INTEGRATION OF SPEECH AND VISION USING MUTUAL INFORMATION

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/

ABSTRACT

We are developing a system which learns words from co-occurring spoken and visual input. The goal is to automatically segment continuous speech at word boundaries without a lexicon, and to form visual categories which correspond to spoken words. Mutual information is used to integrate acoustic and visual distance metrics in order to extract an audio-visual lexicon from raw input. We report results of experiments with a corpus of infant-directed speech and images.

1. INTRODUCTION

We are developing systems which learn words from co-occurring audio and visual input [5, 4]. Input consists of naturally spoken mutliword utterances paired with visual representations of object shapes (Figure 1). Output of the system is an audio-visual lexicon of sound-shape associations which encode acoustic forms of words (or phrases) and their visually grounded referents. We assume that, in general, the audio and visual signals are uncorrelated in time. However, when a word is spoken, its visual representation will sometimes be present in close temporal proximity. The goal is to detect and model such cross-modal structure.

The problem of finding structure from this data may be viewed as both a supervised and unsupervised learning problem. Viewed as unsupervised learning, the only data available to the system is raw acoustic and visual input without any clean training labels. On the other hand, each stream of data may be treated as noisy labels for the other. Speech segments embedded within spoken utterances may be labels for co-occurring images, and images may be labels for segments of co-occurring speech. This paper reports on recent advances in multimodal integration using mutual information as a measure to combine audio and visual distance metrics. The system has been evaluated on a corpus of infant directed audio-visual data.

This work is motivated by two goals. First, automatic language learning systems may be used to create robust human-computer spoken language interfaces which adapt to individual differences and preferences [1]. Second, we are interested in using computational models to gain insights into infant language acquisition [3].

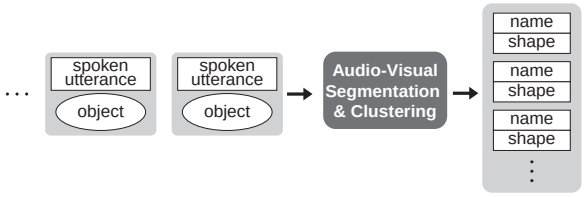

Figure 1: Input consists of spoken utterances paired with images of objects. Audio-visual clustering extracts visual shape categories and corresponding spoken names.

2. THE AUDIO-VISUAL CORPUS

We have gathered a corpus of audio-visual data from infant-directed interactions [3]. Six caregivers and their pre-linguistic infants (aged 7-11 months) 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 within 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 a small CCD camera mounted on a four degree-of-freedom robot shown in Figure 2. A total of 209 images from different perspectives were collected for each of the 42 objects resulting in a database of 8,778 images.

To prepare the corpus for processing, we performed the following steps: (1) The audio was segmented at utterance boundaries. This was done automatically by finding contiguous frames of speech detected by a recurrent neural network (see below), and (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. Video recordings of the caregiver-infant interactions were used to determine the correct object for each utterance. Each utterance-image set is referred to as an *AV-event* (audio-visual event). Input to

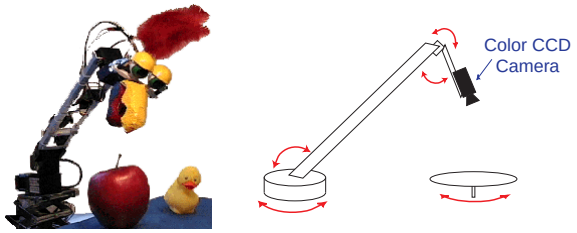

Figure 2: Photograph and drawing of a robot with four degrees of freedom used to capture images of objects. The CCD camera was mounted in the right eyeball. A turntable provided a fifth degree of freedom for viewing objects.

the learning system consists of a sequence of AV-events.

3. SPEECH REPRESENTATION, SEGMENTATION, AND COMPARISON

Spoken utterances were represented as arrays of phoneme probabilities and corresponding hidden Markov models (HMMs). A recurrent neural network processed RASTA-PLP coefficients [2] to estimate phoneme and speech/silence probabilities. The RNN had 12 input units, 176 hidden units, and 40 output units. The 176 hidden units were connected through a time delay and concatenated with the RASTA-PLP input. The RNN was trained off-line using the TIMIT database resulting in 69.4% accuracy using the standard TIMIT training and test datasets.

