LEARNING WORDS FROM NATURAL AUDIO-VISUAL INPUT

Deb Roy *Alex Pentland*

20 Ames Street, Rm. E15-388, Cambridge, MA 01239, USA
http://www.media.mit.edu/ dkroy/toco.html
{dkroy, sandy}@media.mit.edu

ABSTRACT

We present a model of early word learning which learns from natural audio and visual input. The model has been successfully implemented to learn words and their audio-visual grounding from camera and microphone input. Although simple in its current form, this model is a first step towards a more complete, fully-grounded model of language acquisition. Practical applications include adaptive human-machine interfaces for information browsing, assistive technologies, education, and entertainment.

1. INTRODUCTION

Around their first birthday, infants first begin to use words to describe salient aspects of their environment including objects, actions, and people. They learn these words by listening to speech and observing their environment. The acquisition process is complex. The infant must successfully segment connected multiword spoken utterances into acoustic units which correspond to the words of their language. The infant must also categorize the world in order to acquire the proper semantic associations of these acoustic units. Remarkably, the infant is capable of all these processes despite the noisy input provided by his or her perceptual system.

This paper reports on our on-going efforts to develop a computational model of early word learning using input similar to what an infant might receive in certain situations. Our current task focuses on learning words that can be grounded in the visual semantics of static views of objects with relatively easy figure-ground separation. Input to the system consists of naturally spoken multiword utterances and color images. We have implemented a system in a probabilistic framework that is able to learn an *audio-visual lexicon* which can then be used to understand and generate spoken language grounded in visual semantics.

Language is grounded at several levels in our model. First, the surface forms of words are represented using statistical models which account for the inherest acoustic variability of speech (Section 3.1). Second, word meanings are grounded in terms of visual models which are derived from camera input (Section 3.2). Word semantics are grounded in terms of statistically defined subspaces of color and shape. Last, word classes are defined in terms of visual dimensions. All words which are grounded in color dimension subspaces (ex. "red", "greenish", "pink") are considered part of the same class since they are all grounded in terms of only color. Similarly shape terms (ex. "car", "ball", "bottle") form a separate word class.

Many interesting computational models of learning word semantics have been proposed including [3], [1], [10]. Each of these models relies on either human-generated text, phoneme transcripts, or assumptions about pre-existing discrete semantic classes. In contrast, our model learns surface-form and semantic models of words from only audio and visual sensory input.

The word learning model has applications in the design of human-machine interfaces that use spoken language. A significant problem in designing effective speech interfaces is the difficulty in anticipating a person's word choice and associated intent [2]. Our system addresses this problem by learning the vocabulary of each user together with its visual grounding. This approach enables a new type of human-machine interface which can adapt to the preferences and abilities of individual users over time.

2. THE MODEL

The word learning model is best understood by considering three tasks which it performs: learning, understanding, and generating spoken language. This section provides a functional description of the various processes and data structures involved in these tasks. Technical descriptions of algorithms are presented in Section 3.

2.1. Learning from Audio-Visual Input

Figure 1 shows the processes (light gray boxes) and data structures (dark boxes) involved in learning from audio-visual input. Input consists of color images paired with spoken utterances acquired using a CCD camera and a microphone. For example a person might present a red ball to the camera and say "this is a red ball".

The input image is analyzed along shape and color[1] dimensions and represented by a set of histograms (See Section 3.2). Phoneme analysis produces a representation of the spoken utterance consisting of (1) a *phoneme trace* consisting of 40 phoneme class probabilities estimated at

[1]We plan to add analysis of motion, texture, and relational position dimensions in the future.

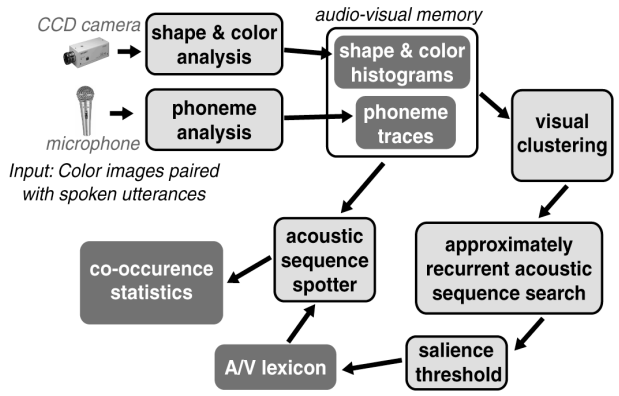

Figure 1: Learning audio-visual models from camera and microphone input

a rate of 100 Hz, and (2) a Hidden Markov Model encoding the most likely phoneme state sequence of the speech (see Section 3.1). Hereafter, we will refer to the data extracted from one audio-visual interaction as an *AV-event*.

