A Unified Model of Early Word Learning: Integrating Statistical and Social Cues

Chen Yu
Department of Psychology and
Cognitive Science Program
Indiana University
Bloomington, IN, 47405
chenyu@indiana.edu

Dana H. Ballard
Department of Computer Science
University of Rochester
Rochester, NY, 14627
dana@cs.rochester.edu

Abstract

Previous work on early language acquisition has shown that word meanings can be acquired by an associative procedure that maps perceptual experience onto linguistic labels based on cross-situational observation. A new trend termed social-pragmatic theory [27] focuses on the effect of the child's social-cognitive capacities, such as joint attention and intention reading. This paper argues that statistical and social cues can be seamlessly integrated to facilitate early word learning. To support this idea, we first introduce a statistical learning mechanism that provides a formal account of cross-situational observation. The main part of this paper then presents a unified model that is able to make use of different kinds of social cues, such as joint attention and prosody in maternal speech, in the statistical learning framework. In a computational analysis of infant data, the quantitative results of our unified model outperforms the purely statistical learning method in computing word-meaning associations.

1 Introduction

What kinds of learning mechanisms underlie language acquisition? One of the central debates concerns whether the innate or environmental contribution plays a vital role in language development. Learning-oriented theories believe that language is learned and the child's environment plays a crucial role [14, 11, 24]. There is growing evidence that babies do possess powerful statistical learning mechanisms [23]. On the other hand, a nativist view sees linguistic universals as a product of the child's linguistic endowment and suggests that they do not need to be learned, which provide an elegant explanation for cross-linguistic similarities between different human languages [10].

In this paper, we first review two theories of language learning: statistical learning theory and social-pragmatic theory. Then Section 3 proposes our unified model that integrates statistical and social cues in a general system. Section 4 describes the implementation of the statistical learning model of word meaning, which provides a probabilistic framework for further study. Section 5 presents the methods to extract prosodic cues from raw speech and joint attention cues from infant-caregiver interactions. Section 6 provides a comparative study of different methods considering different sets of statistical and social cues.

2 Two Theories of Language Learning

This section reviews two well-known learning-oriented theories of language acquisition. The theory of statistical learning suggests that language acquisition is a statistically driven process in which young language learners utilize the lexical content and syntactic structure of speech as well as non-linguistic contextual information as input to compute distributional statistics. The social-pragmatic theory focuses on mind reading (social cognition) as fundamental to the word learning process. Both theories have been supported by various empirical and computational studies.

2.1 The Theory of Statistical Learning

Human language learners possess powerful statistical learning capacities. That is, the cognitive system in both children and adults is sensitive to features of the input (e.g., occurrence statistics). Saffran, Aslin and Newport [23] showed that 8-month-old infants are able to find word boundaries in an artificial language only based on statistical regularities. Later studies [22] demonstrated that infants are also sensitive to transitional probabilities over tone sequences, suggesting that this statistical learning mechanism is more general than the one dedicated solely to processing linguistic data. Furthermore, statistical language learning includes not only statistical computations to identity words in speech but also algebraic-like computations to learn grammatical structures [18].

In the study of word learning, *associationism* claims that word acquisition is based on statistical learning of

co-occurring data from the linguistic modality and non-linguistic context (see a review by [19]). Richards and Goldfarb [21] proposed that children come to know the meaning of a word through repeatedly associating the verbal label with their experience at the time that the label is used. Smith [25] argued that word learning is initially a process in which children's attention is captured by objects or actions that are the most salient in their environment, and then used to associate those objects or actions with acoustic patterns voiced by an adult. Plunkett [19] developed a connectionist model of vocabulary development to associate preprocessed images and linguistic labels. The linguistic behaviors of the network can mimic the well-known vocabulary spurt based on small continuous changes in the connection strengths with and across different processing modalities in the network. In general, the statistical and associative mechanism of word learning divides the word learning task into three subtasks: word discovery, meaning discovery and word-meaning association. The vital part is to use multiple word-meaning pairs collected from different situations to compute co-occurrences and then establish word-to-world mappings [14].

