Spontaneous Speech Recognition Using Visual Context-Aware Language Models

by

Niloy Mukherjee

Bachelor of Technology (Hons.) in Computer Science and Engineering,
Indian Institute of Technology, Kharagpur
June 2001

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2003

Author .....
Program in Media Arts and Sciences,
School of Architecture and Planning
August 8, 2003

Certified by .....
Deb K. Roy
AT&T Career Development Professor
Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Andrew B. Lippman
Chairman, Department Committee on Graduate Students

Spontaneous Speech Recognition Using Visual Context-Aware Language Models

by

Niloy Mukherjee

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on August 8, 2003, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

The thesis presents a novel situationally-aware multimodal spoken language system called Fuse that performs speech understanding for visual object selection. Fuse uses semantic information from immediate visual context to guide spoken language recognition and understanding. An experimental task was created in which people were asked to refer, using speech alone, to objects arranged on a table top. Fuse extracts visual information about objects and their spatial configurations using computer vision techniques. This visual information is used to generate language models, which, together with an adaptive visual attention mechanism, biases spoken language recognition to more likely interpretations of speech signals to select the target object referred to. In an evaluation of the system, visual context-based language modeling is shown to significantly decrease the word error rate of the speech recognition task.

During training, Fuse acquires a grammar and vocabulary from a "show-and-tell" procedure in which visual scenes are paired with verbal descriptions of individual objects. Fuse determines a set of visually salient words and phrases and associates them to a set of visual features. Given a new scene, Fuse uses the acquired knowledge to generate language models conditioned on the objects present in the scene as well as a spatial language model that predicts the occurences of spatial terms conditioned on target and landmark objects. The speech recognizer in Fuse uses a weighted mixture of these language models to search for more likely interpretations of user speech in context of the current scene. During decoding, the weights are updated using a visual attention model which redistributes attention over objects based on partially decoded utterances. The dynamic situationally-aware language models enable Fuse to jointly infer spoken language utterances underlying speech signals as well as the identities of target objects they refer to.

The underlying ideas of context-aware speech understanding that have been developed in Fuse may may be applied in numerous areas including assistive and mobile human-machine interfaces.

Thesis Supervisor: Deb K. Roy
Title: AT&T Career Development Professor
Media Arts and Sciences

Spontaneous Speech Recognition Using Visual Context-Aware Language Models

by

Niloy Mukherjee

Thesis Reader .....

Rosalind W. Picard
Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Thesis Reader .....

Allen Gorin
Distinguished Member of Technical Staff
AT&T Labs Research

Acknowledgments

First of all, I would like to thank MIT for giving me a wonderful opportunity to explore my research and academic capabilities. Thank you, Deb, for bringing me up here in the MIT Media Laboratory to assist your research. Working with you these two years has indeed been an enlightening experience.

Special thanks to Prof. Roz Picard and Dr. Allen Gorin to guide me as my thesis readers. Roz, your course was really one of the best classes I have taken as a part of my engineering curriculum. Allen, I hope that someday I will also build systems that contribute to the benefit of millions.

Finally, thanks to each and everyone of you for helping me out in one way or the other during these two years, one of the best periods of my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The variety of applications of automatic speech recognition (ASR) systems for human computer interfaces, telephony and robotics [18][39][30] has driven the research of a large scientific community in recent decades. In recent years, speech recognition has reached the point of commercial viability realizable on off-the-shelf computers. There have also been recent efforts to develop natural spoken language understanding systems for human machine interaction. Such systems consist of ASR engines, whose output is used as the input of a language understanding component. Such a late-integration framework poses an immediate challenge since the ASR component typically introduces speech recognition errors into the system.

Statistical language modeling [34] techniques have been developed to capture regularities of natural language for the purpose of improving the performance of ASRs. These techniques estimate the probability distributions of various linguistic units such as words, sentences, and whole documents. A typical speech recognizer uses these estimates as background knowledge to predict possible word strings or sentences. Most common language models such as n-grams [15] estimate the probability of a linguistic unit conditioned on n previously seen units using large amounts of texts. They take no particular advantage of the deep semantic structures underlying the language being modeled. Recent attempts have been made to incorporate linguistic theories, semantic knowledge, dialog histories and other rich structures into language modeling [34, 5]. However, such attempts have been limited to restricted domains.

