Article In Press, Computer Speech and Language

Towards Situated Speech Understanding: Visual Context Priming of Language Models

Deb Roy and Niloy Mukherjee

Cognitive Machines Group
The Media Laboratory
Massachusetts Institute of Technology

Abstract

Fuse is a situated spoken language understanding system that uses visual context to steer the interpretation of speech. Given a visual scene and a spoken description, the system finds the object in the scene that best fits the meaning of the description. To solve this task, Fuse performs speech recognition and visually-grounded language understanding. Rather than treat these two problems separately, knowledge of the visual semantics of language and the specific contents of the visual scene are fused into speech processing. As a result, the system anticipates various ways a person might describe any object in the scene, and uses these predictions to bias the speech recognizer towards likely sequences of words. A dynamic visual attention mechanism is used to focus processing on likely objects within the scene as spoken utterances are processed. Visual attention and language prediction reinforce one another and converge on interpretations of incoming speech signals which are most consistent with visual context. In evaluations, the introduction of visual context into the speech recognition process results in significantly improved speech recognition and understanding accuracy. The underlying principles of this model may be applied to a wide range of speech understanding problems including mobile and assistive technologies in which contextual information can be sensed and semantically interpreted to bias processing.

1. Introduction

Modularity is a central principal in the design of complex engineered systems, and is often postulated in theories of human cognition [7, 9]. Modules operate as encapsulated "black boxes" that can only access other modules through well-defined interfaces. Access to internal data structures and processing across modules is usually restricted. Studies of human behavior, however, sometimes reveal surprising breaches of modularity. For example, recent psycholinguistic

experiments have shown that acoustic and syntactic aspects of online spoken language comprehension are influenced by visual context. During interpretation of speech, partially heard utterances have been shown to incrementally steer the hearer's visual attention [25], and conversely, visual context has been shown to steer speech processing [27, 26]. Motivated by these findings, we have developed a spoken language understanding system in which visual context primes early stages of speech processing, resulting in significantly improved speech recognition and understanding accuracy.

The development of robots provides an exemplary problem that suggests modular design. In practically all robots, the perceptual, planning, motor control, and speech systems (if any) operate independently and are integrated through relatively high level interfaces. In this paper, we consider the design of a speech understanding system that will eventually provide speech processing capabilities for an interactive conversational robot [15, 24]. A straight forward approach would be to take an off-the-shelf speech recognition system and connect its output to other modules of the robot. We argue, however, that by treating the speech recognizer as a black box that is unaware of the contents of other modules, valuable contextual information is lost. Since high accuracy speech recognition in natural conditions remains unattainable, leveraging information from other channels can be of immense value in improving performance.

We have addressed the problem of understanding spoken utterances that make reference to objects in a scene. We make the simplification that the system can assume that all utterances contain references to objects in the immediate environment. Clearly, this assumption is not always valid since people often talk about things that are not in the here-and-now, and not all speech acts are descriptive. Thus, our current work represents one component of a larger effort which will eventually incorporate speech act classification to determine when visual context should be used to constrain the analysis of utterances.

Based on our assumption of immediate reference, knowledge of the visual environment is used by the system to anticipate words and phrases that the speaker is likely to choose. A challenge in this approach is that there are typically numerous potential referents in environments of even moderate complexity. Since the system does not know, a priori, which referent the speaker intends to describe, the system must anticipate descriptions of all potential referents. In most situations, many choices of words might fit the same referent. Furthermore, since the contents of the scene are determined by visual analysis, scene information is bound to be noisy and of variable reliability.

Our approach is to jointly infer the most likely words in the utterance along with the identity of the intended visual referent in a unified multimodal stochastic decoding framework. This approach has been implemented in an on-line, real-time multimodal processing system. Visual scene analysis reaches into the core of the speech recognition search algorithm and steers search paths towards more likely word sequences. The semantic content of partially decoded spoken utterances, in complement, feed back to the visual system and drive a dynamic model

of visual attention. As processing proceeds, linguistic and visual information mutually reinforce each other, sharpening both linguistic and visual hypotheses as sensory evidence accumulates. We show that the integration of visual context leads to substantial improvements in speech recognition and understanding accuracy. We believe that the strategic introduction of cross-module bridges may be an important design principle in a wide range of applications beyond the specific system presented.

After providing some background remarks, we introduce the task we used for our experiments. Section 3 provides an overview of our approach. Subsequent sections provide details on aspects of this approach, followed by experimental evaluations.

2. Background

Integration of spoken and visual input has been investigated in a wide range of domains. It is useful to distinguish two broad classes of tasks. Let S and V denote the speech and visual input signals, respectively. The speech signal's primary role is to encode sequences of words. Prosodic aspects of speech also encode affective, syntactic, and stress information. All information in S convey the speaker's intent. In contrast, V may carry two distinct kinds of information, depending on the task. Consider first the problem of audiovisual lipreading. In this task, visual input typically consists of images of the speaker's lips as they speak. In this case, the kind of information carried in V is the same as S. The visual channel provides complimentary or redundant aspects of the surface form of words. This complementarity of encodings of word surface forms can be leveraged to increase speech recognition accuracy. For lipreading, we can say that $V = V_i$, where i reminds us that the purpose of the visual channel is to *indicate intentions*. Lip motions are part of the speaker's way of conveying his/her intentions. A related problem that has received significant attention is the integration of speech with visually observed gestures made by pen or mouse [14, 11]. For example, Johnston and Bangalore [11] developed a speech and gesture understanding system in which a finite state automaton jointly processes speech (S) and gesture (V_i) signals to produce a semantic interpretation of multimodal input. Although gestures are very different in nature from the motion of lips, broadly speaking, both belong to the same class of V_i since gestures also play the role of indicating the speaker's intentions.

In contrast, consider the problem of building a speech understanding system for a robot in which the visual input comes from a camera mounted on the robot, looking out into the robot's environment. The speaker asks the robot to pick up a red block. The visual channel might capture the speaker, complete with lip movements and other body gestures. However, the visual signal will also contain information about the robot's *situational context*, which in this case may include a red block. We indicate this kind of visual information by saying $V = V_i + V_c$ where V_c denotes contextual information captured in the visual signal. If the

speaker is not in view, then $V = V_c$. The contents of V_c are fundamentally different from V_i since S may be *about* aspects of V_c but not, in general, V_i*.

The focus of this paper is for a task in which $V = V_c$, i.e, the visual input contains purely contextual information. In contrast to lipreading and gesture understanding problems, we will instead investigate the semantic referential content of the visual signal and how it can be integrated with S in useful ways for a real-time multimodal understanding system.