2.2 The Social-Pragmatic Theory

The social-pragmatic theory of language acquisition argued that the major sources of constraints in language acquisition are social cognitive skills, such as children's ability to infer the intentions of adults as adults act and speak to them [1, 27, 7]. These kinds of social cognition are called "mind reading" by Baron-Cohen [4]. Kuhl et al. [16] studied whether phonetic learning of 9-10 month children is simply triggered by hearing language. If so, children should be able to learn by being exposed to language materials via digital video without human interaction. However, the results showed that infants cannot learn phonetics through this way, suggesting that the presence of a live person provides not only social cues but also referential information. Butterworth [9] showed that even by 6 months of age, infants demonstrate sensitivities to social cues, such as monitoring and following another person's gaze. In Baldwin's work [1], the 18-month old infant heard the novel word while his/her attention was focused on one toy and the experimenter looked at another toy. When children heard the same word in a testing phase, they chose the object at which the experimenter had been looking. This suggested that the infants were able to follow the speaker's attention and infer the mental state of the speaker to determine the referent of the novel word. Furthermore, Baldwin et al.[2] proposed that infants give a special weight to the cues of indexing the speaker's referential intent when determining the reference of a novel label. Their experiments showed that infants established a stable link between the novel label and the target toy only when that label was uttered by a speaker who concurrently showed his attention toward the target, and such a stable mapping was not established when the label was uttered by a speaker who showed no signs of attention to the target toy, even if the object appeared at the same moment when that label was uttered and the speaker was touching the object. In addition, their results suggested that children not only attend to referential intentions of a speaker but also actively look for the intention of the speaker when determining whether to associate a novel word with an object.

3 A Unified Model

Bloom [6] argued that children's conceptual biases, intentional understanding and syntactic knowledge are not only necessary for word learning but that they are also sufficient. This claim contrasts with the theory that word learning is based on an associative learning mechanism that is sensitive to statistical properties of the input [19]. The statistical and associative theory suggested that the child's sensitivity to spatio-temporal contiguity is sufficient for word learning, as postulated by associationist models of language acquisition with support by computational implementation [11, 20]. The debate on these two theories has been going on for several years.

Associative learning mechanisms make sense because words are typically spoken at the moment when the child looks at the things that those words refer to. In western cultures, parents provide linguistic labels of objects for their kids when the objects are in the kids' visual fields. Thus, no one doubts that humans can learn co-occurrence relationships and that the easiest way to teach language is to provide linguistic labels at the same time that children focus on them. However, parents do not carefully name objects for their kids in many cultures. Even in western cultures, words are not always used at the moment that their referents are perceived. For instance, Gleitman [13] showed that most of the time, the child does not observe something being opened when the verb "open" is used. Nevertheless, children have no difficulty in learning those words. Associative learning, without some further constraints or additional information, cannot explain this observation.

The theory of mind reading is able to explain many phenomena from the perspective of the inference of a speaker's referential intentions, especially for the cases that words and the corresponding meanings are not co-occurring, or words are temporally correlated with irrelevant meanings. However, the environment in which infants develop does contain the information that is useful for statistical learning mechanisms. Meanwhile, empirical studies (e.g. [23] and [18]) showed that infants can utilize the statistical properties of the input in language acquisition. Taken together, it is very plausible that infants perform statistical computations in language learning.

Fortunately, the theory of statistical learning and social-

pragmatic theory are not mutually exclusive. Recently, Hirsh-Pasek, Golinkoff and Hollich [15] proposed a coalition model in which multiple sources, such as perceptual salience, prosodic cue, social eye gaze, social context, syntactic cues and temporal contiguity, are used by children to learn new words. They argued that during the development, the weighting of the cues changes over time while younger children can just detect and make use of only a subset of the cues in the coalition and the older can use a wider subset of cues.

The purpose of this study is to show quantitatively the effects of statistical cross-situational observation and social cues through computational modeling. In early word learning, children need to start by pairing spoken words with the co-occurring possible referents, collecting multiple such pairs, and then figuring out the common elements. Although no one doubts this process, few research has addressed the details of cross-situational observation. This work first introduces a formal model of statistical word learning which provides a probabilistic framework for encoding multiple sources of information. Given multiple scenes paired with spoken words collected from natural interactions between caregivers and their kids, the model is able to compute the association probabilities of all the possible word-meaning pairs. Moreover, we argue that social cues can be naturally integrated in the model as additional constraints in computation. The claim here is that language learners can use social cues, such as gaze direction, head direction, body movement, gesture, intonation of speech and facial expression, to infer speakers' referential intentions. We show how these social cues can be seamlessly integrated in the framework of statistical learning and facilitate word learning. Specifically, we focus on two kinds of social cues: body movement cues indicating the speaker's attention and prosodic cues in speech. This study proposes that those social cues can play a spotlight role (shown in Figure 1) in the statistical learning by causing language learners to focus on certain aspects of a scene. Since every scene is ambiguous and contains multiple possible referents, this spotlight function is crucial in solving the word-to-world mapping problem. The following subsections discuss how those cues might help in detail.

