Grounded Spoken Language Acquisition: Experiments in Word Learning

Deb Roy, *Member, IEEE*

***Abstract*—Language is grounded in sensory-motor experience. Grounding connects concepts to the physical world enabling humans to acquire and use words and sentences in context. Currently most machines which process language are not grounded. Instead, semantic representations are abstract, pre-specified, and have meaning only when interpreted by humans. We are interested in developing computational systems which represent words, utterances, and underlying concepts in terms of sensory-motor experiences leading to richer levels of machine understanding. A key element of this work is the development of effective architectures for processing multisensory data. Inspired by theories of infant cognition, we present a computational model which learns words from untranscribed acoustic and video input. Channels of input derived from different sensors are integrated in an information-theoretic framework. Acquired words are represented in terms of associations between acoustic and visual sensory experience. The model has been implemented in a real-time robotic system which performs interactive language learning and understanding. Successful learning has also been demonstrated using infant-directed speech and images.**

***Index Terms*—Cross-modal, language learning, multimodal, semantic grounding.**

I. Introduction

LANGUAGE is grounded in experience. Unlike dictionary definitions in which words are defined in terms of other words, humans understand basic concepts in terms of associations with sensory-motor experiences (cf. [1]–[4]). To grasp the concepts underlying words such as *red, heavy,* and *above* requires interaction with the physical world. This link to the body and the environment is a fundamental aspect of language which enables humans to acquire and use words and sentences in context.

Although many aspects of human cognition and language processing are not clearly understood, we can nonetheless draw lessons from human processing to guide the design of intelligent machines. Infants learn their first words by associating speech patterns with objects, actions, and people [5]. The primitive meanings of words and utterances are inferred by observing the world through multiple senses. Multisensory grounding of early words forms the foundation for more complex concepts and corresponding linguistic capacities. Syntax emerges as children begin to combine words to refer to relations between concepts. As the language learner's linguistic abilities mature, their speech refers to increasingly abstract notions. However, all words and utterances fundamentally have meaning for humans because of their grounding in multimodal and embodied experience. The sensory-motor basis of semantics provides common ground for people to understand each another.

In contrast, currently most automatic spoken language processing systems are not grounded. Machine training is based on recordings of spoken utterances paired with manually generated transcriptions and semantic labels. Depending on the task, the transcriptions may vary in level of abstraction ranging from low level phonetic labels to high level semantic labels. Various statistical methods including hidden Markov models (HMMs) and neural networks are employed to model acoustic-to-label mappings. In this paper we refer to the general approach of modeling mappings from speech signals to human specified labels as "ungrounded speech understanding," since the semantics of the speech signal are only represented abstractly in the machine. The use of abstract labels isolates the machine from the physical world. The ungrounded approach has lead to many practical applications in transcription and telephony. There exist, however, fundamental limits to the ungrounded approach.

We can anticipate the limitations of ungrounded speech understanding by comparison with human counterparts. At least two interrelated advantages can be identified with the grounded approach. First, the learning problem may be solved without labeled data since the function of labels may be replaced by contextual cues available in the learner's environment. Language does not occur in a vacuum. Infants observe spoken language in rich physical and social contexts. Furthermore, infant-directed speech usually refers to the immediate context [6]; caregivers rarely refer to events occurring in another time or place. This connection of speech to the immediate surroundings presumably helps the infant to glean the meaning of salient words and phrases by observing contexts in which speech occurs. The advantage to this approach is that the learner acquires knowledge from observations of the world without reliance on labeled data. Similar advantages are anticipated for machines.

A second advantage of the grounded approach is that speech understanding can leverage context to disambiguate words and utterances at multiple levels ranging from acoustic to semantic ambiguity. The tight binding of language to the world enables people to integrate nonlinguistic information into the language understanding process. Acoustically and semantically ambiguous utterances can be disambiguated by the context in which they are heard. We use extra-linguistic information so often and so naturally that it is easy to forget how vital its role is in language processing. Similar advantages can be expected for machines which are able to effectively use context

Manuscript received January 11, 2001; revised October 22, 2001. This work was supported in part by AT&T. The associate editor coordinating the review of this paper and approving it for publication was Dr. Thomas R. Gardos.

The author is with the Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: dkroy@media.mit.edu).

Digital Object Identifier 10.1109/TMM.2003.811618

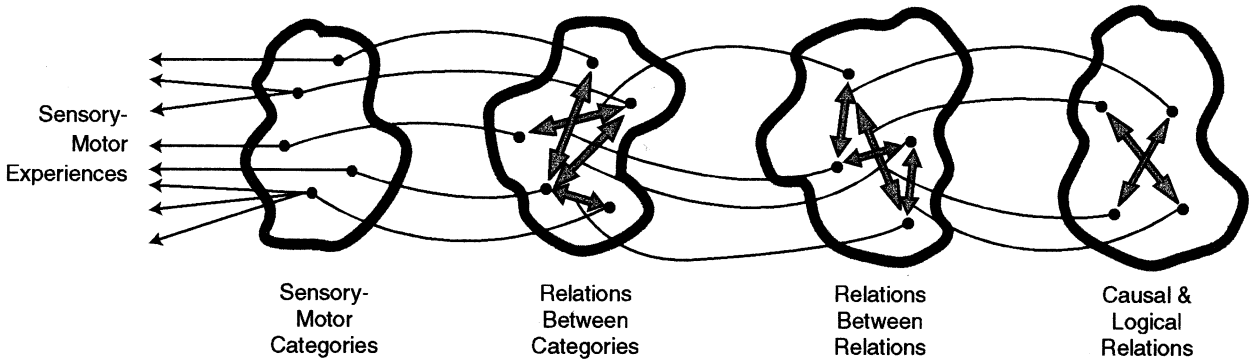


Fig. 1. Levels of conceptual abstraction grounded in sensory-motor experience. Language is acquired by forming concepts, and learning associations from words and utterances to conceptual structures.

when processing language. These advantages motivate us to investigate grounded speech acquisition.

It is illuminating to examine the differences between learning procedures for speech systems and infants. Traditionally, speech understanding systems are trained by providing speech and corresponding transcriptions (which may include semantic labels in addition to phonetic and word labels). This constitutes drastically impoverished input when compared with infants. With such a handicap, infants would be unlikely to acquire much language at all. Training with labeled data does have its advantages. The recognition task is well defined, and mature techniques of supervised machine learning may be employed for parameter estimation of classifiers. We propose new methods which explore more human-like learning from multiple channels of unlabeled data. Although the learning problem becomes more challenging, the potential payoffs are great. Our goal is to build multimodal understanding systems which leverage cross-channel information, leading to more intelligent and robust systems, and which can be trained from untranscribed data.