Most previous work on integrating visual context (V_c) with speech / language understanding have all used modular, late integration across modalities. SAM (Speech Activated Manipulator) [2] is a robotic system with sensory capabilities that interacts with a human conversation partner through spoken language dialog. Speech recognition and visual analysis are integrated at a relatively late stage through an augmented transition network that operates on a frame-based knowledge representation. Crangle and Suppes [4] have proposed an approach to verbal interaction with an instructable robot based on a unification grammar formalism. They have examined the use of explicit verbal instructions to teach robots new procedures and have studied ways a robot could learn from corrective user commands containing qualitative spatial expressions. Although speech may provide linguistic input to their framework, there is no mechanism for propagating semantic information to the speech recognizer due to the modular design of their model. Wachsmuth and Sagerer (2002) presents a probabilistic decoding scheme that takes the speech signal and an image or image sequence as input. The speech signal is decoded independent of the decoding of the image data. A Bayesian network integrates speech and image representations to generate a representation of the speaker's intention. In summary, all of these systems integrate spoken language with visual context, but the conversion of speech to text occurs in a contextual vacuum.

In our own previous work, we developed a trainable spoken language understanding system that selects individual objects on a table top in response to referring spoken language expressions [23]. The system combines speech recognition output and image representations generated by a visual analysis module to point to objects that best fit spoken descriptions. Similar to the other work cited above, speech and visual processing occurred independently. In contrast, through the development of Fuse we have explored tight integration of visual context into speech processing.

3. Overview

To study the role of visual context in spoken language comprehension, we chose a constrained scene description task. Participants in a data collection study were asked to verbally describe objects in scenes consisting of oversized Lego

*One can imagine rare exceptions to this. A person, while waving their arm in some manner, might say, "It hurts when I do *this*", where, "this" refers to the gesture. As a first approximation, we can ignore such cases and treat V_i and V_c as distinct.

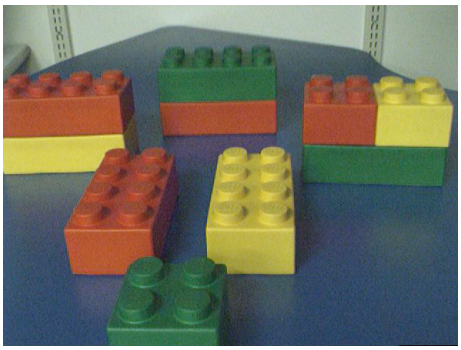

Figure 1: A typical visual scene in the current experimental task.

blocks (Figure 1). No restrictions were placed on the vocabulary, style, or length of description. Typical descriptions ranged from simple phrases such as, "The green one in front" to more complex utterances such as, "The large green block beneath the smaller red and yellow ones". The result of this data collection was a set of images paired with spoken descriptions of objects in the images.

3.1. Language Modeling

Speech recognition is most commonly formulated in a maximum likelihood framework [1]. Given an observed spoken utterance, X, we wish to choose a word string $\widehat{W}$ such that

$$\widehat{W} = \underset{W}{\text{argmax}} \; P(X|W)P(W) \tag{1}$$

The terms $P(X|W)$ and $P(W)$ correspond to an acoustic model and language model, respectively. In conventional speech recognition systems, the acoustic model captures the acoustic properties of speech and provides the probability of a speech observation given hypothesized word sequences. In audio-visual speech recognition systems, speech observations include both acoustic and visual information.

The language model, $P(W)$, provides probabilities of word strings W based on context. In practically all speech recognition systems, this context is a function of the history of words that the speaker has uttered. In contrast, our approach is to dynamically modify $P(W)$ on the basis of visual context (V_c).

Since our focus will be on dynamic language models, we provide a brief review of n-gram statistical language models which will serve as a basis for our cross-modal extension. The n-gram model assigns probabilities to hypothesized word sequences. The probability of a word sequence $W = w_1, w_2, \ldots, w_k$ which we denote as w_1^k, can be expressed as a product of conditional probabilities:

$$P(w_1^k) = P(w_1)P(w_2|w_1)\cdots P(w_k|w_1^{k-1}) \tag{2}$$

Within the term $P(w_k|w_1^{k-1})$, w_1^{k-1} is called the history and w_k the prediction.

In the n-gram approach, two histories are treated as identical when they end in the same $n-1$ words. For example, with $n=2$, we obtain a bigram language model:

$$P(w_1^k) = P(w_1)P(w_2|w_1)\cdots P(w_k|w_{k-1}) \quad (3)$$

Many extensions to basic n-gram language models have been proposed such as variable length histories [13] and long distance dependencies [10, 21] (for a review of these and other methods, see [20]). Our goal is to introduce a form of visually-driven semantic priming into the statistical language model of a real-time speech recognizer. In principal, other n-gram extensions such as those mentioned above can be augmented with visual context in the way that we propose. For simplicity, we have chosen to work with the bigram language model which has sufficient modeling power for the present scene description task.

The parameters of a bigram model are usually estimated from a large text corpus. Given a training corpus of size T words in which word w occurs $|w|$ times, the maximum likelihood estimate of $P(w)$ is $|w|/T$. The maximum likelihood estimates for the conditional terms $P(w_i|w_{i-1})$ are given by $|w_{i-1}, w_i|/|w_i|$ where $|w_{i-1}, w_i|$ is the number of times the sequence w_{i-1}, w_i occurs in the training corpus. Some form of smoothing is necessary since the vast majority of n-grams rarely occur (for an overview of smoothing small sample counts, see [12]).

Words may be clustered into equivalence classes leading to *n-gram class models* [3]. For example, if the distribution of words in the neighborhood of *Monday* and *Tuesday* are believed to be similar, the words can be clustered, and treated as equivalent for language modeling. The principal benefit of creating word classes is that we are able to make better use of limited training data to make predictions for word histories that are not encountered in training. We can partition a vocabulary into word classes using a function which maps each word w_i to its corresponding class $c(w_i)$. For bigram class models,

$$P(w_i|w_{i-1}) = P(w_i|c(w_i))P(c_i|c_{i-1}) \quad (4)$$

Standard word bigrams are a special case of bigram class models in which each word is mapped to a unique word class.

3.2. Visual-Context Sensitive Language Models

Figure 2 illustrates our approach to integrating visual context with speech processing in a model called Fuse. Input consists of a speech signal paired with an image. Figure 1 is representative of images in the current task, captured by a color video camera. The speech signal is recorded from a head-worn microphone. The spoken utterances used for evaluations consisted of naturally spoken, fluent speech.

The visual scene analysis module detects objects in the scene and extracts a set of visual features that represent individual objects, and intra-object spatial

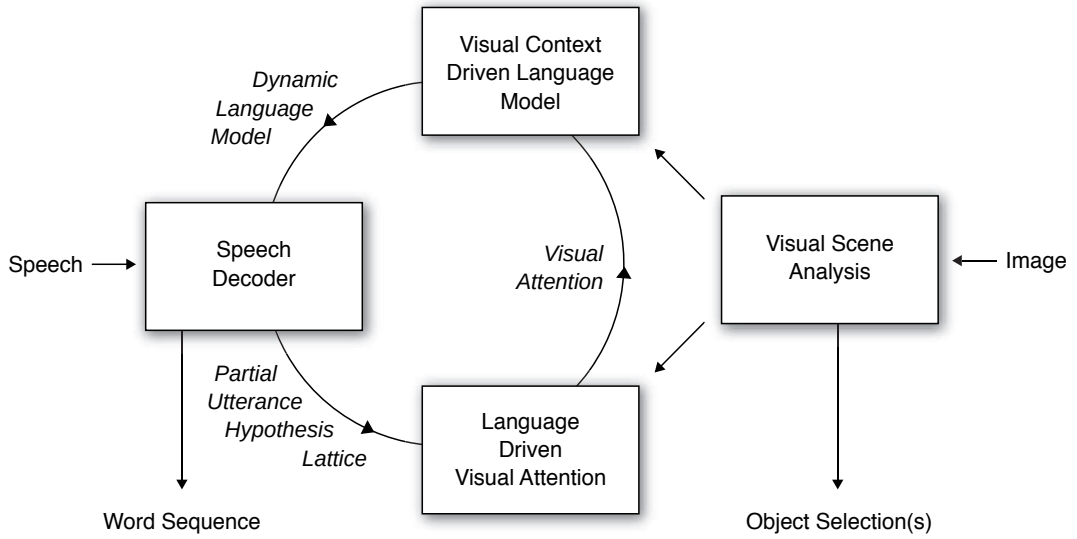

Figure 2: Overview of the Fuse architecture.