3.1 The Role of Body Movement in Language Acquisition

Ballard et al. [3] argued that at time scales of approximately one-third of a second, orienting movements of the body play a crucial role in cognition and form a useful computational level, termed the embodiment level. At this level, the constraints of the body determine the nature of cognitive operations. This computation provides a language that links external sensory data with internal cognitive programs and motor actions through a system of implicit reference termed deictic, whereby pointing movements of the body are used to bind objects in the world to cognitive programs. Examples of sensorimotor primitives at the embodiment level include an eye movement, a hand movement, or a spoken word.

We apply the theory of embodied cognition in the context of language learning. To do so, one needs to consider the role of embodiment from both the perspective of a speaker (language teacher) and that of a language learner. In the study of speech production, Meyer et al. [17] found that speakers' eye movements were tightly linked to their speech output. When speakers were asked to describe a set of objects from a picture, they usually looked at each new object before mentioning it, and their gazes remained on the object until they were about to say the last word about it. From the perspective of a language learner, Baldwin [1] showed that infants actively gathered social information to guide their inferences about word meanings and they systematically checked the speaker's gaze to clarify his/her reference. In the follow-up studies, Baldwin and Baird [2] proposed that humans gradually develop the skill of mind reading so that ultimately they care little about the surface behaviors of others' dynamic action but focus on discerning underlying intentions based on a generative knowledge system. Summarizing all these ideas on embodied cognition, speech production and social development, the speakers' body movements, such as eye movements, head movements and hand movements, can reveal their referential intentions in verbal utterances, which, in turn almost certainly could possibly play a significant role in early language development [29]. A plausible starting point of learning the meanings of words is the deployment of speakers' intentional body movements to infer their referential intentions. To support this idea, we provide a formal account of how the intentions derived from body movements, which we term *embodied intention*, facilitate the early stage of vocabulary acquisition. We argue that infants learn words through their sensitivity to others' intentional body movements in a very

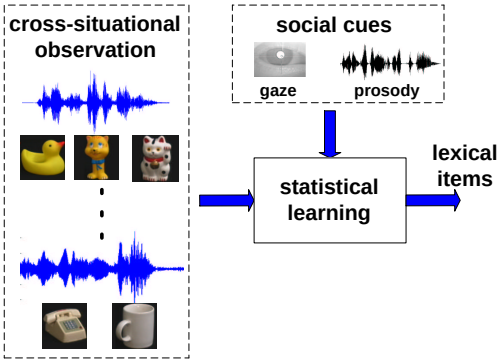

Figure 1. Cross-situational observation and social cues can be seamlessly integrated in a statistical learning model.

specific way: They use temporal synchrony between speech and referential body movements to find the referents of language.

3.2 The Role of Prosodic Cue

When talking to human infants, parents use vocal patterns that are different from normal conversation. They speak slowly and with higher pitch and exaggerated intonation contours. Fernald [12] proposed a model consisting of four developmental functions of intonation in speech to infants. The first function is that infants are attentive to intrinsic perceptual and affective salience in the melodic intonation of mothers' speech. At the second level, the exaggerated intonation patterns of mothers' speech would influence both attentional preference and affective responsiveness of infants. The third function is about the inference of the communicative intents of speakers from maternal intonation of speech. Infants are able to interpret the emotional states of others and make predictions about the future actions of others using information available in vocal and facial expressions, which provide reliable cues to the affective state and intentions of speakers. The fourth level focuses on the role of prosodic cues in early language development. Fernald argued that the prosody of speech helps to identify linguistic units within the continuous speech signal. Thus it serves as an attention-focusing device so that mothers use a distinctive prosodic strategy to highlight focused words. Most often, exaggerated pitch peaks are correlated with lexical stress. In light of this, we investigate the role of prosodic cue in early word learning in this paper. Specifically, we focus on the spotlight function of prosody and provide a formal account of how prosodic cues might be used in word learning.

4 A Statistical Model of Cross-Situational Observation

Our study uses the video clips of mother-infant interactions from the CHILDES standard database. These clips contain simultaneous audio and video data wherein a mother introduces her child to a succession of toys stored in a nearby box.

