A Sensor-Grounded Computational Model of Early Word Learning

Deb Roy
MIT Media Laboratory
20 Ames Street, Rm. E15-384C, Cambridge, MA 02139, USA
dkroy@media.mit.edu
http://dkroy.www.media.mit.edu/people/dkroy/

December 19, 1999

1 Introduction

Around their first birthday, infants first begin to use words[1] which refer to salient aspects of their environment including objects, actions, and people. They learn these words by attending to the sights, sounds, and other sensations. The acquisition process is complex. Infants must successfully segment linguistic input into units which correspond to the words of their language. They must also identify semantic categories which correspond to the meanings of these words. Remarkably, infants are capable of all these processes despite continuous variations of natural phenomena and the noisy input provided by their perceptual systems.

This paper presents a computational model of early lexical learning. Learning is driven by a search for structure across channels of sensory input. The model is accordingly named Cross-channel Early Lexical Learning (CELL). It is a cognitively plausible on-line model which processes data incrementally.

This is a significant result: CELL is the first implemented model of language acquisition which learns words and their semantics from raw sensory input without any human-assisted preparation of data. This paper presents evaluations of the implementation with infant directed data. This work differs from previous computational models of language learning such as [4], [3], [10] in that both linguistic and contextual input are derived from physical sensors.

[1] The term "word" is used throughout this thesis in accordance with Webster's Dictionary: "A speech sound or combination of sounds having meaning and used as a basic unit of language and human communication."

2 Problems of Early Lexical Learning

Problem 1: Linguistic Unit Discovery The first problem is to discover linguistic units which correspond to words of the target language. This is a challenging problem since most infant-directed language consists of multi-word utterances [1]. Furthermore, words will be rendered differently each time they are produced due to variations in the speakers' pronunciation, identity, age, gender, emotional state, background noise, and other factors. The learner must cope with this noisy evidence and infer the underlying linguistic units.

Problem 2: Semantic Categorization A second problem is to learn semantic categories which serve as referents for linguistic units. No knowledge about innate semantic categories is assumed by the model; instead, categories appropriate for the target language must be learned from positive examples. The learner must once again confront the problems of variations of natural phenomena and noise. For example, two shades of red rarely look *exactly* the same. The learner must form a category for the color red based on a finite number of exemplars.

Problem 3: Linguistic-to-Semantic Mappings The learner must associate linguistic units with appropriate semantic categories. These associations must be inferred from collections of experiences since single examples will have multiple interpretations.

These three problems are treated as different facets of one underlying problem: to discover structure across linguistic and contextual sensory input.

3 The Model

The Cross-channel Early Lexical Learning (CELL) model was developed to understand how infants acquire early words from multiple streams of sensory input. Lexical learning is driven by a search for structure across input channels. CELL constructs models of this structure by creating *lexical items*. A lexical item contains a specification of a linguistic unit (for spoken language this would correspond to the acoustic model of a word), and a specification of a corresponding perceptually grounded semantic category to which the linguistic unit refers. CELL is a model of *early* stages of the acquisition process. Later stages which involve feedback from caregivers as the language learner attempts to produce words is not modeled.

An overview of the model is illustrated in Figure 1. Sensors such as cameras, microphones, and other physical sensors provide input for the model. Feature analyzers extract multiple channels of input from the sensors. There is no one-to-one mapping between channels and sensors. Multiple channels may be derived from a single sensor. Conversely, a single channel may combine evidence from multiple sensors.

Sensors generate input signals which encode information about the learner's environment. The feature analyzers extract salient information from the sensory signals. Each channel represents a different aspect of the environment. A subset of input channels convey linguistic information to the learner and are referred to as linguistic channels. The remaining input channels are referred to as contextual channels. They carry information about the environment which encodes the underlying semantics of linguistic input.

Continuous streams of features from input channels are segmented into discrete chunks called *events*. These events are referred to as *L-events* when generated from the linguistic channels. L-events capture activity in all linguistic channels for the period of time corresponding to an utterance. For spoken input, an L-event corresponds to a spoken utterance delimited by silence. Another set of segmentation processes operate on the contextual channels to generate *S-events* which record non-linguistic information. An S-event captures activity in all contextual channels corresponding to a discrete period of time. S-events are generated in response to salient events in the environment.

