ELSEVIER

Cognitive Science 26 (2002) 113–146

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Learning words from sights and sounds: a computational model

Deb K. Roy*, Alex P. Pentland

MIT Media Laboratory, 20 Ames Street, Room E15–488, Cambridge, MA 01242, USA

Received 9 January 2001; received in revised form 24 May 2001; accepted 16 August 2001

Abstract

This paper presents an implemented computational model of word acquisition which learns directly from raw multimodal sensory input. Set in an information theoretic framework, the model acquires a lexicon by finding and statistically modeling consistent cross-modal structure. The model has been implemented in a system using novel speech processing, computer vision, and machine learning algorithms. In evaluations the model successfully performed speech segmentation, word discovery and visual categorization from spontaneous infant-directed speech paired with video images of single objects. These results demonstrate the possibility of using state-of-the-art techniques from sensory pattern recognition and machine learning to implement cognitive models which can process raw sensor data without the need for human transcription or labeling. © 2002 Cognitive Science Society, Inc. All rights reserved.

Keywords: Language acquisition; Cross-modal; Sensor grounded; Learning; Computational model

1. Introduction

Infants are born into a rich and complex environment from which they construct mental representations to model structure that they find in the world. These representations enable infants to understand and predict their surroundings and ultimately to achieve their goals. They accomplish this using a combination of evolved innate structures and powerful learning algorithms.

To explore issues of early language learning, we have developed CELL, a computational

* Corresponding author. Tel.: +1-617-253-0596; fax: +1-617-253-8874.
E-mail address: dkroy@media.mit.edu (D.K. Roy).

PII: S0364-0213(01)00061-1

model which acquires words from multimodal sensory input. CELL stands for Cross-channel Early Lexical Learning. Set in an information theoretic framework, the model acquires a lexicon by finding and statistically modeling consistent intermodal structure. The model was implemented using current methods of computer vision and speech processing. By using these methods, the system is able to process natural speech and images directly without reliance on manual annotation or transcription. Although the model is limited in its ability to deal with complex scenes and noisy acoustic signals, it nonetheless demonstrates the potential of using these techniques for the purpose of modeling cognitive processes involved in language acquisition.

CELL learns by finding and modeling consistent structure across channels of sensor data. The model relies on a set of innate mechanisms which specify how speech and visual input are represented and compared, and probabilistic learning mechanisms for integrating information across modalities. These innate mechanisms are motivated by empirical findings in the infant development literature. CELL has been implemented for the task of learning shape names from a database of infant-directed speech recordings which were paired with images of objects.[1]

2. Problems of early lexical acquisition

CELL addresses three inter-related questions of early lexical acquisition. First, how do infants discover speech segments which correspond to the words of their language? Second, how do they learn perceptually grounded semantic categories? And tying these questions together: How do infants learn to associate linguistic units with appropriate semantic categories?

Discovering spoken units of a language is difficult since most utterances contain multiple connected words. There are no equivalents of the spaces between printed words when we speak naturally; there are no pauses or other cues which separate the continuous flow of words. Imagine hearing a foreign language for the first time. Without knowing any of the words of the language, imagine trying to determine the location of word boundaries in an utterance, or for that matter, even the number of words. Infants first attempting to segment spoken input face a similarly difficult challenge. This problem is often referred to as the speech segmentation or word discovery problem. Our goal was to understand and model the identification and extraction of semantically salient words from fluent contexts.

In addition to successfully segmenting speech, infants must learn categories which serve as referents of words. In the current work we consider object shape categories derived from camera images. No pre-existing shape categories are assumed in the model. Instead, visual categories are formed from observations alone. By representing object shapes, the model is able to learn words which refer to objects based on their shape.

A third problem of interest is how infants learn to associate linguistic units with appropriate semantic categories. Input to the model, as to infants, consists of spoken utterances paired with visual contexts. Each utterance may consist of one or more words. Similarly, each context may be an instance of many possible shape categories. Given a pool of

utterance-context pairs, the learner must infer speech-to-shape mappings (lexical items) which best fit the data.

Within CELL, these three problems are treated as different facets of one underlying problem: to discover structure across spoken and contextual input.

3. Background

The CELL model addresses problems of word discovery from fluent speech and word-to-meaning acquisition within a single framework. Although computational modeling efforts have not explored these problems jointly, there are several models which treat the problems separately.

Several models have been proposed which perform a complete segmentation of an unsegmented corpus. In contrast, our model does not attempt to perform a complete segmentation. Our goal of *word discovery* refers to the problem of discovering *some* words of the underlying language from unsegmented input. Our task is thus a subtask of complete segmentation. Models of complete segmentation are nonetheless interesting as indicators of the extent to which segmentation may be performed by analysis of speech input alone.

Speech segmentation models may be divided into two classes. One class attempts to detect word boundaries based on local sound sequence patterns or statistics. As a side effect of finding segmentation boundaries, words are also found. The idea of finding boundaries by considering probabilities or frequencies of local sounds sequences dates back to Harris (1954) and has been explored more recently with computational models. Harrington, Watson and Cooper (1989) developed a computational algorithm which detected word boundaries based on trigrams of phonemes (sequences of three phonemes). The model was trained by giving it a lexicon of valid words of the language. A list of all within-word trigrams were compiled from this lexicon. To segment utterances, the model detected all trigrams which did not occur word-internally during training. In experiments, the system achieved 37% word boundary detection using a training lexicon of 12,000 common English words. The performance could be improved by using trigram probabilities rather than discrete occurrence tables. Although this work did not address learning word boundaries (since the training phase relied on a presegmented lexicon), the results demonstrate that phonotactic patterns may aid segmentation. This hypothesis is supported by infant research which has shown that 8-month old infants are sensitive to transition probabilities of syllables suggesting that they may use these cues to aid in segmentation (Saffran, Aslin & Newport, 1996).

A second class of segmentation algorithms explicitly model the words of the language. The concept of minimum description length (MDL) (Rissanen, 1978) rovides a natural framework for constructing such algorithms. Within the MDL framework, the objective of the language learner is to acquire a lexicon which is most consistent with the observed linguistic input. A corpus of input text or speech is encoded as sequences of items from the acquired lexicon. A set of utterances may then be represented by the lexicon and a sequence of indices into the lexicon. The encoding of indices into the lexicon is optimized by assigning shorter codes to common lexical items. A trade-off exists between the size of the lexicon and the length of the resulting encoding of a corpus of utterances. The MDL framework provides

a probabilistically sound basis for optimizing this trade-off to arrive at an optimal lexicon. de Marcken (1996) developed an algorithm which obtained a hierarchical decomposition of an unsegmented corpus of text or phoneme transcripts. The decomposition was optimized within the MDL framework. Rather than posit word boundaries, this model generated multiple levels of possible segmentation. The hierarchical design reflects the hierarchical nature of language extending from phonemes, morphemes and words to phrases. Brent (1999) developed an algorithm which generated a prior probability distribution over all possible sequences of all possible words constructed from a given alphabet. A corpus of unsegmented utterances was treated as a single observation sample in this model. The lexicon which was most probable for the observed corpus was selected and could be used to segment the corpus. Brent reported favorable segmentation results in comparison to several alternative schemes. This algorithm also operates according to the minimum description length criterion since choosing a maximally probable model of the language is equivalent to minimizing description length (Cover & Thomas, 1991).

In addition to problems of word discovery from unsegmented speech, CELL also addresses the problem of learning word-to-meaning associations. CELL is concerned with learning words whose referents may be learned from direct physical observations. The current instantiation of the model, however, does not address learning words which are abstractly defined or difficult to learn by direct observation. Nonetheless, acquisition of word meaning in this limited sense is not trivial. There are multiple levels of ambiguity when learning word meaning from context. First, words often appear in the absence of their referents, even in infant-directed speech. This introduces ambiguity for learners attempting to link words to referents by observing co-occurrences. Ambiguity may also arise from the fact that a given context may be interpreted in numerous different ways (Quine, 1960). Even if a word is assumed to refer to a specific context, an unlimited number of interpretations of the context may be logically possible. Further ambiguities arise since both words and contexts are observed through perceptual processes that are susceptible to multiple sources of variation and distortion. The remainder of this section discusses approaches to resolving such ambiguities.

Infant directed speech often refers to the infant's immediate context (Snow, 1977) Thus it is reasonable for the learner to assume that some or all of the words of an utterance will refer to some aspect of the immediate context. Ambiguities inherent in single utterance-context observations may be resolved by integrating evidence from multiple observations.

Siskind (1992) modeled the acquisition of associations of words to semantic symbols using *cross-situational learning*. By considering multiple situations, the most likely word-to-symbol associations were obtained by looking for consistent word-to-context patterns. The model acquired partial knowledge from ambiguous utterance-context pairs which were combined across situations to eliminate ambiguity. In related work, Sankar and Gorin (1993) created a computer simulation of a blocks world in which a person could interactively type sentences which were associated with synthetic objects of various colors and shapes. The system learned to identify words which could be visually grounded and associated them with appropriate shapes and colors. The mutual information between the occurrence of a word and a shape or color type computed from multiple observations was used to evaluate the strength

of association. CELL employs a cross-situational strategy to resolve word-referent ambiguity using mutual information.

Even if we assume that utterances refer to immediate contexts, Quine observed that any feature or combination of features of the context may serve as the referent of a word. To overcome this problem, some prior bias can be assumed to favor some meanings over others. In humans, consistent constraints bias which aspects of the environment are most salient and thus likely to serve as referents for words (for example, the shape bias (Landau, Smith & Jones, 1988)). Computational models similarly may be preprogrammed to attend to specific features of contextual input and ignore others. For example, Sankar & Gorin's model only represented shape and color attributes of synthetic objects, thus constraining their model to only learn words groundable in these input channels. By not representing texture, weight and countless other potential attributes (and combinations of attributes) of an object, implicit constraints were placed on what was learnable.

The choice of representation is equally important in constraining the semantics of acquired words. Regier (1996) developed a model for learning spatial words ("above", "below", "through", etc.) by presenting a neural network with line drawing animations paired with word labels. Regier proposed a simple set of geometrical attributes derived from the relative positions, shapes, and sizes of objects would serve as the grounding for spatial terms. He showed that his choice of attributes were sufficient for learning a variety of spatial terms across several languages. A general purpose learning system which could represent many attributes in addition to those hardwired in Regier's model would likely be much slower to learn and may initially be more prone to incorrect generalizations. For the experiments reported in this paper, CELL does not address Quine's dilemma since only one type of contextual attribute, object shape, is represented.

