in our previous tone learning simulations (Gauthier et al., 2007b) are straightforward, showing no clear substructure of focus, nor any other discernable substructure. This difference indicates that tone learning is likely more primary than focus learning. Thus, we suggest that during the course of language acquisition, more complex linguistic knowledge such as focus might be acquired based on less-complex, lower-level learned knowledge such as tones.

ACKNOWLEDGMENTS

This work was supported by funding from SSHRC, NSERC, and FQRSC to the second author and supported in part by a NIH grant to the third author. We would like to thank Lucie Ménard for her helpful comments on an earlier draft of the paper. We also wish to thank LouAnn Gerken, Susan Goldin-Meadow, and the anonymous reviewers for their helpful comments on the paper.

REFERENCES

Allen, G. D., & Hawkins, S. (1980). Phonological rhythm: definition and development. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Vol. 1: Production* (pp. 227–256). New York: Academic Press.

Bahill, A. T., Kallman, J. S., & Lieberman, J. E. (1982). Frequency limitations of the two-point central difference differentiation algorithm. *Biological Cybernetics*, *45*, 1–4.

Behnke, K. (1998). *The acquisition of phonetic categories in young infants: A self-organizing artificial neural network approach*. Published doctoral dissertation, Universiteit Twente, Enschede, The Netherlands.

Chen, Y. (2003). The phonetics and phonology of contrastive focus in standard Chinese. Unpublished doctoral dissertation, State University of New York at Stony Brook.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, *77*, 2142–2156.

Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 113–121.

Dahan, D., & Bernard, J.-M. (1996). Interspeaker variability in emphatic accent production in French. *Language and Speech*, *39*, 341–374.

DeCasper, A. J., Lecanuet, J.-P., Busnel, M.-C., Granier-Deferre, C., & Maugeais, R. (1994). Fetal reactions to recurrent maternal speech. *Infant Behavior and Development*, *17*(2), 159–164.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, *97*, 491–504.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Bardies, B. D., & Fikui, I. (1989). A cross-language study of prosodic modification in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*, 477–501.

Gauthier, B., Shi, R., & Xu, Y. (2007a). Learning phonetic categories by tracking movements. *Cognition*, *103*(1), 80–106.

Gauthier, B., Shi, R., & Xu, Y. (2007b). Simulating the acquisition of lexical tones from continuous dynamic input. *Journal of the Acoustical Society of America*, *121*(5), EL190–EL195.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*(2), 1111–1121.

Haugen, E., & Joos, M. (1972). Tone and intonation in East Norwegian. In D. Bolinger (Ed.), *Intonation* (pp. 414–436). Harmondsworth, UK: Penguin Ltd.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111).

Hirsh-Pasek, K., Newlson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269–286.

Hornby, P. A., & Hass, W. A. (1970). Use of contrastive stress by preschool children. *Journal of Speech and Hearing Research*, *13*, 395–399.

Johnson, K., & Mullennix, J. W. (1997). *Talker variability in speech processing*. New York: Academic Press.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, *64*(3), 675–687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*(3), 402–420.

Jusczyk, P. W., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Kaplan, E. L. (1969). *The role of intonation in the acquisition of language*. Ithaca, NY: Cornell University.

Kohonen, T. (1989). Self-organization and associative memory. Berlin: Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*, 1–6.

Ladd, R. D. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Lecanuet, J.-P., & Granier-Deferre, C. (1993). Speech stimuli in the fetal environment. In B. deBoysson-Bardies, S. deSchonen, P. W. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 237–248). New York: Kluwer Academic/Plenum Publishers.

Lecanuet, J.-P., Granier-Deferre, C., DeCasper, A. J., Maugeais, R., Andrieu, A. J., & Busnel, M. C. (1987). Perception et discrimination fœtales de stimuli langagiers; mise en évidence à partir de la réactivité cardiaque; résultats préliminaires. *Comptes rendus de l'Académie des sciences. Série III, Sciences de la vie*, *305*(5), 161–164.

Lehiste, I., & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America*, *33*(4), 419–425.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese*. Berkeley: University of California Press.

Liu, F., & Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, *62*, 70–87.

Liu, F., & Xu, Y. (2007). Question intonation as affected by word stress and focus in English. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1189–1192). Saarbrücken, Germany: Universitat des Saarlandes.

Lœvenbruck, H., Vilain, C., Carota, F., Baciu, M., Abry, C., Lamalle, L., et al. (2007). Cerebral correlates of multimodal pointing: An fMRI study of prosodic focus, syntactive extraction, digital - and ocular - pointing. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1861–1864). Saarbrücken, Germany: Universitat Saarlandes.

Mattock, K. J. (2004). *Perceptual reorganisation for tone: Linguistic tone and non-linguistic pitch perception by English language and Chinese language infants.* Unpublished doctoral dissertation, University of Western Sydney, Sydney, Australia.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Ménard, L., Lœvenbruck, H., & Savariaux, C. (2006). Articulatory and acoustical correlates of contrastive focus in French: A developmental study. In J. Harrington & M. Tabain (Eds.), *Speech production: Models, phonetic processes and techniques* (pp. 227–251). New York: Psychology Press.

Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2004). The role of vocal tract morphology in speech development: Perceptual targets and sensori-motor maps for French synthesized vowels from birth to adulthood. *Journal of Speech, Language and Hearing research*, *47*(5), 1059–1080.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*(4), 495–500.

Morse, P. A. (1972). The discrimination of speech and nonspeech stimuli in early infancy. *Journal of Experimental Child Psychology*, *14*(3), 477–492.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 756–766.