Infant directed speech usually refers to the immediate context [7]. Words and their meanings are often perceived together. In CELL, we assume that L-events contain linguistic units whose reference may be found in co-occurring S-events. A natural bias to facilitate early lexical learning is to only attend to L-events and S-events which occur together. To facilitate this bias, S-events and L-events feed into a *co-occurrence filter* which detects events that overlap in time. When co-occurring events are detected, the filter generates an LS-event which consists of the overlapping L-event and S-event.

LS-events are placed in a *short term memory (STM)* which maintains a complete record of recent LS-events. The STM is a first-in-first-out (FIFO) buffer. When a new LS-event is added, the oldest element of the STM is discarded. The STM serves as a rough approximation of human memory limits [6]. It also conserves computational resources. CELL processes contents of the STM using an exhaustive search for recurring subevents. The search space grows factorially with the size of the STM, motivating use of a small STM.

Infant directed speech is highly redundant [11], often containing whole or partial phrase repetitions. Thus, we can expect that salient words will often recur in close temporal proximity. This observation motivates the next stage of processing. The recurrence filter searches exhaustively for matching subsequences of L-events which recur within the STM in similar contextual contexts. The search is invoked each time a new LS-event arrives in STM.

If a sub-sequence of an L-event recurs in similar contexts within the STM, a prototype is generated to represent the {L-subevent, s-subevent} pairing. The L-subevent and S-subevent pair forms a lexical candidate, a hypothesis of a possible linguistic unit and its corresponding referent. Lexical candidates generated by the recurrence filter are stored in the *mid-term memory (MTM)*. The MTM