relations. The results of the scene analysis are accessible by two modules: a language model, and a visual attention model. As the speech signal is processed, both the language and attention models are dynamically updated.

To understand the main processing loop in Figure 2 and the role of the language model and visual attention model, we will work through a simple example. Let us consider a situation in which a speaker says, "The red block on the left" in the context of a scene containing four blocks: a red one and a blue one on the left, and a red one and blue one on the right. As the first portion of the input utterance is processed, let us assume that the speech recognizer correctly recovers the first two words of the utterance, "the red". In actuality, in Fuse, the output of the speech recognizer is a lattice that encodes multiple word hypotheses, but to keep the example simple, we first consider a single word sequence.

The partially decoded word sequence is fed to the visual attention module which also receives the output of the visual scene analyzer. Visual attention is modeled as a probability mass function (pmf) over the set of objects in the scene. Initially, before speech recognition begins, the pmf is non-informative and assigns equal probability to all objects in the scene. When the words "the red" are fed into the visual attention module, the pmf is updated so that most of the probability mass is shifted to the red objects in the scene. In effect, the visual attention of the system shifts to the red objects. The attention module uses a set of visually-grounded semantic models to convert the word sequence into the pmf (Section 6).

The visual attention pmf, which now favors the two red objects in the scene, is transmitted to the language model. The language model may be thought of as a linguistic description generator. For each object in the scene, the model generates a set of referring expressions that a person might use to describe the object.

For the red block on the left, the model might generate a set of descriptions including "the red block", "the large red block", the "the red block on the left", and so forth. Each description is assigned a likelihood that depends on how well the description matches the visual attributes of the object, and also based on syntactic and contextual measures of fitness. The likelihoods of the descriptions for each object are scaled by the probability assigned to that object by the visual attention pmf. The resulting mixture of descriptions is summarized as a bigram language model which is used by the speech recognizer. In effect, visual attention steers the speech recognizer to interpret the input speech signal as a description of objects that have captured more of the system's attention.

As acoustic evidence is incrementally processed, the visual attention pmf evolves. The dynamic pmf in turn biases the language model of the speech recognizer. As more of the utterance is processed, the visual attention becomes progressively sharpened towards potential referents in the scene.

Several details have been simplified in this overview. One complication is introduced with utterances containing relative spatial clauses such as, "The red block to the left of the large blue one". In this class of utterances, visual attention must be refocused mid-way through processing from potential target objects (red blocks) to potential *landmark* objects (large blue blocks to the right of the potential targets). Another complication arises from the fact that the output of the speech recognizer at any moment is not a single word sequence, but rather a lattice that encodes multiple (potentially thousands) of alternative word hypotheses. These and other aspects of Fuse are explained in the following sections which provide detailed descriptions of each component of the system.

4. Visual Scene Analysis

The visual scene analysis module segments objects in an input scene and computes visual properties of individual objects, and spatial relations between pairs of objects. The resulting representation of the scene is used by both the language model and visual attention model.

Objects are segmented based on color. A statistical color model is created for objects by training Gaussian mixture models on sample images of the objects. We assume that objects will be single-colored, greatly simplifying the segmentation process. The Expectation Maximization (EM) algorithm is used to estimate both the mixture weights and the underlying Gaussian parameters for each color model. The color models are used as a Bayes classifier to label each 5x5 pixel region of an input image. Regions of the image that do not match any object color model are classified as background using a fixed threshold. Objects are found by extracting connected foreground regions of consistent color.

A set of visual properties are computed for each object found in the segmentation step, and for spatial relations between each pair of objects. These properties and relations constitute the complete representation of a visual scene. The fea-

tures attempt to capture aspects of the scene that are likely to be referred to in natural spoken descriptions. The following visual features are extracted:

- **Color** is represented by the mean RGB value of the 10x10 pixel region in the center of the object.
- **Shape** is represented by five geometric features computed on the bounding box of each object: height, width, height-to-width ratio, ratio of the larger to the smaller dimension (height / width), and bounding box area.
- **Position** is represented by the horizontal and vertical position of the of center of the region.
- **Spatial relations** are encoded by a set of three spatial features suggested in [18] that are measured between pairs of objects. The first feature is the angle (relative to the horizon) of the line connecting the centers of area of an object pair. The second feature is the shortest distance between the edges of the objects. The third feature measures the angle (relative to the horizon) of the line which connects the two most proximal points of the objects.

To summarize, each object is represented by a ten-dimensional feature vector (3 color features, 5 shape, and 2 position). The spatial relation between each pair of objects is represented by 3 additional spatial features. In real time operation, the visual analysis system captures and processes video frames at a rate of 15Hz. When Fuse detects the onset of a spoken utterances, the visual frame co-occurring with the start of the utterance is captured, and the resulting visual features are used to provide context for processing of the entire spoken utterance. Changes made to the scene once the utterance has begun are ignored.

5. Speech Decoding

The role of the speech decoder is to find word sequences that best explain acoustic input. The decoding strategy and algorithms are based on standard methods. Speech is represented using a 24-band Mel-scaled cepstral acoustic representation [17]. Words are modeled by concatenating context sensitive phoneme (triphone) models based on continuous-density three-state, Hidden Markov Models [16]. Speech decoding is accomplished using a time-synchronous Viterbi beam search [16]. The decoder has been tested on standard speech recognition test corpora and performs competitively with other research platforms, and thus serves as a useful baseline for the experiments presented here [28].

6. Visual Context Driven Language Model

The language model is designed to "second guess" what the speaker is likely to say, assuming he/she will speak a description of an object in the current visual scene. If the language model is able to accurately anticipate the speaker's words,

the model can bias the speech decoder towards more likely interpretations of the incoming speech signal. There are several sources of uncertainty in predicting how a person will describe objects in the scene:

1. The identity of the target item is unknown, so the language model must consider descriptions that fit all objects in the scene.
2. People may use different words to refer to the same attributes. For example, one person might call an object blue, while another speaker will call it purple.
3. Speakers may use different combinations of words to refer to the same object. "The blue one", the "the tall block", and "the cube to the left of the red one" may all refer to the same referent.