Once a sufficient number[2] of AV-events have been accumulated, the system builds an AV-lexicon which can then be used for spoken language understanding and generation. Items in the AV-lexicon can be semantically grounded in terms of color, shape, or both color and shape.

Conceptually the lexicon is created as follows. All AV-events are clustered along each possible combination of visual dimensions. A search for re-occurring speech segments then attempts to find acoustic labels for each visual cluster[3]. An example taken from real input data will be used to describe this process. In the sample task a person spoke several utterances to describe images of simple blocks of various colors and shapes. The visual clustering process generated several AV-event clusters. One such cluster roughly corresponded to the blue region of color space. Some of the spoken utterances and their automatically generated phonetic transcripts are shown in Table 1.

Utterance	Phonetic transcript
"this cone is blue"	i s - k ah m i n - **b l i w**
"that's a blue cube"	b eh - f ah - **b l i** - k y oo
"this ball is blue"	i z - b au l i z - **b l oo**
"this is blue"	i z oo z - **b l i oo**

Table 1: Sample spoken utterances associated with an AV-event generated by clustering in the color dimension. Portions of the phonetic transcripts corresponding to the word "blue" are emphasized for the convenience of the reader.

Once a subset of the AV-events have been selected based on visual clustering, a search procedure locates reoccurring acoustic segments within this set of utterances. In Table 1, this might include segments corresponding to re-occurring words such as "this" and "is" as well as "blue". A filtering process, labeled *salience threshold* in Figure 1, eliminates words which are not strongly correlated with the visual cluster being analyzed. The result is a set of speech segments and associated visual models that are stored in the AV-lexicon. Table 2 shows sample results from analyzing a set of utterances including those listed in Table 1. The first section shows several acoustic segments which were found to re-occur across multiple utterances but were rejected by the salience threshold process. The second section contains utterances which exceeded the salience threshold making them likely acoustic labels for the visual cluster.

Acoustic matches rejected due to low salience	f ah ("a") i z ("is") k y oo ("cube")
Acoustic matches accepted due to high salience	i z - b l oo - ("is blue") b l oo ("blue") b l i oo ("blue")

Table 2: Sample output of the acoustic search generated by clustering along the color dimension. Human generated transcripts of each segment are shown in parentheses for the convenience of the reader. Note that the last two entries are alternate pronunciations of the word "blue".

In a final stage of processing, the *acoustic sequence spotter* searches all spoken utterances in the audio-visual memory for occurrences of each speech segment stored in the AV-lexicon. Co-occurrence statistics are accumulated to determine patterns of word order at the word class level. In English, color terms always precede shape terms when the words are adjacent. If the shape term proceeds the color term, some other sequence (such as the word "is") must be inserted between the shape and color terms. These types of regularities can be learned as word-class co-occurrence statistics. In the current implementation the statistics are limited to recording how often shape terms precede color terms and vice versa. In the future we expect to expand the role of this analysis to address more advanced types of statistical syntax learning.

To summarize, the model takes as input a set of color images of objects paired with descriptive spoken utterances. The learning process produces two data structures. The AV-lexicon contains a set of audio-visual models consisting of speech models and associated visual models. Co-occurrence statistics model higher level word-class regularities. We now describe how these two data structures can be used to perform language understanding and generation.

2.2. Understanding Spoken Language

Figure 2 shows how spoken input is processed for an object recall task. The goal is to take a spoken description of one or more objects as input, and to sort a pool of objects by how well they match the meaning of the spo-

[2]The number of events required depends on the size of vocabulary and the distribution of words used in the input utterances.

[3]Note that the same AV-event will be a member in multiple visual clusters. For example an AV-event with an image of a red ball would be part of the cluster representing the red part of color space, the "round objects" part of shape space, and the "red, round objects" part of color-shape space.