Spoken utterances were segmented along phoneme boundaries, providing hypotheses of potential word boundaries. To locate phoneme boundaries, the RNN outputs were treated as state emission probabilities in an HMM framework. Viterbi search was used to obtain the most likely phoneme sequence for a given phoneme probability array. Viterbi decoding of an utterance obtained: (1) The most likely sequence of phonemes in 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.

We defined a distance metric, $d_A()$, which measured the dissimilarity between two speech segments. One possibility was 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 [7]. A limitation of this method is that it relies on only the single most likely phoneme sequence. To make more complete use the entire phoneme probability array, we developed a novel 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 that λ_j generated α_i.

The Forward algorithm can be used to compute $P(\alpha_i|\lambda_j)$ and $P(\alpha_j|\lambda_i)$, the likelihood that the HMM derived from speech segment α_i (λ_i) generated speech segment α_j and that the HMM from α_j (λ_j) generated α_i. 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. In this method, each likelihood estimate is scaled by the likelihood of a default alternate hypothesis, λ^A:

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

We set the alternative hypothesis to be 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 was defined as:

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

4. VISUAL REPRESENTATION AND COMPARISON

Three-dimensional objects were represented using a view-based approach in which histograms of local image features from multiple two-dimensional images of an object represented the shape and color of the object. Figure 3 shows the stages of visual processing. 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 [1].

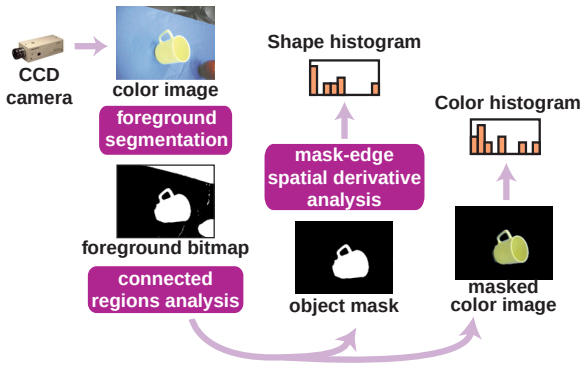

Figure 3: Extraction of object shape and color representations from a raw image.

Based on methods developed by Schiele and Crowley [6], objects were 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

[1] The robot's motion control was learned off-line by automatically creating joint angle tables for centering objects in the camera's field of view [3]

object edge at each point were computed. All pair-wise values were accumulated in a 2D histogram to represent an image. The shape representation was invariant to transformations in position, scale and in-plane rotation. Using multidimensional histograms to represent object shapes allowed the use of information theoretical or statistical divergence functions for the comparison of object models. We used the χ^2–divergence:

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

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 was 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. Histogram sets were compared by summing the divergences of the four best matches between individual histograms.

5. AUDIO-VISUAL INTEGRATION

Integration of audio and visual input consisted of two steps. In the first step, the AV-events were passed through a first-in-first-out (FIFO) 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. If a speaker repeated similar sounding words or phrases at least twice within five contiguous utterances while playing with similar shaped objects, the filter would select that recurrent sound-shape pair as a potential lexical item. The recurrence filter used the audio and visual distance metrics presented earlier to look for matches. It performed an exhaustive search over all possible image sets and speech segments (at phoneme boundaries) in the buffer. By using the FIFO buffer as a front-end for processing, the system exhibited on-line learning since AV-events were discarded once they passed through the buffer. Output from the recurrence filter consisted of speech segments and their hypothesized visual referents represented with phoneme probability arrays, HMMs, and visual histograms.

In the second step, the hypotheses generated by the recurrence filter were clustered, and the most reliable clusters were used to generate audio-visual lexical items. Let us assume that after processing a particular speaker's data, N sound-shape hypotheses were generated. 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:

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

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

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:

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

The probabilities required to calculate $I(A;V)$ can be estimated from smoothed frequency counts of the indicator variables. Note that $I(A;V)$ is a function of the thresholds t_V and t_A. 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.