A final type of ambiguity arises due to natural variations of sensory phenomena. A word may be uttered with an infinite number of variations and yet be recognized. An object's shape may also vary in countless ways and still be identified. A variety of statistical pattern recognition techniques exist for representing and classifying noisy signals. Popular methods including artificial neural networks and probability density estimation[2] (Bishop, 1995). Computational models which learn from examples can acquire central prototypes from multiple observations and exhibit prototypicality effects similar to human subjects (Rosch, 1975). In the domain of word learning, Plunkett, Sinha, Moller and Strandsby (1992) created a connectionist neural network which learned to pair labels with visual patterns of dots. The network exhibited several behaviors which were typical of children including a prototype effect. During training, the network was exposed to a set of randomly perturbed variations of prototype visual patterns. Although the network was not exposed to the unperturbed prototypes, after training, it was able to accurately produce labels when shown unperturbed visual prototypes (a similar finding has been demonstrated with human subjects (Posner & Keele, 1968)). In general, both neural networks and probability density estimation methods produce prototype effects and are suitable approaches for computational models which must deal with sensor ambiguities. In CELL, both neural networks and density estimation are employed to deal with variations in acoustic and visual signals. Both spoken words and visual categories in CELL are represented using prototypes (ideal forms of the category). Each prototype is paired with with a radius of allowable deviation. An newly observed signal

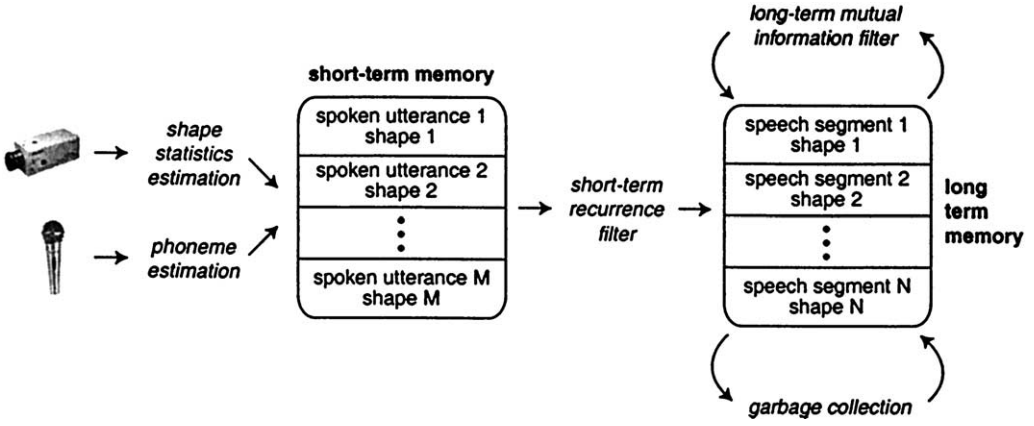

Fig. 1. Overview of the CELL model.

is classified as belonging to a category if it falls with the radius of allowable variation with respect to the prototype.

The CELL model may be differentiated from previous work in two significant ways. First, CELL operates with only sensor-grounded representations of both linguistic and contextual input. In contrast, previous models rely on manually encoded representations of speech and/or context. The representations used in previous work vary widely in level of abstraction. For example, Brent's model processed phonetic transcripts derived from natural infant-caregiver interactions. In contrast, both Regier and Plunkett et al. used binary labels, which completely abstracted away the acoustic form of words.

A second significant difference between CELL and most previous work is that CELL examines the interaction between speech segmentation and word-to-meaning learning. The problem of segmenting fluent speech to discover words is addressed jointly with the problem of associating words with co-occurring referents. CELL is compatible with models which treat each problem separately, but brings to light the advantage of leveraging partial evidence from each task within a joint framework.

4. The CELL model

A schematic of the CELL model is presented in Fig. 1. The model discovers words by searching for segments of speech which reliably predict the presence of visually co-occurring shapes. Input to the system consists of spoken utterances paired with images of objects. This multimodal input approximates the stimuli that infants may receive when listening to caregivers while visually attending to objects in the environment. A short-term memory (STM) buffers recent utterance-shape pairs for a brief period of time. Each time a new observation is entered into the STM, a *short-term recurrence filter* generates hypotheses of word-shape associations by segmenting and extracting speech subsequences from utterances in the STM and pairing them with co-occurring shape observations. These hypotheses are placed in a long-term memory (LTM). A second filter operates on the LTM to consolidate

reliable hypotheses over multiple observations. A garbage collection process eliminates unreliable hypotheses from the LTM. Output of the model is a set of {speech segment, shape} associations.

Learning is an on-line process in CELL. Unprocessed input is stored temporarily in STM. Repeated speech and visual patterns are extracted by the recurrence filter, and the remaining information is permanently discarded. As a result, CELL has only limited reliance on verbatim memory of sensory input.

Our goal was to approximate the type of input that infants receive when listening to caregivers while visually attending to objects. In order to build a computational simulation, we made the following simplifying assumptions:

- The learner possesses a short term memory (STM) which can store approximately 10 s of speech represented in terms of phoneme probabilities. The STM also stores representations derived from co-occurring visual input.
- The visual context of an input utterance is a single object
- Only the shape of objects are represented. Other attributes such as color, size, texture, and so forth are not represented in the current implementation.
- There are built-in mechanisms for generating and comparing phonemic representations derived from the acoustic waveform. In this model, coarse level phonemic representation occurs prior to word learning. Similarly, the model also has built-in mechanisms for extracting and comparing shape representations.

Although these assumptions simplify learning, the current implementation nonetheless provides a first step towards understanding early word learning from real-world data.

4.1. Representing and comparing spoken utterances

Spoken utterances are represented as arrays of phoneme[3] probabilities (Fig. 2). Acoustic input is first converted into a spectral representation using the Relative Spectral-Perceptual Linear Prediction (RASTA-PLP) algorithm (Hermansky & Morgan, 1994). RASTA-PLP is designed to attenuate nonspeech components of an acoustic signal. It does so by suppressing spectral components of the signal which change either faster or slower than speech. First, the critical-band power spectrum is computed and compressed using a logarithmic transform. The time trajectory of each compressed power band is filtered to suppress nonspeech components. The resulting filtered signal is expanded using an exponential transformation and each power band is scaled to simulate laws of loudness perception in humans. Finally, a 12-parameter representation[4] of the smoothed spectrum is estimated from a 20 ms window of input. The window is moved in time by 10 ms increments resulting in a set of 12 RASTA-PLP coefficients estimated at a rate of 100 Hz.

A recurrent neural network analyses RASTA-PLP coefficients to estimate phoneme and speech/silence probabilities. The RNN has 12 input units, 176 hidden units, and 40 output units. The 176 hidden units are connected through a time delay and concatenated with the RASTA-PLP input coefficients. Thus, the input layer at time t consists of 12 incoming RASTA-PLP coefficients concatenated with the activation values of the hidden units from time $t-1$. The time delay units give the network the capacity to remember aspects of old

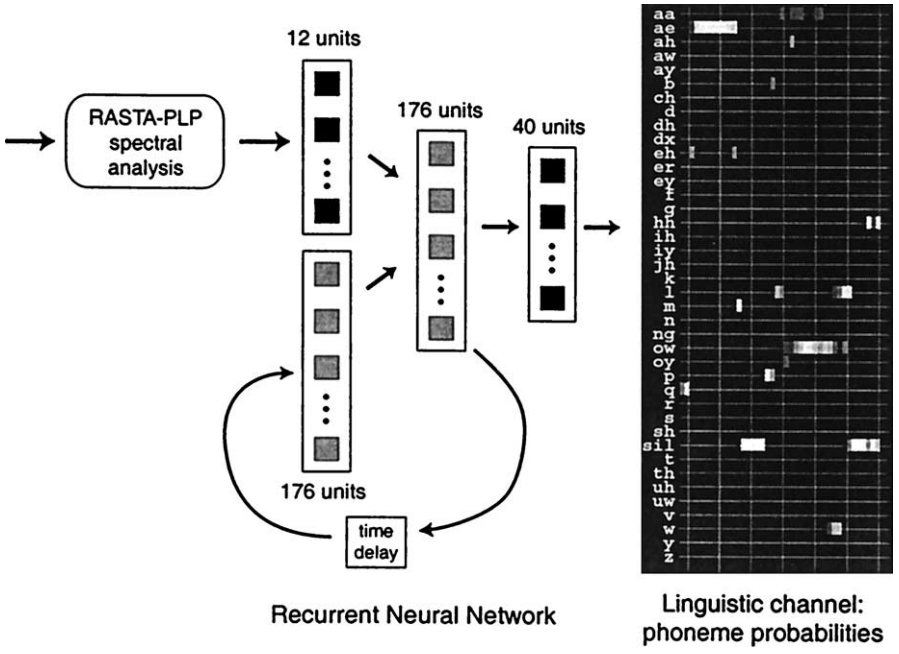

Fig. 2. Input speech is represented as an array of probabilities for the 40 phonemes.

input and combine those representations with fresh data. This capacity for temporal memory has been shown to effectively model coarticulation effects in speech (Robinson, 1994). The RNN was trained off-line using the TIMIT database of phonetically transcribed American English speech recordings (Garofolo, 1988).

To gain insight into the level of accuracy of the recurrent network, the TIMIT test-set sentences[5] were passed through the network in order to generate a confusion matrix (Fig. 3). This test set consisted of 1344 utterances spoken by 168 native English male and female adult speakers. For each 10ms frame of speech, we recorded the activations of all 40 output units of the network. If the RNN was perfect, all activation would be concentrated along the diagonal of the plot. We found, however, that over 40% of the activation mass was off the diagonal. In other words, on a frame-by-frame phoneme classification task, the RNN would correctly classify slightly less than 60% of the frames. We refined the performance of the RNN by adding probabilistic biases based on bigram phoneme transition probabilities and phoneme duration statistics. Nonetheless, the representation of speech used in this model is very noisy. The high level of errors reflects the inherent acoustic confusability of fluent speech. For the purpose of modeling word learning, however, we found this representation to be sufficient for matching speech segments. Our goal was to compare syllable structure and consonant clusters, a task which did not require perfect phoneme recognition.

To process the acoustic input, spoken utterances were first segmented in time along phoneme boundaries which simultaneously provided hypotheses of word boundaries. The output of the RNN given a sample infant-directed spoken utterance is shown in Fig. 4. The probability of each phoneme is indicated by the brightness of the corresponding trace as a function of time.

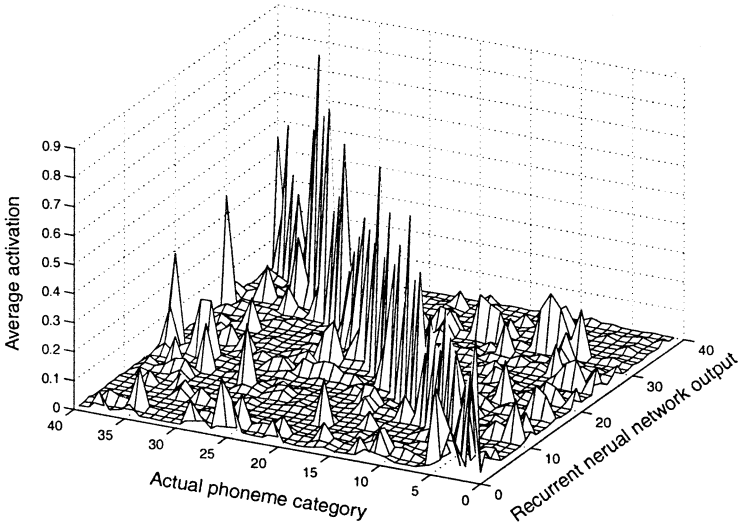

Fig. 3. The performance of the RNN was evaluated by comparing RNN activations to actual phoneme categories using the TIMIT test-set. Correct activations lie on the diagonal with errors on the off-diagonals.