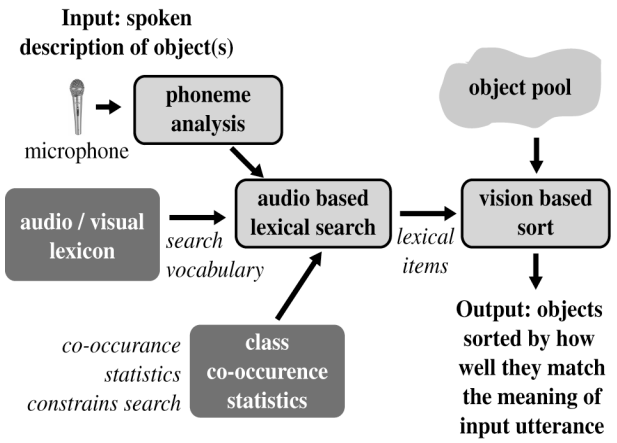

Figure 2: Understanding spoken language

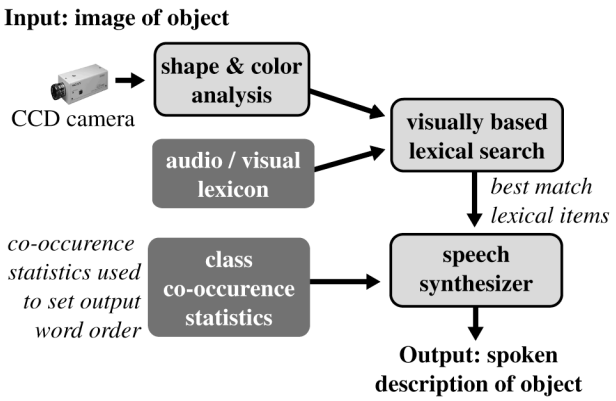

Figure 3: Generating spoken language

ken input. The spoken utterance is recorded and analyzed using the same process as the learning phase. An audio-based lexical search is performed to find occurrences of items from the AV-lexicon in the input speech. The class co-occurrence statistics are used to constrain this search process similar to a statistical grammar in a conventional speech recognizer. The search produces one or more lexical items which were detected in the input speech. The visual models associated with these lexical items are then used to sort a pool of images in descending relevancy. In a typical interaction a person might say "blue balls". In response the system should locate images of objects which are blue in color and round in shape to demonstrate its understanding of the input speech.

2.3. Generating Spoken Language

Figure 3 shows the complementary task of generating a spoken description of an input image. The input image is analyzed in terms of color and shape. The AV-lexicon is searched for items which best match each visual dimension of the input image. The best matching lexical items are then spoken using a commercial phonetic speech synthesizer. The class co-occurrence statistics are used to set the word order of the multiword output utterance. In a typical interaction a person might present a blue ball and in response the system would speak the phonetic sequence "b l oo - b o l" (i.e. "blue ball").

3. IMPLEMENTATION DETAILS

Section 2 gave a functional description of how the word learning model creates and uses an audio-visual lexicon. This section presents selected algorithmic details of the current implementation.

3.1. Speech Processing

Speech is processed in real-time by a recurrent neural network producing phoneme probability estimates at a rate of 100 Hz. When a spoken utterance is detected, a Hidden Markov Model of the most likely phone sequence is generated using the Viterbi algorithm and an all-phone loop grammar with unigram phoneme transition probabilities. The phoneme recognition accuracy is approximately 70% when insertion and deletion errors are balanced. The sequence of phoneme probability estimates associated with the utterance is referred to as the *phoneme trace*. The HMM and the associated phoneme trace forms our underlying representation of a spoken utterance. For more details on the speech processing and segmentation algorithms see [7] and [8].

Speech segments are compared by using forced viterbi alignment as follows. Consider two speech segments A and B. For each segment we will have a phoneme trace and corresponding HMM. If we let L_{AB} be the log likelihood that HMM_A produced $Trace_B$ (computed using standard forced Viterbi alignment), and L_{BA} be the log likelihood that HMM_B produced $Trace_A$, we can then define the distance between A and B to be $(L_{AB} + L_{BA})/2$. Similar to many word spotting algorithms (for example see [4]) we normalize this score by an all-phone garbage model.

3.2. Visual Processing

Objects are segmented from a controlled background using a statistical background color model. The resulting object mask is used to compute a set of local shape features based on spatial relations between derivatives. Color features are computed by sampling pixel values within the object mask region. The shape and color features are accumulated in separate histograms which can later be compared using χ^2 divergence. For further details on the visual processing algorithms see [9].

3.3. CREATING AND USING AN AUDIO-VISUAL LEXICON

Several segmentation and clustering algorithms are used in the learning processes described in Section 2.1. Initial visual clustering is performed by computing pairwise distances between each pair of AV-events using the χ^2 divergence on shape and color histograms. A preset threshold is used to determine cluster membership along each combination of dimensions.

The next stage of the algorithm finds acoustic labels for

each of the visual clusters. The search for re-occurring acoustic segments is based on an exhaustive search within sets of spoken utterances generated by visual clustering. Ideally the algorithm would search for acoustic matches between each possible pair of segments. Unfortunately, the search space becomes immense if we search for speech segment boundaries at a phoneme trace frame rate of 100 Hz for moderate-sized clusters. Even if the search only considers segment boundaries which coincide with phoneme boundaries the search space is often still impractically large. To speed the search we have implemented a syllable hypothesis generator based on consonant-vowel analysis. The algorithm searches for re-occurring segments by exhaustively computing the distance between all one- and two-syllable segments within the cluster using forced alignments at syllable boundaries. Pairs of segments which have a symmetric log likelihood below a predefined threshold are considered an acoustic match.