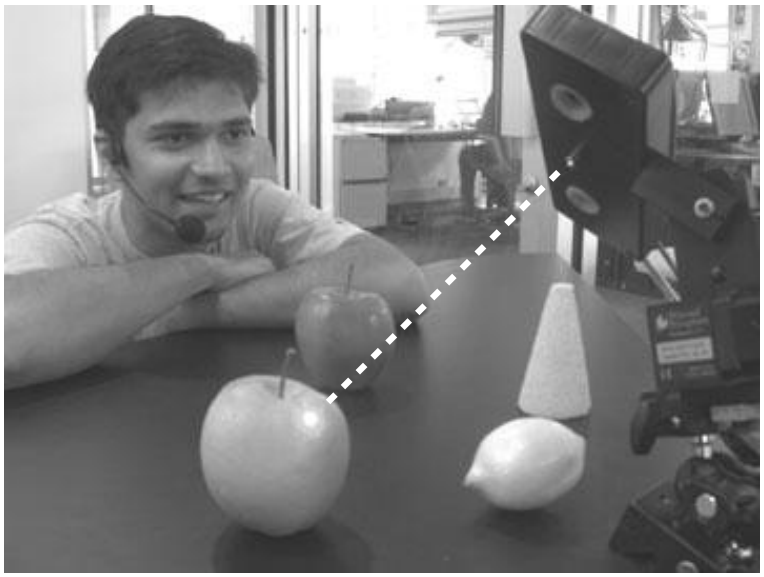

Figure 1-1: Typical Human Robot Interaction.

There have been recent efforts to build situatioanlly-aware applications [2, 8] that support human-computer interaction in personal computing domains. Situationally-aware systems extract, interpret and use situational information and adapt their functionalities to current environments. A few of the current speech understanding systems, such as the AT&T "How May I Help You", [11], use user contexts as well as dialog contexts to assist human-computer spoken language dialog in customer-care domains. However, little work has been done to introduce situational awareness in the ASR language modeling through sensory information channels to assist human robot or human machine interaction. Although audio-visual speech recognition has emerged as an active field in multi modal research in recent years, much focus has been on facial features, to be precise, lip movements, to enhance recognition confidence. The efforts do not address spoken language acquisition to assist speech recognition and understanding for human computer interaction. They deal more with surface forms of speaker features rather than semantics of the shared human machine environment. However, statistical language models that can use situational knowledge obtained from other channels of sensory information to dynamically adjust utterance distributions based on the environment may improve performance of spoken language processing systems.

1.1 Thesis Motivation

One of the essential properties of language is that its grammatical structure can be characterized independently of meaning or reference (universal grammar) [6] This has led to a long tradition in psycholinguistics, which assumes that the brain processes responsible for the rapid syntactic structuring of continuous linguistic input are "encapsulated" from other cognitive and perceptual systems. Much of early visual processing often is claimed to be structured by independent processing modules [10]. The primary empirical evidence that syntactic processing is modular is due to the fact that brief syntactic ambiguities, which arise because language unfolds over time, seem to be initially resolved independently of prior context. However, recent psycholinguistic studies have revealed surprising breaches of modularity. These studies have found that acoustic and syntactic aspects of online spoken language processing are influenced by visual context [44, 43, 9]. During interpretation of speech, partially heard utterances have been shown to incrementally steer visual attention and vice versa, visual context has been shown to influence speech processing. It has also been found that the visual context may affect the resolution of temporary ambiguities within individual words. Motivated by the compelling evidence for the early and nearly seamless integration of visual and linguistic information that emerged from these experiments, the thesis presents a cross-modal spoken language processing (recognition / understanding) system in which visual context affects early stages of speech processing to assist confident speech recognition and understanding. The system design is based on the assumption that valuable cross-modal information is lost if the speech recognition module remains unaware of other sensory modules in a multimodal system.

1.2 Thesis Objective

In this thesis, we consider the problem of recognition / understanding of spoken language utterances that consist of expressions referring to objects in a non-mobile

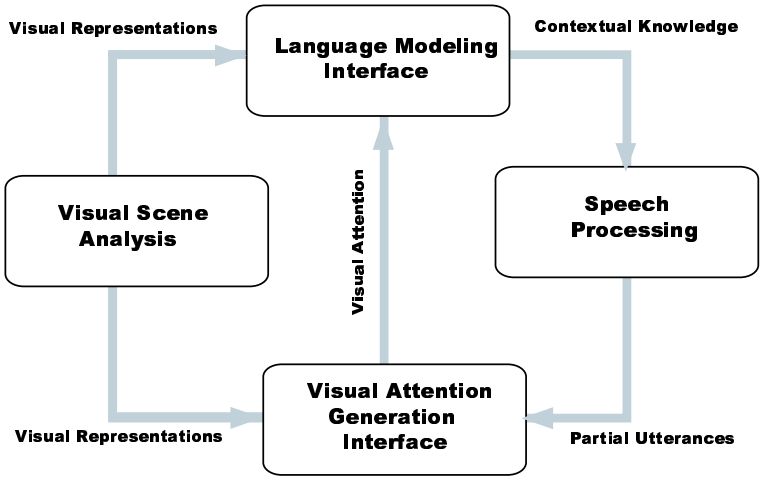

Figure 1-2: Block Diagram of the Proposed Early Integration Approach

visual environment of object configurations shared between a robot and a human participant. Spoken language processing in this domain is challenging since there are numerous potential candidate objects present that may lead to ambiguities in the interpretation of such classes of expressions. Furthermore, speakers may use various description strategies to refer to an object in a visual scene.