In this kind of natural interaction, the vocabulary is rich and varied and the central items (toy names) are far from the most frequent words. This complex but perfectly natural situation can be easily quantified by plotting a histogram of word frequency which shows that none of the key words – toy names make it into the top 15 items of the list. An elementary idea for improving the ranking of key words assumes that the infants are able to weight the toy utterances more by taking advantage of the approximately coincident body cues. For instance, the utterances that were generated when the infant's gaze was fixated on the toys by following the mother's gaze have more weights than the ones the young child just looked around while not paying attention to what the mother said. We examined the transcript and weighted the words according to how much they were emphasized by such cues, but this strategy does little to help spot the toy names.

Next, we manually labeled visual objects in the context when a spoken utterance was produced, and found what *is* helpful is to partition the toy sequences (contextual information when the speech was produced) into intervals where within each interval a single toy or small number of co-occurring toys is the central subject or meaning, and then categorize spoken utterances using the contextual bins labeled by different toys. The hypothesis is that mothers use temporal synchrony to highlight novel word-referent relations for young infants. That is, presenting information across multiple modalities simultaneously serves to highlight the relations between the two patterns of stimulation. Thus, temporal synchrony can facilitate infants' detection of word-referent relations. Formally, associating meanings (toys, etc.) with words (toy names, etc.) can be viewed as the problem of identifying word correspondences between English and a "meaning language", given the data of these two languages in parallel. With this perspective, a technique from machine translation can address the correspondence problem [8]. The probability of each word is expressed as a mixture model that consists of the conditional probabilities of each word given its possible meanings. In this way, an Expectation-Maximization (EM) algorithm can find the reliable associations of object names and their meanings which will maximize the likelihood function of observing the data set.

The general setting is as follows: suppose we have a word set $X = \{w_1, w_2, ..., w_N\}$ and a meaning set $Y = \{m_1, m_2, ..., m_M\}$, where N is the number of words and M is the number of meanings (toys, etc.). Let S be the number of spoken utterances. All word data are in a set $\chi = \{(S_w^{(s)}, S_m^{(s)}), 1 \leq s \leq S\}$, where each spoken utterance $S_w^{(s)}$ consists of r words $w_{u(1)}, w_{u(2)}, ..., w_{u(r)}$, and $u(i)$ can be selected from 1 to N. Similarly, the corresponding contextual information $S_m^{(s)}$ include l possible meanings $m_{v(1)}, m_{v(2)}, ..., m_{v(l)}$ and the value of $v(j)$ is from 1 to M. Assume that every word w_n can be associated with a meaning m_m. Given a data set χ, We use the machine translation method proposed by Brown et al. [8] to maximize the likelihood of generating the meaning strings given English descriptions:

$$\begin{aligned} & P(S_m^{(1)}, S_m^{(2)}, ..., S_m^{(S)} | S_w^{(1)}, S_w^{(2)}, ..., S_w^{(S)}) \\ = & \prod_{s=1}^{S} \sum_{a} p(S_m^{(s)}, a | S_w^{(s)}) \\ = & \prod_{s=1}^{S} \frac{\epsilon}{(r+1)^l} \prod_{j=1}^{l} \sum_{i=0}^{r} p(m_{v(j)} | w_{u(i)}) \end{aligned} \quad (1)$$

where the alignment a indicates which word is aligned with which meaning. $p(m_{v(j)}|w_{u(i)})$ is the association probability for a word-meaning pair and ϵ is a small constant. The expected number of times that any particular word w_n in a language string $S_w^{(s)}$ generates any specific meaning m_m in the co-occurring meaning string $S_m^{(s)}$ is given by

$$c(m_m|w_n, S_m^{(s)}, S_w^{(s)}) = \frac{p(m_m|w_n)}{p(m_m|w_{u(1)}) + ... + p(m_m|w_{u(r)})} \times \sum_{j=1}^{l} \delta(m_m, v(j)) \sum_{i=1}^{r} \delta(w_n, u(i)) \quad (2)$$

where δ is equal to one when both of its arguments are the same and equal to zero otherwise. Accordingly, the association probabilities are given by

$$p(m_m|w_n) = \frac{\sum_{s=1}^{S} c(m_m|w_n, S_m^{(s)}, S_w^{(s)})}{\sum_{m=1}^{M} \sum_{s=1}^{S} c(m_m|w_n, S_m^{(s)}, S_w^{(s)})} \quad (3)$$

The method sets an initial $p(m_m|w_n)$ to be flat distribution, and then successively compute the counts of all word-meaning pairs $c(m_m|w_n, S_m^{(s)}, S_w^{(s)})$ using Equation (2) and the association probabilities using Equation (3). The technical details of our method can be found in [28]. The results of this statistical learning model are reported in Section 6.

