Learning Audio-Visual Associations using Mutual Information

Deb Roy, Bernt Schiele, and Alex Pentland
Perceptual Computing Section, MIT Media Laboratory
20 Ames Street, Cambridge, MA 02139 USA
{dkroy, bernt, sandy}@media.mit.edu

Abstract

This paper addresses the problem of finding useful associations between audio and visual input signals. The proposed approach is based on the maximization of mutual information of audio-visual clusters. This approach results in segmentation of continuous speech signals, and finds visual categories which correspond to segmented spoken words. Such audio-visual associations may be used for modeling infant language acquisition and to dynamically personalize speech-based human-computer interfaces for various applications including catalog browsing and wearable computing. This paper describes an implemented system for learning shape names from camera and microphone input. We present results in an evaluation of the system for the domain of modeling language learning.

1 Introduction

We are developing multimodal systems which learn from audio-visual data. We are interested in situations which provide input of the form illustrated in Figure 1. Input consists of mutliword naturally spoken utterances paired with visual representations of objects. In this paper we present a system which can automatically learn shape categories and their corresponding spoken names. The output of the system is a set of sound-shape associations which encode words (or phrases) and their visually grounded meanings. The approach we describe in this paper may be generalized to other types of inputs to expand the range of semantics beyond object shape.

1.1 Motivations

This work is motivated by various domains we have been investigating in which speech and visual input is available. The common denominator across each domain is that we expect linguistically meaningful correlations between the audio and visual streams. The audio stream contains words, and the visual stream carries potential meanings of these words. In each case the goal is to discover sound units which correspond to words, and visual categories which correspond to the meaning of the sound units. Situations we are considering include:

Modeling language acquisition Infants learn words based on noisy perceptual input. We are interested in modeling the process of early word learning based on audio-

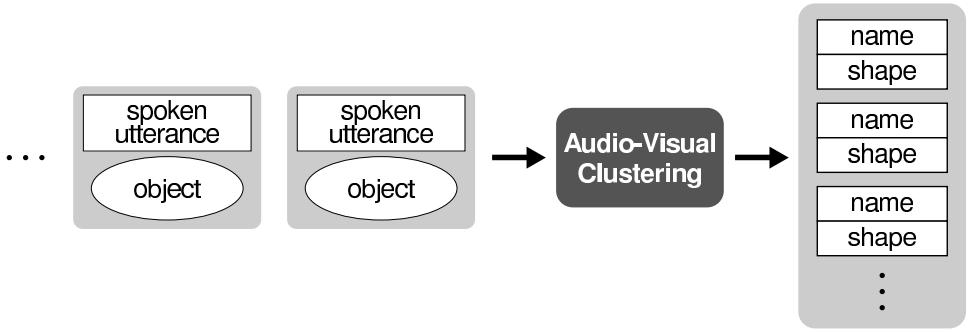

Figure 1. Input consists of multiword spoken utterances paired with images of objects. Audio-visual clustering extracts visual shape categories and corresponding spoken names.

visual input. The system presented in this paper has been used to process raw infant-directed speech paired with camera images of objects. The system has successfully learned words from this input. Sample results from this domain are presented in this paper. For more details see [22].

Wearable computing Wearable computers provide a unique opportunity to record and index the audio-visual environment of a user from a "first-person" perspective. Cameras and microphones may be added to a wearable computer enabling a system to watch everything a user sees, and record everything the user says. Over time, an audio-visual learning system can learn correspondences between the user's speech and visual input, assuming that at least some of the user's speech refers to the immediate visual context. To date, we have implemented wearable systems which record audio-visual data, and which cluster and analyze the visual environment [26, 27].

Catalog browsing In the future we plan to create a multimodal catalog browsing system in which users may use both speech and mouse clicks to browse and search images of clothing and other items. To train the system, the user may click on images and speak descriptions of them (eg. click on a shirt and say "this is a funky blue summer shirt"). Such interaction leads to a corpus of images associated with spoken utterances. By learning speech-visual correspondences, the system is able to learn specific words and phrases which the user chooses and their intended visual semantics. For the domain of clothing catalogs, the visual representations might naturally include shape, color and texture. Once trained, speech may be used to retrieve items which are not in view.

All of the above domains share the property that the speech may refer to aspects of co-occurring visual input. In each case, any word or phrase within an utterance may refer to any co-occurring visual feature.

1.2 Problems

In this paper we specifically deal with the problem of finding sound-shape associations as a special case of extracting audio-visual associations. In order to achieve this goal, the system must address three problems:

Word Discovery We assume that spoken utterances contain multiple connected words.

There are no equivalents of spaces between printed words when we speak naturally; There are no pauses or other acoustic cues which separate the continuous flow of words. To acquire appropriate audio-visual clusters, the system must segment connected speech in order to discover word boundaries.