The proposed approach of early cross-modal integration has been implemented in an on-line, real-time, multimodal processing system that jointly infers the most likely word sequences in spoken language utterances as well as identities of the referent target objects. Figure 1-2 illustrates a block diagram of the proposed approach. Unlike state-of-the-art language models that are derived from textual transcriptipons, our system generates dynamic visually driven language models in the form of statistical distributions of words and phrases using semantics extracted from visual representations of an input image. These visually driven language models are used in the speech recognition search algorithm to steer search paths towards more likely word sequences. Partially decoded spoken utterances during the speech recognition search update a dynamic model of visual attention spread over the input scene, which, in turn, re-estimates the possible words and phrases that a speaker is likely to say next. As processing continues, linguistic and visual information mutually reinforce each other, sharpening both linguistic and visual hypotheses. The cross-modal linguistic

and visual integration approach takes care of the challenges highlighted above. Hence, the system constitutes four broad components:

- A medium vocabulary speech recognition system.
- A machine vision system to process simple objects and spatial relations.
- A trainable language-modeling component that generates predictions of words and phrases from visual representations.
- An adaptive visual attention model to integrate the language model into the speech recognition system on-the-fly to bias the recognition process towards more likely interpretations of acoustic signals.

An overview of the design and the main components are described in chapter 3. Details of each of the system components are described in chapters 4, 5 and 6.

1.3 Application Scenario in Human Computer Interaction

The research may find various applications which involve spoken language recognition / understanding accompanied by user context. Portable devices that are context-aware through visual as well as other sensory modalities may apply them to perform better user speech understanding. For example, a personal travel assistant may integrate other non-speech sensory situational awareness such as, user location, current time, user schedule, to the core speech processing in order to improve task understanding. Other domains may include context-sensitive automatic medical transcription systems, and assistive robots [40].

1.4 Thesis Outline

The rest of the thesis describes our approach of early cross-modal integration as well as the design and implementation details of the proposed system. The thesis is

organized as follows:

- Chapter 2 highlights some of the prior work related to spoken language understanding, language modeling and related research in these areas.
- Chapter 3 gives an overview of the system highlighting the main components.
- Chapter 4 details the visual analysis module of the proposed system.
- Chapter 5 describes the speech recognition module of the proposed system.
- Chapter 6 details the early cross modal integration through dynamic language modeling interfaces.
- Chapter 7 summarizes performance evaluation of Fuse
- Chapter 8 concludes the thesis with suggestions for future work.

Chapter 2

Background

Automatic speech recognition has been an active field of research in recent decades. Most of the research has focused on building automatic transcription systems and voice command-driven telephony systems. These systems address four broad research components

1. Robust signal processing of the acoustic waveforms
2. Statistical modeling of sub word units such as phonemes
3. Statistical modeling of language regularities
4. Fast speech decoding over a large vocabulary, and
5. Language understanding to enable human-to-computer conversations

The thesis makes contributions to (3) and (5) by introducing a novel method for generating trainable language models using visual representations.

This chapter gives a brief overview of some of the research that has taken place in the areas of statistical language modeling and spoken language understanding. The chapter also briefs some of the recent cognitive studies of cross-modal speech interpretation that show early integration of linguistic and visual information. Due to the broad range of topics, selected works have been highlighted.

2.1 Statistical Language Modeling (SLM)

A survey of major SLM techniques is listed in the following subsections.

2.1.1 n-grams

N-grams are the most successful language models used in current speech recognition technology. An n-gram model estimates the probability of occurrence of the n^{th} word, phrase or any arbitrary symbol based on a history of previously encountered $n-1$ symbols. For large training corpora, a trigram (n=3) language model is a common choice whereas a bigram (n=2) language model is often used with smaller ones. The problem in estimating trigrams as well as bigrams is a sparse data problem because there may be several bigrams and trigrams combinations that never occur in the training corpora. For example, [33] mentions that even after training a trigram language model using a 38 million words' worth of newspaper articles, one-thirds of trigrams in new articles from the same source are novel.

