SITUATION-AWARE SPOKEN LANGUAGE PROCESSING

D.K. Roy MIT Media Lab, Cambridge MA (dkroy@media.mit.edu)

1 INTRODUCTION

Words and utterances are understood in context. Context may take many forms including shared knowledge of current and recent topics of discourse, knowledge of the communication partner's beliefs, and knowledge of relevant facts in the world to which the speaker might refer. The focus of this paper is the construction of speech processing systems which leverage this third type of context.

Our goal is to create *situation-aware* speech systems which integrate knowledge of relevant aspects of the current state of the world into spoken language learning, understanding and generation. Awareness of the situation is achieved by extracting salient information about the speaker's world from sensors including cameras, touch sensors, and microphones. A key challenge in this approach is to design and integrate linguistic and non-linguistic representations. This paper presents an implemented system which learns to understand and generate spoken utterances *grounded* in camera-derived visual semantics [18, 19]. Current work on a more sophisticated cross-modal language learning system is also introduced.

To motivate the development of situation-aware speech systems, we can consider some sample applications. Speech interfaces are useful when the user's eyes and hands are busy. Consider a person working in a warehouse. This person may need to control a forklift, read stock labels, and interact via speech with a portable information system to enter/query current stock information. Understanding the user's speech and generating contextualized responses would be simplified if the system had access to the current stock of the warehouse, and knowledge of where the worker is attending. Situational information could be leveraged in at least two ways in this example. First, the language model used by the speech recognizer (e.g., n-grams) to predict the user's words could be dynamically updated based on knowledge of likely referents (i.e., by knowing what is currently in view of the person). Semantic resolution may be achieved by combining linguistic and situational knowledge to understand requests such as "Do we have more of these on order?" where the meaning of the word "these" could be resolved through situational-awareness.

As a second example consider a speech interface to a robotic assistant which can move objects around in a person's workspace (we are currently working on this task). Such a machine might be useful if the person is engaged in an assembly task and could use a helping hand, or if the user has a motor disability and needs help in completing daily tasks. If the helping hand was another human, spoken requests would often refer to objects, locations, and actions in terms of the physical situation. Sentences such as "Take the blue one and put it next to the other one" can only understood through shared common ground between speaker and listener. Thus the machine must be situation-aware

so that it can understand context-dependent semantics of speech.

Before presenting our work on multisensory language learning, the following section expands on the notion of grounding semantics and its implications for speech processing systems.

1.1 Grounding Semantics in Sensors

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 sensory-motor experiences [10, 8]. To grasp the concepts underlying words such as *red, heavy* and *above* requires that the language user have a body and bodily experiences of interacting with the physical world. This link to the senses, 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 [13]. 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. We might 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 is achievable without labeled data since labels are replaced by context. Language does not occur in a vacuum. Infants observe spoken language in a rich physical and social contexts. Furthermore, infant-directed speech usually refers to the immediate context [25]; 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 non-linguistic information into the language understanding process. Acoustically and semantically ambiguous utterances can be disambiguated by the context in which they are heard. People 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 when processing language. These advantages motivate us to investigate grounded speech acquisition and processing.

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 favor 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.

In this paper we summarize our results of a visually grounded word learning system. The system learns from untranscribed spontaneous speech paired with unannotated video images. In a form of unsupervised learning, spoken utterances serve as noisy labels for video sequences, and vice versa. Once trained, the system is able to understand and generate one and two word descriptions of objects. We conclude by introducing our current efforts to expand on this work to address more sophisticated forms of situated language processing including treatment of verbs, spatial relations, and limited forms of syntax.

2 LEARNING WORDS FROM AUDIO-VISUAL INPUT

To explore issues of grounded language, we have developed a model of Cross-channel Early Lexical Learning (CELL) [18, 19]. CELL learns spoken words and their visual semantics by integrating visual and acoustic input. The system learns to segment continuous speech without a lexicon and forms associations between acoustic word forms and their visual semantics. This effort represents a step towards 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.

The major components of CELL are summarized in Figure 1. The 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. Output of CELL consists of a lexicon of audio-visual items. Each lexical item includes a statistical model (based on Hidden Markov Models) of a 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 assumed to be extremely noisy. The model must be able to "fish out" salient cross-channel associations from noisy input.