Shape Categorization Shape categories must be established from noisy camera input. The number of categories is unknown and must be learned from the data. In the experiments described in this paper the segmentation of objects is facilitated by using a uniform background. In [27] we describe a system which uses color, texture and motion in order to extract object hypotheses from a continous video signal.

Speech-to-Visual Association Inference The third problem is to establish associations between speech sounds and corresponding shape categories.

In our formulation, all three problems are considered as different facets of one underlying problem of audio-visual association inference.

1.3 Approach

We have developed a model of audio-visual processing which learns audio-visual associations from a set of images paired with spoken utterances. Processing in the system consists of two stages. In the first stage, utterances are segmented into possible sounds units of the language. Each segment serves as a hypothesis of the name of the co-occurring object shape. Thus each utterance-shape input pair results in a set of sound-shape hypotheses. This initial set is reduced using one or more attentional filters which retain hypotheses which are visually or acoustically salient.

In the second stage, the hypothesis sets generated from a corpus of object-utterance pairs are clustered using a mutual information metric. Clusters which score highest using this metric are used to generate sound-shape represenations as final output of the system. In the following sections, we first describe how representations of objects and utterances are extracted from sensory input. Similarity metrics used to compare representations within each input modality are presented. Mutual information based clustering across modalities is then described.

An important aspect of this work is the combination of information from different modalities. In our approach, we assume that similarity metrics are avaliable for comparing representations within each modality. Mutual information is used to combine similarity metrics without the need for ad hoc heuristics for combining similarity metrics. This approach has several advantages over using some fixed combination of similarity metrics, as discussed later in this paper.

2 Background

Several models of word learning have been proposed to explain how words are learned from linguistic and contextual input including [30, 10, 7, 23, 11]. The work of Gorin and his colleagues on models of language learning from raw speech is particulary relevant to our work. Gorin has developed an language acquisition system which learns from raw acoustic speech and semantic annotations [11]. Input consisted of connected word spoken utterances paired with one of 20 semantic class identification tags. The system is able to learn which

words are useful for a classification task using a measure of semantic relevance based on mutual information. In Gorin's framework, the *salience*, or semantic importance of a word is estimated using the weighted mutual information between the occurrence of the word, and the the occurrence of classification tags. Gorin's system does not attempt to learn semantic classes, and also does not address the problem of speech segmentation (a large vocabulary speech recognizer is assumed to exist). In contrast, our work address both of these problems.

The use of multiple sensor modalities is meaningful in many different contexts. The related research areas are therefore highly diverse and interdisplinary. In the robotics literature *sensor fusion* is an active research area [16, 8, 9, 6]. The typical approach in robotics is to use a common representation such as a 3D-model of the environment or a certainty grid in order to integrate the information coming from different sensor modalities. The uncertainty of each sensor modality is modeled explicitly and used e.g. in a Kalman filter based framework.

Mulimodal processing has been addressed for the problem of audio-visual lip reading [32, 3]. The speech signal and co-occurring visual lip signal are combined to increase accuracy of speech recognition. This approach has been effective in situations with high acoustic noise levels.

The most straightforward way to combine information coming from multiple sensors is to use a single feature vector compromising information coming from different sensors (eg. [4]). However, such an approach assumes that the sensor information is synchronized. Another way of combining information is to use Bayesian networks [34] where the uncertainity is propagated from the individual sensors all the way to the semantic interpretation level.

In most of the above approaches the input signals from different sensors are assumed to be synchronized and simultaneous. Also, the method of combining different input channels is coded explicitly and often manually. In this paper we propose to combine audio and visual signals by maximizing mutual information of clustered instances. This allows the system to deal with features which are not synchronized in time, and also enables the combination of arbitrary distance metrics.

3 Representing and Comparing Object Shapes

Three-dimensional objects are represented using a view-based approach [33, 17, 19, 25]. In this approach, no explicit three dimensional model is recovered. Instead, multiple two-dimensional images of an object captured from multiple viewpoints collectively form a model of the object. Figure 2 shows the stages of visual processing which are used to extract representations of object shapes and colors. The shape representation is invariant to similarity transformations i.e. to changes in position, scale and in-plane rotation. In this paper we limit our presentation to shape categories only. The ultimate goal however is to use additional feature representations of the visual information simultaneously such as color, motion and texture.