This paper presents a model of grounded language learning called CELL (Cross-Channel Early Lexical Learning). CELL leverages cross-modal structure to segment and discover words in continuous speech, and to learn visual associations for those words. Rather than rely on transcriptions or labels, speech provides noisy and ambiguous labels for video, and vice versa. We describe new algorithms which have been developed to implement this model in a real-time audio–visual processing system. The system has been embedded into a robotic embodiment enabling language learning and understanding in "face-to-face" interactions. We also present experimental evaluations with infant-directed speech and co-occurring video in which word learning was achieved in the face of highly spontaneous speech.

II. Grounding: Connecting Meaning to the World

Grounding in its most concrete form is achieved by giving machines the capacity to sense and act upon the physical world. Since humans also sense and act upon the same world, this shared physical context provides a common ground which mediates communication between humans and machines. Fig. 1 illustrates how abstract concepts can emerge from sensory-motor experience through layers of analysis. At the left side of the figure, interactions with the physical world give rise to sensory and motor (or action) categories. Structures which represent relations between these categories are inferred at increasing levels of abstraction to the right. Ultimately, causal and logical relations may be inferred if appropriate types of structured learning are employed. Our current work is restricted to the first two levels shown in the figure but the framework leads naturally to higher levels of conceptual and linguistic learning. Based on this philosophy, we have built communication systems which ground all input in physical sensors.

Humans are endowed with similar sensory and motor capacities. This shared endowment results in similar semantic representations at least at the lowest levels of abstraction. No person is able to perceive infrared or ultraviolet rays, and thus no young child will naturally acquire words grounded in these referents. Young children's first nouns label small objects [7], probably since those are the objects they are able to manipulate with their hands and thus build up sufficiently accurate models. Names of larger objects are only acquired in later stages of development. The design of sensors and manipulators regulates the type of concepts which a machine can acquire. We argue that machines must at a functional level share the abilities and limits of our physiology if they are to acquire human-like semantics.

An emphasis is placed on grounding all learning in sensors, avoiding any reliance on human-generated labels or transcriptions. This ensures that the machine will develop representations which capture the richness inherent in continuous variations of the physical world. From an engineering perspective, sensory grounding forces us to adopt statistical approaches which are robust to various types of noise encountered in sensory signals.

Although this paper focuses on grounding language in the physical world, in many situations it may also be useful to ground semantics in virtual worlds [8]–[11]. For example, in [11], we created a video game in which a synthetic character could "see" objects in a virtual world using synthetic vision. The semantics of spoken words were grounded in attributes of virtual objects enabling speech-based human–machine interaction in the course of playing the video game. In many situations, the level of semantic abstraction required in a communication task might render direct physical grounding impractical. In such cases, a virtual representation of the task may serve as a useful proxy to ground human–machine communication. The common denominator across virtual and physical grounding is that both humans and machines have perceptual access to shared nonlinguistic referents.

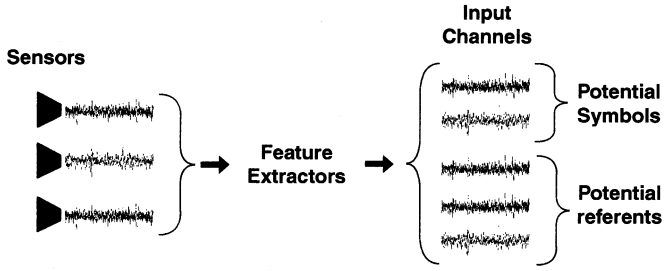


Fig. 2. Framework for learning from untranscribed sensory data. Feature detectors extract channels of input from sensors. The input channels are divided into two sets. The first carries symbolic information such as words and signed gestures. The second set carries representations of referents which may be associated with symbols. For example, visual channels may represent the shape or color of objects which are associated with shape and color symbolic terms.

III. Learning Cross-Channel Structure

The world does not provide infants with transcribed data. Instead, the environment provides rich streams of continuously varying information through multiple modes of input. Infants learn by combining information from multiple modalities. A promising path of research is to build machines which similarly integrate evidence across modalities to learn from naturally occurring data without supervision [12], [13]. The key advantage to this approach is that potentially unlimited new sources of untapped training data may be utilized to develop robust recognition technologies. Ultimately, we envision machines which actively explore their world and acquire knowledge from sensory-motor interactions.

Fig. 2 shows our framework for learning from multisensory input. A set of sensors provides input. Feature detectors extract channels of input from the sensors. In general the number of input channels is greater than the number of sensors. For example, shape, color, texture, and motion channels might be extracted from a camera. Phonemes, speaker identity, and prosody (e.g., pitch, loudness) are examples of channels which might be extracted from acoustic input. A subset of the input channels are assumed to represent symbolic information (words and phrases). The remaining channels represent the referents of these symbols. The goal of learning is to appropriately segment and cluster incoming data in the input channels in order to identify and build associations between symbols and referents.

Recent models of language acquisition include models of speech segmentation based on minimum description length encoding of acoustic representations [14], [15], and cross-situational learning from text coupled with line drawings representing simple visual semantics [8]–[10]. Algorithms for acquiring syntactic structure and semantic associations for acoustic words based on semantic transcriptions have been demonstrated [16]. This work has lead to tabula rasa learning of acoustic vocabularies and higher level language structures from speech recordings transcribed at only the semantic level [17]. Physical grounding of concepts has been explored in the context of robotics as an alternative to the symbol processing view of artificial intelligence [18], [19]. The model presented in this paper departs from previous work in language learning in that both words and their semantics are acquired from sensor input without any human-assisted transcription or labeling of data.

IV. CELL: Cross-Channel Early Lexical Learning

To explore issues of grounded language, we have created a system which learns spoken words and their visual semantics by integrating visual and acoustic input [20]. The system learns to segment continuous speech without an a priori lexicon and forms associations between acoustic words and their visual semantics. This effort represents a step toward introducing grounded semantics in machines. The system does not represent words as abstract symbols. Instead, words are represented in terms of audio-visual associations. This allows the machine to represent and use relations between words and their physical referents. An important feature of the word learning system is that it is trained solely from untranscribed microphone and camera input. Similar to human learning, the presence of multiple channels of sensory input obviates the need for manual annotations during the training process. In the remainder of this paper we present the model of word learning and describe experiments in testing the model with interactive robotics and infant-directed speech.