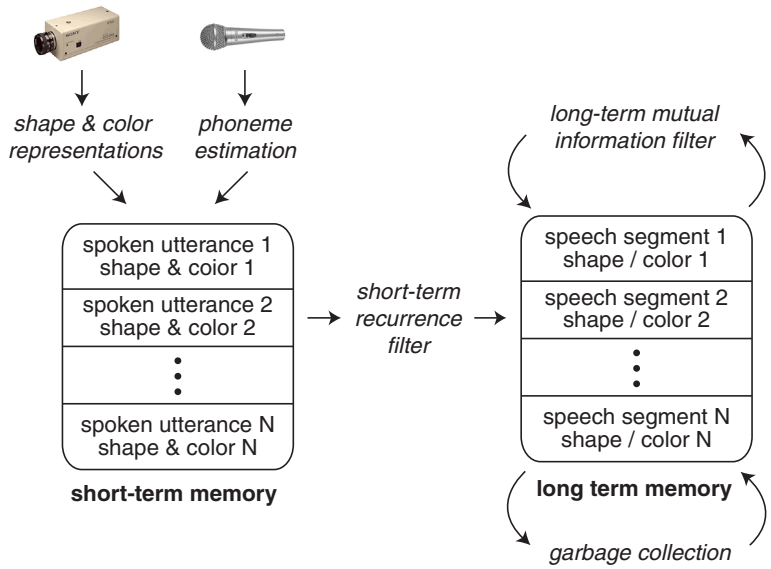

Figure 1: The CELL model. A layered memory architecture combined with recurrence and mutual information filters (see text) are used to acquire an audio-visual lexicon from unlabelled input.

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 sub-sequences of speech which occur in matching visual contexts. The resulting pairs of speech segments and shapes 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 pairings are retained in the LTM, and rejected pairings are periodically discarded by a garbage collection process.

2.1 Representing and Comparing Spoken Utterances

Spoken utterances are represented as arrays of phoneme probabilities. A recurrent neural network similar to [16] processes RASTA-PLP coefficients [9] 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 [27] with the TIMIT database of phonetically transcribed speech recordings [23][1]. The RNN recognizes phonemes with 69.4% accuracy using the standard TIMIT training and test datasets. The RNN includes a silence output which is used to perform speech/silence segmentation.

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 Hidden Markov Model (HMM) framework. The Viterbi dynamic programming search 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 [28]. 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 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 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 [14] can be used to compute $P(Q_i|\lambda_j)$ and $P(Q_j|\lambda_i)$, the likelihood that the HMM derived from speech segment α_i generated speech segment α_j and vice versa. However, these likelihoods 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 [17]. In this method, each likelihood estimate is scaled by the

[1] 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 without transcriptions.

likelihood of a default alternate hypothesis, λ^A:

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

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:

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

2.2 Representing and Comparing Object Shapes and Colors

Three-dimensional objects are represented using a view-based approach in which multiple two-dimensional images of an object captured from multiple viewpoints collectively form a model of the object. The two-dimensional 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.

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 indicate the presence of an object.

Based on methods developed by Schiele and Crowley [22], objects are represented by histograms of local features derived from multiple 2D 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. All pair-wise values are accumulated in a 2D histogram to represent an image. Using multidimensional histograms to represent object shapes allows the use of information theoretical or statistical divergence functions [22] directly for the comparison of object models. We use the χ^2–divergence:

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

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 3D shapes is based on a collection of 2D shape histograms, each corresponding to a particular view of the object. Each 3D object is represented by 15 shape histograms. 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.

Color of objects is also represented using histograms. For each image, a two-dimensional color histogram is generated by accumulating illumination-normalized red and green values for each pixel specified by the region of the object mask. The normalized red and green chromaticity values are divided into 8 bins leading to $8 \times 8 = 64$ bin histograms.

2.3 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, the goodness of a lexical candidate is measured using the mutual information between a spoken sequence and a visual category. The use of mutual information effectively integrates information from multiple channels. 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. Each time a new event is inserted into the buffer, a *recurrence filter* searches for approximately repeating audio and visual patterns within the buffer. For example, 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. It performed 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 may be repeated across multiple 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 randomly 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:

$$A = \begin{cases} 0 & if\ d_A(X, h_i) > t_A \\ 1 & if\ d_A(X, h_i) \leq t_A \end{cases} \quad (3)$$

$$V = \begin{cases} 0 & if\ d_V(X, h_i) > t_V \\ 1 & if\ d_V(X, h_i) \leq t_V \end{cases} \quad (4)$$

$$(5)$$

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:

$$I(A;V) = \sum_i \sum_j P(A=i, V=j) \log \left[\frac{P(A=i, V=j)}{P(A=i)P(V=j)} \right] \quad (6)$$

The probabilities required to calculate $I(A;V)$ can be estimated from smoothed frequency counts. 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 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.