Clustering methods have been developed to tackle data sparseness in statistical language modeling [3]. A typical vocabulary-clustering algorithm searches for equivalence classes (word classes) from the training data. It assigns equal probability distributions to all members in a word class. This accounts for comparable weighting of syntactically similar words in the language model. [33] reports that good results are obtained by manual clustering of words [45] in narrow discourse domains (e.g. ATIS, [28]). [20] suggests an automatic iterative clustering method using information theoretic criteria that can be applied to large corpora to increase the performance of the resulting model. Class-based n-gram models form the basis of the language modeling component in the system presented in this thesis and are described in details in section 6.1

While n-gram models do reflect the semantics of the language being modeled, theymainly model syntactic regularities in a language. Every language possesses a deep set of structures contributing to its comprehensibility [6]. N-gram models do not take advantage of these semantic characteristics in language modeling.

2.1.2 Linguistically Motivated Models

Several SLM techniques, however, are directly derived from grammars commonly used by linguists. Context free grammars (CFGs) [25], as the name suggests, are context free models [14] of natural language. A typical context-free grammar consists of an initial non-terminal symbol, a set of non-terminal symbols and terminals such as words, and a set of transition rules called "rewrite rules". Sentences are generated by applying these transition rules that fan out non-terminal symbols in terms of terminals and non-terminals, until a sequence of terminals is achieved. Specific CFGs have been created based on parsed and annotated corpora such as [22]. A stochastic context free grammar (SCFG) [4] puts a probability distribution on the transitions fanning out from each non-terminal, thereby inducing a distribution over the set of all sentences.

Link grammar is a lexicalized grammar proposed by [42]. In a link grammar, each word is associated with one or more ordered sets of typed links with each such link connected to a similarly typed link of another word in the sentence. Probabilistic forms of link grammars have also been attempted [16].

Although these models provide more sophisticated language information, their main disadvantage is that they suffer from data fragmentation, in that more detailed modeling results in estimating numerous parameters with less and less data. Estimation of these models is data-driven and often results in generating locally optimal models. Moreover, they do not account for actual context sensitive behavior of the language.

2.2 Spoken Language Understanding

This section briefs some of the research in developing spoken language understanding systems in recent decades to assist human-computer interaction in various domains. A particular domain that has been in focus is the Airline Travel Information System (ATIS) [28] domain. Many research laboratories have developed spoken language understanding systems in this domain. One such system is the SRI International

ATIS spoken language understanding system [25]. The SRI ATIS system uses sophisticated natural language processing algorithms that transform speech recognition output into well-posed database queries. The natural language processing component incorporates information from natural language and lexicon into a statistical language model to assist both recognition and understanding. It also uses context-modeling mechanism for air travel planning that constructs user dependent models to assist speaker-dependent speech understanding.

The Berkeley Restaurant Project (BeRP) [17] is a medium-vocabulary spontaneous speech understanding system developed at ICSI. The system constitutes a medium-vocabulary Viterbi [31] speech decoder called Y_0 that uses simple bigram language models as well as stochastic context free grammar (SCFG) parsers / generators. The BeRP natural language processing backend transforms candidate word strings from the decoder to well-formed database queries. The backend uses bottom-up stochastic chart parser on a stochastic context free attribute grammar to compute all possible parses and semantic interpretations of the input.

GALAXY [41], a spoken language system architecture developed by the Spoken Language Systems group at MIT, performs spoken language understanding to enable human-to-computer conversations in domains such as weather forecasts, travel reservations and city guides. Like BeRP, Galaxy follows a modular design in which a speech recognition component, SUMMIT, generates a list of candidate sentences from the acoustic signals. The top choice sentence selected by SUMMIT is represented in a logical, meaningful structure. Based on stored rules, the language understanding component, TINA, parses each sentence into grammatical components, such as subject, verb, object, and predicate. TINA then formats this information in a semantic frame, a command-like structure consisting of a clause, topic, and predicate. The language generation component, GENESIS, converts the semantic frames to well-formed messages.

2.3 Spoken Language Acquisition

The systems described in the previous section use built-in databases and rules to perform spoken language understanding. In recent years, there has been some research on automatic acquisition of linguistic information to assist speech understanding. One of the first systems to be deployed that performs spoken language acquisition in real-world customer-care scenario is the AT&T "How May I Help You" spoken language understanding system [11, 12]. The system learns salient phrases and grammar fragments from human-operator conversations related to telephony customer care services. During spoken language processing, the system takes in the output of the AT&T Watson speech recognition system [13] in the form of word graphs or lattices [21], retrieves salient parts of the dialog and maps it to one of many possible subsequent dialog actions. The system performs late integration of both customer as well as dialog context to guide the speech understanding.