We have developed a model of CELL, summarized in Fig. 3 [20], [21]. This model discovers words by searching for segments of speech which reliably predict the presence of co-occurring visual categories. Input consists of spoken utterances paired with images of objects. In experiments presented later in this paper, we present results using spoken utterances recorded from mothers as they played with their infants in natural settings. The play centered around everyday objects such as shoes, balls, and toy cars. Images of those objects were paired with the spontaneous speech recordings to provide multisensory input to the system. Our goal was to approximate the input that an infant might receive when listening to a caregiver and simultaneously attending to objects in the environment. The output from CELL consists of a lexicon of audio-visual items. Each lexical item includes a statistical model (based on HMMs) of an acquired spoken word, and a statistical visual model of either a shape or color category. To acquire lexical items, the system must (1) segment continuous speech at word boundaries, 2) form visual categories, and 3) form appropriate correspondences between word and visual models.

The correspondence between speech and visual streams is extremely noisy. In experiments with infant-directed speech described in Section IX, the majority of spoken utterances in our corpus contained no direct reference to the co-occurring visual context. Thus the learning problem CELL faces is extremely challenging since the system must "fish out" salient cross-channel associations from noisy input.

Camera images of objects are converted to statistical representations of shapes. Spoken utterances captured by a microphone are mapped onto sequences of phoneme probabilities. A *short term memory* (STM) buffers phonetic representations of recent spoken utterances paired with representations of co-occurring visual input. A short-term recurrence filter searches the STM for repeated subsequences of speech which occur in matching visual contexts. The resulting pairs of speech segment and shape representations are placed in a *long term memory* (LTM). A filter based on mutual information searches the LTM for speech-shape or speech-color pairs which usually occur together, and rarely occur apart within the LTM. These

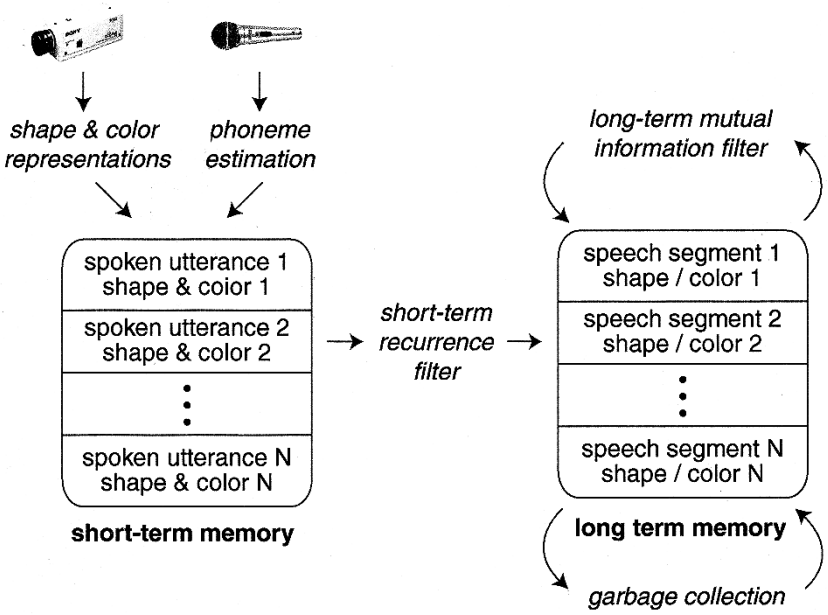


Fig. 3. CELL model. A layered memory architecture combined with recurrence and mutual information filters (see text) are used to acquire an audio–visual lexicon from unlabeled input.

pairings are retained in the LTM, and rejected pairings are periodically discarded by a garbage collection process.

V. Representing and Comparing Spoken Utterances

Motivated by the fact that infants at the age of six months[1] possess language-specific phonemic discrimination capabilities [22], [23], the system is "endowed" with pretrained English phoneme feature extraction. Spoken utterances are represented as arrays of phoneme probabilities. A recurrent neural network similar to [24] processes RASTA-PLP coefficients [25] 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 input coefficients. The RNN was trained off-line using back-propagation in time [26] with the TIMIT database of phonetically transcribed speech recordings [27].[2] The RNN recognizes phonemes with 69.4% accuracy using the standard TIMIT training and test datasets. Session recordings are segmented into utterances by detecting contiguous segments of speech in which the probability of silence estimated by the RNN are low.

Spoken utterances are segmented in time along phoneme boundaries, providing hypotheses of word boundaries. To locate phoneme boundaries, the RNN outputs are treated as state emission probabilities in a HMM framework. The Viterbi dynamic programming search [28] 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 in the utterance and 2) the location of each phoneme boundary for the sequence (this information is recovered from the Viterbi search). Each phoneme boundary can serve as a speech segment start or end point. Any subsequence within an utterance terminated at phoneme boundaries can form a word hypothesis.

We define a distance metric, $d_A()$, which measures the similarity between two speech segments. One possibility is 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 [29]. A limitation of this method is that it relies on only the single most likely phoneme sequence. A sequence of RNN output is equivalent to an unpruned phoneme lattice from which multiple phoneme sequences may be derived. To make use of this additional information, we developed the following distance metric.

Let $Q = \{q_1, q_2, \ldots, q_N\}$ be the best-path 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 within the states of a phoneme 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 decoded as 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 [28] 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

[1]As with any learning system, certain structures must be made "innate" to support data-driven learning. Given our goal of word learning, we chose to start at the "six-month-old" stage, a point at which infants are able to discern phonemic speech sound differences but have not begun word learning. To model different stages of language acquisition such as phonological or syntactic learning, different choices of what to make innate would have been made.

[2]Note that the use of transcribed data was strictly for the purpose of training the RNN to serve as a feature detector for generating phoneme probabilities. Word learning was performed by CELL on our new experimental database without transcriptions.

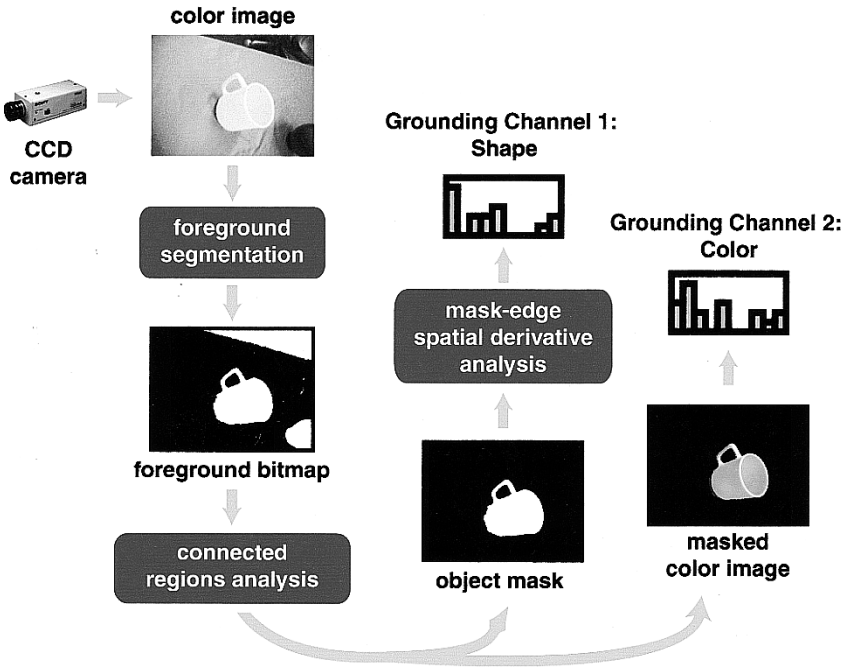


Fig. 4. Extraction of object shape and color channels from a CCD camera.