A final step is to threshold the MMI score of each hypothesis and select those which exceed the threshold. Determination of this threshold is beyond the scope of our current system. It might be set manually, or by reinforcement feedback from higher levels in the system.

2.4 Related Work

Recent models of language acquisition include models of speech segmentation based on minimum description length encoding of acoustic representations [2, 4], and cross-situational learning from text coupled with line drawings representing simple visual semantics [24, 21, 15]. Algorithms for acquiring

syntactic structure and semantic associations for acoustic words based on semantic transcriptions have been demonstrated [6]. 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 [11]. Physical grounding of concepts has been explored in the context of robotics as an alternative to the symbol processing view of artificial intelligence [3, 26]. The model presented this paper departs from previous work in that both words and their semantics are acquired from sensor input, and the goal of the system is to acquire spoken human language.

The idea of learning from evidence across multiple modalities used in CELL has been explored in the machine learning community, for example in [1] and [5].

3 AN INTERACTIVE WORD-LEARNING ROBOT

To experiment with human-machine spoken interactions, CELL has been incorporated into a real-time speech and vision interface embodied in a robotic system. The robot can be taught by showing it objects and verbally describing them using natural speech. CELL extracts an audio-visual lexicon from a set of training pairs. 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).

3.1 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 (Figure 2). 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.

3.2 Acquiring a Lexicon

The robot has three modes of operation: acquisition, generation, and understanding. In the 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.

The robotic system has successfully learned a vocabulary of color and shape terms based on in-

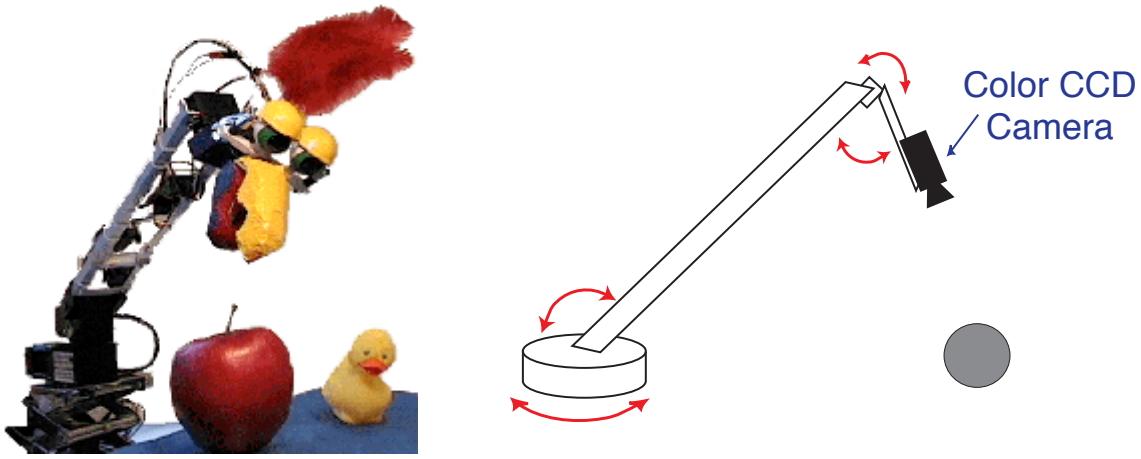

Figure 2: A robot with four degrees of freedom used to capture images of objects. A small CCD camera is mounted in the right eyeball.

teractions with several people and everyday objects. Typical lexical entries include words such as "green" and "cup". Typical errors include associating phrases with single visual categories ("green cup" or "the cup" associated with cup shapes), and segmentation errors ("yell" associated with the color yellow). For detailed performance analysis with a corpus of spontaneous speech and video, see [18, 20].

3.3 Acquiring Lexical Order: A First Step Towards 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, with approximately 100 spoken utterances describing ten objects of varying shape and color, 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 towards 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 [12, 7]. 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.

3.4 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 sequence of phonemes specified in the retrieved lexical item is resynthesized using a phoneme synthesizer. 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 for both aspects 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”) when it had already learned the words “red” and “ball”, even if it had never seen an apple before.

3.5 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.