Roy and Pentland [35, 38] modeled the early stages of word acquisition from sensor-grounded speech and visual signals. They demonstrated the utility of learning algorithms based on cross-modal mutual information in discovering words and their visual associations from untranscribed speech paired with images of three-dimensional everyday objects. An embodied system was able to learn object names from a corpus of spontaneous infant-directed speech. The system discovered and segmented word-like acoustic units from spontaneous speech driven by cross-modal analysis. An acquired lexicon of visually grounded words served as the basis for a small vocabulary speech understanding and generation system. The system was able to process single and two-word phrases which referred to the color and shape of objects.

Roy [36] developed a spoken language generation system that performs language acquisition to describe objects in computer-generated visual scenes. The system was trained by a "show-and-tell" procedure in which visual scenes are paired with natural language descriptions. Learning algorithms acquire probabilistic structures that encode the visual semantics of phrase structure, word classes, and individual words. The system integrates syntactic, semantic, and contextual constraints to generate

natural and unambiguous descriptions of objects in novel scenes. The system generates syntactically well-formed phrases phrases, as well as relative spatial clauses to describe target objects through simple utterances as well as complex utterances involving landmark objects and spatial relations. A similar research reported by Roy et.al. [37] presents a trainable, visually grounded, spoken language understanding system that acquires a grammar and vocabulary from a "show-and-tell" procedure in which visual scenes are paired with verbal descriptions. During training, a set of objects is placed in front of the vision system. Using a laser pointer, the system points to objects in random sequence, prompting a human teacher to provide spoken descriptions of the selected objects. The descriptions are used to automatically acquire a visually-grounded vocabulary and grammar. Once trained, the system late-integrates speech recognition and visual analysis outputs and points, in real-time, to the object which best fits the visual semantics of the spoken description.

2.4 Psycholinguistic Experiments on Cross-modal Processing

The early cross-modal integration of linguistic and visual input to guide language comprehension forms the basis of the design of our system. This section reviews a couple of psycholinguistic experiments that have been carried out in the past few years to observe whether human spoken language comprehension follows the same principles of cross-modal behavior. In one such study on cross-modal behavior of human language comprehension, Tanenhaus et. al.[44] demonstrated that individuals process spoken language instructions incrementally using visual cues from the environment. Eye movements were recorded with a head-mounted eye-tracking system while subjects followed instructions to manipulate real objects to test the effects of relevant visual context on the rapid mental processes that accompany spoken language comprehension. The subjects were found to make saccadic eye movements to objects immediately after hearing relevant words in the instruction. For example,

during processing an utterance "touch the starred yellow square", the subjects made eye movements to the target block after hearing the word "starred" when only one starred object was presenting the environment. The authors found strong evidence for seamless integration of visual and linguistic information to hypothesize that a relevant visual context, if available for the listener to query as the linguistic input unfolds, may influence its initial syntactic analysis.

Spivey et.al. [43] conducted a linguistically mediated visual search experiment to support theories about a connection between the language comprehension and visual perception portions of the brain and that language comprehension can constrain visual perception. In this task, a subject was asked whether a target object was present on the computer screen then instructed to press a key on the keyboard that corresponded to their 'yes' or 'no' answer. The reaction time increased with the increase in the number of competitive objects in the scene. The experiments led to a theory of how mental processes are able to respond to the question of whether a target object is present or absent on the screen.

2.5 Summary

To summarize, statistical language modeling and speech understanding form the core components of the system described by this thesis. The chapter highlights some of the research that has taken place in these areas over the past few years. Some of the language modeling techniques reviewed in this chapter, such as n-grams and vocabulary clustering, have been modified in the system design to generate visually-steered language models that use visual context to assist spoken language understanding. Unlike most of the spoken language understanding systems described in this chapter that perform late integration of speech recognition output with natural language processing, this thesis presents a system that performs early cross-modal integration of linguistic and visual information to assist speech recognition as well as understanding simultaneously. The chapter also presents some of the recent studies in human language comprehension that form the basis of our experiments to compare human visual

attention patterns to the proposed visual attention mechanism used in the system.

Chapter 3

Fuse: A Model of Cross-modal Spoken Language Processing

This chapter provides an overview of our approach to speech recognition / understanding, which is integrated with visual context through dynamic language models. The chapter starts with a brief description of the speech recognition task used in our experiments. The rest of the chapter introduces the main components of the developed system named Fuse.

3.1 The Scene Description Task

A simple scene description task was developed to study the role of visual context in spoken language recognition / understanding. A typical visual scene used in the task consists of a camera image consisting of ten lego blocks in average with random spatial configurations. During data collection, the task interface randomly selects one of the ten lego bricks as the target object. Participants were asked to verbally describe target objects in scenes (Figure 3-1) without any restrictions 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 referential utterances such as, "The large green block above the large red and yellow blocks". The result of this data collection was a set of images and spoken descriptions to objects in the images.