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 [30]. In this method, each likelihood estimate is scaled by the likelihood of a default alternate hypothesis, λ^A:

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

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

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

In practice, we have found this metric to robustly detect phonetically similar speech segments embedded in spontaneous speech. It is used as the basis for determining acoustic matches between segments in the recurrence filter used by the STM, and by the mutual information filter used to build lexical items from LTM (see Section VII).

VI. Visual Processing

Motivated again by the visual abilities of preverbal infants [31], [32], the system is "endowed" with color and shape feature extractors. Three-dimensional (3-D) objects are represented using a view-based approach in which multiple two-dimensional (2-D) images of an object captured from multiple viewpoints collectively form a model of the object. The 2-D representations were designed to be invariant to transformations in position, scale and in-plane rotation. The representation of color is invariant under changes in illumination. Fig. 4 shows the stages of visual processing used to extract representations of object shapes and colors.

Figure-ground segmentation is accomplished by assuming that the background has uniform color. A Gaussian model of the illumination-normalized background is estimated from a set of 20 images. Given a new image, the Gaussian model is evaluated at each pixel and thresholded (using an empirically determined threshold value) to classify pixels as either background or foreground. Large connected regions of pixels classified as foreground indicate the presence of an object.

The 3-D shape of an object is represented using a set of histograms, each of which represents the silhouette of the object from a different viewpoint.[3] We assume that with sufficient stored viewpoints, a novel viewpoint of an object may be matched by interpolation. Given the pixels of an image which correspond to an object using figure-ground segmentation, the following steps are used to build a representation of the object's silhouette.

- Locate all outer edge points of the object by finding all foreground pixels adjacent to background pixels. Edge points in the interior of the object are ignored.
- For each pair of edge points, compute two values: 1) the Euclidean distance between the points, normalized by the largest distance between any two edge points of that silhouette, and 2) the angle between the tangents to the edge of the object at the two edge points.
- Accumulate a 2-D histogram of all distance-angle measurements.

The resulting histogram representation of the object silhouette is invariant under rotation (since all angles are relative) and object size (since all distances are normalized). Using multi-dimensional histograms to represent object shapes enables the

[3]Schiele and Crowley have shown that histograms of local images features are a powerful representation for object recognition [33].

use of information theoretical or statistical divergence functions for the comparison of silhouettes. Through experimentation we found the χ^2-divergence to be most effective:

$$d_V(X,Y) = \chi^2(X,Y) = \sum_{\mathbf{i}} \frac{(x_{\mathbf{i}} - y_{\mathbf{i}})^2}{x_{\mathbf{i}} + y_{\mathbf{i}}} \tag{2}$$

where $X = \cup_{\mathbf{i}} x_{\mathbf{i}}$ and $Y = \cup_{\mathbf{i}} y_{\mathbf{i}}$ are two histograms indexed by $\mathbf{i}$ and $x_{\mathbf{i}}$ and $y_{\mathbf{i}}$ are the values of a histogram cell.

The representation of three dimensional shapes is based on a collection of 2-D shape histograms, each corresponding to a particular view of the object. For all results reported in this paper, each three dimensional object is represented by 15 histograms. The 15 viewpoints are chosen at random. We found that for simple objects, 15 views are sufficient to capture basic shape characteristics. We refer to a set of histograms as a *view-set*. View-sets are compared by summing the divergences of the four best matches between individual histograms.

The color of objects is also represented using histograms. To compensate for lighting changes, the red (R), green (G), and blue (B) components of each pixel are divided by the sum of all three components $(R+G+B)$ resulting in a set of "illumination-normalized" values. Since all triplets of illumination-normalized values must add to 1.0, there are only two free parameters for each pixel. For this reason, the normalized blue value of all pixels are not stored (any one of the three colors could have been dropped). For each image, a 2-D color histogram is generated by accumulating illumination-normalized red and green values for each foreground pixel in the object. The normalized red and green values are divided into eight bins leading to and 8×8 histogram. Similar to the representation of shape, 15 color histograms are recorded for each image to capture color difference from different viewpoints. Also similar to shape comparisons, the sum of the χ^2-divergences of the four best matching views is used to compare the color of object.

VII. Audio-Visual Lexical Acquisition

The heart of the CELL model is a cross-channel learning algorithm which simultaneously solves the problems of speech segmentation, visual categorization, and speech-to-vision association. A key problem in clustering across different representations is the question of how to combine distance metrics which operate on distinct representations. In CELL, mutual information is used to quantify cross-channel structure. This section describes CELL's cross-channel lexical learning architecture; the following two sections provide results of using this algorithm for learning from robot-directed and infant-directed speech and images of objects.

Input to CELL consists of a series of spoken utterances paired with view-sets. We refer to an {utterance, view-set} pair as an *audio-visual event, or AV-event*. AV-events are generated when an object is in view while an spoken utterance is detected.

Lexical acquisition is comprised of two steps. In the first step, AV-events are passed through a first-in-first-out short term memory (STM) buffer. The buffer has a capacity of five AV-events.[4] When a new event is inserted into the buffer, a *recurrence filter* searches for approximately repeating audio and visual patterns within the buffer. If a speaker repeats a word or phrase at least twice within a five contiguous utterances while playing with similar shaped objects, the recurrence filter would select that recurrent sound-shape pair as a potential lexical item. The recurrence filter uses the audio and visual distance metrics presented earlier to determine matches. The distance metrics are applied independently to the visual and acoustic components of AV-events. When matches are found simultaneously using both metrics, a recurrence is detected. The recurrence filter performs an exhaustive search over all possible image sets and speech segments (at phoneme boundaries) in the five most recent AV-events. To summarize, output from the recurrence filters consists of a reduced set of speech segments and their hypothesized visual referents.

