GROUNDED SPEECH COMMUNICATION

Deb Roy

20 Ames Street, Rm. E15-384C, Cambridge, MA 01242, USA
http://www.media.mit.edu/ dkroy
dkroy@media.mit.edu

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, machines which process text and spoken language are not grounded in human-like ways. Instead, semantic representations in machines are highly abstract 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 understanding by machines. Inspired by theories of infant cognition, we present a computational model which learns from untranscribed multisensory input. Acquired words are represented in terms associations between acoustic and visual sensory experience. The system has been tested in a robotic embodiment which supports interactive language learning and understanding. Successful learning has also been demonstrated using infant-directed speech and images.

1. INTRODUCTION

Language is grounded in experience. Unlike dictionary definitions in which words are defined in terms of other words, humans understand many basic concepts in terms of associations with sensory-motor experiences [1]. To grasp the concepts underlying words such as *red, heavy* and *above* requires interaction with the physical world. Grounding is a fundamental aspect of spoken language which enables humans to acquire and use words and sentences in context.

Infants learn their first words by associating speech patterns with objects, actions, and people. The 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 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.

Current spoken language recognition and understanding systems are not grounded. During training, systems are presented with recordings of spoken utterances paired with manually generated transcriptions. Acoustic waveforms are the only sensor signal available to the system during training. Once trained, these systems convert novel speech input into transcriptions or machine actions. Non-acoustic signals are typically not integrated into the recognition or understanding process. Although ungrounded speech systems have led to practical applications for transcription and telephony, numerous difficult problems of speech communication cannot fully be addressed until multisensory grounding is introduced to speech systems.

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 [3]. 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. In this paper we first present the model of word learning, and then describe experiments in testing the model with interactive robotics and infant-directed speech.

2. CELL: A COMPUTATIONAL MODEL OF WORD LEARNING

Inspired by theories of infant language acquisition, we have developed a model of cross-channel early lexical learning (CELL), summarized in Figure 1 [3]. This model discovers words by searching for segments of speech which reliably predict the presence of visually co-occurring shapes. Input consists of spoken utterances paired with images of objects. This approximates the input that an infant might receive when listening to a caregiver while visually attending to objects in the environment. Output 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 an object

class. To acquire lexical items, the system must (1) segment continuous speech at word boundaries, (2) form visual categories corresponding to objects, and (3) form appropriate correspondences between word and object models.

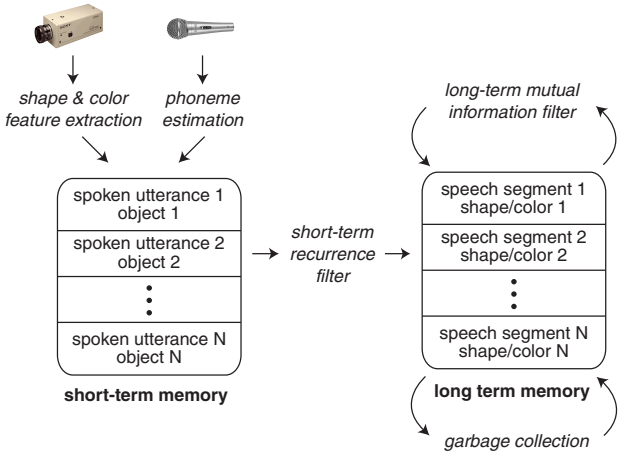

Figure 1: The CELL model. Camera images of objects are converted to statistical representations of shapes and colors. Spoken utterances captured by a microphone are mapped onto sequences of phoneme probabilities. A layered memory architecture consisting of short and long term memory supports search for consistent cross-modal patterns based on mutual information.

A speech processor converts spoken utterances into sequences of phoneme probabilities. At a rate of 100Hz, this processor computes the probability that the past 20 milliseconds of speech belonged to each of 39 English phoneme categories or silence. The phoneme estimation was achieved by training a recurrent neural network using the TIMIT database. Utterance boundaries are automatically located by detecting stretches of speech separated by silence.

A visual processor was developed to extract statistical representations of shapes and colors from images of objects. The visual processor uses 'second order statistics' to represent object appearance. Color is represented by computing a two-dimensional histogram of illumination-normalized RGB pixel values from the area of the image occupied by the target object. To represent shape, the edge pixels of the viewed object are first located. For each pair of edge points, the normalized distance between points and the relative angle of edges at the two points are computed. All distances and angles are accumulated in a two-dimensional histogram representation of the shape (the 'second order statistics'). The chi-squared divergence statistic is used to compare shape histograms, a measure that has been shown to work well for object comparison [4]. Three-dimensional shapes are represented with a collection of two-dimensional shape histograms, each derived from a particular view of the object. Sets of images are compared by summing the chi-square divergences of the four best matches between individual histograms.