To address these sources of uncertainty, multiple descriptions are generated, in turn, for each object in the current scene to account for variations due to factors (2) and (3). The potentially large set of resulting descriptions are then weighted and combined to create a bigram language model that is used by the speech decoder. Although the descriptions stay fixed during the processing of an utterance, the relative weighting of individual descriptions is dynamically updated using the visual attention model that is described in Section 7. As a result, the bigram language model is not only influenced by visual context as recorded at the onset of the utterance, but further evolves online as the utterance is processed.

The method for generating descriptions is adapted from the trainable object description system called Describer that was reported in [22]. In this work, we developed learning algorithms that take as input synthetic visual scenes paired with natural language descriptions of objects. The output of the system consists of a set of visually grounded word models that are grouped into word classes, and a set of class bigrams that model transitions between word classes. Word classes are formed on the basis of both visual (semantic) and syntactic properties of words. Each word is associated with an acquired visual model that consists of a multidimensional Gaussian distribution defined over a subset of the 10 visual features described in Section 4. The learning algorithm automatically associates visual features with word classes. Complete details of the learning algorithm are provided in [22].

All parameters of the description model are learned from examples of objects embedded in scenes that are labeled with descriptive phrases. For our experiments with Fuse, a set of 60 training examples were collected from eight participants, resulting in a total of 480 examples in the training dataset. Since the training methods have been previously described [22], here we describe the data structures created by that learning algorithm and then show how the structures are used to generate descriptions.

Figure 3 shows the visual models associated with the members of an acquired word class. The learning algorithm decided to cluster these four words, and to

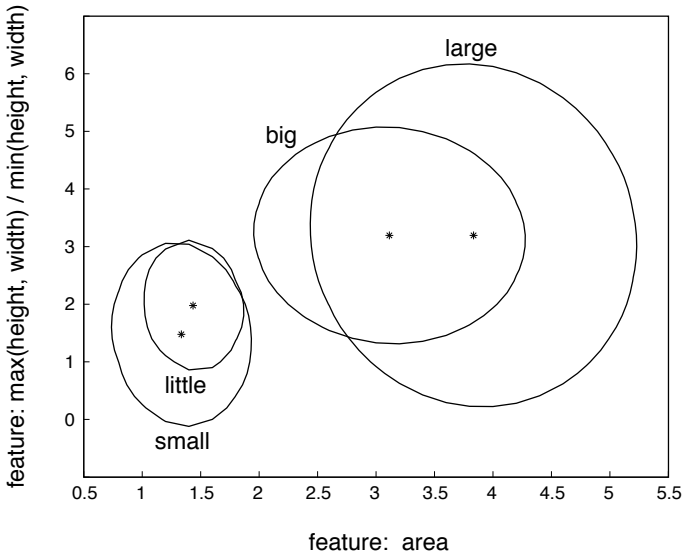

Figure 3: Example of a word class with four members. Each ellipse indicates an equal probability density contour associated with a visual model (a full-covariance Gaussian distribution), centered on its mean which is indicated with a small asterisk. An automatic feature selection algorithm determined the two visual features used for defining this set of four words.

ground them in terms of the two visual features (from a choice of 10). Two geometric features (area, and ratio of dimensions) have been selected as the salient visual attributes for this cluster of words. The overlapping distributions show the relation between the words *big* and *large*, and their antonyms *little* and *small*. As we shall see, word classes and their associated visual models are used as Bayes classifiers in order to generate labels for novel objects.

Word order is modeled through bigrams that specify transition probabilities between words and word classes. Figure 4 shows a subset of phrase level bigrams in the form of a transition network. Each arc is labeled with the transition probability between pairs of words / word classes (bigram transitions with probability less than 0.10 have been pruned for readability). Word classes with single members are labeled with that word. The six classes with multiple visually grounded words are listed in Table 1. Many words that occur in the training corpus such as *the* and *and* appear in the grammar but are not visually grounded. As we explain below, those words play a role in predicting words during speech recognition, but do not effect semantic analysis. Any path through the network in Figure 4 constitutes a possible description of an object. For instance, *the red block* and *the leftmost large one* are word sequences that may be generated by this network. A higher-order phrase network (Figure 5) models relative spatial phrases. The automatic acquisition of higher level grammars is described in [22]. The phrase nodes in this network, marked "TARGET OBJECT" and "LANDMARK OBJECT", each embed a copy of the phrase network and are connected

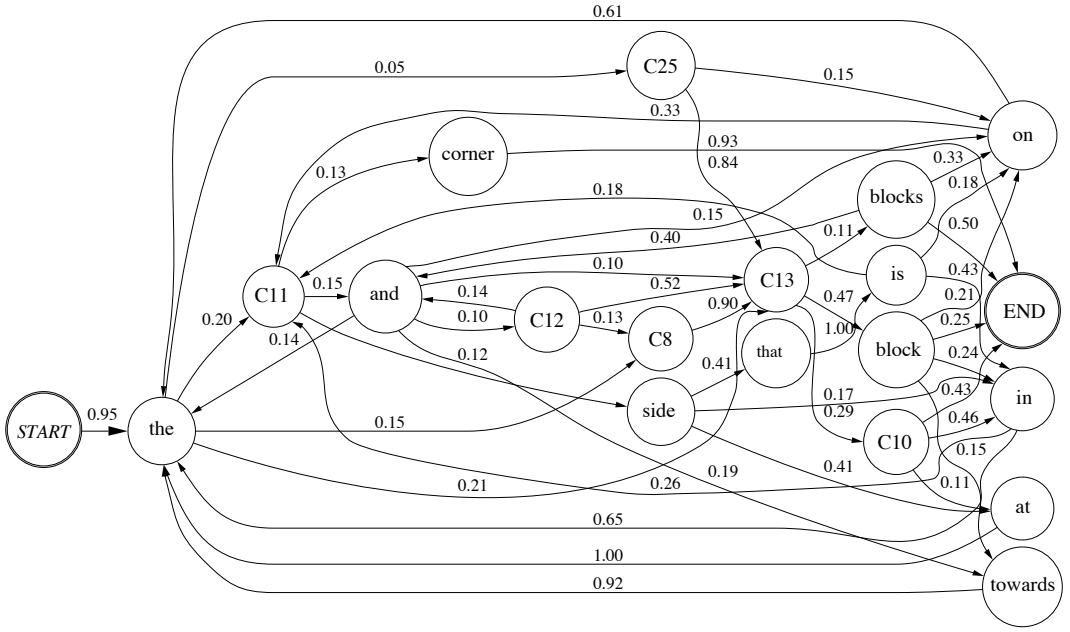

Figure 4: The probabilistic grammar used to generate descriptions of objects. Nodes include individual ungrounded words and grounded word classes. To allow legibility, the full grammar used in experiments has been pruned for the figure (18 of 55 nodes are shown).

by relative spatial terms. Spatial terms are grounded (defined) by Gaussian distributions over the three spatial relation features described in Section 4. This phrase network can generate sequences such as *the large green block beneath the red one.*

Word Class	Members
C8	large, big, small, little
C10	rectangle, square
C11	front, back, left, right, top, bottom, rear, upper
C12	frontmost, topmost, bottommost, leftmost, rightmost, centermost
C13	red, blue, yellow, green
C25	horizontal, vertical

Table 1: Visually-grounded words that are grouped into word classes in the bigram network in Figure 4. Each word class is assigned a set of visual features, and the semantics of each word is grounded in a Gaussian probability distribution over the set of features assigned to its class.