Each hypothesis is taken as a reference point and the maximum mutual information (MMI) is found by searching all values of t_V and t_A. The hypotheses which result in the highest MMI are are selected as output of the system. The result is a set of audio-visual prototypes (the selected hypotheses) and radii (t_V and t_A) which specify allowable divergence from these prototypes.

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

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

A final step is to threshold the MMI score of each hypothesis and select those which exceed the threshold. The threshold was set manually for experiments reported below. In the future, reinforcement feedback from higher levels in the system could be used to learn the threshold value.

6. RESULTS

The audio-visual data corresponding to each of the six speakers was processed separately. For each dataset, the 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 sound-shape hypothesis generated by the recurrence filter, a two-dimensional space of acoustic and visual radii is searched to locate the point of maximum mutual information. Figure 4 presents two examples of mutual information surfaces from the corpus. 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 incorrectly paired with images of a shoe. The resulting surface is relatively low for all values of radii. The lexical candidate on the

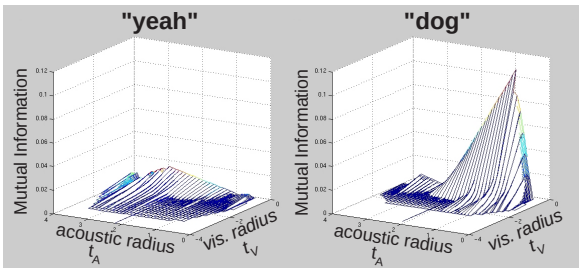

Figure 4: Mutual information as a function of L-radius and S-radius for two lexical candidates.

right correctly paired a speech segment of the word "dog" with images of a dog. The result is a strongly peaked surface form indicating that this pairing captures useful cross-modal structure.

We evaluated the six lexicons (one for each speaker) 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. M2 accepted single-phoneme segmentation errors. The third measure, M3, was the percentage of lexical items which satisfied M2 and additionally 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 so instead the system searched for globally recurrent speech patterns, i.e. speech segments which were most often repeated in the entire set of recordings of each speaker. This acoustic-only model may be thought of as a rough approximation to a minimum description length approach in which the underlying assumption is that stable and often-repeated sound patterns are likely to be words of the language.

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 of the evaluation are shown in Table 1. The second measure, M2, showed that 72% of the lexical items were English words with possible single-phoneme segmentation errors. This is a significant result given the extremely natural style of speech collected from the infant interactions. Typical entries in the lexicons corresponded to the names of all six objects in the study, as well as onomatopoetic sounds such as "ruf-ruf" for dogs, and "vroooom" for cars. The third measure, M3, indicated that 57% of the lexical items associated acoustic forms with correct visual semantics. In some cases, words with no concrete visual grounding (ex. "good") were acquired which passed M2 but not M3.

The comparison with the audio-only system clearly demonstrates the greatly improved performance in all three measures when visual context is combined with acoustic evidence in the lexical learning process. The segmentation accuracy, though relatively low at 28%, is impressive for such natural speech, and surprisingly, is four times higher than the acoustic-only output of only 7%.

7. CONCLUSIONS

Systems which process multiple input modalities typically rely on the fact that correlated features of the input streams are synchronized in time. This assumption may hold in certain cases such as lip reading. In many situations, however, precise time synchronized input cannot be assumed. We have presented an approach which combines multiple modalities without reliance on time synchronization. Although each modality is noisy, integrating them significantly improves performance for the word learning task. The maximization of mutual information effectively combines arbitrary distance metrics without need for an ad hoc method for directly combining metrics.

Future directions include the application of these methods to building adaptive human-computer spoken interfaces. The learning mechanisms may be used to provide adaptive vocabularies in speech systems which not only acquire personalized acoustic forms of words, but in certain domains where semantics may be grounded in input, the system may also learn person-specific semantics of words.

8. ACKNOWLEDGEMENTS

This work has benefited from collaborations with Alex Pentland, Bernt Schiele, Rupal Patel and Allen Gorin.

9. REFERENCES

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

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

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

[4] D.K. Roy and A. Pentland. Learning words from natural audio-visual input. *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[5] D.K. Roy and A. Pentland. Word learning in a multimodal environment. In *Proceedings of ICASSP*, Seattle, Washington, May 1998. IEEE Computer Society Press.

[6] 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.

[7] 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.