Phonemic representations of multi-word utterances and co-occurring visual representations are temporarily stored in a short term memory (STM). The STM has a capacity of five utterances, corresponding to approximately 20 words of infant-directed speech. As input is fed into the model, each new [utterance,object] entry replaces the oldest entry in the STM. A short-term recurrence filter searches the contents of the STM for recurrent speech segments which occurred in matching visual contexts. The STM focuses initial attention to input which occurred closely in time. To determine matches, an acoustic distance metric [2] is used to compare each pair of potential speech segments drawn from the utterances stored in STM. This metric estimates the likelihood that the segment pair in question are variations of similar underlying phoneme sequences and thus represent the same word. The chi-squared divergence metric is used to compare the visual components associated with each STM utterance. If both the acoustic and visual distance are small, the segment and shape are copied into the LTM. Each entry in the LTM represents a hypothesized prototype of a speech segment and its visual referent.

Infant-directed speech usually refers to the infant's immediate context [5]. When speaking to an infant, caregivers rarely refer to objects or events which are in another location or which happened in the past. On this premise, a long-term mutual information filter assesses the consistency with which speech-shape pairs co-occurred in the LTM. The mutual information (MI) between two random variables measures the amount of uncertainty removed regarding the value of one variable given the value of the other. Mutual information is used to measure the amount of uncertainty removed about the presence of a specific shape or color in the learner's visual context given the observation of a specific speech segment. Since MI is a symmetric measure, the converse was also true: it measured the uncertainty removed about the co-occurrence of a particular speech segment given a visual context. Speech-shape or speech-color pairs with high MI are retained, and periodically a garbage collection process removes hypotheses from LTM which do not encode associations with high MI.

3. EXPERIMENTS WITH AN INTERACTIVE ROBOT

CELL was incorporated into a real-time speech and vision interface embodied in a robotic character (Figure 2). We chose a robotic form to enable natural human-machine conversation in which the machine is able to use body and facial gestures to indicate internal state.

Input consisted of continuous multiword spoken utterances and images of objects acquired from a CCD camera mounted on the robot. To teach the system, a person places an object in front of the robot and describes it. After accumulating multiple audio-visual observations, the system acquires a lexicon of color and shape terms grounded in microphone and camera input. Once a lexicon has been acquired, the robot can be engaged in an object labeling task (i.e., lexical generation), and an object selection task (i.e., lexical understanding).

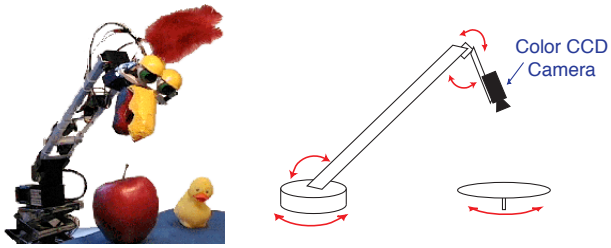

Figure 2: The robot consists of a four degree-of-freedom armature with a CCD camera mounted at the tip. The camera may be moved to obtain images of a target image from various perspectives. Target objects are placed on a turn table providing an additional degree of freedom.

3.1. 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 the viewing surface. When an object is detected, the system gathers images of the object from various random perspectives. If a spoken utterance is detected while the images are being gathered, a [utterance, object] event is generated and processed by CELL.

To teach the system, the user might place a cup in front of the robot and say, "Here's my coffee cup". To verify that the system has received contextualized spoken input, the robot "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. In informal tests, the system was able to learn and use a lexicon of 12 shape and color terms from 70 input utterances.

The system acquires word order statistics for learning the order of shape and color terms in adjacent positions without intervening words. The shape and color channels serve to ground primitive categories of speech enabling higher level distributional analysis of word category ordering. Lexical items are assigned to either the shape or color class depending on their contextual grounding. The system tracks the distribution of color-shape and shape-color terms for input utterances. In experiments, the system learned that color terms precede shape terms in English.

3.2. Object Description

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 grounding of the best match. Speech is generated by sending phoneme sequences to a model-based speech synthesizer.

To use word order statistics, a second generation mode finds the best matching lexical 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 a tennis ball, the robot would say "yellow ball" when it had already learned the words "yellow" and "ball".

3.3. Speech Understanding

In the speech understanding mode, the system waits for the user to name objects in terms of shape and color. The input utterance is assumed to contain only lexical items in LTM. The input utterance is 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 is determined by the acquired word order statistics.

In a second step, 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 task, it selects the best match and returns the robot's gaze to that object. To provide additional feedback, the selected object is used to index back into the lexicon and generate a spoken description. This feedback leads to revealing behaviors when an incorrect or incomplete lexicon has been acquired. The nature of errors provides the user with guidance for subsequent training interactions.

4. GROUNDED WORD LEARNING: EXPERIMENTS WITH INFANT-DIRECTED SPEECH

To evaluate CELL with natural spoken input, we gathered a corpus of audio-visual data from infant-directed interactions [3]. Six caregivers and their pre-linguistic (7-11 months) infants were asked to play with objects while being recorded. We selected 7 classes of objects commonly named by young infants: balls, shoes, keys, toy cars, trucks, dogs and horses. A total of 42 objects, six objects for each class, were obtained. The objects of each class varied in color, size, texture, and shape.