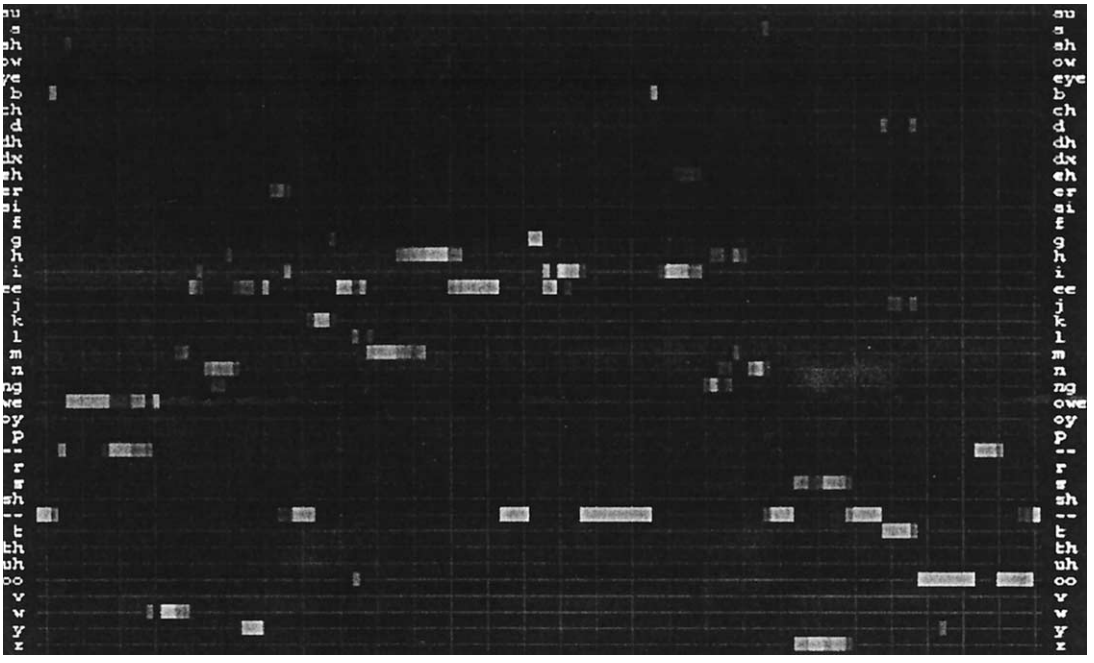

Fig. 4. Sample output from the recurrent neural network for the utterance “Oh, you can make it bounce too!” taken from the infant-directed speech corpus. Time runs along the horizontal axis and phoneme labels are displayed vertically on the left and right edges.

To locate approximate phoneme boundaries, the RNN outputs are treated as state emission probabilities in a Hidden Markov Model (HMM) framework. The Viterbi dynamic programming search[6] (Rabiner, 1989) is used to obtain the most likely phoneme sequence for a given phoneme probability array. After Viterbi decoding of an utterance, the system obtains (1) a phoneme sequence, the most likely sequence of phonemes which were concatenated to form the utterance and (2) the location of each phoneme boundary for the sequence. Each phoneme boundary serves as a speech segment start or end point. Any subsequence within an utterance terminated at phoneme boundaries is used to form word hypotheses. Additionally, any word candidate is required to contain at least one vowel. This constraint prevents the model from hypothesizing consonant clusters as word candidates. Instead, each candidate is guaranteed to consist of one or more syllables consisting of a vowel and consonant or consonant cluster on either side of the vowel. We refer to a segment containing at least one vowel as a *legal segment*.

A distance metric, $d_A(\,)$, measures the similarity between two speech segments. It is possible to treat the phoneme sequence of each speech segment as a string and use string comparison techniques. This method has been applied to the problem of finding recurrent speech segments in continuous speech (Wright, Carey & Parris, 1996). A limitation of this method is that it relies on only the single most likely phoneme sequence. A sequence of RNN output contains additional information which specifies the probability of all phonemes at each time instance. To make use of this additional information, we developed the following distance metric.

Let $Q = \{q_1, q_2, \ldots, q_N\}$ be a sequence of N phonemes observed in a speech segment. This sequence may be used to generate a HMM model λ by assigning an HMM state for each phoneme in Q and connecting each state in a strict left-to-right configuration. State transition probabilities are inherited from a context-independent set of phoneme models trained from the TIMIT training set. Consider two speech segments, α_i and α_j with decoded phoneme sequences Q_i and Q_j. From these sequences, we can generate HMMs λ_i and λ_j. We wish to test the hypothesis that λ_i generated α_j (and vice versa).

The Forward algorithm (Rabiner, 1989) can be used to compute $P(\alpha_i|\lambda_j)$ and $P(\alpha_j|\lambda_i)$, the probability that the HMM derived from speech segment α_i generated speech segment α_j and vice versa. However, these probabilities are not an effective measure for our purposes since they represent the joint probability of a phoneme sequence and a given speech segment. An improvement is to use a likelihood ratio test to generate a confidence metric (Rose, 1996). In this method, each likelihood estimate is scaled by the likelihood of a default alternate hypothesis, λ^{A}:

$$L(\alpha, \lambda, \lambda^{\mathrm{A}}) = \frac{P(\alpha|\lambda)}{P(\alpha|\lambda^{\mathrm{A}})}$$

The alternative hypothesis is that the HMM was derived from the speech sequence itself, i.e., $\lambda_i^{\mathrm{A}} = \lambda_j$ and $\lambda_j^{\mathrm{A}} = \lambda_i$. The symmetric distance between two speech segments is defined in terms of logarithms of these scaled likelihoods:

$$d_A(\alpha_i, \alpha_j) = -\frac{1}{2}\left\{\log\left[\frac{P(\alpha_i|\lambda_j)}{P(\alpha_i|\lambda_i)}\right] + \log\left[\frac{P(\alpha_j|\lambda_i)}{P(\alpha_j|\lambda_j)}\right]\right\} \quad (1)$$

The speech distance metric defined by Eq. (1) measures the similarity of phonetic structure between two speech sounds. The measure is the product of two terms: the probability that the

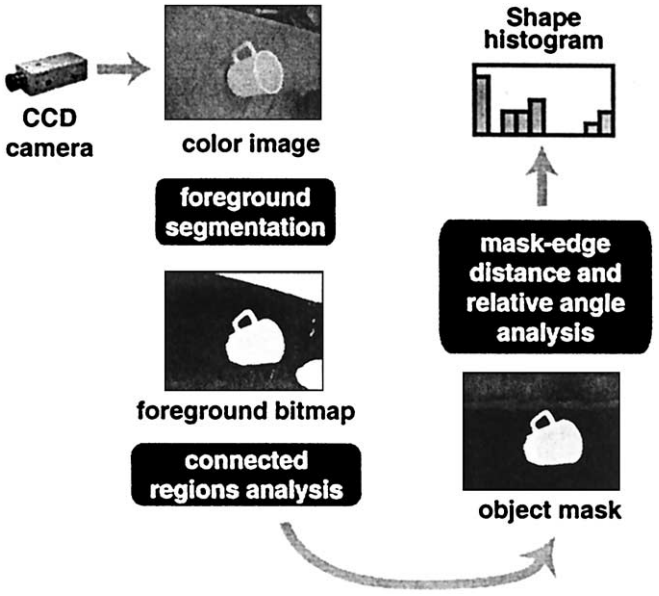

Fig. 5. Object shapes are represented in terms of histograms of features derived from the location of object edges.

HMM extracted from observation A produced observation B, and vice versa. Empirically, this metric was found to return small values for words which humans would judge as phonetically similar.

4.2. Representing and comparing visual input

Similar to speech input, the ability to represent and compare shapes is also built into CELL. Three-dimensional objects are represented using a view-based approach in which two-dimensional images of an object captured from multiple viewpoints collectively form a visual model of the object. Fig. 5 shows the stages of visual processing used to extract representations of object shapes. An important aspect of the shape representation is that it is invariant to transformations in position, scale and in-plane rotation.

Figure-ground segmentation is simplified by assuming a uniform background. A Gaussian model of the illumination-normalized background color is estimated and used to classify background/foreground pixels. Large connected regions near the center of the image indicated the presence of an object.

Based on methods developed by Schiele and Crowley (1996), objects are represented by histograms of local features derived from multiple two-dimensional views of an object. Shape is represented by locating all boundary pixels of an object in an image. For each pair of boundary points, the normalized distance between points and the relative angle of the object edge at each point are computed. Each two-dimensional (relative angle, relative distance) data point is accumulated in a two-dimensional histogram to represent an image.

Using multidimensional histograms to represent object shapes allows for direct comparison of object models using information theoretic or statistical divergence functions (Schiele

& Crowley, 1996). In practice, an effective metric for shape classification is the χ^2-divergence:

$$d_{V_{2D}}(X, Y) = \chi^2(X, Y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i} \quad (2)$$

where $X = \mathrm{U}_i x_i$ and $Y = \mathrm{U}_i y_i$ are two histograms indexed by i while x_i and y_i are the values of each histogram cell. This measure is used to compare single view points of two shapes.

Representations of three-dimensional shapes are based on a collection of two-dimensional shape histograms, each corresponding to a particular view of the object. Each three-dimensional object is represented by 15 shape histograms. A set of two-dimensional histograms representing different view points of an object is referred to as a *view-set.* In practice, a series of images of a single object are converted into a corresponding series of view-sets. The bundle of view-sets serves as a representation of the object.

View-sets are compared by summing the divergences of the four best matches between individual histograms. If we denote $X(i)$ and $Y(j)$ as the ith and jth histograms in view-sets X and Y, then the distance between the view-sets is defined as:

$$d_{V_{2D}}(X, Y) = d_V(X, Y) = \sum \chi^2(X(i), Y(j)) \quad (3)$$

where the summation is over values of i and j which select the four best matching histograms in X and Y.

Fig. 6 shows images of four objects used in evaluations along with the corresponding objects masks and shape histograms. The symmetrical shape of a ball leads to a near diagonal activation of the histogram whereas more complex shapes result in complex histogram patterns. Using the chi-squared distance metric, the two toy dogs at the bottom of the figure are closer to one another than to either the shoe or ball.

In the experiments reported here, visual input to CELL was highly simplified since only single unoccluded objects were presented to the system. In contrast to the speech input which is taken directly from the infant-caregiver recordings, the visual data are generated off-line since processing the original video from the infant-caregiver interactions would be a much more difficult computer vision problem. Even with the simplified visual input, shape representations are nonetheless susceptible to various forms of error including shadow effects and foreground/background segmentation errors. In addition, some shape classes such as dogs and horses are highly confusable, and other classes such as trucks (which include pickup trucks and fire trucks) are quite varied. Working with sensor-derived data are motivated by our goal of developing computational learning systems which do not rely on manual annotation of any input data.