The video signal from the CCD camera is sampled at a resolution of 160x120 pixels at a rate of 10Hz. The CCD camera has been mounted on a four DOF robotic armature enabling active positioning of the camera. The mechanical platform is shown schematically on the left side of Figure 3. The robot lets us move the camera and view objects from multiple perspectives. A photograph of the robot is shown on the right. Controllable facial

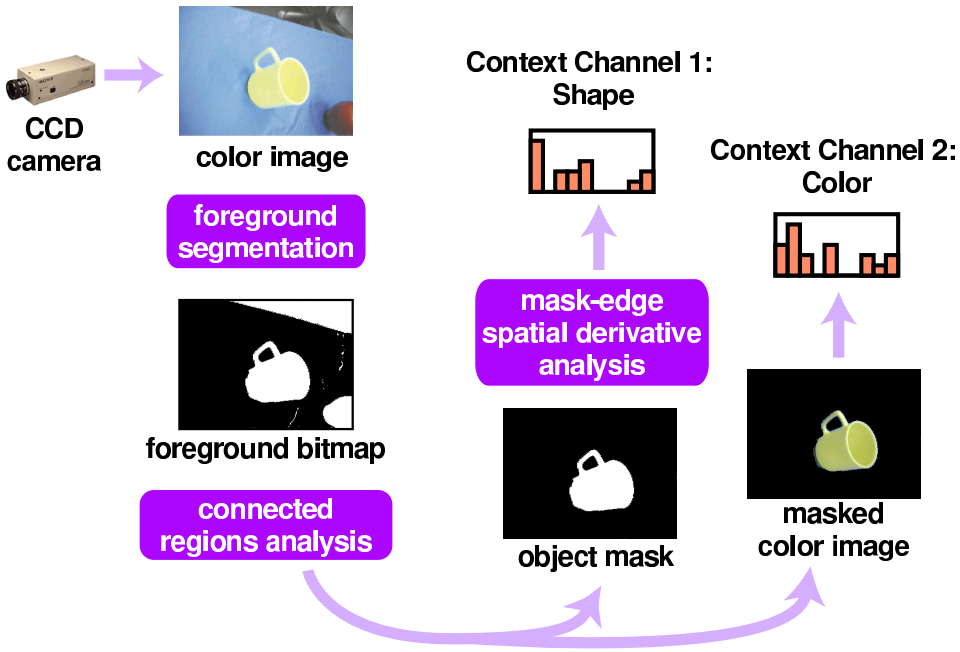

Figure 2. Extraction of object shape and color channels from a CCD camera.

features were added to create a life-like character based on the robot for live interaction with the system (see [22]). Control of the robot is achieved by a tight feedback loop with the vision sensor which ensures that the object is kept in the center of view as the robot changes vantage points.

In the current version of the system figure-ground segmentation is simplified by assuming a uniform background. A simple Gaussian model of the normalized background color is estimated and used to classify background/foreground pixels. Connected regions in the image are candidate objects and referred to as object masks in the following. The middle column of figure 4 shows the largest object masks of the images in the first column.

Schiele and Crowley [25, 28] have proposed a general framework where objects are represented by joint statistics of local properties such as Gaussian derivatives. This provides a robust means to represent objects and allows the reliable recognition of over 100 objects independent of viewing position. Building on their results we propose to represent 2D-shapes of objects based on the joint statistics of two-point features. More specifically, we propose to use 2D-shape histograms which are invariant to similarity transformations (translation, scale and in-plane rotation). In order to obtain a description of the 3D-shape of an object we use sets of 2D-shape histograms.

In this paper we use pairs of points on the border of the object mask in order to calculate the shape histograms. The first feature is the distance between the two points normalized by the average distance of all border point pairs of the object mask. The second feature is the angle between the tangents at the two points. These two features correspond to the two axes of the shape histograms. Both features are invariant to in-plane rotation, scale changes and translation. Since the size of the mask varies with scale the number of entries in the shape histograms have to be normalized to a fixed number of entries. The resulting 2D-shape histograms are invariant to similarity transformations of the object mask.

The third column of figure 4 shows such 2D-shape histograms which have been calculated

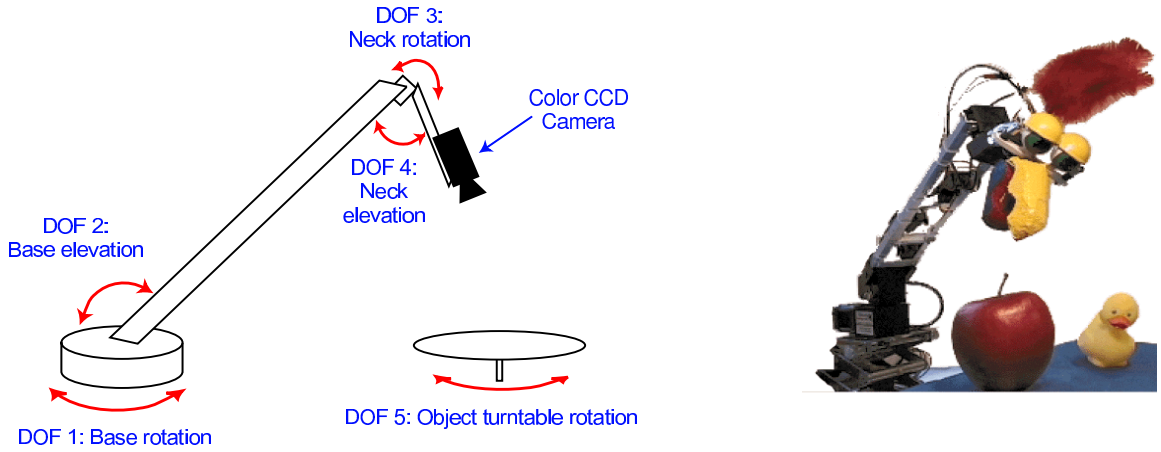

Figure 3. The robot has four degrees of freedom: two at the base and two at the neck. A turntable provides a optional fifth degree of freedom for viewing objects from various perspectives. Servo controlled facial features were added to the robot for life-like real time interaction.