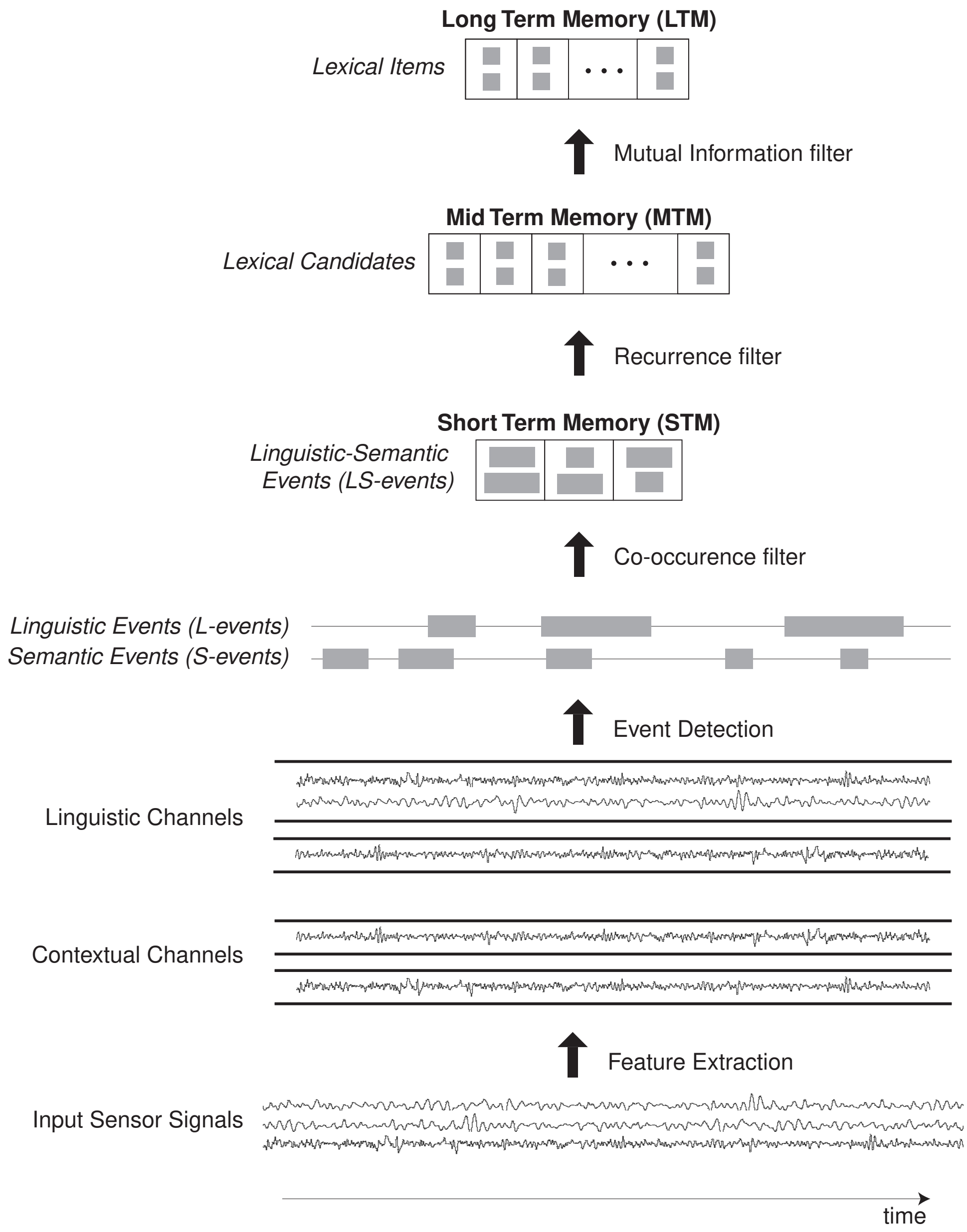

Figure 1: Overview of data flow through the CELL model.

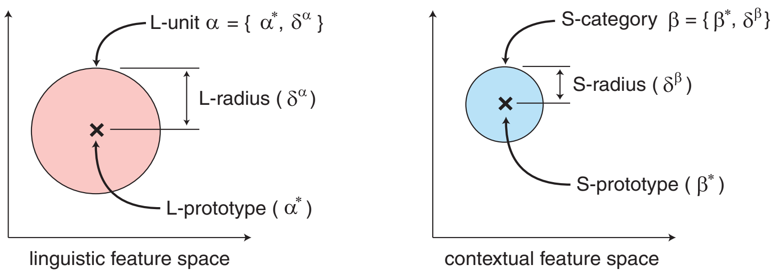

Figure 2: A lexical item consists of a linguistic unit (L-unit) and an associated semantic category (S-category).

is a buffer of lexical candidates being considered by the learner. Similar to the STM, the MTM is also a FIFO buffer of limited but significantly larger size.

CELL uses a model of linguistic units and semantic categories based on prototypes. Consider a linguistic unit in the target language. Each time an instance of this unit is produced it will be realized with natural variations, presumably centered around some idealized form. A simple model of the unit is to posit a central exemplar, or prototype, and a error tolerance to allow for variation. A prototype and radius define a range of inputs which may be grouped together for the goal of forming lexical items. These concepts are illustrated in Figure 2. Each plot represents an abstract space in which linguistic units and semantic categories may be defined. Any point in the two-dimensional space corresponds to an instance of a linguistic unit or semantic category. The distance between two points in each space is determined by the distance metrics $d_L()$ and $d_S()$. A new observation matches an S-category or L-unit if it falls within the model's radius.

The stages of CELL presented thus far can be viewed as a series of attentional filters. A continuous stream of sensory input is processed by feature analyzers, event detectors, a co-occurrence filter and a recurrence filter to produce lexical candidates.

The next stage of processing combines evidence from multiple lexical candidates to create hypotheses of lexical items. To create lexical items, CELL first calculates an optimal L-radius and S-radius for each {L-prototype, S-prototype} candidate in MTM. A measure of goodness based on mutual information is generated for each hypothesized lexical item in the process of radii optimization. This process is invoked each time a new lexical candidate is added to MTM.

Mutual information is used to evaluate the degree of cross-channel structure captured by pairing an L-unit and S-category. Mutual information is a measure of the reduction in uncertainty of one variable due to knowledge about a second variable [2]. We assume that the mutual information between L and S will be high if the associated prototypes correspond to an actual lexical item of the target language. This assumption is based on the fact that infant-directed speech often refers to the immediate context [11]; knowledge about the existence of a word will greatly reduce uncertainty about the presence of the word's referent, and vice versa.

For each lexical candidate, a search is performed to determine the optimal radii around each prototype pair. Candidates which lead to high maximum mutual information are selected as output and placed in a long term memory (LTM). These constitute the final output of CELL.