4.3. Word learning

Word learning is achieved by processes which operate on two levels of memory, short term memory (STM) and long term memory (LTM). Input representations of spoken utterances paired with visual objects are temporarily stored in the STM. A {spoken utterance, object pair} is referred to as an *audio-visual event* or *AV-event.* Each entry of the STM

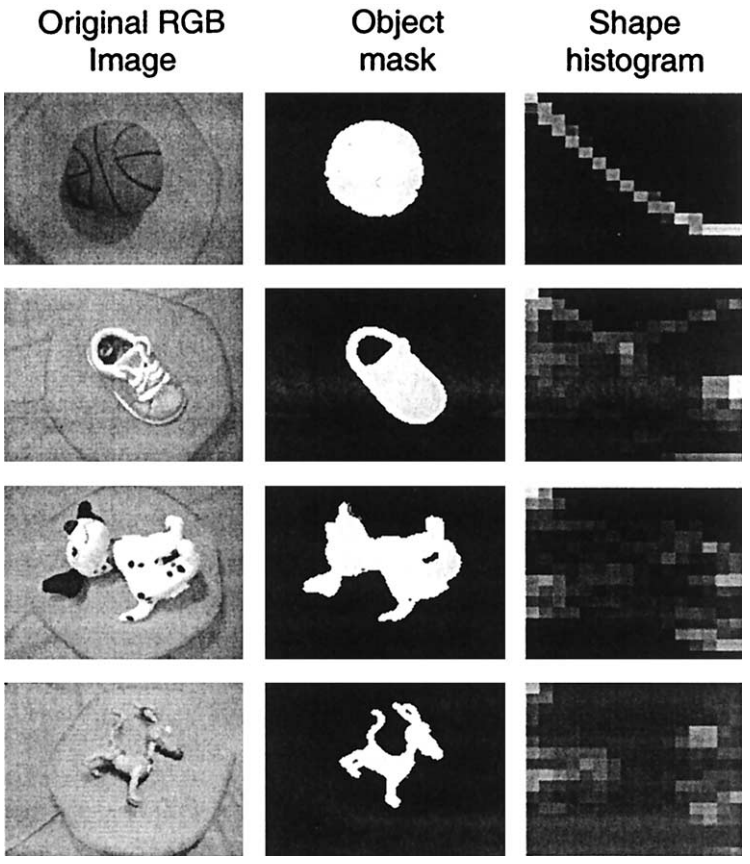

Fig. 6. Sample shape and color histograms]{Examples of several toys (balls, shoes, and dogs) and their corresponding object masks (center) and histograms (right).

contains a phoneme probability array representing a multiword spoken utterance, and a set of histograms which represent an object. For experiments reported in this paper, the STM capacity was limited to five spoken utterances and co-occurring objects. The STM functioned as a first-in-first-out buffer. As input is passed through the buffer, each new utterance-object pair replaces the oldest STM entry.

The STM is critical for efficient learning. CELL processes the contents of STM using an exhaustive search for recurrent speech segments. The space for this search grows factorially with the size of the STM, motivating a low upper bound on buffer size. The STM focuses the initial scope of learning on a relatively small amount of unanalyzed observations so that subsequent search mechanisms are not overwhelmed.

A *recurrence filter* acts on the contents of the STM. This filter searches exhaustively or repeating subsequences of speech which occur in similar visual contexts. For each legal speech segment (legal segments contain at least one vowel) in the newly received AV-event, the filter searches for matches with each legal segment in the remainder of the STM which also have matching visual contexts. The distance metrics $d_A(\,)$ and $d_V(\,)$ are used to determine matches between pairs of speech segments and objects.[7] Each recurring {speech

segment, object} pair is copied into LTM, forming an *audio-visual prototype* or *AV-prototype*. In experiments, multiple utterances are typically observed in the presence of the same object. A unique view-set is generated for each utterance so that natural variations due to continuously varying viewpoints is captured in the input to the model.

The output of the recurrence filter constitutes the system's first attempt to segment continuous speech at word boundaries. The segmentation is guided by several constraints:

- Start and end points of hypothesized speech segments must coincide with phoneme boundaries (as determined by the recurrent neural network).
- Segments must contain at least one vowel.
- Segments must *recur in close temporal proximity* since the recurrence filter only processes the contents of STM.

The validity of the third constraint rests on the assumption that at least some target words recur within the span of the STM. In addition to positive experimental results with the corpus reported later in this paper, additional cross validation of this assumption were performed on a corpus of 20 caregivers' speech transcripts (Warren-Leubecker, 1982; Warren-Leubecker & Bohannon, 1984) taken from the CHILDES database (MacWhinney, 2000). The findings of this study support our hypothesis and are summarized in Appendix A.

The LTM contains two types of data structures. The first are AV-prototypes which represent hypotheses of possible words of the target language. Each AV-prototype specifies an instance of a speech segment which was observed in STM, and the hypothesized referent of that segment, a representation of an object. Since AV-prototypes are generated based on local recurrency within STM, they are prone to errors. For example, the spoken word "the" might occur repeatedly in the STM in the context of a ball. The recurrence filter would generate an AV-prototype based on this input and erroneously pair an instance of "the" with an observation of a ball. The same spoken word may be paired with multiple visual prototypes (or vice versa). For example, if the word "the" and "ball" both recur in similar visual contexts, each word may be separately paired with visual prototypes derived from the same object.

The second type of data structure found in LTM are *lexical items*. Lexical items are created by consolidating AV-prototypes based on a mutual information criterion. This consolidation process identifies clusters of AV-prototypes which may be merged together to model consistent intermodal patterns across multiple observations.

To understand the consolidation process, consider a situation in which the LTM has n AV-prototypes in store. When a new AV-prototype x is placed in LTM by the recurrency filter, the prototype is evaluated to determine the likelihood that it is a reliable lexical item. First, a pair of radii are assigned to the AV-prototype. A speech or shape prototype without an associated radius simply marks a point in the space of possible speech or shape observations. By adding a radius of allowable deviation, the prototype-radius pair defines a subspace centered on the prototype. A subspace representation is necessary to account for variations in pronunciation of the same word, or differences in appearance of objects belonging to the same class. By adding a radius to both the speech and shape prototypes, the result is a model of an association between a speech subspace and a shape subspace. This speech-shape association constitutes a lexical item in CELL (Fig. 7).

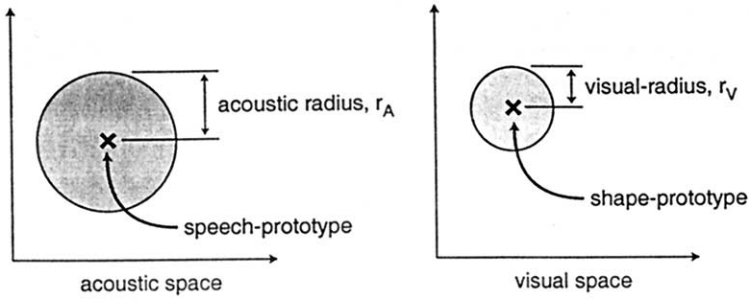

Fig. 7. A lexical item models a speech-shape association. A speech prototype specifies the ideal or canonical form of the speech sound. The acoustic radius specifies the allowable acoustic deviation from this ideal form. Similarly, the shape-prototype specifies the ideal shape to serve as a referent for the lexical item and the visual-radius specifies allowable deviation from this shape. The radii are a necessary component of the representation to account for natural variations inherent in acoustic and visual input.

A pair of random variables A and V may be defined in terms of the prototype x. For each prototype y in LTM (i.e., y is one of n AV-prototypes in LTM), these variables are set as:

$$A = \begin{cases} 0 & if \quad d_A(x, y) > r_A \\ 1 & if \quad d_A(x, y) \leq r_A \end{cases}$$

$$V = \begin{cases} 0 & if \quad d_V(x, y) > r_V \\ 1 & if \quad d_V(x, y) \leq r_V \end{cases} \tag{4}$$

where r_A and r_V are determined by a search procedure that maximizes the mutual information. A and V are indicator variables which are set to 1 if the acoustic or visual distance from x to y fall within the radius associated with x. To evaluate a prototype, the mutual information between A and V is computed (Cover & Thomas, 1991):

$$I(A; V) = \sum_i \sum_j P(A = i, V = j) \log \left[\frac{P(A = i, V = j)}{P(A = i)\, P(V = j)} \right] \tag{5}$$

The summations variables i and j are binary and $P(A = i, V = j)$ denotes the joint probability of $A = i$ and $V = j$. Note that mutual information is a symmetric measure, that is, $I(A; V) = I(V; A)$.

The probabilities comprising Eq. (5) are estimated using relative frequencies of all n prototypes in LTM:

$$P(A = i) = \frac{|A = i|}{n} \tag{6}$$

$$P(V = j) = \frac{|V = j|}{n} \tag{7}$$

$$P(A = i, V = j) = \frac{|A = i, V = j|}{n} \tag{8}$$

The vertical bars denote the count operator (the counts are performed over all n AV-prototypes in LTM). For experiments, smoothing was incorporated into the count operator to avoid noise due to small count values.

The mutual information I is a function of the radii r_A and r_V. A search is performed to find the radii which maximize I. A smoothing factor is added to insure nonzero values of both radii. The two radii define a two-dimensional space which is searched to locate the point of maximum mutual information.

For clarity, we summarize how AV-prototypes lead to lexical items in LTM. Given a set of AV-prototypes in LTM, any specific AV-prototype may evaluated as the basis for acquiring a new lexical item. To convert an AV-prototype to a lexical item, a radius of allowable variation is added to both the acoustic and visual prototype. The lexical item consists of an association between a visual category and an acoustic category and represents the hypothesis that the acoustic category (a speech sequence) refers to the visual category. For a particular setting of the acoustic and visual radii, the equations stated above provide a means for computing the mutual information between the visual and acoustic category. The mutual information quantifies the amount of information gained about the presence (or absence) of one category given that we know whether the associated category is present or not. It is assumed that an association with high mutual information indicates a good lexical item. The mutual information depends directly on the values of the radii. Since we have no a priori knowledge of how to set these radii, a search is performed to find the settings of both radii which maximizes the mutual information. A lexical item is formed when the maximum mutual information exceeds an empirically set threshold. If a lexical item is formed, all other AV-prototypes which match the lexical item acoustically and visually are removed from LTM since they are similar to the lexical item and contribute no new information. Fig. 8 presents two examples of mutual information surfaces from one of our data sets. In each plot, the height of the surface shows mutual information as a function of the acoustic and visual radii. On the left plot, the speech prototype corresponding to the word “yeah” was paired with visual representation of a shoe. The resulting surface is relatively low for all values of radii. The AV-prototype on the right correctly paired a speech segment of the word “dog” with a visual representation of a dog. The result is a strongly peaked surface indicating a reliable prototype which should be retained. Thus mutual information is used to learn speech-to-meaning mappings across multiple observations.

Each time a new AV-prototype is added to LTM, all AV-prototypes in LTM are evaluated. Any AV-prototype x which results in high mutual information is converted into a lexical item by (1) recording its optimal radii, and (2) removing all AV-prototypes y which match x, that is, remove y if $A = 1$ and $V = 1$ according to Eq. (4).

A final process, as shown in Fig. 1, is a *garbage collection filter* which removes unused AV-prototypes. This filter tracks the number of AV-prototypes in LTM and removes old prototypes when the number of unconsolidated prototypes exceeds a memory capacity limit.