for the object masks of the second column. In particular, the histogram of the basketball (first row) can be interpreted easily: the object mask is a circle and therefore the angle between the tangents increases with increasing distance between the points. More specifically the circular object mask produces a sine curve in the histogram. The sinusoidal curve also appears in the histogram of the shoe since the lower right part of the shoe corresponds to a half-circle. The last histogram we can see that many entries correspond to approximately the same distance but with varying angle which correspond to the body and other parts of the dog.

Using multidimensional histograms to represent object shapes allows the use of information theoretical or statistical divergence functions [1] directly for the comparison of object models. Among these are e.g. the KL-divergence and the χ^2-divergence. We have experimentally compared such divergences to several histogram matching functions [24]. From experiments we found the following calculation of the χ^2–divergence provides the best results in most cases:

$$d_V(X,Y) = \chi^2(X,Y) = \sum_{\mathbf{i}} \frac{(x_\mathbf{i} - y_\mathbf{i})^2}{x_\mathbf{i} + y_\mathbf{i}} \tag{1}$$

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. Using the χ^2-divergence the third and forth histogram in figure 4 are more similar than any other pair of histograms in that figure.

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 is therefore represented by N shape histograms which form a view-set. In order to compare two such view-sets we calculate the M best out of N matches between individual histograms and sum their divergences[1]. By choosing only a subset of views for comparing view-sets, the system does not require all views of one view-set to correspond to a matching view in the other set. In this paper we use $N = 15$ and $M = 4$.

[1]Each individual histogram is only allowed to match another histogram once

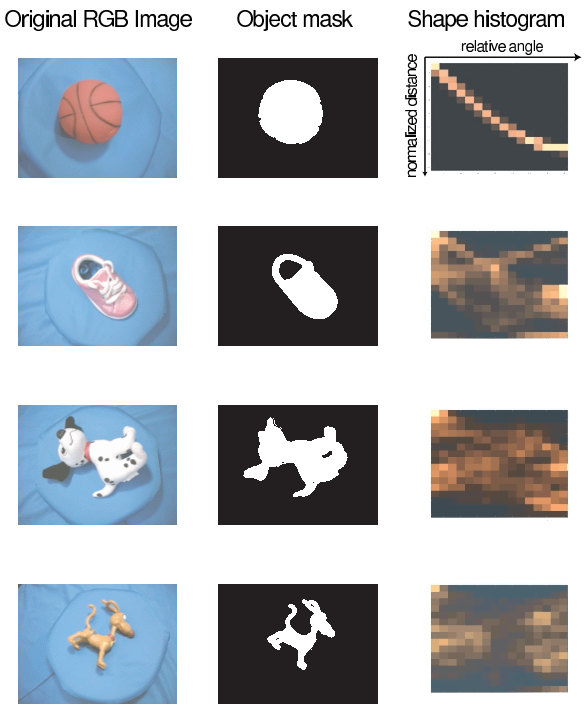

Figure 4. Sample images of objects, their object masks, and shape histograms.

4 Representing and Comparing Spoken Utterances

4.1 Representing Spoken Utterances

Spoken utterances are represented as arrays of phoneme probabilities (Figure 5). We use the Relative Spectra Perceptual Linear Predictive (RASTA-PLP) analysis to extract a spectral representation of the acoustic signal which is robust to background noise which contains temporal dynamics significantly different from speech [13]. A recurrent neural network takes RASTA-PLP coeffients as input and estimates phoneme and speech/silence probabilities. The RNN is similar to the system developed by Robinson [20]. We trained the RNN using the TIMIT database which contains phonetically transcribed speech samples of 630 native English speakers [2] [29]. For each incoming set of RASTA-PLP coefficients, the RNN estimates the probability of 40 English phonemes (including silence).

In our implementation, 12 RASTA-PLP coefficients are computed on a 20ms (320 sample) windows of input. A window step size of 10ms (160 samples) is used, resulting in a set of 12 RASTA-PLP coefficients estimated every 10ms. The RNN re-estimates phoneme probabilities at a rate of 100Hz. An utterance is thus represented as an array of phoneme probabilities. Utterance end points are found by detecting contiguous segments of speech in which the probability of silence estimated by the RNN is low. Figure 6 depicts the RNN output for an utterance produced by a mother directed to her 10-month old infant while playing with a ball. Symbols for each RNN output are printed along the left and right edges of the plot. The strength of each RNN output determines the brightness of the associated trace as a function of time.