4 Implementation

Figure 3 summarizes the components of the implementation in terms of the CELL architecture. Input is provided by two sensors. A color CCD video camera provides images of objects presented to the system, and a microphone senses acoustic speech signals. Three channels of features are extracted from these sensors. A phoneme analyzer produces time-varying estimates of 39 English phoneme probabilities. These 39 features are grouped to form the linguistic channel. A visual analyzer detects objects in the scene. When an object is present, two contextual channels of features which represent the object's shape and color are extracted.

4.1 Representing and Comparing Spoken Utterances

Spoken utterances were represented as arrays of phoneme probabilities. A recurrent neural network processed RASTA-PLP coeffients [5] to estimate phoneme and speech/silence probabilities. Session recordings were segmented into utterances by detecting contiguous segments of speech in which the probability of silence estimated by the RNN was low.

Spoken utterances were segmented in time along phoneme boundaries, providing hypotheses of word boundaries. To locate phoneme boundaries, the RNN outputs were 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 would obtain: (1) A phoneme sequence: the most likely sequence of phonemes which were concatenated to form the utterance., and (2) The location of each phoneme boundary for the sequence. Any subsequence within an utterance terminated at phoneme boundaries could form a word hypothesis.

A distance metric, $d_A()$, is a measure of the similarity between two speech segments. A sequence of RNN output is equivalent to an unpruned phoneme lattice from which multiple phoneme sequences may be derived. The following distance metric was developed to compare pairs of RNN output sequences.