For the next stage of processing, we need to define the *salience* of each speech segment similar to [3]. If we let C be the subset of AV-events associated with the visual cluster of interest and $\overline{C}$ be all remaining AV-events which are not part of the cluster, we can define the salience of a speech segment s to be $\frac{Pr(s|C)}{Pr(s|\overline{C})}$, where $Pr(s|C)$ is the probability of detecting s within the AV-events in C and $Pr(s|\overline{C})$ is the probability of detecting s in the remaining AV-events, $\overline{C}$. A highly salient word will occur often within the visual cluster and rarely in the context of other visual input. Words such as "red" or "ball" will likely be salient for appropriate clusters in color or shape dimensions respectively. In contrast words such as "the" and "is" will not be salient for any cluster when using sufficiently large input data sets.

Class co-occurrence statistics are computed by acoustically searching for all entries of the final AV-lexicon in the audio-visual memory. Different thresholds are set for the syllable-based search algorithm described above to find occurrences of lexical items in AV-memory.

Continuous multiword speech recognition is performed in a standard HMM framework using Viterbi decoding. The search lattice is constructed using phoneme state sequences of lexical items. Word transition probabilities within the search lattice are determined using class co-occurrence statistics.

4. CONCLUSIONS AND FUTURE DIRECTIONS

We have presented a computational model of early word learning that clearly defines the processing of audio and visual input leading to the formation of a lexicon of words with acoustic and visual grounding. This model has been implemented and initial tests show that the model works with natural data from interactions with people in controlled situations. We are currently running experiments on larger tasks with multiple subjects and will report our findings in the future.

We believe that building models of this type might be a first step towards a better understanding how infants come to master language with such remarkable speed and consistency. Defining semantics in terms of visual input allows for grounding word semantics and word classes in terms of the ontology of the world as projected through the system's perceptual system. We believe that further exploration of this theme may lead to powerful new algorithms for syntax acquisition.

This work has practical applications for adaptive human-machine interfaces in numerous domains including information browsing (for example browsing and searching catalogs), command-and-control, entertainment [5], and disability aids [6].

5. REFERENCES

1. 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.
2. G.W. Furnas, T.K. Landauer, L.M. Gomez, and S.T. Dumais. The vocabulary problem in human-system communications. *Communications of the Association for Compuring Machinery*, 30:964–972, 1987.
3. Allen L. Gorin. On automated language acquisition. *Journal of the Acoustic Society of America*, 97(6):3441–3461, 1995.
4. Richard Rose. *Word Spotting from Continuous Speech Utterances*, chapter 13, pages 303–329. Kluwer Academic, 1996.
5. Deb Roy, Michal Hlavac, Marina Umaschi, Tony Jebara, Justine Cassell, and Alex Pentland. Toco the toucan: A synthetic character guided by perception, emotion, and story. In *Visual Proceedings of Siggraph*, Los Angeles, CA, August 1997. ACM Siggraph.
6. Deb Roy and Rupal Patel. Adaptive user interfaces for individuals with speech and physical impairments. Technical Report (upcoming), MIT Media Lab Vision and Modeling Group, 1998.
7. Deb Roy and Alex Pentland. Multimodal adaptive interfaces. Technical Report 438, MIT Media Lab Vision and Modeling Group, 1997.
8. Deb Roy and Alex Pentland. Word learning in a multimodal environment. In *Proceedings of ICASSP*, Seattle, Washington, May 1998. IEEE Computer Society Press.
9. B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV96-I*, pages 610–619, 1996.
10. Jeffrey Siskind. A computational study of cross-situational techniques for learning word-to-meaning mapping. *Cognition*, 61(1-2):39–91, 1996.

LEARNING WORDS FROM NATURAL AUDIO-VISUAL INPUT

Deb Roy and Alex Pentland

20 Ames Street, Rm. E15-388, Cambridge, MA 01239, USA
http://www.media.mit.edu/ dkroy/toco.html
{dkroy, sandy}@media.mit.edu

We present a model of early word learning which learns from natural audio and visual input. The model has been successfully implemented to learn words and their audio-visual grounding from camera and microphone input. Although simple in its current form, this model is a first step towards a more complete, fully-grounded model of language acquisition. Practical applications include adaptive human-machine interfaces for information browsing, assistive technologies, education, and entertainment.