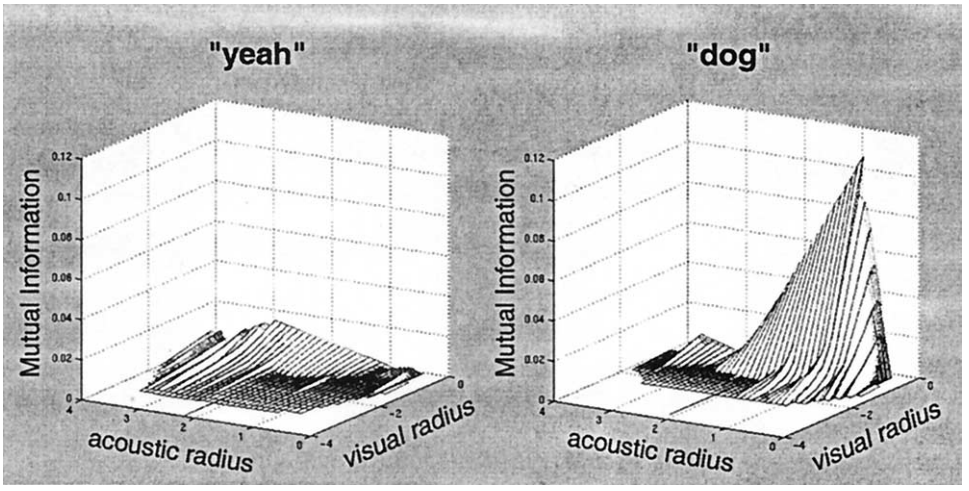

Fig. 8. Mutual information plotted as a function of acoustic and visual radii for two speech segments which were both paired with view-sets of a toy dog. In the example on the left, the word "yeah" was linked to the dog since the word recurred in the STM in the presence of a dog. However this hypothesis found little support from other AV-prototypes in LTM which is indicated by the low flat mutual information surface. In contrast, in the example on the right, the word "dog" was correctly paired with a dog shape. In this case there was support for this hypothesis as indicated by the strongly peaked structure of the mutual information surface. CELL detects peaks such as this one using a fixed threshold. The point at which the surface peaks is used to determine the optimal settings of the radii. These radii along with the AV-prototype lead to a new lexical item. This is how CELL learns words from sights and sounds.

5. Experimental results

This section describes an evaluation of CELL using infant-directed speech and visual images. We gathered a corpus of infant-directed speech from six caregivers and their preverbal infants. Participants were asked to engage in play centered around toy objects. Caregiver speech recordings and sets of camera images of toys were used as input to CELL.

5.1. Participants

All caregivers were native speakers of Canadian English and ranged in age from 17 to 44 years. Infants ranged in age from eight to eleven months. Each participant confirmed that their child could not yet produce single words. However, they reported varying levels of limited comprehension of words (e.g., their name, *no, dog, milk, wave,* etc.).

5.2. Objects

Caregivers were asked to interact naturally with their infants while playing with a set of age-appropriate objects. We chose seven classes of objects commonly named by young children (Huttenlocher & Smiley, 1994): balls, toy dogs, shoes, keys, toy horses, toy cars, and toy trucks. A total of 42 objects, six objects from each class, were used (Fig. 9).

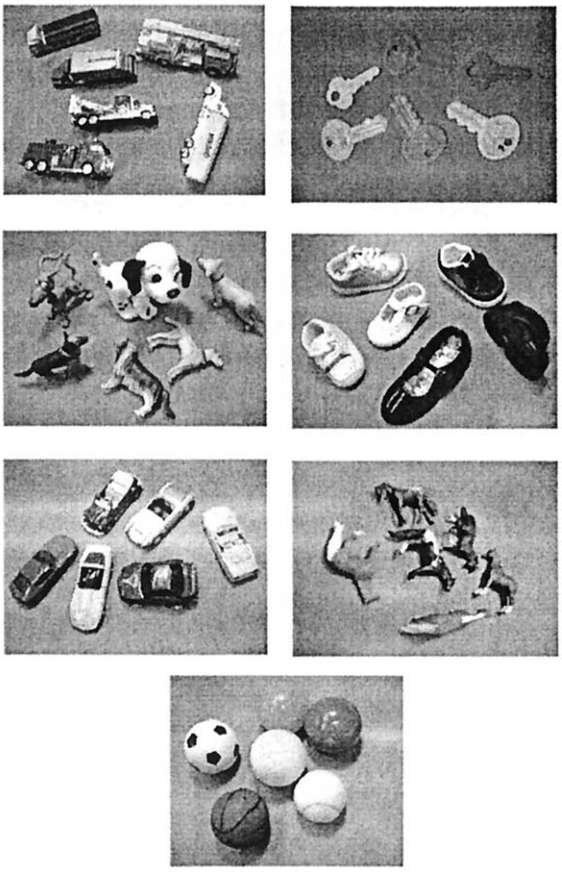

Fig. 9. Objects used in the infant-directed speech experiments.

5.3. Protocol

To ensure controlled experimental conditions, collection of speech samples took place in a sound-treated child-appropriate room. To elicit natural interactions, caregivers and their infants were left alone in the room during sessions. The room was equipped with one-way observational windows and video cameras. Caregivers wore a noise-canceling head-worn microphone and a wireless transmitter. All caregiver speech was recorded on a digital audio recorder for off-line analysis by CELL. Interactions were video taped for annotation purposes.[8] Each caregiver participated in six sessions of play with their infants over a two day period (i.e., three sessions per day). For each of the six sessions, participants were provided with a different set of seven objects, one from each of the seven object classes. The order in which object sets were provided as randomized across participants. The objects were placed in a box marked "in-box" at the start of each session. Participants were asked to take out one object at a time, play with it, and then return it to an "out-box." They were *not* told to teach their infants words. This resulted in very natural speech including singing, laughing, and various vocal effects generated to attract the attention of the infant. Participants were free to

Table 1
Transcription of automatically extracted spoken utterances from a sample session. The left column shows the object in play at the time of each utterance

Object	Utterance
dog	He's gonna run and hide
dog	He's gonna hide behind my shoe
dog	Look, Savanah
dog	See his eyes?
dog	You like anything with eyes on it, eh?
dog	Just like you he has eyes
dog	Ruf ruf ruf
car	That's what your daddy likes, look!
car	Doors open vroom!
car	The seats go forward, and they go back!
shoe	You're always climbing into the shoes at home
shoe	Savanah! (infant's name)
truck	OK, you want it to drive?
truck	The wheels go around
truck	Your uncle Pat drives a truck like that
dog	He has a red collar
key	Let me see it
key	Do the keys have teeth?
key	You only have two teeth

choose the order in which objects were selected for play, and the duration of play with each object.

5.4. Speech data

A total of 36 sessions of speech recordings were obtained (6 participants, 6 sessions per participant). Utterances were automatically segmented based on the silence/nonsilence probability estimate generated by the recurrent neural network. On average, each speaker produced a total of about 1,300 utterances. Each utterance contained approximately five words. About one in 13 words referred directly to an object and could thus be grounded in terms of visual shapes. This provides a rough indication of the level of filtering which had to be performed by the system to extract meaningful speech segments from the recordings since the large majority (92%) of words were not directly groundable in terms of visual input. Some sample utterances from one of the participant are shown in Table 1. As would be expected, many utterances contain words referring to the object in play. Note, however, that many utterances do not contain direct references to the object in view, and occasionally even contain references to other objects used in the experiment but not in play at the time that the utterance was spoken.

The use of recurrence filtering in CELL is based on the assumption that that infant-directed speech is redundant and that salient words will often be repeated in close temporal proximity (Snow, 1972). This assumption was confirmed in our speech corpus. Repetition of words occurred throughout all data sets, despite the fact that participants were not specifically instructed to teach their infants, or to talk exclusively about the objects. They were simply

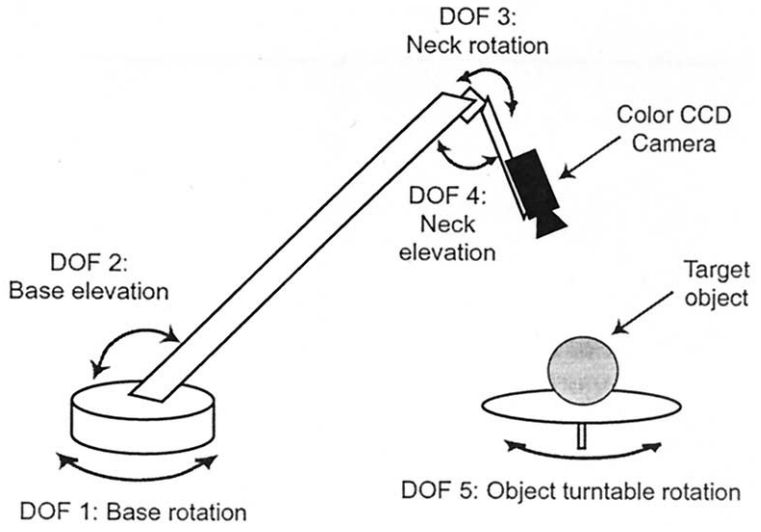

Fig. 10. A robotic armature was constructed to capture images of objects from a variety of perspectives. The object was placed on a motorized turn table for an added degree of freedom in generating viewpoints.

asked to play naturally. A temporal "clumping" effect for salient words was evident. For example, the word *ball* would appear several times within the span of half a minute because of the focused and repetitive nature of the interaction. Although we did not carefully examine the interaction between focus of attention and recurrence, it seemed that salient words were repeated even more when caregivers and infants were engaged in joint attention with respect to an object. The STM in CELL may be thought of as a buffer which is large enough to capture such temporal clumps of repeated patterns.

5.5. Visual data

A camera was mounted on a robotic armature to capture images of all 42 objects (Fig. 10). The robot armature had four degrees of freedom enabling it to move the camera through a large range of vantage points. An addition degree of freedom was created by placing the target object on a motorized turn-table. A program was written to sweep the camera and turn-table through a range of positions while capturing images of an object. A set of 209 images were captured of each of the 42 objects from varying perspectives resulting in a database of 8,778 images. From each pool of 209 images, unique sets of 15 images were randomly selected to generate distinct view-set representations of the objects.

The histogram-based visual representation results in significant ambiguity reflecting inherent difficulties of visual shape categorization. Shapes belonging to the same class were easily confused with shapes from other classes. To help characterize the image database and understand the visual confusability between objects, we generated a pair of histograms of divergences between view-sets. In the first histogram, view-sets of all objects belonging to the same class were compared to one another. For example, each truck was compared to every other truck including itself (note there were multiple view-sets for each object). All

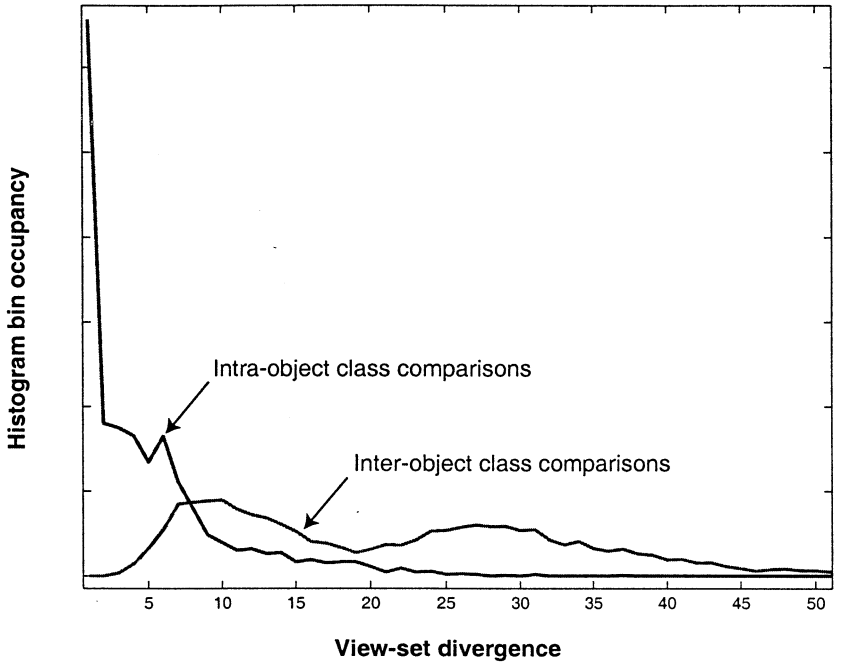

Fig. 11. Histogram of distances between all inter- and intraclass comparisons of object view-sets. Since there are more interobject class comparisons than intraobject class comparisons, the range of bin occupancies of the former is much greater than the latter. To aid in visualizing the difference in divergence distributions, the bin occupancy values of the histograms have been linearly scaled so that both histograms are displayed in a single plot. The absolute values of the vertical axis are not relevant, only the difference in distribution shapes (see text for an analysis of the distributions).

intraclass comparisons for all seven object classes are summarized in one histogram and plotted in Fig. 11. The strong peak of the plot at the very left of the plot is due to comparisons of different view-sets derived from the same object. The second histogram was generated by comparing view-sets of each object to all other objects in the study which did not belong to the same object class. For example, each ball was compared to each truck, shoe, dog, horse, key and car, but not to other balls. The superposition of the histograms shows significant overlap between the distributions. The overlap shows that many intraclass comparisons result in values which are greater than interclass comparisons. For example, two trucks may be farther from each other than a horse from a dog. The overlap may be quantified by considering a decision boundary placed at the point of cross-over between the distributions (at $x = 7$ in Fig. 11). If this decision boundary were used to classify each divergence score as either inter- or intraclass, an error rate of 30.2% would be obtained.