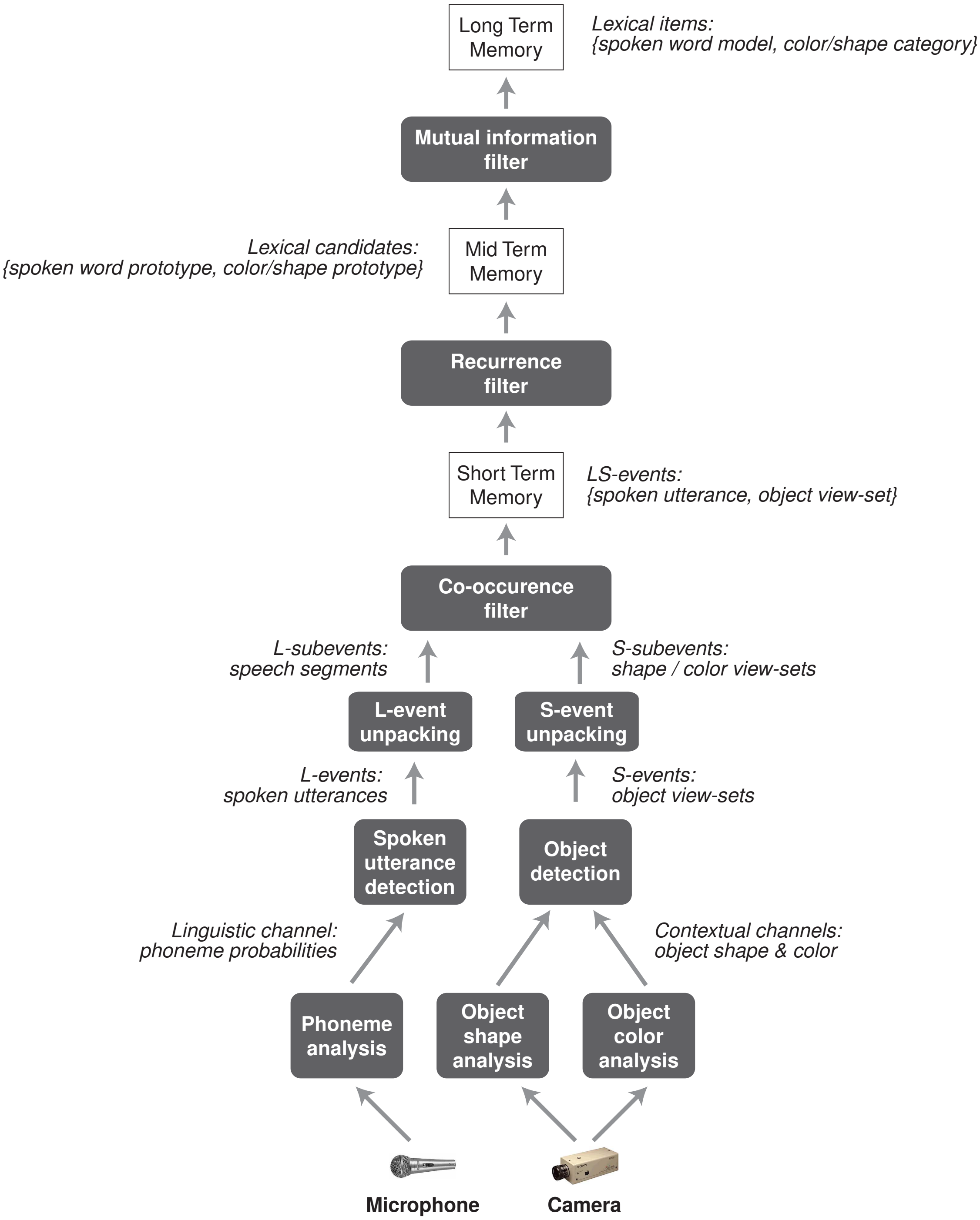

Figure 3: Overview of an implementation of CELL for color and shape name learning based on acoustic and visual input.

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). A confidence metric is generated by scaling each likelihood estimate by the 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:

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

4.2 Visual Processing

Three-dimensional objects were represented using a view-based approach in which two-dimensional images of an object captured from multiple viewpoints. Figure-ground segmentation was simplified by assuming a uniform background. A Gaussian model of the illumination-normalized background color was estimated and used to classify background/foreground pixels. Large connected regions near the center of the image indicated the presence of an object [2].

Based on methods developed by Schiele and Crowley [9], objects are represented by histograms of local features derived from multiple 2D views of an object. Shape was 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 were computed. All pair-wise values were 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 [9] directly for the comparison of object models. We used the χ^2–divergence for this purpose. The representation of 3D shapes are based on a collection of 2D shape histograms, each corresponding to a particular view of the object. Each 3D object was represented by 15 shape histograms. Image sets were compared by summing the divergences of the four best matches between individual histograms.

[2] The robot was used a learning algorithm to control its motors in order to center an object in its field of view from varying perspectives [8]

4.3 Joint Audio-Visual Clustering

Clustering consisted of two steps. In the first step, the AV-events were passed through a first-in-first-out buffer. The buffer had a capacity of five events. Each time a new event was inserted into the buffer, a *recurrence filter* searched for repeating audio and visual patterns within the buffer. For example, if a speaker repeated 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 to perform an exhaustive search over all possible image sets and speech segments in the n most recent AV-events where n is the size of the *short term memory*. Thus, the output of the recurrence filters consist 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, and the most reliable clusters are used to generate output of the system. Assume that there are N sound-shape hypotheses. The clustering process proceeds by considering each hypothesis as a reference point, in turn. Assume hypothesis X has been chosen randomly as a reference point. Each remaining $N-1$ hypotheses may be compared to X using $d_V()$ and $d_A()$. Two thresholds, t_V and t_A are used to compare the hypotheses. Two indicator variables are defined with respect to X:

$$A = \begin{cases} 0 & if\ d_A(X, h_i) > t_A \\ 1 & if\ d_A(X, h_i) \leq t_A \end{cases} \tag{2}$$

$$V = \begin{cases} 0 & if\ d_V(X, h_i) > t_V \\ 1 & if\ d_V(X, h_i) \leq t_V \end{cases} \tag{3}$$

$$\tag{4}$$

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:

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

The probabilities required to calculate $I(A;V)$ can be estimated from smoothed frequency counts. $I(A;V)$ is a function of the thresholds t_A and t_V which are determined by searching for values that maximize the mutual information between A and V.

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

This process 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.

5 The Audio-Visual Corpus

We have gathered a corpus of audio-visual data from infant-directed interactions [8]. 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, dog, horses. A total of 42 objects, six objects for each class varying in colow, size, texture, and shape were obtained.

Each caregiver-infant pair participated in 6 sessions of play with 7 objects, one at a time. All caregiver speech was recorded using a wireless headset microphone onto DAT. In total we collected approximately 7,600 utterances comprising 37,000 words accross 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 objects were imaged from various perspectives using a camera mounted on a motorized armature. A total of 209 images were collected for each of 42 objects resulting in a database of 8,778 images.

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 a 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.

6 Results

Audio-visual data corresponding to each of the six speakers was processed separately. For each dataset, AV-events were processed in the order they were generated in the sessions. The top 15 items resulting from the MMI maximization step were assessed for each speaker.