6.1. Mixtures of Descriptions for Language Modeling

The speech recognizer requires a language model consisting of a set of word bigram transition probabilities. As Equation 4 shows, the word bigram can be obtained from the product of word class transition probabilities $P(c_i|c_{i-1})$ and

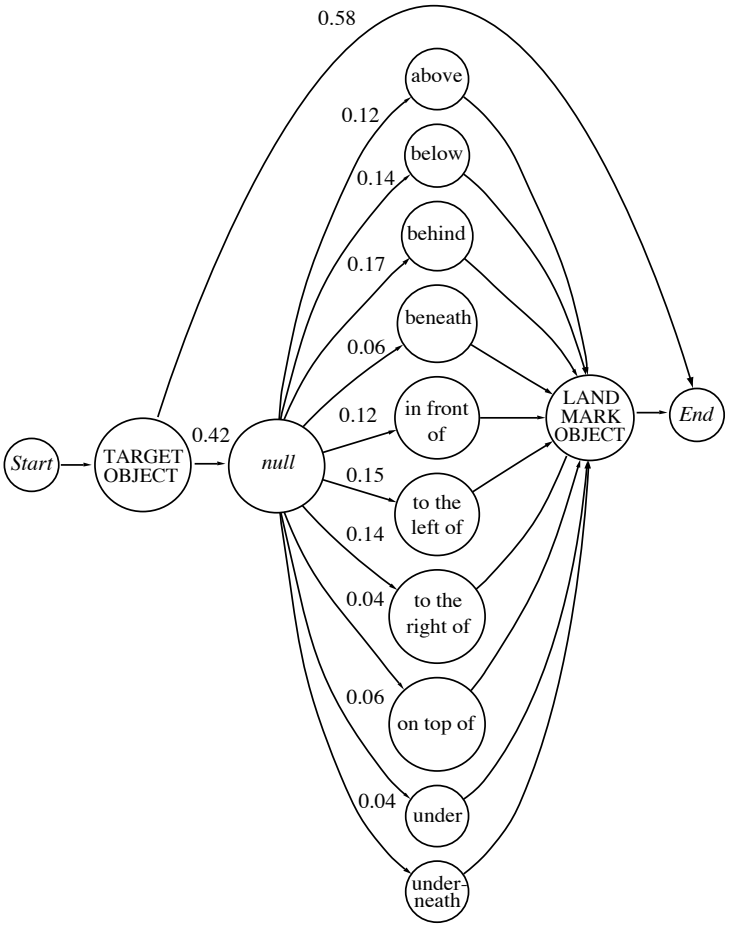

Figure 5: The probabilistic grammar used to generate descriptions with relative spatial clauses.

class conditional word probabilities $P(w_i|c_i)$. The word class transition probabilities are fully determined from training data (Figure 4) and remain static during speech processing. Thus, the expected order of word classes, and transition probabilities between classes is not expected to change as a function of visual context since these capture syntactic regularities of the language. The probabilities of words *within* each word class, on the other hand, do depend on context. As a simple example, if there are no blue objects in the scene, the probability for the word *blue* should be reduced relative to other words in its class given our assumption that the utterance refers to some object in the scene. To capture this intuition, class conditional word probabilities are dynamically estimated as a function of the scene and the pmf model of visual attention using a five-step process:

1. *Enumerate all left-to-right paths through the object description grammar*
   All distinct paths connecting the *start* and *end* nodes of the transition

network (Figure 4) are enumerated. Loops are avoided, resulting in only left-to-right paths. This process leads to a set of N sequences, $\{C_1, C_2, \ldots, C_N\}$. Each sequence C_i consist of a ordered set of T_i word classes:

$$C_i = c_i^1, c_i^2, \ldots, c_i^{T_i} \tag{5}$$

These sequences constitute the set of syntactic frames embedded in the transition network.

2. *Map word classes to words*

   Each word in a class may be grounded in a visual model (Gaussian distribution). The models associated with the words of each class are used as a Bayes classifier [6] to classify objects based on their measured visual attributes. For example, consider the word class shown in Figure 3. To use this word class as a Bayes classifier to label an object, the two features of the object associated with visual models must be measured. Each of the visual models of this class are then evaluated at the measured values, and the model with the highest value (probability density) is selected as the best match to the object. The word associated with that model is thus the best choice within the word class for describing the object. The mapping from word class to word is thus object dependent; different words may become most activated within a class depending on the visual properties of the object. We denote the word sequence generated by using the word class sequence C_i to describe object O_j as:

$$W_i^j = w_{ij}^1, w_{ij}^2, \ldots, w_{ij}^{T_i} \tag{6}$$

   For a scene with M objects, this mapping process results in $N \times M$ word sequences (N descriptions for each of M objects).

3. *Compute the descriptive fitness of each description*

   Each description can be evaluated on how well it visually matches its target object by computing the product of the word conditional probabilities of the observed object properties, which is equivalently expressed as a sum of log probabilities:

$$fit(W_i^j, O_j) = \frac{\Sigma_{t=1}^{T_i} \log p(O_j|w_{ij}^t)}{G(C_i)} \tag{7}$$

   Where $G(C_i)$ is the number of visually grounded word classes in the sequence C_i (i.e., one of the classes listed in Table 6). The denominator term normalizes effects due to the length of the description. $p(O_j|w_{ij}^t))$ evaluates the visual model associated with word w_{ij}^t for the visual features of object O_j . For ungrounded words, $p(O_j|w_{ij}^t))$ is set to 1.0.

   This fitness function measures how well a descriptive phrase matches the

properties of the target object, but it does not account for contextual effects due to other objects in the scene. A description that matches the target well may also describe a non-target equally well. To capture contextual effects, we define a context-sensitive fitness:

$$\psi(C_i, O_j) = fit(W_i^j, O_j) - max_{k \neq j} fit(W_i^j, O_k) \tag{8}$$

This measure subtracts the fitness of the competing object in the scene that best fits the description intended for the target and tends to favor contextually-unambiguous descriptions.

4. *Compute object-conditional word predictions*
   For a given object and word class sequence, object-conditional probabilities are assigned to each visually grounded word:

$$P(w|O_i, c(w)) = \frac{p(O_i|w) \sum_{all\ C_j\ s.t.\ c(w)\ \epsilon\ C_j} \psi(C_j, O_i)}{\sum_{k=1}^{M} p(O_k|w) \sum_{all\ C_j\ s.t.\ c(w)\ \epsilon\ C_j} \psi(C_j, O_k)} \tag{9}$$

   Where $c(w)$ is the word class to which w belongs. The context-sensitive fitness scores $\psi(C_j, O_i)$ scale each visually based probability density $p(O_i|w)$ depending on how well the syntactic frame C_j is able to generate an unambiguous description of O_i. Note that if two words both describe an object well, Equation 9 will assign relatively large probabilities to both words. On the other hand, for words that tend to increase ambiguity due to other objects in the scene that also fit the semantics of the term, Equation 9 will obtain relatively low probability estimates due to the use of the context-sensitive evaluation based on $\psi()$.