5.6. Combining speech and visual data to create AV-events

Caregivers were asked to play with one object at a time. All utterances that were produced between the time when the object was removed from the in-box and placed in the out-box

were paired with that object. Spoken utterances were paired with randomly selected view-sets of the corresponding object to generate a set of AV-events.

This procedure was based on a simplifying assumption about the infant-interaction. We assumed that all utterances occurred while the infant was looking at the object. In reality, however, infants were not always watching the object. Although this assumption allowed the use of all the recorded speech for evaluation, it may have increased the difficulty of lexical acquisition for the system since caregivers may have been less likely to refer to an object if they were aware that the infant was not attending to it.

5.7. Processing the audio-visual corpus

CELL was used to process each participant's data in a separate experiment. The AV-events generated from each participant were presented to the system in the sequence in which the original acoustic events were recorded. Recurrence filtering resulted in a series of AV-prototypes in LTM which were consolidated by the mutual information filter to create a set of lexical items for each caregiver.

5.8. Baseline acoustic-only model

To provide a baseline for comparing intermodal to unimodal learning, an *Acoustic-only Model* which ignored visual input was implemented and tested on the same data set. Although the system was presented with AV-events, only the acoustic portion of the input was used to generate lexical items. The Acoustic-only Model acquires a lexicon by identifying speech segments which recur most often in the input. The model assumes that some underlying language source concatenates words according to a set of unknown rules. Segments of speech which are found to repeat often are assumed to correspond to words of the target language. This is similar to previously suggested models of speech segmentation which are based on detection of recurrent sound patterns (Brent, 1999; de Marcken, 1996).

The Acoustic-only Model utilized many of the same components implemented in CELL. The recurrency filter was modified to ignore the visual channel. Recurrent speech segments were extracted from STM and copied into LTM based on only acoustic matching. The STM and acoustic recurrence threshold were configured identically to the experiments with CELL. The mutual information filter was replaced with a second recurrence filter which searched for speech segments which occurred often in LTM. In effect, this second recurrence filter identifies speech segments which occurred often across long time spans of input.

5.9. Evaluation measures

For each set of speaker data, we ran the CELL model and extracted the 15 highest scoring lexical items.[9] Each lexical item acquired by the system was manually evaluated using three measures:

5.9.1. Measure 1: segmentation accuracy

Do the start and end of each speech prototype correspond to word boundaries in English?

Table 2
Contents of LTM using CELL to process one participant's data

Rank	Phonetic transcript	Text transcript	Shape category	Segment. accuracy	Word Disc.	Semantic accuracy
1	ʃu	shoe	shoe E	1	1	1
2	fair ə	fire*	truck D	0	1	1
3	rək	*truck	truck C	0	1	1
4	dɑg	dog	dog D	1	1	1
5	Iŋəʃ	in the*	shoe A	0	0	0
6	ki	key	key C	1	1	1
7	ki	key	key E	1	1	1
8	dɑggi	doggie	dog C	1	1	1
9	bɑl	ball	ball C	1	1	1
10	bɑl	ball	ball A	1	1	1
11	kiə	key*	key C	0	1	1
12	ʌʃu	a shoe	shoe B	0	1	1
13	ənðIsIz	*and this is	shoe B	0	0	0
14	(ono.)	(engine)	truck A	—	—	—
15	(ono.)	(barking)	dog A	—	—	—
Total				54%	85%	85%

5.9.2. Measure 2: word discovery

Does the speech segment correspond to a single English word? We accepted words with attached articles and inflections, and we also allowed initial and final consonant errors. For example the words/dɑg/(*dog*), /ɑg/(*dog*, with initial /d/ missing), and /ðədɑg/ (*the dog*), would all be accepted as positive instances of this measure. However /dɑgIz/ (*dog is*) would be counted as an error. Initial and final consonant errors were allowed in this measure since we were interested in measuring how often single words were discovered regardless of exact precision of segmentation.

5.9.3. Measure 3: semantic accuracy

If the lexical item passes the second measure, does the visual prototype associated with it correspond to the word's meaning? If a lexical item fails on Measure 2, then it automatically fails on Measure 3.

It was also possible to apply Measure 3 to the Acoustic-only Model since the visual prototype was carried through from input to output. In effect, this model assumes that when a speech segment is selected as a prototype for a lexical candidate, the best choice of its meaning is whatever context co-occurred with the speech prototype.

5.10. Results

Table 2 lists the contents of the lexicon generated by CELL for one of the participants. A phonetic and text transcript of each speech prototype has been manually generated. For the text transcripts, asterisks were placed at the start and/or end of each entry to indicate the presence of a segmentation error. For example "dog*" indicates that either the /g/ was cutoff, or additional phonemes from the next word were erroneously concatenated with the target

Table 3
Contents of LTM using the Acoustic-only Model to process one participant's data

Rank	Phonetic transcript	Text transcript	Shape category	Segment. accuracy	Word Disc.	Semantic accuracy
1	(ono.)	(engine)	car C	—	—	—
2	dʒudʒudʒu	do do do	shoe A	0	0	0
3	(ono.)	(engine)	truck C	—	—	—
4	(ono.)	(engine)	truck C	—	—	—
5	wʌyugonnʌd	what you gonna do*	shoe A	0	0	0
6	nawhirk	now here okay*	ball B	0	0	0
7	lʌmiyuz	*amuse	car E	0	1	0
8	beybi	baby	horse A	1	1	0
9	ahhiʔ	ah he's*	horse E	0	0	0
10	iah	*be a	ball A	0	0	0
11	wʌyugonnʌd	what you gonna do*	key A	0	0	0
12	iligʊd	*really good	shoe F	0	0	0
13	iv	—	ball F	0	0	0
14	yulbiə	you'll be a	ball A	0	0	0
15	ʔey	*today	dog D	0	1	0
Total				8%	25%	0%

word. For each lexical item we also list the associated object based on the visual information. The letters A-F are used to distinguish between the six different objects of each object class.

Several phoneme transcripts have the indicator "(ono.)" which indicate onomatopoeic sounds such as "ruf-ruf" for the sound of a dog, or "vroooommmm" for a car. The corresponding text transcript shows the type of sound in parentheses. We found it extremely difficult to establish accurate boundaries for onomatopoeic words in many instances. For this reason, these lexical items were disregarded for all measures of performance. It is interesting to note that CELL did link objects with their appropriate onomatopoeic sounds. They were considered meaningful and groundable by CELL in terms of object shapes. This finding is consistent with infant learning; young children are commonly observed using onomatopoeic sounds to refer to common objects. The only reason these items were not processed further is due to the above stated difficulties in assessing segmentation accuracy.

The final three columns show whether each item passed the criterion of each accuracy measure. In some cases a word such as *fire* is associated with a fire truck, or *lace* with a shoe. These are accepted as valid by Measure 3 since they are clearly grounded in specific objects. The confusion between laces and shoes is a classic example of the part-whole distinction (Quine, 1960) which CELL is susceptible to since only whole objects are nameable. At the bottom of the table, the measures are accumulated to calculate accuracy along each measure.[10]

For comparison, the lexical items acquired by the Acoustic-only Model are shown in Table 3. These results are derived from the same participant's data as Table 2. In cases where no discernible words were heard, the text transcript is left blank. CELL out-performed the

Table 4
Summary of results using three measures of performance. Percentage accuracy of CELL for each caregiver is shown. Performance by the Acoustic-only Model is shown in parentheses

Participant	Segmentation accuracy	Word Discovery	Semantic Accuracy
PC	54 (8)	85 (25)	84 (0)
SI	25 (0)	75 (10)	42 (10)
CL	20 (33)	87 (60)	80 (20)
TL	17 (7)	50 (35)	25 (14)
CP	17 (0)	50 (8)	42 (8)
AK	33 (0)	92 (45)	67 (27)
Average	28 ± 6 (7 ± 5)	72 ± 8% (31 ± 8%)	57 ± 10% (13 ± 4%)

Acoustic-only Model on all three measures. This pattern was observed consistently across all six participants.[11] Table 4 presents the results for each subject using bCELL and the Acoustic-only Model. Figs. 12, 13, and 14 plot average scores across all six participants for each measure. Note, these results were achieved without making any person-dependent parameter adjustments to the systems.

Raw acoustic data posed a significant challenge for Measure 1, segmentation accuracy. The Acoustic-only Model correctly located word boundaries only 7% of the time. In contrast, 28% of lexical items produced by CELL were correctly segmented single words. The additional information provided by the visual channel improved segmentation accuracy. It is interesting to note that of these 28%, half of the lexical items were not grounded in the contextual channel (i.e., they failed on Measure 3). For example, the words *choose* and *crawl* were successfully extracted by CELL and associated with car A and ball E respectively. These words did not directly refer to shape categories and thus failed on Measure 3. Yet, there seems to have been some consistent co-occurrence patterns between these words and shapes which aided the system in producing these segmentations. The low accuracy levels

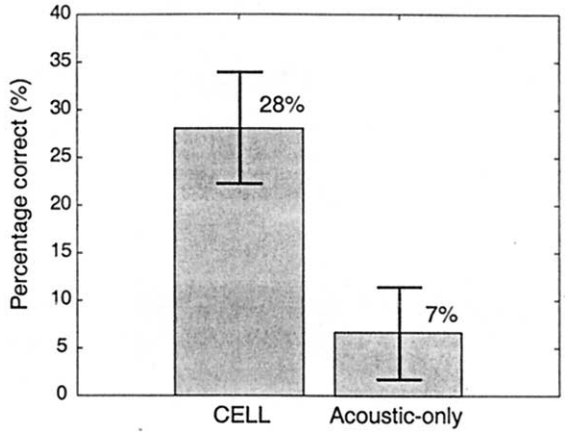

Fig. 12. Segmentation accuracy (Measure 1) for 15 best lexical items. Results indicate CELL's average performance on all six caregivers. Error bars indicate standard deviation about the mean.