[2] The RNN recognizes phonemes with 69.4% accuracy using the standard TIMIT training and test datasets

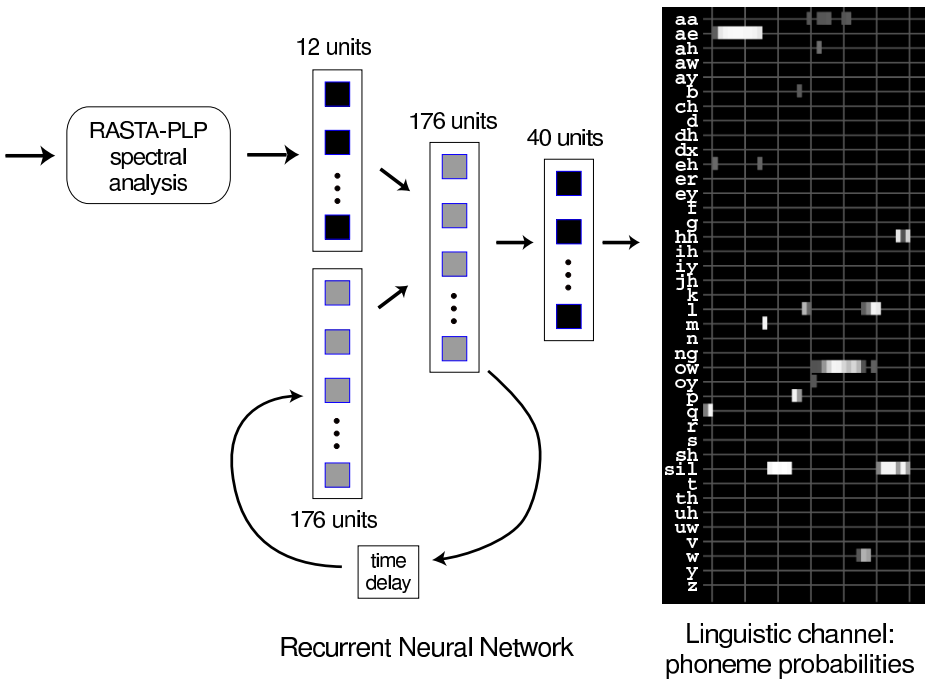

Figure 5. Extracting the linguistic channel from microphone input.

4.2 Speech 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 [2]. The Viterbi dynamic programming search is used to obtain the most likely phoneme sequence for a given phoneme probability array. A context-independent HMM has been trained for each of 40 phonemes using the TIMIT training data set. After Viterbi decoding of an utterance, the system obtains:

- A phoneme sequence. This is the most likely sequence of phonemes which were concatenated to form the utterance.
- The location of each phoneme boundary for the sequence (this information is recovered from the Viterbi search).

Each phoneme boundary may serve as a speech segment start or end point. Any subsequence of an utterance terminated at phoneme boundaries may form a word hypothesis.

4.3 Comparing Speech Segments

We now 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 [15]. This method has been applied to the problem of finding recurrent speech segments in continuous speech [35]. 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

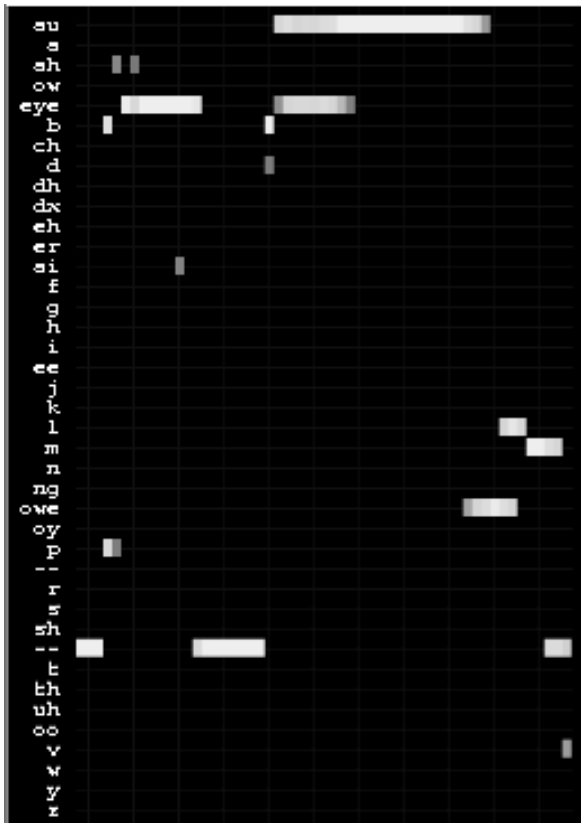

Figure 6. RNN output for the utterance "Bye, ball!".

sequences may be derived. To make use of this additional information, we have devised a new method for comparing speech segments.