5. *Mix word predictions using visual attention*
   The final step is to mix the influences of all objects in the scene to obtain class conditional word probability estimates:

$$P(w|c(w)) = \sum_{i=1}^{M} P(w|O_i, c(w)) P(O_i) \tag{10}$$

   The degree to which each object biases word predictions depends on Fuse's visual attention state, $P(O_i)$ (Section 7).

Using these five steps, a set of class conditional word probabilities are generated that represent the system's anticipation of words the speaker will use, given the contents of the visual scene, and the system's current visual attention state. Referring back to Equation 4, we can see that the dynamic formulation of class conditional probability estimates $P(w|c(w))$ in Equation 10 can be directly inserted into the computation of bigrams that feed into the speech recognizer. As certain objects in the scene capture more of Fuse's attention, the words that better describe those objects become more probable and thus steer the speech recognizer towards those parts of the vocabulary.

6.2. Relative Spatial Clauses

The spatial grammar (Figure 5) is used to model the use of relative spatial clauses. For example, "The red block beneath the small green block" contains references to two objects, the target (the red block) and a landmark ("the small green one"). The spatial relation "beneath" describes the relation between target and landmark.

Spatial connective terms may consist of multiple words (e.g., "to the left of") that are automatically tokenized [22] and treated as a single acoustic unit during speech decoding. A description consists either of a single phrase describing the target, or descriptive phrases of target and landmarks connected by an appropriate spatial relation. The probability of using spatial relative phrases is encoded in the probability transitions from the "TARGET OBJECT" node of the spatial grammar. This pair of transition probabilities is estimated based on the ratio of training utterances that contained spatial relations versus total training utterances [22].

After describing the visual attention pmf update process in the next section, we explain how spatial relations are handled during speech processing.

7. Language Driven Visual Attention

As Fuse processes incoming speech and generates partial word sequences, a model of visual attention is incrementally updated to reflect the system's current "belief" of the intended referent object. Attention consists of a probability mass function (pmf) distributed over the objects in the current scene. The pmf is used to mix object-dependent description bigrams into a single weighted bigram (Equation 10). Thus, as speech is processed, the evolving distribution of attention shifts the weight of bigrams to favor descriptions of objects that capture more attention. The visual attention model enables the early integration of visual context to provide dynamic incremental estimation of the priors associated with the interpolated class conditional probabilities.

As we mentioned earlier, the speech decoder used in Fuse is based on a single pass Viterbi beam search [16]. In this strategy, multiple word sequences within a search beam are considered during a forward pass, and in a backward pass the best word sequence is selected. In the following, we show how the visual attention model, $P(O_i)$, is computed for a partial word sequence. Separate attentional pmf's are maintained for each parallel word sequence hypothesis. The average pmf over all search paths of the decoder may be interpreted as the system's overall attention at any given point of time.

At the start of each utterance, before any words have been processed, visual attention is shared equally by all M objects in the scene:

$$P(O_i)[0] = \frac{1}{M} \tag{11}$$

The index in square parentheses indicates that this is the attention pmf when 0

words have been processed. As each new word w_n is posited in one of the search paths of the speech decoder, the path-dependent attention pmf is incrementally updated using one of three update rules depending on the type of the new word:

1. w_n is a visually-grounded word. In this case, the update rule is:

$$P(O_i)[n] = \frac{p(O_i|w_n)P(O_i)[n-1]}{\sum_{j=1}^{M} p(O_j|w_n)P(O_j)[n-1]} \tag{12}$$

   That is, the product of the visual models corresponding to modifier terms of an object.

2. w_n is a visually-grounded spatial relation (e.g., "above", "beneath", etc.). The update rule is:

$$P(O_i)[n] = \frac{\sum_{j=1,j\neq i}^{M} p(O_i|w_n, O_j)P(O_j)[n-1]}{\sum_{k=1}^{M}\sum_{j=1,j\neq k}^{M} p(O_k|w_n, O_j)P(O_j)[n-1]} \tag{13}$$

   where $P(O_j|w, O_i)$ is derived from visual models of spatial relations in which O_i is the target object, O_j is the landmark object, and w is the relative spatial term. This update rule causes the attention of the system to shift to objects that hold the spatial relation indicated by w_n relative to whatever object has been described by the partial word sequence $w_1 \ldots w_{n-1}$.

3. w_n is a visually ungrounded word (e.g., "the", "by", etc.). In this case, the update rule is:

$$P(O_i)[n] = P(O_i)[n-1] \tag{14}$$

   Thus, visually ungrounded words have no effect on visual attention.

Using these three update rules, Fuse maintains separate attentional state pmf's for each path of the decoder's search lattice.

8. Visually-Grounded Speech Recognition and Understanding

Processing in Fuse is initiated by the detection of a spoken utterance. A forward search pass of the Viterbi algorithm maintains multiple word sequence hypotheses in a search lattice. Following standard speech recognition methods, a beam is used to limit the number of active paths at any point in the forward pass. Early integration of context effects the paths which are retained within the beam search. Although we believe our approach is advantageous to late integration (e.g., n-best recognition output rescoring using visual context), we have not experimentally compared these two approaches. The attention model biases the search to word sequences that semantically match the properties and spatial configurations of objects in the co-occurring visual scene. Once the entire

utterance has been processed (i.e., the forward pass is complete), backchaining is used to recover the most likely word sequence.

Fuse is able to understand two classes of referring expressions which we refer to as simple and complex [22]. Simple expressions refer to single objects without use of spatial relations, and are fully modeled by the transition network in Figure 4. Complex expressions include relative spatial clauses and are modeled by the network in Figure 5.

Once the forward pass of the beam search is complete, the best word sequence is extracted. We denote this word string as $W = w_1 \ldots w_N$. In the case of a simple referring expression, Fuse selects the object with greatest visual attention:

$$\underset{i}{\text{argmax}} \; P(O_i)[N] \tag{15}$$

For complex referring expressions, W is segmented into three sub-sequences, $W = w_1 \ldots w_{m-1}, w_m, w_{m+1} \ldots w_N$ where w_m is a relative spatial term, $w_1 \ldots w_{m-1}$ describes the target object, and $w_{m+1} \ldots w_N$ describes a landmark object. Fuse selects O_i based on:

$$\underset{i}{\text{argmax}} \; P(O_i)[m-1] \sum_{j=1, j \neq i}^{M} p(O_j | w_m, O_i) P(O_j)[N] \tag{16}$$

where $p(O_j|w_m, O_i)$ is derived from the visual model associated with the relative spatial term w_m. By using Equation 16, a distribution of possible landmarks are combined to determine the single most likely target object.

8.1. A Detailed Example of Visually-Steered Speech Processing

To illustrate the interaction between visual attention and speech processing, we now work through a detailed example. Table 2 shows the transcription of a sample utterance from our test corpus, the output of the speech decoder using standard bigrams without use of the visual context, and the decoder's output using context.