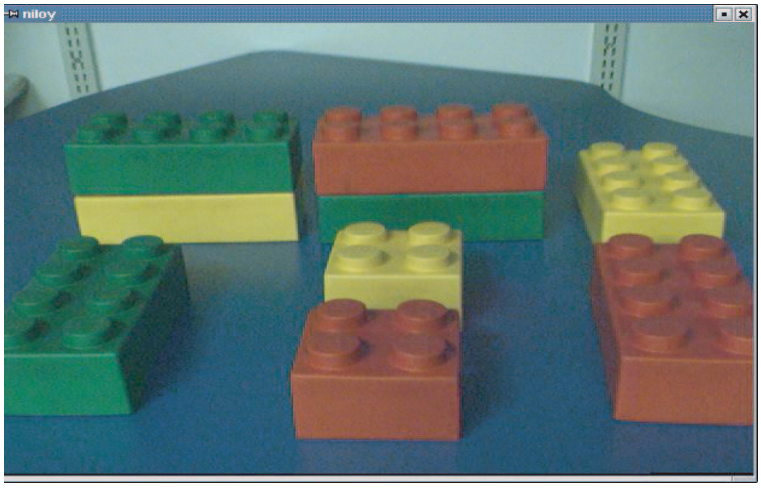

Figure 3-1: An example scene.

The performance of the system was evaluated based on two criteria:

- Speech recognition performance: given a visual scene, compare the speech recognition output with and without the knowledge of visual context.

- Spoken language understanding: given a spoken description, find the object in the scene that best fits the description.

Details of the data collection and evaluation are presented in chapter 7

3.2 The System Overview

Figure 3-2 provides an overview of our approach to integrating visual context with speech recognition and understanding. The remainder of this section highlights the main ideas underlying the approach.

As shown in Figure 3-2, input consists of a speech signal and an image. Figure 3-1 is representative of images in the current task, captured by a color video camera. The speech signal is recorded from a head-worn microphone.

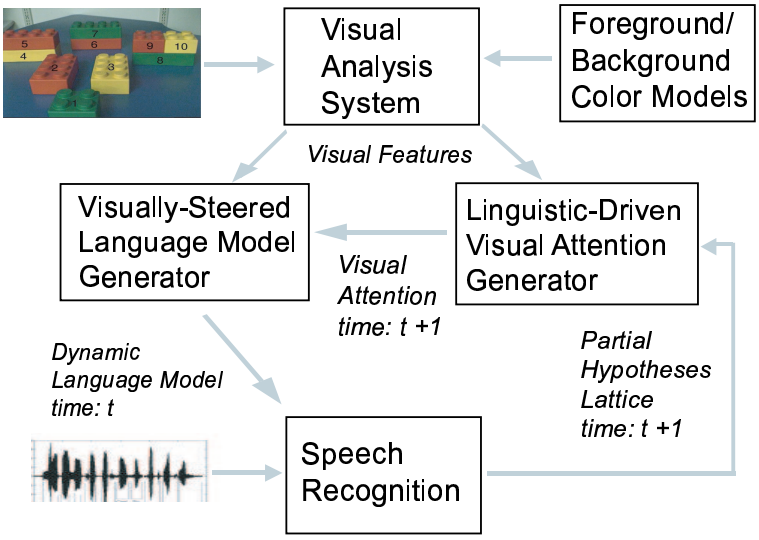

Figure 3-2: The overview of the system architecture.

3.2.1 The Main System Components

Fuse is implemented as a distributed system that allows system components to communicate with each other through network message passing protocols. Processing in Fuse is initiated by the detection of spoken utterance by the **Speech Recognition** module. At this stage, the utterance segmenter sends a request to the **Visual Scene Analysis** module to capture the current video frame. The visual scene analysis component segments individual objects in the scene and extracts visual properties of objects and spatial relations between object pairs. The resulting representation is sent to both the **Dynamic Language Model Generation** module and the **Visual Attention Generation** module. The visual attention generation module keeps the representation in memory, generates a vector of equally distributed attention over all the objects in the scene and sends it to the language model generation module. The language model generation module estimates the probability distribution of occurrences of visually relevant words and phrases given the object configuration in the current scene and the visual attention. The generated language model provides contextual knowledge to the **Speech Decoder** during speech processing.