In the second step, the hypotheses generated by the recurrence filter are clustered using an information-theoretic measure, and the most reliable clusters are used to generate a lexicon. Let us assume that there are N sound-shape hypotheses in LTM. For simplicity we ignore the color channel in this example, but the same process is repeated across both input channels. The clustering process would proceed by considering each hypothesis as a reference point, in turn. Let us assume one of these hypotheses, X, has been chosen as a reference point. Each remaining $N-1$ hypotheses may be compared to X using $d_V()$ and $d_A()$. Let us further assume that two thresholds, t_V and t_A are defined (we show how their values are determined below). Two indicator variables are defined with respect to X:

$$A = \begin{cases} 0, & \text{if } d_A(X,h_i) > f_A \\ 1, & \text{if } d_A(X,h_i) \le f_A \end{cases} \tag{3}$$

$$V = \begin{cases} 0, & \text{if } d_V(X,h_i) > f_V \\ 1, & \text{if } d_V(X,h_i) \le f_V \end{cases} \tag{4}$$

where h_i is the i^{th} hypothesis, for $i = 1 \ldots N-1$. For a given setting of thresholds, the A and V variables indicate whether each hypothesis matches the reference X acoustically and visually, respectively. The mutual information between A and V is defined as [34]

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

The probabilities required to calculate $I(A;V)$ are estimated from frequency counts. To avoid noisy estimates, events which occur less than four times are disregarded. Note that $I(A;V)$ is a function of the thresholds t_A and t_V. To determine t_V and t_A, the system searches for the settings of these thresholds which maximizes the mutual information between A and V. Smoothing of frequencies avoids the collapse of thresholds to zero.

Each hypothesis is taken as a reference point and its point of maximum mutual information (MMI) is found. The hypotheses

[4]The size of the STM was determined experimentally and represents a balance between learning performance and speed. Smaller STMs lead to poor learning performance; larger STMs did not significantly improve learning, but dramatically increased learning speed.

which result in the highest MMI are selected as output of the system. For each selected hypothesis, all other hypotheses which match both visually and acoustically are removed from further processing. In effect, this strategy leads to a greedy algorithm in which the hypotheses with best MMI scores are extracted first.

The process we have described effectively combines acoustic and visual similarity metrics via the MMI search procedure. The mutual information metric is used to determine the goodness of a hypothesis. If knowledge of the presence of one cluster (acoustic or visual) greatly reduces uncertainty about the presence of the other cluster (visual or acoustic), then the hypothesis is given a high goodness rating and is more likely to be selected as output by the system.

An interesting aspect of using MMI to combine similarity metrics is the invariance to scale factors of each similarity metric. Each metric organizes sound-shape hypotheses independently of the other. The MMI search finds structural correlations between the modalities without directly combining similarity scores. As a result, the clusters which are identified by this method can locally and dynamically adjust allowable variances in each modality. Locally adjusted variances cannot be achieved by any fixed scheme of combining similarity metrics.

A final step is to threshold the MMI score of each hypothesis and select those which exceed the threshold. Automatic determination of this MMI threshold is not addressed in this work. In current experiments, it is set manually to optimize performance.

VIII. Interactive Robotic Implementation

To support human-machine interactions, CELL has been incorporated into a real-time speech and vision interface embodied in a robotic system. Input consists of continuous multiword spoken utterances and images of objects acquired from a video camera mounted on the robot. The visual system extracts color and shape representations of objects to ground the visual semantics of acquired words. To teach the system, a person places objects in front of the robot and describes them. Once a lexicon is acquired, the robot can be engaged in an object labeling task (i.e., speech generation), or an object selection task (i.e., speech understanding).

A. Robotic Embodiment

A four degree-of-freedom robotic armature has been constructed to enable active control of the orientation of a small video camera mounted on the end of the device (Fig. 5). An animated face has been designed to give the robot the appearance of a synthetic character. Facial features including eyelids, mouth and feathers are used to convey information about the state of the system to the user in a natural manner.

Direction of gaze: A miniature camera is embedded in the right "eyeball" of the robot. The direction of the camera's focus is apparent from the physical orientation of the robot and provides a mechanism for establishing joint attention.

Facial Expressions: Several servo-controlled facial features are used to convey information about the internal state of CELL. The eyes are kept open when the vision system is in use. Feathers mounted on the head are extended to an attentive pose when the audio processing system detects the start of an utterance. The robot's mouth (beak) moves in synch with output speech.

Spoken Output: A phoneme-based speech synthesizer[5] is used to convey internal representations of speech segments. The Viterbi decoder is used to extract the most likely phoneme sequence for a given segment of speech. This phoneme sequence is resynthesized using the phoneme synthesizer. Naturalness of output is improved by controlling the duration of individual phonemes based on observed durations in the Viterbi decoding.

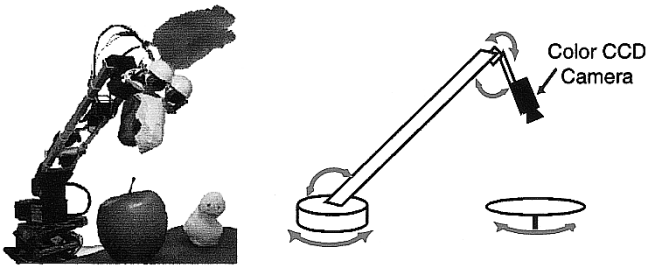


Fig. 5. A robot with four degrees of freedom used to capture images of objects. A small CCD camera is mounted in the right eyeball. A turntable provides a fifth degree of freedom for viewing objects from various perspectives. The turntable was only used for collecting images for the infant-directed speech experiments described in Section IX.

B. Acquiring a Lexicon

The robot has three modes of operation: acquisition, generation, and understanding. The mode is toggled manually through a software switch. In acquisition mode, the robot searches for the presence of objects on a viewing surface. When an object is detected, the system gathers multiple images to build a view-set of the object. If a spoken utterance is detected while the view-set is being gathered, an AV-event is generated and processed by CELL.

To teach the system, the user might, for example, place a cup in front of the robot and say, "Here's my coffee cup." To verify that the system received contextualized spoken input, it "parrots" back the user's speech based on the recognized phoneme sequence. This provides a natural feedback mechanism for the user to understand the nature of internal representations being created by the system.

C. Acquiring Lexical Order: A First Step Toward Syntax

To learn word order, a language learner must have some method of clustering words into syntactic categories. A syntax can then be used to specify rules for ordering word classes. In CELL, acquired lexicons are divided into two natural classes: words grounded in shape, and words grounded in color. Distributional analysis is used to track the ordering of word classes in utterances that contain both color and shape words in adjacent position (i.e., spoken with no intervening words).

In a pilot experiment, a single user provided the robot with 100 spoken utterances describing eight objects of varying

[5]The TrueTalk speech synthesizer made by Entropic Research Laboratory, Inc., 600 Pennsylvania Ave. SE, Suite 202, Washington, DC 20003.