4 CURRENT WORK: LEARNING TO UNDERSTAND SPOKEN COMMANDS IN A SPATIAL MANAGEMENT TASK

We are now working on a language learning task which introduces several new problems including lexical acquisition of action verbs (*put, move*, etc.), spatial relations (*above, beside, to the left of*, etc.) and limited syntactic processing (*place a magnet to the left of the blue ball*, etc.) Clearly this task is more challenging than those addressed with CELL. We expect, however, that mutual information based cross-modal learning algorithms will continue to be useful for lexical acquisition once additional representations are extracted from the visual scene. For syntactic processing we are investigating new learning schemes.

Figure 3 shows a screen shot from a video game we are using to collect multimodal data. The game consists of a set of synthetic objects including bricks, balls, magnets and repellers. Given a configuration of bricks and balls, the player must position magnets and repellers to solve a puzzle. A physics simulator sets the balls into motion which interact with the magnets and repellers. For the puzzle shown in the figure, the goal is to position magnets and repellers such that the blue and red balls fall into separate bins when gravity is switched on. The puzzle is difficult and requires multiple iterations (of moving magnets and repellers around) to solve. The player must verbally instruct a human assistant to move objects on her behalf. This results in speech and action pairs that are used as input to the languge learning system

The initial goal of this learning task is to replace the human assistant with an automatic one so that a person can play the game simply by speaking to the machine with natural speech. This work will serve as a step towards our longer term objective of building a physical robotic manipulator which can learn to understand verbal instructions. The virtual world of the game allows us to focus on spoken language learning by simplifying problems of vision and motor control. Once we have robust algorithms for speech processing, we will transfer to the robotic platform and confront the remaining issues.

Initial results from this new task will be presented at the WISP workshop.

5 CONCLUSIONS

People generate and understand spoken language in context. Non-lingustic information is used to resolve semantic, syntactic and acoustic ambiguities. This paper has introduced our efforts to develop systems which infer the user's world state from sensors and harness this situation-awareness to learn, understand, and generate grounded spoken language.

We have demonstrated a word learning system which can acquire names of object shapes and colors from natural interactions with people. In contrast to conventional speech systems, this system represents the semantics of words in terms of camera-grounded representations. When put to use, this system can acquire information about the physical state of the world (i.e., what objects are in view)

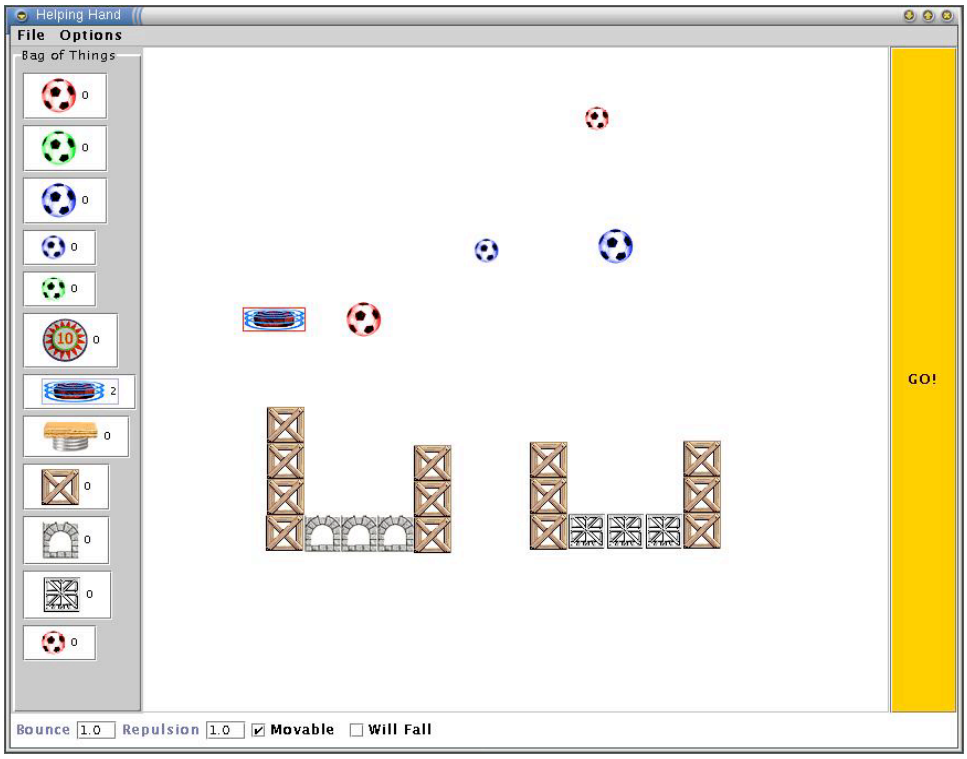

Figure 3: Screen shot of a video game designed to record speech-action data. A player instructs a human assistant to place, remove, and move objects within the game's playing area. A record is kept of each spoken utterance, and the before and after configuration of the world.