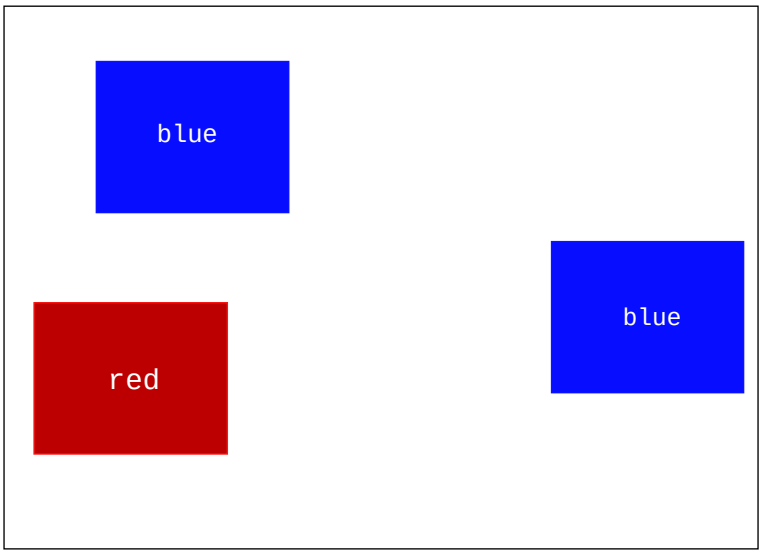

Figure 3-3: Visual Configuration for Case 1: Simple Utterance

As the spoken utterance is processed by the speech decoder on line, multiple partial hypotheses of utterances up to the current time (t) feed back to the visual attention module in the form of a partially ordered word graph [21]. The visual attention module redistributes attention over the objects based on the partial utterances. The language model is updated based on the reestimated visual attention and is accessed by the decoder at time $t + 1$ The process continues until the utterance segmenter detects speech endpoint. At this point, the **Speech Understanding** component uses the decoder results to highlight the object that best fits the utterance. Changes made to a scene during the processing of individual utterances are not taken into account.

We will work through a couple of examples to explain the main processing loop in Figure 3-2 and the role of the language model and visual attention model. The details of these components are described in chapter 6.

3.2.2 Handling Simple Referent Expression

Let us consider a situation in which a speaker utters the expression "The red block on the left" in the context of a scene as shown in the figure 3-3. As the first portion of the utterance is processed, let us assume that the system has correctly pulled out

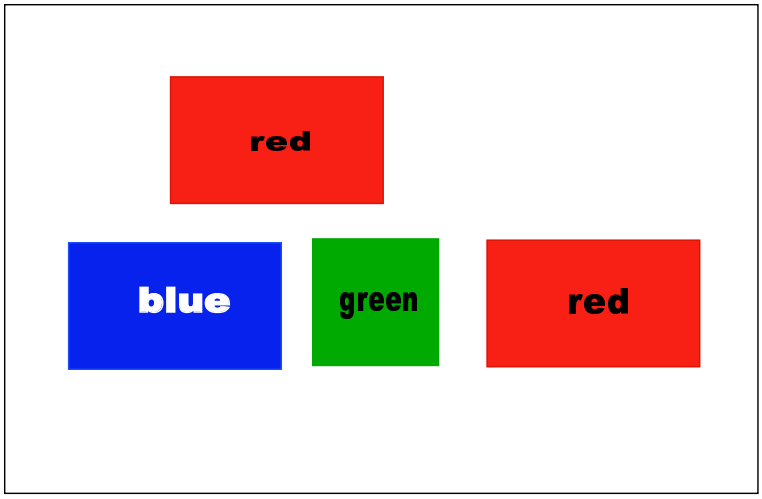

Figure 3-4: Visual Scene Configuration for Case 2: Complex Utterance

a hypothesis "the red". We consider a single hypothesis for this example. In reality, there are thousands of hypotheses active at a given time frame. The visual attention module at this point receives the output from the speech recognizer as well as the set of visual features from the visual analysis module. Visual attention is modeled as a probability mass function (pmf) vector whose elements are identical during the onset of an utterance. When the words "the red" are fed into the visual attention module, it updates the probability mass function vectors so that visual attention is shifted to the the red things in the scene, which results in a shift to the single red block present in the scene.

The language model generator module generates a probabilistic mixture of descriptions for each object in the scene. At this point, the updated pmf s cause the generator module to favor the words "left" and "bottom" rather than "right" or "top" since there is only one red block in the scene. In this way, the updated language model biases the speech recognizer to decode the rest of the utterance more confidently.