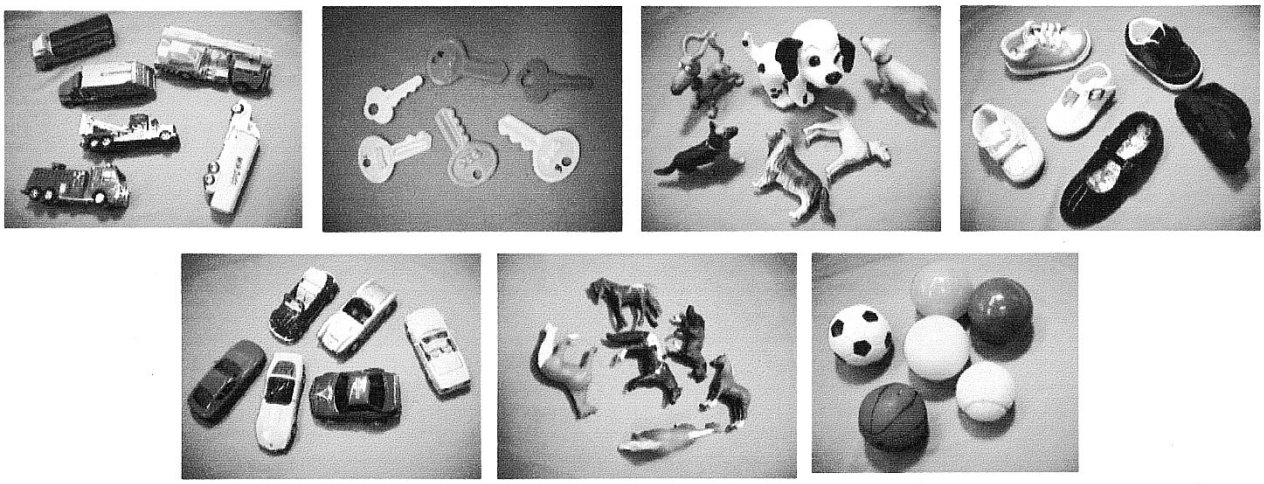


Fig. 6. Objects used during play in the infant-caregiver interactions.

shapes and colors. Approximately equal numbers of utterances were produced to describe each object. The speech was gathered in a spontaneous "face-to-face" setting with the robot running in its acquisition mode. From this small data set, the system learned that color terms precede shape terms in English. This information was encoded by a single statistic: a higher probability of shape-color compared to color-shape word pairs. This statistic was used to determine the sequence of words for speech generation, and to build a simple language model for speech understanding.

This experiment in word order learning represents a first step toward semantically grounded syntax acquisition. This method of linking early lexical learning to syntax acquisition is closely related to the semantic bootstrapping hypothesis which posits that language learners use semantic categories to seed syntactic categories [35], [36]. According to this theory, perceptually accessible categories such as objects and actions seed the syntactic classes of nouns and verbs. Once these seed categories have been established, input utterances are used to deduce phrase structure in combination with constraints from other innate biases and structures. In turn, the phrase structure can be used to interpret input utterances with novel words. Distributional analysis can be used to expand syntactic classes beyond initial semantically bootstrapped categories. In future work we plan to expand CELL to enable more complex aspects of grounded syntax learning.

D. Speech Generation

Once lexical items are acquired, the system can generate spoken descriptions of objects. In this mode, the robot searches for objects on the viewing surface. When an object is detected, the system builds a view-set of the object and compares it to each lexical item in LTM. The acoustic prototype of the best matching item is used to generate a spoken response. The spoken output may describe either shape or color depending on the best match.

To use word order statistics, a second generation mode finds the best matching LTM item for the color and shape of the object. The system generates speech to describe both features of the object. The order of concatenation is determined by the acquired word order statistics. When presented with an apple, the robot might say "red ball" (as opposed to "ball red") assuming it has already learned the words "red" and "ball," even if it had never seen an apple or heard that specific word sequence before.

E. Speech Understanding

When in the speech understanding mode, input utterances are matched to existing speech models in LTM. A simple grammar allows either single words or word pairs to be recognized. The transition probabilities between word pairs are determined by the acquired word order statistics.

In response to speech, the system finds all objects on the viewing surface and compares each to the visual models of the recognized lexical item(s). In a forced choice, it selects the best match and returns the robot's gaze to that object. In effect, the person can speak a phrase such as "brown dog," or "brown," or "dog," and the robot will find the object best matching the visual semantics of the spoken word or phrase.

To provide additional feedback, the selected object is used to index into LTM and generate a spoken description. This feedback leads to revealing behaviors when an incorrect or incomplete lexicon had been acquired. The nature of the errors provides the user with guidance for subsequent training interactions.

IX. Experiments With Infant-Directed Spontaneous Speech

To evaluate CELL on natural and spontaneous spoken input, experiments were conducted with a corpus of audio–visual data from infant-directed interactions [20]. Six caregivers and their prelinguistic (seven to 11 months) infants were asked to play with objects while being recorded. We selected seven classes of objects commonly named by young infants: balls, shoes, keys, toy cars, trucks, dog, horses [7]. A total of 42 objects, six objects for each class, were obtained (see Fig. 6). The objects of each class vary in color, size, texture, and shape.

Each caregiver-infant pair participated in six sessions over a course of two days. In each session, they played with seven objects, one at a time. All caregiver speech was recorded using

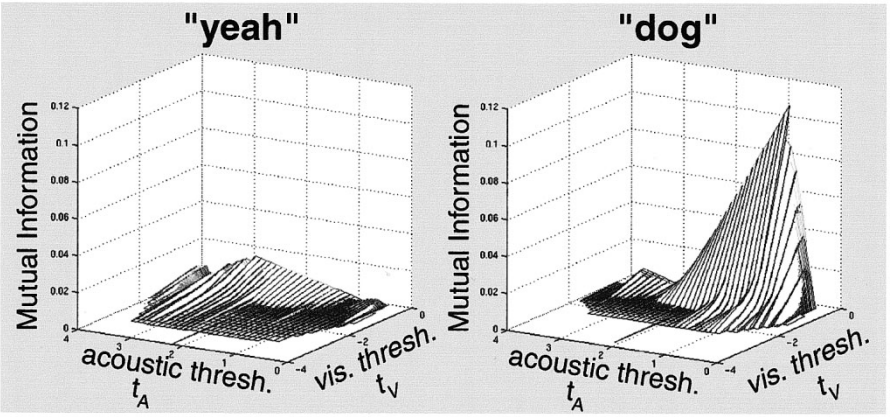


Fig. 7. Mutual information as a function of the acoustic and visual thresholds for two lexical candidates.

a wireless headset microphone onto DAT. In total we collected approximately 7600 utterances comprising 37 000 words across all six speakers. Most utterances contained multiple words with a mean utterance length of 4.6 words.