5 The Integration of Social Cues in Statistical Learning

The communication of infants and their caregivers is multisensory. It involves visual information, tactile information as well as auditory information. Besides linguistic information, we believe that social cues encoded in multimodal interaction highlight target word-referent relations for young language learners. In a bidirectional relationship between maternal multimodal communication styles and infants' perception of word-referent relations, mothers synchronize their verbal references and nonverbal body movements (eye gaze, gesture, etc.) for infants. At the same time, infants are able to rely on observing mother's eye gaze and other pointing motions to detect their's referential intentions in speech. Thus, both mothers and infants actively involve into multimodal communication to solve the mapping problem in lexical acquisition. This study provides a quantitative account of how those multimodal social cues can facilitate word learning. Specifically, we focus on two cues: joint attention cues as deictic reference and prosodic cues in maternal speech.

5.1 Visual Spotlight

Children as young as 12-18 months spontaneously check where a speaker is looking when he/she utters a word, and then link the word with the object the speaker is looking at. This observation indicates that joint visual attention (deictic gaze) is a critical factor that should be considered in word learning. When presenting information, that visual spotlight gives maximal processing to that part of the visual field. During natural infant-caregiver interactions, joint visual attention involves detecting a spotlight of a mother's attention to the object in the scene, and then moving the body, head and eyes to acquire the target object with high-resolution focal vision, which is one of the crucial steps to deal with the mapping problem.

transcriptions	attended objects	other objects
– the kitty-cat go meow meow	kitty-cat	baby, big-bird, rattle, book
– ah and a baby	baby	kitty-cat, big-bird, rattle, book
– there's a baby just like my David	baby	kitty-cat, big-bird, rattle, book
– a baby	baby	kitty-cat, big-bird, rattle, book
that's a nice book	book	kitty-cat, big-bird

Table 1. Examples of transcriptions and contextual labels.

In our experiment, we coded visual contexts to study the role of joint attention. As shown in Table 1, we provided two labels to describe visual contextual information for each spoken utterance. One label indicated the objects of joint attention which were attending by both the mother and the kid. The second label represented all the other objects in the visual field of the kid. Figure 3 illustrates two examples of speech-scene pairs in which the shaded meanings are attentional objects and non-shaded meanings are other objects in the scene. In Section 5.3, we describe our method that makes use of this attentional information in word learning.

5.2 Prosodic Spotlight

Snedeker and Trueswell [26] showed that speakers produce and listeners use prosodic cues to disambiguate alternative meanings of a syntactic phrase in a referential communication task. Moreover, previous research suggests that mothers adapt their verbal communication to infants in order to facilitate their language learning. In this work, we analyze maternal speech by extracting low-level acoustic features and using those features to spot the words emphasized by adults. We proposed that perceptually salient prosodic patterns may serve as "spotlights" on linguistic information conveyed by speech. Thus, we focus on the role of prosodic features in word learning, which might help language learners to identify key words from the speech stream.

Fernald [12] suggested that the exaggerated acoustic patterns have evolved to elicit and sustain infants' attention to speech as well as highlight the important parts of the speech stream. In the context of word learning, we observe that prosodically salient words in maternal speech can be categorized into two classes. One group of words serve as communication of intention and emotion. One important role of those words is to attract the kid's attention so that the child would follow what the mother talks about and what she looks at. In this way, both the mother and the language learner share the visual attention, which is a cornerstone in social and language development. The right column in Figure 2 illustrates an example in the video clips in which the mother used high pitch to say *you* to attract the kid's attention. Some other common words and phrases frequently used by the mother are *yeah*, *oh*, *look* and *that's*. The other group of words contain the most important linguistic information that the mother wants to convey. In the context of word learning, most of those words refer to the concepts that are related to visual objects in the physical environment, such as object names, their colors, sizes and functions. An example of words in the second group is the object name *baby* shown in the left column of Figure 2.

In implementation, CMU sphinx speech recognition system was used to align maternal speech and transcriptions. As a result, the timestamps of the beginning and end of each spoken word was extracted. Next, we made three kinds of low-level acoustic measurements on each utterance and word. The prosodic features were extracted based on pitch (f0) information. For each feature, we extracted the values over both an utterance and each word within this utterance.

- **75 percentile pitch** p_{75}: the 75 percentile pitch value of all voiced part of the speech unit.
- **Delta pitch range** p_r: the change in pitch between frames (20ms) was calculated as delta pitch. This measure represents the difference between the highest and the lowest delta pitch values within the unit (utterance or word).
- **Mean delta pitch** p_m: the mean delta pitch of the voiced part of the spoken unit.