Errors from the decoder are underlined, and omitted words are indicated by square parentheses. Corrections due to visual context are shown in italics. The introduction of visual context in this case makes two important differences. First, the word *lower* is corrected to *large*, and the incorrectly decoded words *to me* are changed to *beneath*. Both of these word substitutions have semantic significance on the interpretation of the utterance. Two occurrences of *the* are also correctly recovered as a result of improved language modeling.

The evolution of visual attention is illustrated for this example in Figure 6. Each graph along the right shows the distribution of attention across the ten objects after integrating the words shown to the left of each graph. The most likely word sequence found by the Viterbi search is shown in the figure. Ungrounded words are shown in parentheses and do not effect the attention pmf. Attention vectors are normalized within each graph. As evidence for the target

Transcript	The large green block on the far right beneath the yellow block and the red block.
No visual context	[The] lower green block in the far right to me [the] yellow block in the red block
Visual context	The *large* green block in the far right *beneath* the yellow block in the red block

Table 2: A example of speech transcription without the use of visual context, and improved output from Fuse with visual context. Deletion errors are marked in square parentheses and substitution errors are underlined.

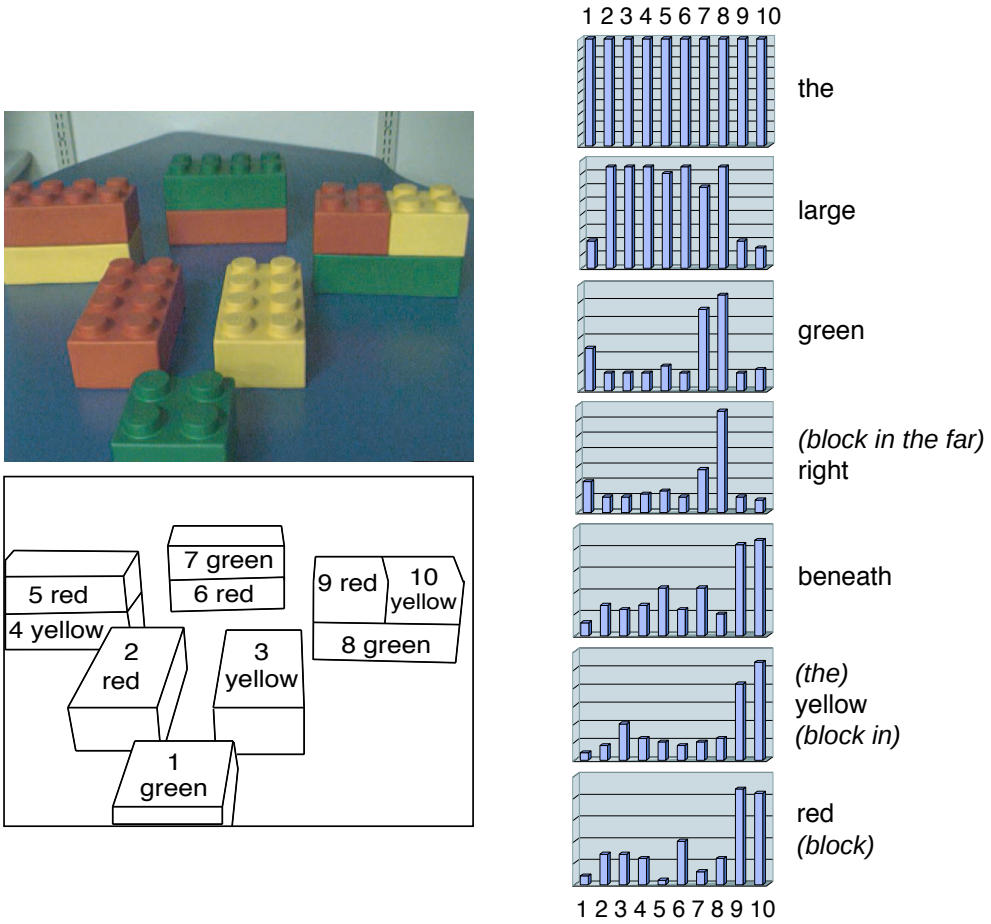

Figure 6: Evolution of attention during processing of the utterance, "The large green block in the far right beneath the yellow block and the red block".

object accumulate from the first part of the utterance, "The large green block in the far right", the pmf becomes progressively sharper with most probability mass focused on Object 8 (fourth graph from the top). When the relative spatial term "beneath" is incorporated (third graph from the bottom), visual attention is captured almost equally by Objects 9 and 10 which are the two smaller blocks above Object 8. Thus, the grounded model associated with "beneath" has caused attention to shift appropriately. The remainder of this utterance refers to two objects. Fuse is designed on the assumption that the remaining phrase will refer to only a single object. Due to the soft assignment of visual attention, however, Fuse is able to robustly deal with the phrase "the yellow block and the red block" by assigning roughly equal attention to both landmark objects. To understand the utterance, Equation 16 is applied and correctly selects Object 8.

8.2. Experimental Evaluation

Eight male and female speakers participated in an evaluation study. The speakers were all students at MIT and had no specific technical background related to this project. Participants were seated at a table, wore a headset microphone, and were asked to produce unambiguous spoken descriptions for selected objects amongst configurations of objects placed on the table. We did not instruct speakers on style of speech, resulting in natural spontaneous speech recordings. Of course, due to the highly constrained nature of the task, the degree of spontaneity was less than would occur in other more natural conversational situations.

A corpus of 990 spoken utterances paired with corresponding visual camera images was collected from the eight speakers. To evaluate Fuse, a leave-one-speaker-out train and test procedure was employed. Each speaker's data was held out and the remaining data was used to train models that were then tested on the held out speaker.

Speech recognition and understanding errors on this corpus are shown in Tables 3 and 4, respectively. Averaged across all eight speakers, the word recognition error rate is reduced by 31% when visual context is used. This result demonstrates that integration of visual context has significant impact on the recognition of speech that refers to the contents of the scene in our experimental task. Although we did not directly compare early versus late integration, we believe that for larger tasks early integration strategies may be preferred since search lattices can be kept smaller while obtaining equivalent overall recognition results.

The effects of visual context on speech understanding are even greater. A speech understanding error occurs when the system selects the incorrect object in response to a description. Since each visual scene had 10 objects, random selection would lead to an average error rate of 90%. The first column of Table 4 shows that without visual context, i.e., using a speech recognizer with static bigrams, the system works quite well, with an average error rate of 24% (i.e., the system chooses the correct object 76% of the time). This system is similar

Speaker	No Visual Context	With Visual Context
1	28.2	21.7
2	24.6	14.3
3	26.9	17.2
4	23.7	16.6
5	19.2	14.5
6	21.3	13.3
7	24.3	17.1
8	26.0	18.8
Ave	24.3	16.7

Table 3: Speech recognition errors (%). Averaged across all eight speakers, the introduction of visual context reduced the word error rate by 31%.