The robot described in Section VIII was used to gather images of each object from various randomly determined viewpoints. These images are a simple approximation of the first-person perspective views of the object which the infants had during play. In total, 209 images were captured of each object resulting in a database of 8778 images. View-sets of objects were generated from these images as described below. For these infant-directed speech evaluations, only the shape channel was extracted from images, so color terms were unlearnable (ungroundable).

To prepare the corpus for processing, we performed the following steps. 1) Segment audio at utterance boundaries. This was done automatically by finding contiguous frames of speech detected by the recurrent neural network. 2) For each utterance, generate a view-set of the object in play by taking 15 randomly chosen images from the 209 available images of the object. Video recordings of the caregiver-infant interactions were used to determine the correct object for each utterance.

Each utterance-image set constituted an AV-event. Input to the learning system consists of a sequence of AV-events, presented to the system in the same order that the utterances were observed during infant interactions. The audio-visual data corresponding to each of the six speakers was processed separately. The top 15 items resulting from the MMI maximization step were evaluated for each speaker. As noted earlier, the learning problem posed by this data set is extremely challenging: less than 30% of the spoken utterances contain words which directly refer to the object in play. For example, the caregiver often said phrases such as "Look at it go!" while playing with a car or ball. CELL had to identify reliable lexical items such as "ball" or "car" despite such poor correspondences.

As described in Section VII, lexical hypotheses are analyzed by searching for maximum mutual information across channels. Fig. 7 presents two examples of mutual information surfaces for two actual lexical hypotheses generated from one of the speakers in this experiment. In each plot, the height of the surface shows mutual information as a function of the thresholds t_V and t_A. On the left, the speech segment corresponding to the word "yeah" was paired with the images of a shoe. The resulting surface is relatively low for all values of the thresholds. The lexical candidate on the right paired a speech segment of the word "dog" with images of a dog. The result is a strongly peaked surface form. The thresholds were selected at the point where the surface height, and thus mutual information, was maximized.

X. Results

Results of the experiments were evaluated using three measures. For each acoustic and visual prototype used to generate a lexical item, a pointer to the source speech recording and view-set were maintained. An interface was built to allow an evaluator to listen to the original speech recording from which a prototype was extracted. The interface also displayed the images of the corresponding view-set. The used this tool to assess the results.

Each lexical item was evaluated using three different measures:

> *Measure 1–Segmentation Accuracy:* Do the start and end of each speech prototype correspond to word boundaries in English?
>
> *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 /dag/ (*dog*), /ag/ (**dog*, with initial /d/ missing), and /ðdag/ (*the dog*), would all be accepted as positive instances of this measure. However /dagIz/ (*dog is*) would be counted as an error.
>
> *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 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 visual association is whatever co-occurred with it.

For comparison, we also ran the system with only acoustic input. In this case it was not meaningful to use the MMI method

TABLE I
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	ɪŋəʃ	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	ənðɪsɪz	*and this is	shoe B	0	0	0
14	(ono.)	(engine)	truck A	-	-	-
15	(ono.)	(barking)	dog A	-	-	-
Total				**54%**	**85%**	**85%**

so instead the system searched for globally recurrent speech patterns, i.e. speech segments which were most often repeated in the entire set of recordings for each speaker. This acoustic-only model may be thought as a rough approximation to a minimum description length approach to finding highly repeated speech patterns which are likely to be words of the language [14], [15].

Table I 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 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. At the bottom of the table, the measures are accumulated to calculate accuracy along each measure.

For comparison, the lexical items acquired by the Acoustic-only Model are shown in Table II. These results are derived from the same participant's data as Table I. In cases where no discernible words were heard, the text transcript is left blank. CELL out-performed the Acoustic-only Model on all three measures. Similar results were found for all subjects. Table III summarizes the performance of CELL

TABLE II
CONTENTS OF LTM USING THE ACOUSTIC-ONLY MODEL TO PROCESS THE DATA FROM THE SAME PARTICIPANT AS TABLE I

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	wʌyugonnd	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	?ey	*today	dog D	0	1	0
Total				**8%**	**25%**	**0%**

and the acoustic-only model for all six speakers in the study. Cross-modal learning achieved higher scores almost without exception.[6]

Measure 1, segmentation accuracy, poses an extremely difficult challenge when dealing with raw acoustic data. The acoustic-only model produced lexical items which corresponded perfectly to English words in only 7% of the lexical items. In contrast, 28% of lexical items produced by CELL were correctly segmented single words. Of these 28%, half of the accepted items were not correctly grounded in the visual channel (i.e., they fail on Measure 3). For example, the words *choose* and *crawl* were successfully extracted by CELL and associated with a car and ball respectively. These words do not directly refer to objects and thus failed on Measure 3. Yet, there was some structural consistency between the word and the shape which aided the system in producing these segmentations.

For Measure 2, word discovery, approximately three out of four lexical items (72%) produced by CELL were single words (with optional articles and inflections). 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. The cross-channel structure lead to a 2.3-fold increase in accuracy compared with analyzing structure within the acoustic channel alone.

On Measure 3, the large difference in performance between CELL and the acoustic-only system is not surprising since visual input is not used during lexical formation in the latter. CELL's performance is very promising since 57% of the hypothesized lexical candidates are both valid English words and linked to semantically relevant visual categories.

For all three measures, we found that cross-channel structure is leveraged to improve learning performance. By looking for agreement between different channels of input, CELL is able to find lexical candidates effectively through unsupervised learning.

The Acoustic-only Model performed well considering the input it received consisted of unsegmented speech alone. In fact it learned some words which are not acquired by CELL including "go," "yes," "no," and "baby. This result suggests

[6]The only exception was that the segmentation accuracy for Participant CL was 33% using the acoustic-only compared with 20% using CELL.

TABLE III
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%)

that in addition to cross-channel structure, within-channel structure is useful and should also be leveraged in learning words. Using other processes, the learner may later attempt to determine the associations of these words.

XI. CONCLUSIONS AND FUTURE DIRECTIONS

We have successfully implemented and evaluated CELL, a computational model of sensor-grounded word learning. The implemented system learns words from natural video and acoustic input signals. To achieve this learning, three difficult problems are simultaneously solved: 1) segmentation of continuous spontaneous speech without a pre-existing lexicon, 2) unsupervised clustering of shapes and colors, and 3) association of spoken words with semantically appropriate visual categories. Mutual information is used as a metric for cross-channel comparisons and clustering. This system demonstrates the utility of mutual information to combine modes of input for multisensory learning.

The results with CELL show that it is possible to learn to segment continuous speech and acquire statistical models of spoken words by providing a learning system with untranscribed speech and co-occurring visual input. Visual input serves as extremely noisy labels for speech. The converse is also true. The system learns visual categories by using the accompanying speech as labels. The resulting statistical models may be used for speech and visual recognition of words and objects. Manually trained data is replaced by two streams of sensor data which serve as labels for each other. This idea may be applied to a variety of domains where multimodal data is available, but human annotation is expensive.