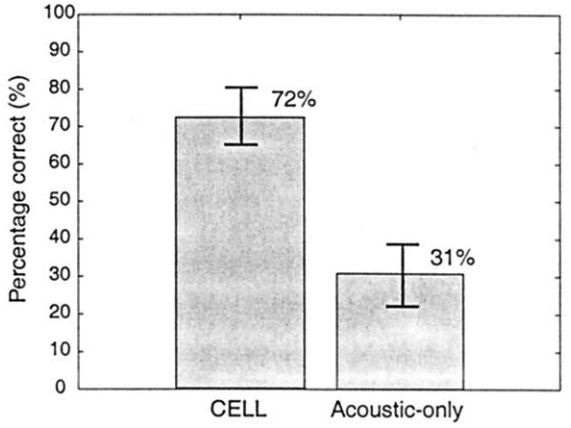

Fig. 13. Word discover (Measure 2) for 15 best lexical items. Results indicate CELL's average performance on all six caregivers. Error bars indicate standard deviation about the mean.

achieved in Measure 1 reflect the inherent difficulty of perfectly segmenting raw acoustic signals.

For Measure 2, word discovery, almost three out of four lexical items (72%) produced by CELL were single words (with optional articles and inflections) (Fig. 13). In contrast, using the Acoustic-only Model, performance dropped to 31%. These results demonstrate the benefit of incorporating cross-channel information into the word learning process. Cross-channel structure leads to a 2.3-fold increase in accuracy compared to analyzing structure within the acoustic channel alone.

On Measure 3, semantic accuracy, we see the largest difference in performance between CELL and the Acoustic-only Model (Fig. 14). With an accuracy of 57%, CELL out-performs the Acoustic-only Model by over a factor of four. The Acoustic-only Model acquires a lexicon in which only 13% of the items are semantically accurate. Through manual analysis

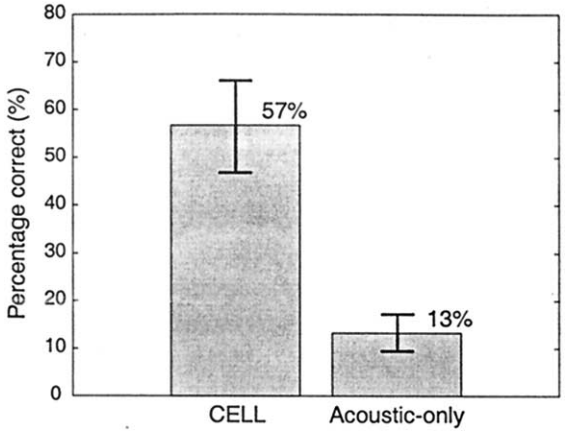

Fig. 14. Semantic accuracy (Measure 3) for 15 best lexical items. Results indicate CELL's average performance on all six caregivers. Error bars indicate standard deviation about the mean.

we found that in the input speech, less than 8% of words are grounded in shape categories. In the output, this ratio increases over seven times.

The Acoustic-only Model performed well considering the input it received consisted of unsegmented speech alone. It also learned many words which are not acquired by CELL including "go", "yes", "no", and "baby". These are plausible words to enter a young infant's vocabulary. This finding suggests that in addition to cross-channel structure, the learner may also use within-channel structure to hypothesize words for which the meaning is unknown (this is in line with the speech segmentation models discussed in the Background section). Using top down processes, the learner may only later attempt to determine the meaning of these words. Our findings suggest that mechanisms which identify within-channel structure may operate in parallel with cross-modal processes to learn early words.

6. Discussion

Learning is often conceptualized as consisting of a series of discrete stages. In terms of early word learning two theoretic vantage points are often posited. First, perhaps infants learn early concepts and then look for spoken labels to fit the concepts. On the other hand, they might first learn salient speech sequences and then look for their referents. Our model and experiments suggest that a more closely knit process in which these two stages in fact occur together is advantageous for the learner. Attending to the co-occurrence patterns within the visual context may help infants segment speech. Spoken utterances simultaneously act as labels for visual context, enabling the learner to form visual categories which ultimately serve as referents for words. By taking this approach, the learner is able to leverage information captured in the structure between streams of input.

In our experiments, intermodal structure led to a 2.3-fold increase in word discovery accuracy compared with analyzing structure within the acoustic channel alone. For speech segmentation, the improvement using CELL was even larger, four-fold. These results do not contradict models which propose that segmentation occurs by speech analysis alone. Our findings provide evidence, however, that the additional structure from contextual channels may accelerate the overall process of early lexical acquisition.

Empirical data suggests that infants may be able to achieve a high level of performance in the segmentation task by analyzing various structural cues within the speech signal alone (Cutler & Mehler, 1993; Jusczyk, Cutler & Redanz, 1993; Morgan, 1996; Echols, Crowhurst & Childers, 1997; Houston, Jusczyk, Kuijpers, Cooler & Cutler, in review; Saffran et al., 1996). The Acoustic-only model attempted to discover words from fluent speech using only recurrence patterns within the speech corpus. The relatively poor performance may have been increased by adding analysis components which take advantage of prosodic and phonotactic cues.

Recent findings suggest that infants' phonetic discrimination performance drops during word learning (Stager & Werker, 1997). Assuming that increased phonetic discrimination aids in speech segmentation, this finding seems to indicate that word learning may in fact interfere with segmentation. Although the net effect of phoneme discrimination and segmentation is unclear, the results with CELL also show that the intermodal structure helps

segmentation. Whether combining the two activities results in a net gain or loss remains to be determined. Simulations of decreased phonetic discrimination within the CELL framework may shed light on this issue of language acquisition.

CELL relies on structure across channels of input, but is not tied to the specific channels of speech and shape discussed in this paper. Thus findings that, for example, blind children are able to acquire sight related words (Landau & Gleitman, 1985) does not pose a problem for the model. The underlying processes of CELL should work equally well with other channels and modalities of input[12] although this remains to be tested.

The CELL model is based on techniques originally developed for automatic speech recognition, computer vision, and machine learning. Lessons learned from modeling infant learning using these techniques may in turn be applied to developing new approaches for machine learning. In current approaches to training speech and vision systems, a necessary step is to collect and annotate a corpus of speech and images. Infants, on the other hand, learn without annotations or transcriptions. Infant-inspired strategies may be applied to develop recognition systems which mimic human learning strategies and avoid the need for costly manual preparation of training input (Roy, 2000a, 2000b).

CELL is not meant to be a complete model of word learning. CELL represents our effort to explore a word learning strategy which leverages intermodal structure. Other complementary strategies will likely improve CELL's performance. For example, analysis of isolated words and segment boundaries of utterances (which coincide with word boundaries) will provide additional segmentation cures. Unsupervised clustering of visual data may speed formation of labeled visual categories. Prosodic processing may provide cues for both segmentation, and help locate semantically salient words and phrases.

6.1. Assumptions in the model

The CELL model rests on a set of simplifying assumptions which were made in order to implement a computational model of word learning. These assumptions may have had implications on the performance of the model.

We assumed that the STM is able to hold approximately 10 s of speech represented in terms of phoneme probabilities. This may be a somewhat optimistic estimate of preverbal working memory capacity. The STM duration may be reduced without significantly reducing learning rates by filtering out some portions of input speech so that only salient portions of the signal enter the STM. Such an approach would require designing filters which are able to reliably select portions of utterances which are more likely to contain semantically salient words. One approach may be to use filters based on prosodic and or nonspeech cues.

Spoken utterances are typically heard in the context of visually complex environments. To simplify the visual processing, however, we assumed that each spoken utterance is paired with a single object. Selection from competing potential referents is not required since a fixed context is paired with each utterance. To cope with more realistic input containing multiple objects and complex backgrounds,

CELL would need to include a model of visual selection to decide which object to pair with which spoken input. Visual selection is an extremely difficult problem which needs to address many complex issues including the analysis of caregiver intent. The supporting

visual system would also grow in complexity since it would need to parse complex visual scenes. The single object assumption used in the experiments reported in this paper by-pass these problems thereby greatly simplifying the problem of acquiring visual categories.

During data collection, it was observed that infants were not always visually attending to the target object when an utterance was spoken by the caregiver. During data processing, however we assumed that the infant was attending to the object during each utterance. The generalization was made to simplify data preparation and to minimize experimenter preprocessing of the data. Performance may have been improved if we discarded utterances that occurred in cases when the infant was not attending to the target object. In these instances, spoken utterances were less likely to contain direct references to the object and thus acted as noise for the learning system.

A third simplification to the visual input was that only object shape information was available to the model. This forced the model to learn words which could be grounded in shape, and avoided potential problems of trying to simultaneously learn groundings in alternate context channels (or combinations of context channels). This simplification is somewhat justified since young infants are known to have a "shape bias" in that they prefer to assign names to shapes rather than colors and other visual classes (Landau et al., 1988). In previous experiments (Roy, 1999), CELL has simultaneously learned words referring to both shape and color categories. No significant interference problems were found across contextual channels.

A fourth assumption inherent in the visual representation is that the correspondence between different views of the same physical object is given. In other words, when a view-set is generated from an object, CELL assumes that all views in the set belong to the same object. This seems to us to be a reasonable assumption since in real situations, the learner may smoothly shift his or her perspective to gather a set of views of a single object. Note that the correspondence between two separate view-sets of the same object are *not* given. Since two view-sets of a shared underlying object will never be identical (due to variations in camera pose, lighting, etc.), a correspondence problem exists for the system at this level. The correspondence problem at the *object class* level, that is, establishing the correspondence between a view-set representing Car A with that of Car C is yet more difficult and also addressed by CELL.

CELL assumes that built-in mechanisms for representing speech in terms of phonemes and for extracting statistical representations of shapes are available prior to word learning. The representation of speech in terms of phonemes is derived from acoustic input using a recurrent neural network. Although the resulting representation of speech often contains phoneme confusions, most confusions are within broad classes of phonemes so that sound patterns may still be effectively compared. At least coarse phonemic representations may reasonably be assumed since infants as young as 6 months are able to make language-dependent phonemic discriminations (Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992). The CELL model does not attempt to account for how initial phoneme discrimination abilities arise prior to word learning. Experimental evidence also supports the assumption that infants have crude object representations and comparison mechanisms (Milewski, 1976; Buschnell, 1979). The shape representations employed in CELL are based on these findings.

6.2. Sensor grounded input

All input processed by CELL is derived from raw sensory signals. In the current implementation, linguistic input comes from a microphone, and contextual information is obtained from a color video camera. The sensory grounded nature of CELL differs significantly from other models of language acquisition which typically provide human-generated representations of speech and semantics to the system. In these models, speech is often represented by text or phonetic transcriptions. As a result, each time a word appears in the input, the model receives a consistent sequence of tokenized input (e.g., Harrington et al., 1989; Aslin, Woodward, LaMendola & Bever, 1996; Brent & Cartwright, 1996; Brent, 1999). Similarly, semantics are usually encoded by a predefined set of symbols or structured sets of symbols (e.g., Siklossy, 1972; Sankar & Gorin, 1993; Siskind, 1992; de Marcken, 1996). In contrast CELL receives noisy input both in terms of raw acoustic speech and unannotated camera images. The problems of word learning would have been greatly simplified if CELL had access to such consistent representations.