to help it understand or generate spoken language. We are now developing a second generation language learning system to address acquisition of action verbs, spatial terms, relative size terms, and aspects of syntax for understanding commands in a spatial management task.

The goal of building grounded speech systems is to endow machines with the ability to leverage contextual information to process speech in more human-like ways. We believe that this approach will lead to natural human-machine interaction through spoken language.

6 ACKNOWLEDGEMENTS

Peter Gorniak implemented the game simulator described in Section 4.

References

[1] S. Becker and G.E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, 1992.

[2] M.R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–106, 1999.

[3] R.A. Brooks. Elephants don't play chess. *Robotics and Autonomous Systems*, 6:3–15, 1990.

[4] C. de Marcken. *Unsupervised Language Acquisition*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996.

[5] V.R. de Sa and D.H. Ballard. Category learning through multi-modality sensing. *Neural Computation*, 10(5), 1998.

[6] A.L. Gorin. On automated language acquisition. *Journal of the Acoustic Society of America*, 97(6):3441–3461, 1995.

[7] J. Grimshaw. Form, function, and the language acquisition device. In C. L. Baker and J. J. McCarthy, editors, *The logical problem of language acquisition*, pages 165–182. MIT Press, 1981.

[8] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

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

[10] G. Lakoff. *Women, fire, and dangerous things*. The University of Chicago Press, Chicago, IL, 1987.

[11] D. Petrovska-Delacretaz, A.L. Gorin, J.H. Wright, and G. Riccardi. Detecing acoustic morphemes in lattices for spoken language understanding. *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[12] S. Pinker. *Language learnability and language development*. Harvard University Press, Cambridge, MA, 1984.

[13] S. Pinker. The bootstrapping problem in language acquisition. In Brian MacWhinney, editor, *Mechanisms of language acquisition*, pages 399–441. Erlbaum, Hillsdale, NJ, 1987.

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

[15] T. Regier. *The human semantic potential*. MIT Press, Cambridge, MA, 1996.

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

[17] R. Rose. Word spotting from continuous speech utterances. In C.H. Lee, F. K. Soong, and K.K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, chapter 13, pages 303–329. Kluwer Academic, 1996.

[18] D.K. Roy. *Learning Words from Sights and Sounds: A Computational Model*. PhD thesis, Massachusetts Institute of Technology, 1999. (full text available at http://dkroy.www.media.mit.edu/people/dkroy/publications.html).

[19] D.K. Roy. Integration of speech and vision using mutual information. In *Proc. of ICASSP*, Istanbul, Turkey, 2000.

[20] D.K. Roy. Learning from multimodal observations. In *Proceedings of the IEEE International Conference on Multimedia*, New York, NY, 2000.

[21] A. Sankar and A. Gorin. *Adaptive language acquisition in a multi-sensory device*, pages 324–356. Chapman and Hall, London, 1993.

[22] B. Schiele and J.L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR'96 Proceedings of the 13th International Conference on Pattern Recognition, Volume B*, pages 50–54, August 1996.

[23] S. Seneff and V. Zue. Transcription and alignment of the timit database. In J.S. Garofolo, editor, *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

[24] J. Siskind. *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition*. PhD thesis, Massachusetts Institute of Technology, 1992.

[25] C.E. Snow. Mothers' speech research: From input to interaction. In C. E. Snow and C. A. Ferguson, editors, *Talking to children: language input and acquisition*. Cambridge University Press, Cambridge, MA, 1977.

[26] L. Steels and P. Vogt. Grounding adaptive language games in robotic agents. In *Proceedings of the 4th European Conference on Artificial Life*, 1997.

[27] P. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78:1150–1160, 1990.

[28] J.H. Wright, M.J. Carey, and E.S. Parris. Statistical models for topic identification using phoneme substrings. In *Proceedings of ICASSP*, pages 307–310, 1996.