We want to obtain prosodically highlighted words in each spoken utterance. To do so, we compare the extracted features from each word with those from each utterance, which indicates whether a word sounds like "highlighted" in the acoustic context. Specifically, for the word w_i in the spoken utterance u_j, we form a feature vector: $[p_{75}^{w_i} - p_{75}^{u_j} \quad p_r^{w_i} - p_r^{u_j} \quad p_m^{w_i} - p_m^{u_j}]^T$, where $p_m^{u_j}$ is the mean delta pitch of the utterance and $p_m^{w_i}$ is that of the word and so on. In this way, the prosodic envelope of a word is represented by 3-dimensional feature vector. We use the support vector clustering (SVC) method [5] to group data point into

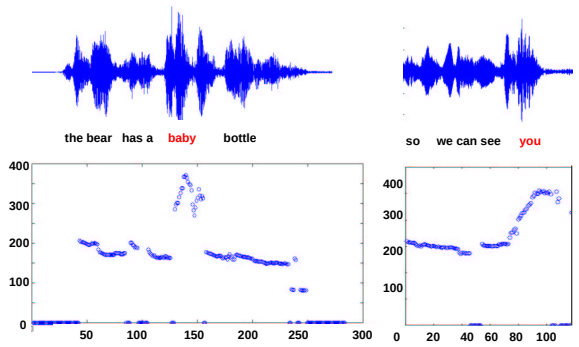

Figure 2. Speech and intonation. The prosodic cues highlight several words. The first column represents speech signals and the second column shows the profiles of fundamental frequency (f0). The word *baby* is highlighted in the left utterance and the word *you* is prosodically distinctive from others in the right utterance.

two categories. One consists of prosodically salient words and the other one includes non-emphasized words. In SVC algorithm, data points are mapped from the data space to a high dimensional feature space using a Gaussian kernel. In this feature space, the algorithm looks for the smallest sphere that encloses the data, and then maps the data points back to the data space and forms a set of contours to enclose them. These contours can be interpreted as cluster boundaries.

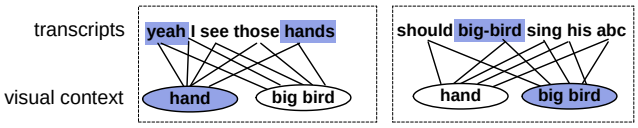

Figure 3. Cross-situational word-meaning association with social cues. The prosodic cues highlight some words in speech and the cues of joint attention highlight attentional objects in visual contexts.

5.3 Modeling the Role of Social Cues in Statistical Learning

We encode social cues in the framework of the statistical learning model as shown in Figure 3. Each word $u(i)$ is assigned with a weight $w_p(i)$ based on its prosodic category. Similarly, each visual object v_j is set with a weight $w_v(j)$ based on whether it is attended by the speaker and the learner. In this way, the same method described in previous section is applied and the only difference is that the estimate of $c(m_m|w_n, S_m^{(s)}, S_w^{(s)})$ now is given by:

$$c(m_m|w_n, S_m^{(s)}, S_w^{(s)}) = \frac{p(m_m|w_n)}{p(m_m|w_{u(1)}) + ... + p(m_m|w_{u(r)})}$$

$$\times \sum_{j=1}^{l} \delta(m_m, v(j) * w_v(j)) \sum_{i=1}^{r} \delta(w_n, u(i) * w_p(i)) \quad (4)$$

In practice, we set the values of $w_v(j)$ and $w_p(i)$ to be 3 for highlighted objects and words. The weights of all the other words and objects are set to be 1.