Each caregiver-infant pair participated in 6 sessions over a course of two days. In each session, they played with 7 objects, one at a time. All caregiver speech was recorded using a wireless head-worn microphone onto DAT. In total we collected approximately 7,600 utterances comprising 37,000 words across all six speakers. Most utterances contained multiple words with a mean utterance length of 4.6 words. Speech segmentation could not rely on the existence of isolated words since these were rare in the data.

The 42 objects were imaged from various perspectives using the robot described in the previous section. A total of 209 images from varying perspectives were collected for each of 42 objects resulting in a database of 8,778 images. Only shape histograms were computed for these experiments.

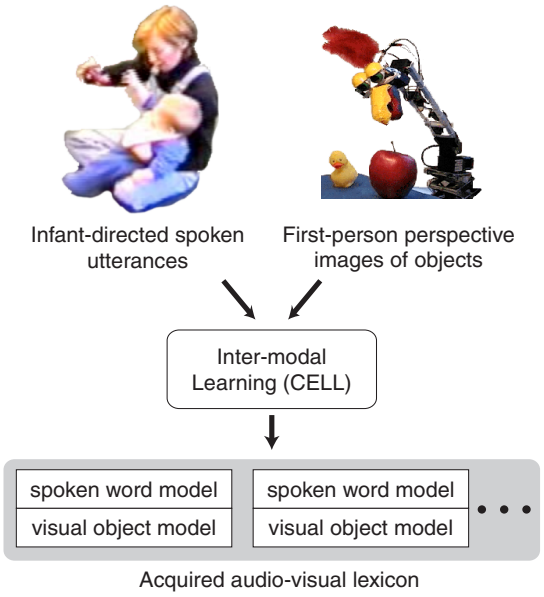

Figure 3: Speech was recorded from natural caregiver-infant interactions centered around objects. A robot was used to capture images from various first-person perspectives of the objects from the infant interactions. Speech recordings and images provided input to the learning experiment. The system learned statistical models of words and objects without manual annotations.

Speech recordings from caregiver-infant play were combined with the images taken by the robot to provide multimodal input to the learning system (Figure 3). 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, we selected a random set of 15 images of the object which was in play at the time the utterance was spoken.

We evaluated the lexicons extracted from the corpus using three measures. The first measure, M1, was the percentage of lexical items with boundaries at English word boundaries. The second, M2, was the percentage of lexical items which were complete English words with an optional attached article. The third measure, M3, was the percentage of lexical items which passed M2 and were paired with semantically correct visual models.

For comparison, we also ran the system with only acoustic input. In this case it was not meaningful to use the MI maximization. Instead the acoustic-only 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.

Table 1: Results of evaluation on three measures averaged across all six speakers.

	M1	M2	M3
audio only	7±5%	31±8%	13±4%
audio-visual	28±6%	72±8%	57±10%

Results of the evaluation shown in Table 1 indicate that the audio-visual clustering was able to extract a large proportion of English words from this very difficult corpus (M2), many associated with semantically correct visual models (M3). Typical speech segments in the lexicons included names of all six objects in the study, as well as onomatopoeic sounds such as "ruf-ruf" for dogs, and "vroooom" for cars. The comparison with the audio-only system clearly demonstrates the improved performance when visual context is combined with acoustic evidence in the clustering process. For word boundary detection (M1), multimodal input lead to a four-fold improvement over acoustic-only processing. Inter-modal structure enabled the system to find and extract useful knowledge without the aid of manual annotations or transcriptions.

5. CONCLUSIONS

The CELL model demonstrates acquisition of a sensory grounded lexicon from untranscribed input. The system is a step towards creating machines with rich semantic representations which move beyond symbolic "dictionary definitions" which lack connections to the physical world. Such representations may allow machines to integrate contextual information when understanding and generating speech, leading to more intelligent linguistic behavior in real-world situations.

From an engineering perspective, CELL demonstrates a method of training pattern recognition systems without manually transcribed data. By mimicking infant learning, we envision systems which can explore and interact with the world in order to discover and model useful patterns without direct human intervention. In the future we plan to extend our models of infant learning, and explore applications in medical domains, assistive aids, entertainment, and mobile computing.

6. REFERENCES

1. G. Lakoff. *Women, fire, and dangerous things*. The University of Chicago Press, Chicago, IL, 1987.
2. D. Roy. Integration of speech and vision using mutual information. In *Proc. of ICASSP*, Istanbul, Turkey, 2000.
3. D.K. Roy. *Learning Words from Sights and Sounds: A Computational Model*. PhD thesis, Massachusetts Institute of Technology, 1999.
4. 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.
5. 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.

GROUNDED SPEECH COMMUNICATION

Deb Roy
20 Ames Street, Rm. E15-384C, Cambridge, MA 01242, USA
http://www.media.mit.edu/ dkroy
dkroy@media.mit.edu

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, machines which process text and spoken language are not grounded in human-like ways. Instead, semantic representations in machines are highly abstract 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 understanding by machines. Inspired by theories of infant cognition, we present a computational model which learns from untranscribed multisensory input. Acquired words are represented in terms associations between acoustic and visual sensory experience. The system has been tested in a robotic embodiment which supports interactive language learning and understanding. Successful learning has also been demonstrated using infant-directed speech and images.