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. In Figure 7 the RNN output for the word *ball* leads to the phoneme sequence /bal/. This sequence is used to generate a 3-state left-to-right HMM. 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 [18, page 335] 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 [21, page 318]. In this method, each likelihood estimate is scaled by the likelihood of a default alternate hypothesis, λ^A:

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

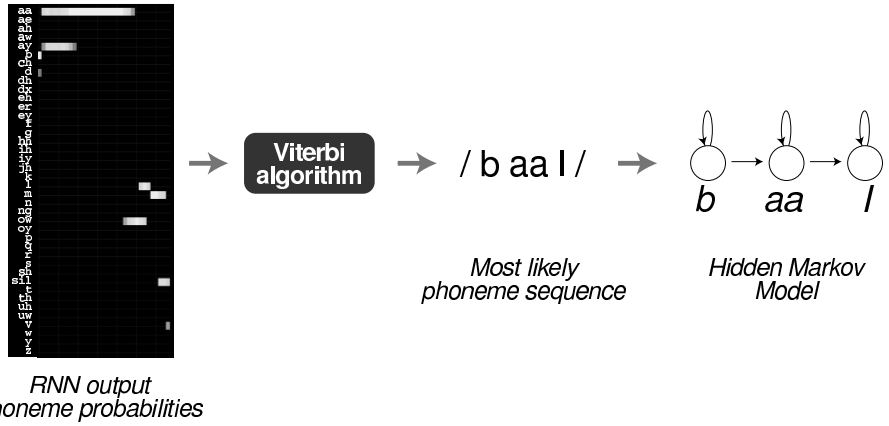

Figure 7. The Viterbi algorithm finds the most likely phoneme sequence for a sequence of RNN output. A left-to-right HMM is constructed by assigning a state for each phoneme in the sequence.

In our metric, 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 distance between two speech segments is defined in terms of logarithms of these scaled likelihoods:

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

This metric is symmetric, i.e. $d_A(\alpha_i, \alpha_j) = d_A(\alpha_j, \alpha_i)$. Logarithms are used to avoid floating point mathematical underflow problems in the implementation. The negative sign converts the likelihood score (which is a measure of similarity) into a score of dissimilarity, i.e. a distance.

5 Combining Audio and Visual Input

5.1 Filtering Hypotheses

In the previous sections, we have described methods for extracting representations of object shapes and spoken utterances from camera and microphone input. As shown in Figure 1, input consists of spoken utterance paired with images of objects. For each utterance, a set of word hypotheses are generated by extracting speech segments at phoneme boundaries and linking them with the co-occurring object.

For a set of object-utterance pairs, the hypothesis set will be huge since each spoken utterance may lead to hundreds or even thousands of word hypotheses. To reduce the number of hypotheses, one or more attentional filters may be used. Some cognitively inspired filters are mentioned below:

Prosodic highlight filter Select speech segments which are prosodically emphasized. Prosodic emphasis may be due to sudden changes in fundamental frequency, increased energy, or lengthened syllables.

Recurrency filter Select speech segments and object shapes which recur often over time.

Visual salience filter Built-in biases may be used to select "interesting" aspect of a visual scene. Visual salience can be based on features such as motion and color.

Object persistence filter Using a head-mounted camera (perhaps in a wearable computer setup), the duration an object stays in the visual field of view may be a good indicator for the importance of object to the user.

Object tracking filter Also in a wearable computing environment, by extracting and tracking objects in the visual field of the user, we can hypothesize objects which the user is following with his or her gaze. Objects which are tracked are likely to be important and more likely to be correlated with concurrent speech signals.

These are a sampling of potential attentional filters. Depending on the problem domain, these and other filters might be used in isolation and in combination. For example, a word-object hypothesis may be selected when the word is prosodically highlighted and the corresponding object is visually salient using color and brightness features.

5.2 Evaluation of Sound-Shape Hypotheses

Output from attentional filters consist of a reduced set of speech segments and their hypothesized visual referents. In the final stage of processing, these hypotheses are clustered, and the most reliable clusters are used to generate output of the system.

Let us assume that there are N sound-shape hypotheses. The clustering process proceeds 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:

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

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

$$\tag{5}$$

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 [5]:

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

The probabilities required to calculate $I(A;V)$ can be estimated from smoothed frequency counts (similar to [12]). 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 maximize 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.

For each output sound-shape pair, the system generates optimized t_V and t_A values. Each selected audio-visual hypothesis serves as a prototype of an associated speech sound and visual category. The thresholds define the radii of the sound and visual categories.

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.

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

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

6 Preliminary Results

This section presents sample results of using the audio-visual learning system to process infant-directed speech and camera images. In doing so, we are able to evaluate the hypothesis that infants might use a clustering algorithm to help learn early words from perceptual input [22].