We are now exploring several applications of this work for robust and adaptive human-computer interfaces. Current spoken language interfaces process and respond only to speech signals. In contrast, humans also pay attention to the context in which speech occurs. These side channels of information may serve to ground the semantics of speech, leading to reduced ambiguity at various levels of the spoken language understanding problem. Based on ideas presented in this paper, we are exploring the use of grounded speech learning and understanding to create systems which are able to resolve ambiguities in the speech signal.

Learning in CELL is driven by a bottom-up process of discovering structure observed in sensor data. In the future, we plan to experiment with learning architectures which integrate top-down purpose driven categorization with bottom-up methods. In doing so, cross-channel clusters and associations can be acquired which are optimized to achieve high level goals.

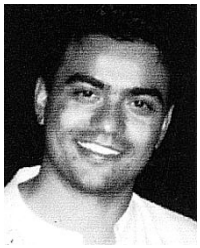
ACKNOWLEDGMENT

The authors would like to acknowledge A. Pentland, A. Gorin, R. Patel, B. Schiele, and S. Pinker, and also the feedback from anonymous reviewers.

REFERENCES

[1] M. Johnson, *The Body in the Mind*. Chicago, IL: Univ. of Chicago Press, 1987.
[2] G. Lakoff, *Women, Fire, and Dangerous Things*. Chicago, IL: Univ. of Chicago Press, 1987.
[3] S. Harnad, "The symbol grounding problem," *Phys. D*, vol. 42, pp. 335–346, 1990.
[4] L. Barsalou, "Perceptual symbol systems," *Beh. Brain Sci.*, vol. 22, pp. 577–609, 1999.
[5] S. Pinker, *The Language Instinct*. New York: HarperPerennial, 1994.
[6] C. E. Snow, "Mothers' speech research: From input to interaction," in *Talking to Children: Language Input and Acquisition*, C. E. Snow and C. A. Ferguson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1977.
[7] J. Huttenlocher and P. Smiley, "Early word meanings: the case of object names," in *Language Acquisition: Core Readings*, P. Bloom, Ed. Cambridge, MA: MIT Press, 1994, pp. 222–247.
[8] J. Siskind, "Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1992.
[9] A. Sankar and A. Gorin, *Adaptive Language Acquisition in a Multi-Sensory Device*. London, U.K.: Chapman & Hall, 1993, pp. 324–356.
[10] T. Regier, *The Human Semantic Potential*. Cambridge, MA: MIT Press, 1996.
[11] D. K. Roy, M. Hlavac, M. Umaschi, T. Jebara, J. Cassell, and A. Pentland, "Toco the toucan: a synthetic character guided by perception, emotion, and story," in *Visual Proceedings of Siggraph*. Los Angeles, CA: ACM Siggraph, Aug. 1997.
[12] S. Becker and G. E. Hinton, "A self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, pp. 161–163, 1992.
[13] V. R. de Sa and D. H. Ballard, "Category learning through multi-modality sensing," *Neural Comput.*, vol. 10, no. 5, 1998.
[14] M. R. Brent, "An efficient, probabilistically sound algorithm for segmentation and word discovery," *Mach. Learn.*, vol. 34, pp. 71–106, 1999.
[15] C. de Marcken, "Unsupervised Language Acquisition," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1996.
[16] A. L. Gorin, "On automated language acquisition," *J. Acous. Soc. Amer.*, vol. 97, no. 6, pp. 3441–3461, 1995.
[17] D. Petrovska-Delacretaz, A. L. Gorin, J. H. Wright, and G. Riccardi, "Detecing acoustic morphemes in lattices for spoken language understanding," in *Proc. Int. Conf. Spoken Language Processing*, 2000.
[18] R. A. Brooks, "Elephants don't play chess," *Robot. Auton. Syst.*, vol. 6, pp. 3–15, 1990.
[19] L. Steels and P. Vogt, "Grounding adaptive language games in robotic agents," in *Proc. 4th Eur. Conf. Artificial Life*, 1997.
[20] D. K. Roy, "Learning Words From Sights and Sounds: A Computational Model," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1999.
[21] ——, "Integration of speech and vision using mutual information," in *Proc. ICASSP*, Istanbul, Turkey, 2000.

[22] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science*, vol. 255, pp. 606–608, 1992.
[23] J. F. Werker and C. E. Lalonde, "The development of speech perception: initial capabilities and the emergence of phonemic categories," *Develop. Psychol.*, vol. 24, pp. 672–683, 1999.
[24] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, Mar. 1994.
[25] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[26] P. Werbos, "Backpropagation through time: what it does and how to do it," *Proc. IEEE*, vol. 78, pp. 1150–1160, 1990.
[27] S. Seneff and V. Zue, "Transcription and alignment of the timit database," in *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, J. S. Garofolo, Ed. Gaithersburg, MD: NIST, 1988.
[28] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.
[29] J. H. Wright, M. J. Carey, and E. S. Parris, "Statistical models for topic identification using phoneme substrings," in *Proc. ICASSP*, 1996, pp. 307–310.
[30] R. Rose, "Word spotting from continuous speech utterances," in *Automatic Speech and Speaker Recognition*, C. H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Norwell, MA: Kluwer, 1996, ch. 13, pp. 303–329.
[31] M. H. Bornstein, W. Kessen, and S. Weiskopf, "Color vision and hue categorization in young human infants," *J. Exper. Psychol.: Human Percept. Perf.*, vol. 2, pp. 115–129, 1965.
[32] A. E. Milewski, "Infant's discrimination of internal and external pattern elements," *J. Exper. Child Psychol.*, vol. 22, pp. 229–246, 1976.
[33] B. Schiele and J. L. Crowley, "Probabilistic object recognition using multidimensional receptive field histograms," in *Proc. 13th Int. Conf. Pattern Recognition (ICPR'96)*, vol. B, August 1996, pp. 50–54.
[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.
[35] S. Pinker, *Language Learnability and Language Development*. Cambridge, MA: Harvard Univ. Press, 1984.
[36] J. Grimshaw, "Form, function, and the language acquisition device," in *The Logical Problem of Language Acquisition*, C. L. Baker and J. J. McCarthy, Eds. Cambridge, MA: MIT Press, 1981, pp. 165–182.



Deb Roy (S'96–M'99) received the B.S. degree in computer engineering from the University of Waterloo, Waterloo, ON, Canada, and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge.

He is Assistant Professor of media arts and sciences at the MIT Media Laboratory, where he directs the Cognitive Machines Group.