Nazzi, T., Newson, D. G. K., Jusczyk, P. W., & Jusczyk, A. M. (2000). Six-month-olds' detection of clauses embedded in continuous speech: Effects of prosodic well-formedness. *Infancy*, *1*(1), 123–147.

Perkell, J. S., & Klatt, D. H. (1986). *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum.

Peters, A. M. (1997). Language typology, *prosody,* and the acquisition of grammatical morphemes. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition* (Vol. 5) (pp. 136–197). Hillsdale, NJ: Lawrence Erlbaum.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184.

Ritter, H., & Schulten, K. (1986). On the stationary state of Kohonen's self-organizing sensory mapping. *Biological Cybernetics*, *54*, 99–106.

Rump, H. H., & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech*, *39*, 1–17.

Shi, R. (1996). Perceptual correlates of content words and function words in early language input (Doctoral dissertation, Brown University, 1995). *Dissertation Abstracts International*, *56*, 3108A.

Shi, R. (2006). Basic syntactic categories in early language development: Evidence from neural network simulation and infant perceptual experiments. In L. Ping, E. Bates, L. Hai Tan, & O. Tseng (Eds.), *Handbook of East Asian psycholinguistics* (pp. 90–102). Cambridge, UK: Cambridge University Press.

Shi, R., Morgan, J., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language*, *25*, 169–201.

Slater, A. (1998). *Perceptual development: Visual, auditory, and speech perception in infancy*. London: Psychology Press.

Soderstrom, M., Nelson, D. G. K., & Jusczyk, P. W. (2005). Six-month-olds recognize clauses embedded in different passages of fluent speech. *Infant Behavior & Development*, *28*(1), 87–94.

Soderstrom, M., Seidl, A., Nelson, D. G. K., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *49*, 249–267.

Spring, D. R., & Dale, P. S. (1977). Discrimination of linguistic stress in early infancy. *Journal of Speech and Hearing Research*, *20*(2), 224–232.

Turk, A. E., & White, L. (1999). Structural influences on accentual lengthening. *Journal of Phonetics*, *27*, 171–206.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, *27*, 55–105.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*, 220–251.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*, 159–197.

APPENDIX

The SOM Algorithm

Architecture. The SOM maps a high-dimensional input space onto a discrete lower-dimensional array of topologically ordered processing units. A one-dimensional SOM is illustrated

in Figure 7 (adapted from Ritter & Schulten, 1986). The input and output layers are fully interconnected to each other. The output space N is a lattice on which units are labeled by a position vector *r* indicating their physical position on that lattice (filled dots). The input space X is mapped onto the output space N by a set of adaptive receptive field centers (empty dots), or connection weights $w_r \varepsilon$ X, for which correspond a typical $x_i \varepsilon$ X. The subset of X closer to a unit's receptive field center than to any other w_r constitutes the receptive field of that unit (dark bars). In the present study, a two-dimensional map of 30 × 30 (900) units is used.

Transmission rule. The transfer function of the network contains two steps. First, the distance between the receptive field center of each unit to that of the input vector x_i is evaluated according to:

$$u_r = (\Sigma\ (x_i\ -\ w_r)^2)^{1/2} \tag{2}$$

where u_r represents the net value of unit *r*. The unit with the shortest Euclidean distance between w_r and reference input pattern x_i is selected to be the winner according to:

$$v = \min\ (u_r) \tag{3}$$

where *v* corresponds to the position of the winner, or Best Matching Unit (BMU). The net value is further transformed to yield the final response, given by the nonlinear Gaussian function:

$$\eta r = \mathrm{Exp}(-\ ((v\ -\ r)^2\ /\ \sigma)) \tag{4}$$

where η_r corresponds to each unit's activation. The Gaussian is peaked at *v* so that the winning unit is the most activated ($\eta_v = 1$). Units falling into neighborhood radius σ get activated by means of lateral activity, although to a lesser degree than the BMU depending on their position relative to the winner. The transmission rule can be conceived as a basic perceptual discriminative function that computes the distance between a perceived signal and a signal stored in a list of prototypes.

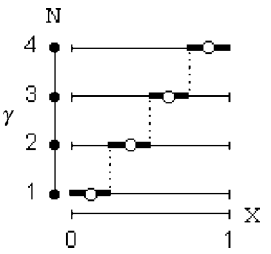

FIGURE 7 Architecture of a one-dimensional SOM: linear array N of 4 output units *r* (filled dots), their receptive field centers (empty dots) and receptive fields (bold horizontal lines) for input space X = [0,1] (adapted from Ritter & Schulten, 1986).

Learning rule. The SOM implements a regression algorithm for mapping an input distribution P(x), $x_i \varepsilon$ X, onto the output space. The lateral connections between output space nodes allow for topological ordering to be preserved in the map during the learning period. Receptive field centers w_r are adapted during a stochastic learning procedure in which a random sequence of data points x_i is presented repeatedly for a predefined number of times. Each time an input vector is presented, the winning unit and its neighbors shift their receptive field centers toward the data point according to:

$$\Delta w_r = \alpha \cdot \eta_r (x_i - w_r) \tag{5}$$

where η_r is the value output according to the transmission rule and α is the learning step size. The weight matrix is then updated according to:

$$w_r (t + 1) = w_r (t) + \Delta w_r \tag{6}$$

The learning rule can be conceived as a basic perceptual learning function that transforms the internal organization to reflect the environmental characteristics.

Initialization of the map. Before the learning phase, the weight matrix is initialized by randomly assigning a value to the weight vector's elements that ranges between the minimum and maximum Hertz (or D_1) value of the input corpus.