6.1 Evaluation Study

A study involving six caregivers and their prelinguistic infants was conducted to gather a corpus of infant-directed speech. The participants were asked to engage in play centered around seven types of objects commonly named in early infant speech. The speech was then coupled with sets of images of these objects (taken by the robot) and used as input for the system. In this section, sample results from this task for one of the subjects is presented. Complete findings of the study are currently being prepared and will be presented in [22].

We present the results of a female participant with an 11-month-old infant. She was asked to interact naturally with her infants while playing with a set of age-appropriate objects. Huttenlocher and Smiley [14] identified a list of object classes commonly named in early infant speech. We chose seven classes from the top of their list: balls, toy dogs, shoes, keys, toy horses, toy cars, and toy trucks. A total of 42 objects, six objects for each class, were obtained and are shown in Figure 8. The objects of each class vary in color, size, texture, and shape.

Figure 8. Objects used in the infant-directed speech experiments. Six examples of seven different objects: trucks, keys, dogs, shoes, cars, horses and balls.

6.2 Attention Filtering for Language Acquisition

For this evaluation, we employed a cognitively inspired *recurrency filter*. This filter looks for sound-shape hypotheses which recur in close temporal proximity, assuming the input pairs are provided as an ordered sequence. This filter relies on the assumption that infant-directed speech is highly redundant such that repeated instances of salient words will be found in close proximity [31]. To accomplish this, we have implemented a search procedure which operates on the n most recent input pairs, where n is 7 ± 2. For the experiment reported below, we used $n = 5$.

To decide whether two speech segments or shapes match, thresholds must be set for each distance metric. These thresholds are set relatively low so that many candidates are generated at the expense of increased false hypotheses. Later stages of processing are designed to remove erroneous candidates.

The search looks for pairs of speech segments and shape view-sets which occur at least

twice in the n most recent input pairs. When multiple hypotheses containing matching segments and matching shapes are found, a representative word-shape pair is extracted. The representative is a copy of the "central" member of a set. In the case of only two matching items in a set, one of them is chosen at random. Recurrent shape-name hypotheses generated by the recurrence filter provide input to the MMI clustering process.

6.3 The Data

Over the course of two days, the caregiver played with each of 42 objects for varying periods of time. This resulted in 1655 spoken utterances.

The robot shown in figure 3 was used to create a database of images for the 42 objects. The motivation was to generate a set of images of each object from a first-person perspective. A set of 209 images were captured of each object from varying perspectives resulting in a database of 8,778 images. From each pool of 209 images, we created view-sets of each object by randomly selecting sets of 15 images. View-sets were compared using a match size of 4 views (see Section 3)[3].

6.4 Results

To process data using the system, each spoken utterance was paired with a view-set of the object in play for that utterance. A total of 1655 utterance-view-set pairs were processed by the system. The top 15 lexical items generated by the clustering process are shown in table 1. The first column shows phonetic transcriptions (manually generated) of the original speech segment which lead to the corresponding cluster. Onomatopoetic sounds were not transcribed and are marked "(ono.)". The next column shows text transcripts for each speech segment. For the text transcripts, asterisks were placed at the start and/or end of each entry to indicate the presence of a segmentation error. For example "dog*" indicates that either the /g/ was cutoff, or additional phonemes from the next word were erroneously concatenated with the target word. For each lexical item we also list the associated object based on the visual information. The letters A-F are used to distinguish between the six different objects of each object class. The final two columns indicate the acoustic and visual radius which were computed during the mutual information maximization step.

6.5 Discussion

The results presented in the previous section are promising. Of the top 15 sound-shape pairs, 13 correspond to English words which are linked to appropriate visual semantics. From the table we can also see that only 5 segmentation errors are made; 25 of the end speech segment endpoints correspond precisely to word boundaries in English. Examination of the visual thresholds reveal that system learned to expect small variances (1000 and 4000 units) for balls, but large visual variances in shoes (15000) and dogs (16000). The acoustic thresholds vary indicating different ranges of acoustic variation in pronunciations. For example the word "key" has relatively little acoustic variation between instances, whereas "ball" had much greater variations. These differences in variance are determined automatically by the MMI search.

[3]The sampling was designed to minimize re-use of images in different view-sets. No two images were used more than twice across multiple view-sets. When view-sets were compared, we used the sum of the best 4 views so the effects of one shared image between view-sets is not significant.

Table 1. Output of the audio-visual clustering.