For a selected reference candidate, the space of acoustic and visual radii were searched to locate the point of maximum mutual information. Figure 4 presents two examples of mutual information surfaces. In each plot, the height of the surface shows mutual information as a function of the radii. 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 radii. 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. Radii were selected at the point where the surface height, and thus mutual information, was maximized.

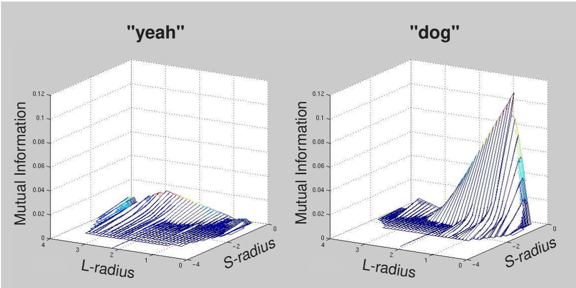

Figure 4: Mutual information as a function of L-radius and S-radius for two lexical candidates.

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 MMI maximization instead the system searched for globally recurrent speech patterns. Speech segments which were most often repeated in the entire set of recordings for each speaker determined a lexical item. This acoustic-only model was a rough approximation of a minimum description length approach to finding highly repeated speech patterns.

Table 1: Results of evaluation on three measures (see text) 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 (Table 1) indicate that the audio-visual clustering was able to extract a large proportion of English words from this very difficult corpus. Typical speech segments in the lexicons included names of all six objects in the study, as well as onomatopoetic 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 combine with acoustic evidence in the clustering process.

7 Conclusions

Our results represent the first successful effort to automatically acquire linguistic knowledge from raw infant directed speech and natural raw visual context. CELL acquired a lexicon of visually grounded words across six different participants without any changes to system parameters. The success of the model on this database serves as an existence proof that the strategies proposed in CELL might be employed by an infant to learn from similar types of input.

Comparisons with an acoustic-only model demonstrate the importance of cross-channel structure in the lexical acquisition process. Semantic accuracy increased over four fold when the visual channel was used. Most surprisingly, cross-channel constraints dramatically increased word segmentation and discovery performance in comparison to the acoustic-only model.

An underlying assumption of modularity is typical in current models of speech segmentation, word discovery, and lexical acquisition. Models of segmentation and word discovery typically operate on the acoustic channel alone. These models assume that sufficient information is available within the structure of a single channel. Similarly, models of lexical learning assume that linguistic units have already been identified by the language learner.

Our results bring the assumption of strict modularity into question. The model demonstrates improved learning by leveraging information across channels at an early stage. Infants are known to possess all the necessary capabilities of sensory processing and correlational analysis necessary to employ the strategies proposed in CELL. Our results suggest that infants harness cross-channel structure on the path to language.

8 Acknowledgements

This work has benefited greatly from discussions with Alex Pentland, Allen Gorin, Steven Pinker, Rupal Patel, and Bernt Schiele.

References

[1] R.N. Aslin, J.Z. Woodward, N.P. LaMendola, and T.G. Bever. Models of word segmentation in fluent maternal speech to infants. In James L. Morgan and Katherine Demuth, editors, *Signal to Syntax*, chapter 8, pages 117–134. Erlbaum, Mahwah, NJ, 1996.

[2] T.M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 1991.

[3] J. Feldman, G. Lakoff, D. Bailey, S. Narayanan, T. Regier, and A. Stolcke. Lzero: The first five years. *Artificial Intelligence Review*, 10:103–129, 1996.

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

[5] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, October 1994.

[6] G.A. Miller. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956.

[7] J.M. Pine. The language of primary care givers. In C. Gallaway and B.J. Richards, editors, *Input and Interaction in Language Acquisition*, chapter 1, pages 15–37. Cambridge University Press, Cambridge, UK, 1994.

[8] D.K. Roy. *Learning Words from Sights and Sounds: A Computational Model.* PhD thesis, Massachusetts Institute of Technology, 1999.

[9] 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.

[10] J. Siskind. A computational study of cross-situational techniques for learning word-to-meaning mapping. *Cognition*, 61(1-2):39–91, 1996.

[11] C.E. Snow. Mother's speech to children learning language. *Child Development*, 43:549–565, 1972. speech is redundant in motherese.