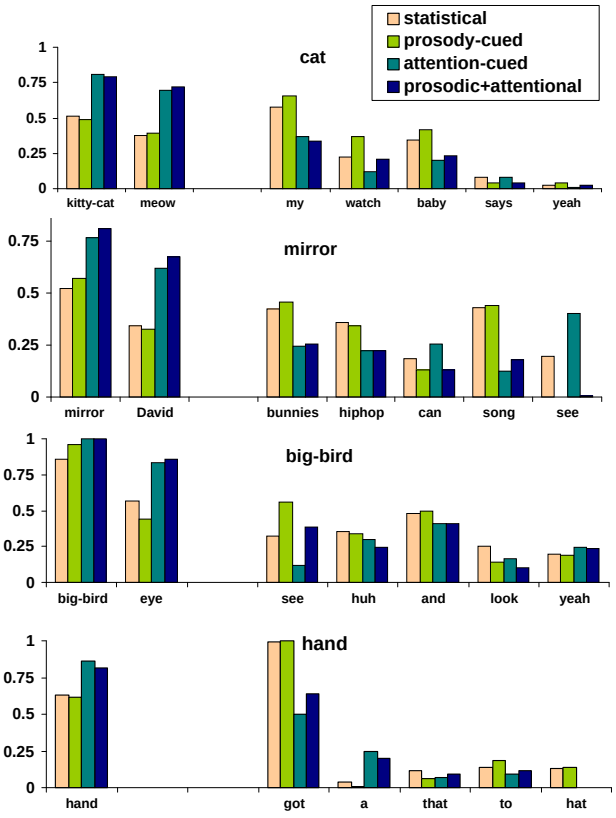

Figure 4. The comparative results of the methods considering different sets of cues. Each plot shows the association probabilities of several words to one specific meaning labeled on the top. The first one or two items are correct words that are relevant to the meanings and the following words are irrelevant.

6 Experimental Results

Our model was evaluated by using two video clips from CHILDES database. We labeled visual contexts in terms of 12 objects that occurred in the video clips. For each object, we selected the correctly associated words based on general knowledge. For instance, both the word *kitty-cat* and *meow* are positive instances because both of them are relevant to the object "cat". Overall, there were 26 positive words for all of the 12 objects. The computational model estimated the association probabilities of all the possible word-meaning associations and then selected lexical items based on a threshold. Two measures were used to evaluate the performance: (1) word-meaning association accuracy (precision) measures the percentage of the words spotted by the model which actually are correct. (2) lexical spotting accuracy (recall) measures the percentage of correct words that the model learned among all the 26 words.

Four methods were applied on the same data and the results are as follows (precision and recall): (1) purely statistical learning (75% and 58%). (2) statisical learning with prosodic cues (78% and 58%). (3) statistical learning with the cues from visual attention (80% and 73%). (4) statistical learning with both attentional and prosodic cues (83% and 77%). Figure 4 shows the comparative results of these four approaches for specific instances. Ideally, we want the association probabilities of the first or second words to be high and others to be low. For instance, the first plot represents the meaning of the object "cat". Both the spoken word *kitty-cat* and the spoken word *meow* are closely relevant to this meaning. Therefore, the association probabilities are high for these two words and are low for all the others words, such as *my*, *watch* and *baby*, which are not correlated with this context. Note that in the meaning of the object "bird", we count the word *eye* as a positive one because the mother uttered it several times during the interaction when she presented the object "bird" to her kid. Similarly, when she introduced the object "mirror", she also mentioned the name of the kid *David* whose face appeared in the mirror.

The results of the statistical learning approach (the first bars) are reasonably good. For instance, it obtains *big-bird* and *eye* for the meaning bird, *kitty-cat* for the meaning "cat", *mirror* for the meaning "mirror" and *hand* for the meaning "hand". But it also makes wrong estimates, such as *my* for the meaning "cat" and *got* for the meaning "hand". We expect that attentional and prosodic constraints will make the association probabilities of correct words higher and decrease the association probabilities of irrelevant words. The method encoding prosodic cues moves toward this goal although occasionally it changes the probabilities on the reverse way, such as increasing the probability of *my* in the meaning "cat". What is really helpful is to encode the cues of joint attention. The attention-cued method significantly improves the accuracy of estimate for almost every word-meaning pairs. Of course, the method including both joint-attention and prosodic cues achieves the best performance. Compared with purely statistical learning, this method highlights the correct associations (e.g., *kitty-cat* with the meaning "cat"), and decreases the irrelevant associations, such as *got* with the meaning "hand". In this method, we can simply select a threshold and pick the word-meaning pairs which are overlapped with the majority of words in the target set. We need to point out that the results here are obtained from very limited data. Without any prior knowledge of the language (the worst case in word learning), the model is able to learn a significant amount of correct word-meaning associations.

7 Conclusion

We believe that in a natural infant-caregiver interaction, the mother provides non-linguistic signals to the infant through her body movements, the direction of her gaze, and the timing of her affective cues via prosody. Previous experiments have shown that some of these non-linguistic signals can play a critical role in infant word learning, but a detailed estimate of their relative weights has not been provided. Based on statistical learning and social-pragmatic theories, this work proposed a unified model of early word learning, which integrates statistical and social cues to enable the word-learning process to function effectively and efficiently. In our model, we explored the computational role of non-linguistic information, such as joint attention and prosody in speech, and provided the quantitative results to compare the effects of different statistical and social cues. We need to point out that the current unified model does not encode any syntactic properties of the language, which definitely play a significant role in word learning, especially in the later stage. Therefore, one natural extension of the current work is to add the syntactic constraints in the current probabilistic framework to study how this knowledge can help the lexical acquisition process and how multiple sources can be integrated in a general system.