Rank	Phonetic Transcript	Text Transcript	Shape Category	Acoustic Threshold	Visual Threshold
1	ʃu	shoe	shoe E	-0.206597	15000
2	fair ə	fire*	truck D	-0.130583	5000
3	rək	*truck	truck C	-0.123627	6000
4	dɑg	dog	dog D	-0.038675	6000
5	ɪŋəʃ	in the*	shoe A	-0.023423	7000
6	ki	key	key C	0.194981	2000
7	ki	key	key E	0.001978	8000
8	dɑggi	doggie	dog C	-0.049451	16000
9	bɑl	ball	ball C	-0.404322	1000
10	bɑl	ball	ball A	-0.503436	4000
11	kiə	key*	key C	-0.121211	8000
12	ʌʃu	a shoe	shoe B	-0.175509	14000
13	ənðɪsɪz	*and this is	shoe B	-0.092787	6000
14	(ono.)	(engine)	truck A	0.261401	30000
15	(ono.)	(barking)	dog A	0.150821	7000

7 Conclusions and Future Directions

Systems which process multiple input modalities typically rely on the fact that correlated features of the input streams are synchronized in time. However, in many application areas this cannot be assumed. In this paper we proposed an approach based on the maximization of mutual information of audio-visual clusters. This approach results in the simultaneous segmentation of the acoustic stream and creation of corresponding visual categories. The maximization of mutual information effectively combines arbitrary distance metrics in the audio domain and the visual domain. The resulting approach is invariant to scaling of the distance metrics. In this paper we described the application of the approach to model language acqusition of infants.

We will apply the same framework to dynamically personalized human-computer interfaces for various application domains such as catalog browsing. In the context of wearable computers we are currently investigating the use of a head-mounted camera combined with a wearable microphone in order to extract meaningful associations of the audio-visual environment. The output of the current system may be analyzed to learn higher level structures which encode relations between low level audio-visual associations. These higher level structures may encode abstract knowledge which is not directly tied to sensory input.

References

[1] M. Basseville. Information: entropies, divergences et moyennes. Technical Report 1020, IRISA, Mai 1996. in French.

[2] H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach.* Kluwer Academic Publishers, 1994.

[3] C. Bregler, H. Hild S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1993.

[4] B.P. Clarkson and A. Pentland. Unsupervised clustering of ambulatory audio and video. In *Proceedings of the International Conference of Acoustics, Speech and Signal Processing*, 1999.

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

[6] I.J. Cox and J.J. Leonard. Probabilistic data association for dynamic world modelling: A multi-hypothesis approach. In *Proc. International Conference on Advanced Robotics*, Pisa, 1991.

[7] Carl de Marcken. *Unsupervised Language Acquisition.* PhD thesis, MIT Artificial Intelligence Laboratory, 1996.

[8] H.F. Durrant-Whyte. Sensor models and multisensor integration. *International Journal on Robotics Research*, 7(6):97–113, 1988.

[9] H.F. Durrant-Whyte and J. Manyika. *Data fusion and sensor management: A decentralized information theoretic approach.* Prentice Hall, 1994.

[10] 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.

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

[12] A.L. Gorin, S.E. Levinson, A.N. Gertner, and E. Goldman. Adaptive acquisition of language. *Computer Speech and Language*, 5:101–132, 1991.

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

[14] J. Huttenlocher and P. Smiley. Early word meanings: the case of object names. In Paul Bloom, editor, *Language acquisition: core readings*, pages 222–247. MIT Press, Cambridge, MA, 1994.

[15] J.B. Kruskal and D. Sankoff. An anthology of algorithms and concepts for sequence comparison. In D. Sankoff and J.B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules.* Addison-Wesley, 1983.

[16] H. Moravec. Sensor fusion in certainty grids for mobile robots. *AI Mag*, 9(2):61–74, 1988.

[17] H. Murase and S.K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[18] L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[19] R.P.N. Rao and D.H. Ballard. Object indexing using an iconic sparse distributed memory. In *International Conference of Computer Vision*, pages 24–31, 1995.

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

[21] R. Rose. Word spotting from continuous speech utterances. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, chapter 13, pages 303–329. Kluwer Academic, 1996.

[22] D.K. Roy. *Learning words from sights and sounds: A computational model.* PhD thesis, MIT Media Laboratory, 1999.

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

[24] B. Schiele. *Object Recognition using Multidimensional Receptive Field Histograms.* PhD thesis, I.N.P.Grenoble, July 1997. English translation.

[25] 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.

[26] B. Schiele, N. Oliver, T. Jebara, and A. Pentland. An interactive computer vision system, dypers: Dynamic personal enhanced remembrance system. In *International Conference on Vision Systems*, pages 51–65, Jan 1999.

[27] B. Schiele and A. Pentland. Attentional objects for visual context understanding. Technical Report 500, Vision and Modeling, MIT Media Laboratory, 1999.

[28] B. Schiele and A. Pentland. Probabilistic object recognition and localization. In *ICCV'99 International Conference on Computer Vision*, 1999.

[29] S. Seneff and V. Zue. Transcription and alignment of the timit database. In *Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language*, Oahu, Hawaii, November 1988.

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

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

[32] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. In *Proceedings of the 1992 International Joint Conference on Neural Networks*, 1992.

[33] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[34] S. Wachsmuth, H. Brandt-Pook, G. Socher, F. Kummert, and G. Sagerer. Multilevel integration of vision and speech understanding using bayesian networks. In *International Conference on Vision Systems*, pages 231–254, 1999.

[35] 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.