We believe that using raw sensory input bears closer resemblance to the natural conditions under which infants learn. Infants only have access to their world through their perceptual systems. There is no teacher or trainer who provides consistent and noiseless data for the infant. Similarly, there should be no equivalent teacher or trainer to help the computational model.

Consider the difference between raw audio and phonetic transcriptions of speech. In raw speech, pronunciations vary dramatically due to numerous factors including phonetic context, syllable stress, the speaker's emotional state, the speaker's age, and gender. On the other hand, a trained transcriptionist will abstract away all these factors and produce a clean phonetic transcription. Transcriptions are not mere equivalents of raw audio with some "noise" removed. Transcriptionists leverage their knowledge of language to overcome ambiguities in the acoustic signal and thus have the potential to influence the model. Pre-existing knowledge is bound to trickle into any model which relies on human-prepared input. In addition, raw speech contains prosodic information which may also provide cues for segmentation and determining points of emphasis. Such information is also lost when the speech signal is reduced to only a phonetic transcript.

Given that CELL operates on sensor data, we expected performance to be somewhat degraded in comparison to computational models which process symbolic input. By using robust signal representations and statistical modeling techniques, however, we were nonetheless able to obtain promising performance.

7. Conclusions

The CELL model represents an important step towards applying methods from signal processing and pattern recognition for the purpose of modeling language acquisition. By using these techniques, models may be implemented which process sensory data without the aid of human annotations or transcriptions. Complex models which involve the interaction of several processes and which are intimately tied to the nature of input may be tested in such

computational frameworks. We plan to expand the types of contextual representations available to the model enabling acquisition of richer classes of words and relations between words.

Notes

1. For a general description of the CELL model which learns from an arbitrary set of sensory channel, see (Roy, 1999).
2. Density estimation is typically performed by assuming the form of the distribution for classes of data. Popular choices of distributions include the Gaussian or mixtures of Gaussians. A set of observations from a known class are used to estimate the parameters of the density (for example, the mean and covariance of a Gaussian density). Once trained, a set of class-conditional densities can be used to classify new unseen observations by selecting the density which most likely generated the observation (weighted by the prior probability of observing that class).
3. We use the set of 40 English phonemes defined in (Lee, 1988).
4. An all-pole model of the spectrum is estimated using linear predictive coding (Oppenheim & Schafer, 1989).
5. The TIMIT corpus is divided into train and test sets for evaluation purposes. Recognizers are trained using only the training data, and then evaluated on (unseen) test data.
6. The Viterbi algorithm is commonly used in speech recognition applications to efficiently find the most likely HMM state sequence corresponding to an observed observation sequence.
7. A threshold was empirically set to determine matches for each metric. We found little sensitivity to this threshold as long as it erred on the side of extracting of too many recurrent segments.
8. CELL did not process these video recordings. They were used to keep track of which object was in play for each recorded spoken utterance. We used this information to pair speech with appropriate object images in order to train CELL.
9. Performance of the model for lower ranking lexical items deteriorates. (Roy, 1999) suggests a remedy to this problem which uses reinforcement feedback to determine a cutoff point automatically
10. Onomatopoeic items do not contribute to the denominator of the sums.
11. The only exception was that one of the caregiver's performance on Measure 1 (segmentation accuracy) increased from 20% to 33% using CELL versus the Acoustic-only Model, respectively.
12. See (Roy, 1999) for a modality-independent presentation of CELL.

Acknowledgments

The authors would like to thank the following people for helping shape this work: Allen Gorin, Steven Pinker, Rupal Patel, Bernt Schiele. The presentation of this paper has also benefited from recommendations of anonymous reviewers.

Appendix A: cross-validation of STM size to support recurrence filtering

The first stage of speech segmentation in CELL occurs in the recurrence filter which searches for recurrent speech segments within STM. In all experiments reported in this paper, the STM size was set to five utterances. In other words, in this work we assume that at least some target words will recur within five contiguous utterances spoken by the caregiver. The results of our experiments verify that this was the case in our corpus. Since this corpus was collected in a structured environment, we additionally performed a recurrence analysis of the text transcripts of a larger study from the CHILDES database (MacWhinney, 2000).

The Warren-Leubecker (Warren-Leubecker, 1982; Warren-Leubecker & Bohannon, 1984) corpus contains text transcripts of caregiver-child interactions from 20 different families. The children in this study ranged in age from 1;6 to 3;1. Recordings were made in the homes of the children. Caregivers were told to play naturally with their children.

We were particularly interested in words whose referents could be directly perceived in the environment (since this is the class of words which CELL is able to learn with appropriate contextual channels). These words might be thought of as *groundable words.* To study the effect of STM on the detection of groundable words, the total vocabulary across all 20 sets of transcripts was compiled. This resulted in a list of 2940 unique words. From this list, we manually extracted a list of 615 groundable words. Most groundable words belonged to one of four broad classes: object/people names (ball, mommy, etc.), adjectives (red, shiny, itsy, etc.) and perceivable verbs (eat, throw, etc.), and spatial relations (above, below).

We wrote a program which simulates the STM recurrence filter in CELL for text processing. This program slides a “window” over the Warren-Leubecker transcripts and searches exhaustively for recurring words only within this window. The search is based on exact string matching since the Warren-Leubecker is a text-only corpus. All 20 transcripts were processed by this program for various sized STMs. For each setting of STM size, we measured the percentage of groundable words which were deteced by the recurrence filter.

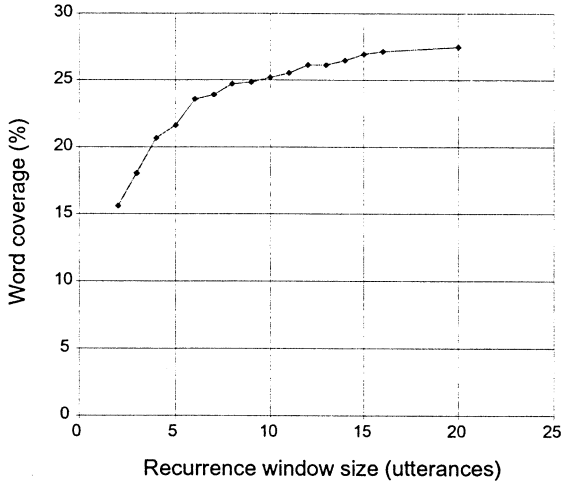

Fig. 15. Detection of groundable words (see text) as a function of STM size.

Fig. 15 shows the results of the recurrence analysis. With an STM which only holds pairs of adjacent utterances from the caregiver (size = 2), 15.6% of the 615 groundable words are detected. With a STM of 20 utterances the detection rate increases to 27.5%. Recall that the computational cost of exhaustively searching the STM grows exponentially with the size of the STM. Thus a balance between detection rate and STM size must be achieved. The "knee" in the curve in Fig. 15 suggests that choosing an STM which holds five or six utterances may be a reasonable compromise. With an STM of five utterance as was used in all experiments with CELL reported in this paper, 21.6% of groundable words are detected in the Warren-Leubecker corpus. This result shows that in even more natural infant-directed speech, windowed recurrence analysis is an effective approach for bootstrapping the speech segmentation process while limiting computational load.

References

Aslin, R., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax* (p. 117–134). Mahwah, NJ: Erlbaum.

Bishop, C. (1995). *Neural networks for pattern recognition.* Oxford: Clarendon Press.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34,* 71–106.

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

Buschnell, I. (1979). Modification of the externality effect in young infants. *Journal of Experimental Child Psychology, 28,* 211–229.

Cover, T., & Thomas, J. A. (1991). *Elements of information theory.* New York, NY: Wiley-Interscience.

Cutler, A., & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics, 21,* 103–108.

de Marcken, C. (1996). *Unsupervised language acquisition.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Echols, C., Crowhurst, M., & Childers, J. (1997). Perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language, 36,* 202–225.

Garofolo, J. (1988). *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database.* Gaithersburgh, MD: National Institute of Standards and Technology (NIST).

Harrington, J., Watson, G., & Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. *Computer Speech and Language, 3,* 367–382.

Harris, Z. (1954). Distributional structure. *Word, 10,* 146–162.

Hermansky, H., & Morgan, N. (1994). Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing, 2* (4) 578–589.

Houston, D., Jusczyk, P., Kuijpers, C., Coolen, R., & Cutler, A. (in review). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin and Review.*

Huttenlocher, J., & Smiley, P. (1994). Early word meanings: the case of object names. In P. Bloom (Ed.), *Language acquisition: core readings* (pp. 222–247). Cambridge, MA: MIT Press.

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

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

Landau, B., & Gleitman, L. (1985). *Language and experience: evidence from the blind child.* Cambridge, MA: Harvard University Press.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3,* 299–321

Lee, K. (1988). *Large-vocabulary speaker-independent continuous speech recognition: the sphinx system.* Unpublished doctoral dissertation, Computer Science Department, Carnegie Mellon University.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

Milewski, A. (1976). Infant's discrimination of internal and external pattern elements. *Journal of Experimental Child Psychology, 22,* 229–246.

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

Oppenheim, A., & Schafer, R. (1989). *Digital signal processing.* Englewood Cliffs, New Jersey: Prentice Hall.

Plunkett, K., Sinha, C., Moller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science, 4* (3&4), 293–312

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Quine, W. (1960). *Word and object.* Cambridge, MA: MIT Press.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77* (2), 257–285.

Regier, T. (1996). *The human semantic potential.* Cambridge, MA: MIT Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14,* 465–471.

Robinson, T. (1994). An application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks, 5* (3).

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology, 104,* 192–233.

Rose, R. (1996). Word spotting from continuous speech utterances. In C. Lee, F. K. Soong, & K. Paliwal (Eds.), *Automatic speech and speaker recognition* (pp. 303–329). Kluwer Academic.

Roy, D. (1999). *Learning words from sights and sounds: A computational model.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Roy, D. (2000a). Integration of speech and vision using mutual information. In *Proceedings of ICASSP.* Istanbul, Turkey.

Roy, D. (2000b). Learning from multimodal observations. In *Proceedings of the IEEE international conference on multimedia.* New York, NY.

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

Sankar, A., & Gorin, A. (1993). Adaptive language acquisition in a multi-sensory device. In *Artificial neural networks for speech and vision* (pp. 324–356). London: Chapman and Hall.

Schiele, B., & Crowley, J. (1996). Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR'96 proceedings of the 13th international conference on pattern recognition, volume b* (pp. 50–54).

Siklossy, L. (1972). Natural language learning by computer. In H. A. Simon & L. Siklossy (Eds.), *Representation and meaning: experiments with information processing systems* (pp. 288–328). Englewood Cliffs, NJ: Prentice-Hall.

Siskind, J. (1992). *Naive physics, event perception, lexical semantics, and language acquisition.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

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

Snow, C. (1977). Mothers' speech research: From input to interaction. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: language input and acquisition.* Cambridge, MA: Cambridge University Press.

Stager, C., & Werker, J. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388,* 381–382.

Warren-Leubecker, A. (1982). *Sex differences in speech to children.* Unpublished doctoral dissertation, Georgia Institute of Technology.

Warren-Leubecker, A., & Bohannon, J. (1984). Intonation patterns in child-directed speech: Mother-father speech. *Child Development, 55,* 1379–1385.

Wright, J., Carey, M., & Parris, E. (1996). Statistical models for topic identification using phoneme substrings. In *Proceedings of ICASSP* (pp. 307–310).