References

[1] D. Baldwin. Early referential understanding: Infant's ability to recognize referential acts for what they are. *Developmental psychology*, 29:832–843, 1993.

[2] D. A. Baldwin and J. A. Baird. Discerning intentions in dynamic human action. *Trends in Cognitive Sciences*, 5(4), 2001.

[3] D. H. Ballard, M. M. Hayhoe, P. K. Pook, and R. P. N. Rao. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20:1311–1328, 1997.

[4] S. Baron-Cohen. *Mindblindness: an essay on autism and theory of mind*. MIT Press, Cambridge, 1995.

[5] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of machine Learning Research*, 2:125–137, 2001.

[6] P. Bloom. Intentionality and word learning. *Trends in Cognitive Sciences*, 1(1):9–12, 1997.

[7] P. Bloom. *How children learn the meanings of words*. The MIT Press, Cambridge, MA, 2000.

[8] P. F. Brown, S. Pietra, V. Pietra, and R. L. Mercer. The mathematics of statistical machine translation:parameter estimation. *Computational Linguistics*, 19(2), 1993.

[9] G. Butterworth. The ontogeny and phylogeny of joint visual attention. In A. Whiten, editor, *Natural theories of mind: Evolution, development, and simulation of everyday mindreading*. Blackwell, Oxford, England, 1991.

[10] N. Chomsky. *Aspects of the theory of syntax*. MIT Press, 1965.

[11] J. Elman, E. Bates, M. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. *Rethinking innateness: A connectionist perspective on development*. MIT Press, 1996.

[12] A. Fernard. Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In *The Adaptive Mind*. Oxford University Press, 1992.

[13] L. Gleitman. The structural sources of verb meanings. *Language Acquisition*, 1:1–55, 1990.

[14] L. Gleitman, K. Cassidy, R. Nappa, A. Papafragou, and J. Trueswell. Hard words. *Language Learning and Development*, 1, in press.

[15] K. Hirsh-Pasek, R. M. Golinkoff, and G. Hollich. An emergentist coalition model for word learning: mapping words to objects is a product of the interaction of multiple cues. In *Becoming a word learner: a debate on lexical acquisition*. Oxford Press, New York, 2000.

[16] P. K. Kuhl, F.-M. Tsao, and H.-M. Liu. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *PNAS*, 100(15):9096–9101, July 22 2003.

[17] A. S. Meyer, A. M. Sleiderink, and W. J. Levelt. Viewing and naming objects: eye movements during noun phrase production. *Cognition*, 66:B25–B33, 1998.

[18] M. Pena, L. L. Bonatti, marina Nespor, and J. Mehler. Signal-driven computations in speech processing. *Science*, 2002.

[19] K. Plunkett. Theories of early language acquisition. *Trends in cognitive sciences*, 1:146–153, 1997.

[20] T. Regier. Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences*, 7:263–268, 2003.

[21] D. Richards and J. Goldfarb. The episodic memory model of conceptual development: An integrative viewpoint. *Cognitive Development*, 1:183–219, 1986.

[22] J. R. Saffran, E. Johnson, R. Aslin, and E. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 1999.

[23] J. R. Saffran, E. L. Newport, and R. N. Aslin. Word segmentation: The role of distributional cues. *Journal of memory and language*, 35:606–621, 1996.

[24] M. S. Seidenberg. Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275:1599–1603, March 1997.

[25] L. Smith. How to learn words: An associative crane. In R. Golinkoff and K. Hirsh-Pasek, editors, *Breaking the word learning barrier*, pages 51–80. Oxford: Oxford University Press, 2000.

[26] J. Snedeker and J. Trueswell. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, 48:103–130, 2003.

[27] M. Tomasello. Perceiving intentions and learning words in the second year of life. In M. Bowerman and S. Levinson, editors, *Language Acquisition and Conceptual Development*. Cambridge University Press, 2000.

[28] C. Yu and D. H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. In *Fifth International Conference on Multimodal Interface*. ACM Press, 2003.

[29] C. Yu, D. H. Ballard, and R. N. Aslin. The role of embodied intention in early lexical acquisition. In *Proceedings the Twenty Fifth Cognitive Science Society Annual Meetings*, Boston, MA, July 2003.