Speaker	No Visual Context	With Visual Context
1	27.4	17.6
2	25.5	12.1
3	27.8	14.8
4	23.3	17.0
5	23.0	13.2
6	23.5	13.9
7	23.8	13.1
8	21.2	12.6
Ave	24.4	14.3

Table 4: Speech understanding errors (%). Averaged across all eight speakers, the integration of visual context reduced the language understanding error rate by 41%.

to that described previously in [23]. The second column of Table III shows the change in understanding errors once visual attention is integrated into the speech decoding process. On average, the number of understanding errors drops by 41%, i.e., Fuse chooses the correct object 86% of the time. The influences of vision on speech processing flow through the system and have substantial effects on overall understanding performance since recognition errors often involve semantically salient words.

8.3. Future Directions

We have presented an implemented model that integrates visual context into the speech recognition and understanding process. In contrast to previous work, Fuse makes use of context at the earliest stages of speech processing, resulting in improved performance in an object selection task. The main idea that this work demonstrates is the payoff of strategically breaking modular boundaries in language processing. A key to achieving this cross-module integration is a model of how natural language semantics relates to visual features of a scene.

We have observed several significant causes if speech understanding errors in Fuse, each of which suggests extensions to the current architecture:

- Speech end point detection errors: The speech segmentation module in our

real time speech recognition system occasionally merges utterances that should have been processed separately. Later stages of Fuse are designed on the assumption that only one referring expression is contained in the utterance. A possible extension is to integrate speech segmentation with semantic analysis for more accurate boundary detection.

- Descriptions with more than one landmark object: We assume that a complex referring expression consists of a target object description, and optionally a landmark object description with connective relative spatial term or phrase. Thus, Fuse cannot consistently handle cases where the referring expressions contain descriptions of more than one landmark object in conjunction or groups of landmark objects (although the example in Section 8.1 demonstrates that sometimes this problem can be overcome in the current approach). This shortcoming suggests the use of more complex grammars, and treatment of semantic composition that goes beyond the multiplication of probability densities. For some steps in this direction, see [8].
- Error Propagation: Due to the feed-forward design of the visual attention update algorithm, errors that enter during initial stages of decoding are propagated throughout the remainder of the utterance. To remedy this, and other related problems, the notion of confidence can be introduced to the visual attention model. For example, the number of active search paths within the Viterbi beam search, which is often used as a source for estimating acoustic confidence in speech recognizers [19], might similarly be used as the basis for estimating confidence of the visual attention pmf. When confidence is low, the effects of attention could be discounted.
- Visual Segmentation Errors: Some errors in understanding occur due to imperfect image segmentation performed by the visual analysis system. Such segmentations may merge more than one objects or divide an object into two or more parts. These cause mismatches among descriptions and the corresponding objects. This problem suggests early integration of speech into visual processing, the complement of the integration we have explored in Fuse. Referring back to Figure 2, this suggests that the visual scene analysis module might be brought into the processing loop. If the speech decoder confidently reports the phrase "the two blue blocks on the right", this might help the visual analyzer decide between interpreting a stack of blocks as a single block versus two.

To implement Fuse, we made strong simplifying assumptions about the task. In addition to assuming that each spoken utterance would in fact be a referring expression to an in-view object, we also assumed that no other modalities are available in parallel to speech for selecting objects. Of course in many natural settings it would be preferable simply to point, or to combine speech and gesture. As we mentioned earlier (Section 2), our goal in this work was to explore multimodal integration in which the non-speech channel encoded information about

context rather than the speaker's intentions. A useful future direction would be to bring these different kinds of multimodal information together.

Looking ahead, we plan to expand this work along two directions. First, Fuse will be integrated into an interactive manipulator robot[15, 24]. Fuse will have access to representations in the robot's visual system and also its planning and memory systems, leading to an enriched encoding of context to help guide speech processing. Second, we plan to extend Fuse to work with non-visual context cues such as geographical position and time of day in order to build context-aware assistive communication devices [5].

Acknowledgments

We thank the anonymous reviewers for their comments. This material is based upon work supported by the National Science Foundation under Grant No. 0083032.

References

[1] L.R. Bahl, F. Jelinek, and R.L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Journal of Pattern Analysis and Machine Intelligence*, 2(5):179–190, 1983.

[2] Michael K. Brown, Bruce M. Buntschuh, and Jay G. Wilpon. SAM: A perceptive spoken language understanding robot. *IEEE Transactions on Systems, Man, and Cybernetics*, 22 . IEEE Transactions 22:1390–1402, 1992.

[3] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[4] C. Crangle and P. Suppes. *Language and Learning for Robots*. CSLI Publications, Stanford, CA, 1994.

[5] E. Dominowska, D. Roy, and R. Patel. An adaptive context-sensitive communication aid. In *Proceedings of the CSUN International Conference on Technology and Persons with Disabilities*, Northridge, CA, 2002.

[6] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[7] Jerry Fodor. *The Modularity of Mind*. MIT Press, 1983.

[8] Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.

[9] Lawrence A. Hirschfeld and Susan A. Gelman, editors. *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, 1994.

[10] Rukmini Iyer and Mari Ostendorf. Modeling long distance dependence in language: Topic mixture vs. dynamic cache models. *IEEE Transactions on Speech and Audio Processing*, 7:30–39, 1999.

[11] M. Johnston and S. Bangalore. Finite-state multimodal parsing and understanding. In *Proceedings of COLING-2000*, 2000.

[12] Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2000.

[13] Thomas Niesler and Philip Woodland. Variable-length category n-gram language models. *Computer Speech and Language*, 21:1–26, 1999.

[14] Sharon Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of Conference on Human Factors in Computing Systems: CHI '99*, pages 576–583. ACM Press, 1999.

[15] Kai-yuh Hsiao, Nikolaos Mavridis, and Deb Roy. Coupling perception and simulation: Steps towards conversational robotics. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, 2003.

[16] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

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

[18] Terry Regier. *The human semantic potential*. MIT Press, Cambridge, MA, 1996.

[19] R. Rose. Word spotting from continuous speech utterances. In C.H. Lee, F. K. Soong, and K.K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, chapter 13, pages 303–329. Kluwer Academic, 1996.

[20] R. Rosenfeld. Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278, 2000.

[21] Ronald Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, pages 187–228, 1996.

[22] Deb Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.

[23] Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. A trainable spoken language understanding system for visual object selection. In *International Conference of Spoken Language Processing*, Denver, 2002.

[24] Deb Roy, Kai-Yuh Hsiao, and Nick Mavridis. Mental imagery for a conversational robot, In press, IEEE Transactions on Systems, Man, and Cybernetics.

[25] Michael J. Spivey, Melinda J. Tyler, Kathleen M. Eberhard, and Michael K. Tanenhaus. Linguistically mediated visual search. *Psychological Science*, 12(4):282–286, 2001.

[26] Michael J. Spivey-Knowlton, Michael K. Tanenhaus, Kathleen M. Eberhard, and Julie C. Sedivy. Integration of visuaspatial and linguistic information: Language comprehension in real time and real space. In Patrick Oliver and Klaus-Peter Gapp, editors, *Representation and processing of spatial expressions*. Erlbaum, 1998.

[27] M.K. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information during spoken language comprehension. *Science*, 268:1632–1634, 1995.

[28] Benjamin Yoder. Spontaneous speech recognition using hidden markov models. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.