3.2.3 Handling Complex Referent Expression

One complication is introduced with utterances with relative spatial clauses such as, "The red block above the small green one". Let us consider a situation in which a speaker says so in the context of a scene as shown in the figure 3-4 . In this class of

utterances, the visual attention must be reset mid-way through processing to refocus from one object to another. Let us assume that the speech recognizer recovers a partially decoded utterance, "the red block". The visual attention module updates the probability mass function so that most of the attention is shifted to the two red blocks on the scene. The dynamic language model generator favors "above" as well as "left" more than "below" or "under" as the red blocks are either above or to the left of some other landmark object in the scene. Let us assume that the recognizer ends up with a partial hypothesis "the red block above". At this stage, the visual attention pmf is updated such that most of the attention is spread over objects that may serve as the possible landmark objects satisfying the spatial relation "above" with respect to the objects probabilistically referred to by the partial utterance "the red block". In our case, the system ends up paying more attention to the green and the blue blocks that are below the red block. This gradually drives the speech decoder to interpret the landmark expression " the small green one" more effectively.

3.3 Summary

To summarize, the system generates a visually steered language model as a mixture of descriptions when presented a new scene. 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. The approach is different from a modular approach, in which speech recognition and visual analysis would be performed separately and combined by an integrator that does not affect the internal operations of earlier stages of processing. The design and implementation aspects of the system are explained in the following chapters that provide detailed descriptions of each component of the system.

Chapter 4

Visual Scene Analysis

The chapter details the visual scene analysis module of Fuse. The module segments colored objects in an input camera image in real time, extracts visual properties of objects and computes inter-object spatial relations. The resulting representation is passed to the language model generator and the visual attention generator modules.

4.1 Object Segmentation

Object segmentation is performed on the basis of color. The color distribution of objects is modeled using mixtures of Gaussian distributions. Although all objects are constrained to be single-colored, shadow effects of three-dimensional objects necessitate the use of a mixture of Gaussian distributions. Three mixtures have been used for each color distribution to capture these effects. The density functions for the color distributions are given by:

$$f_{color}(x) = \Sigma_{j=1}^{3} w_j N(x|\mu_j, C_j); \tag{4.1}$$

where

$$N(x|\mu_j, C_j) = \frac{1}{[2\pi]^{N/2}|C_j|^{1/2}} exp[-\frac{1}{2}(x-\mu_j)^T C_j^{-1}(x-\mu_j)] \tag{4.2}$$

w_j are the mixture weights ($\Sigma_{j=1}^{3} w_j = 1, w_j \geq 0$) μ_j are the mixture means, and

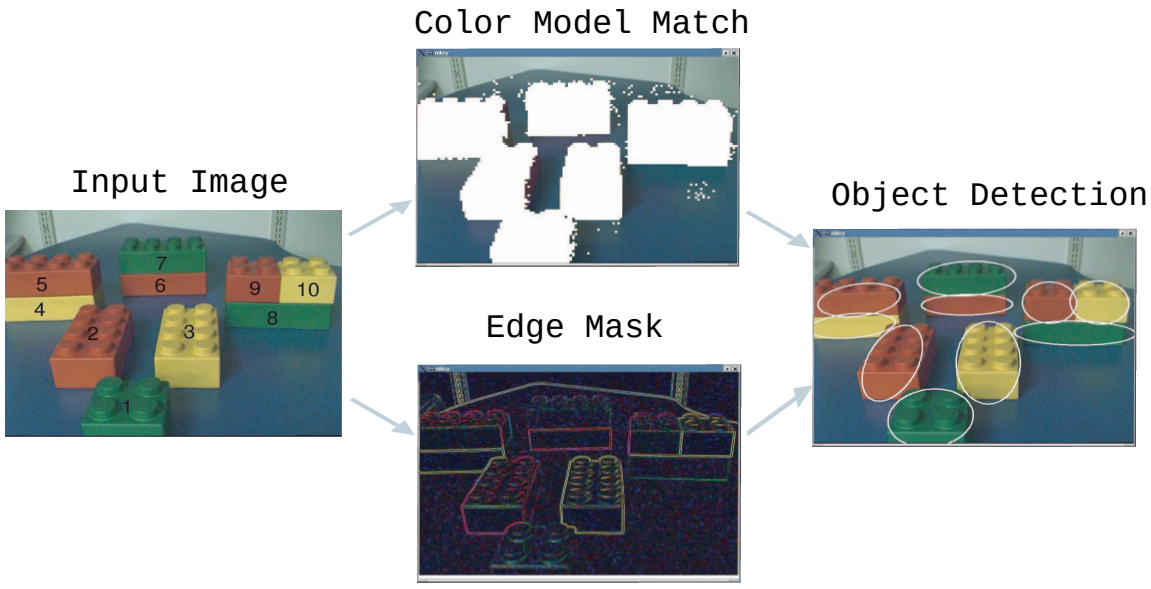

Figure 4-1: Object detection procedure.

C_j are the mixture covariance matrices.

For each type of object used in the experiments, a color model is created by collecting sample images of each object and manually specifying the region within each image that corresponds to the object. The Expectation Maximization (EM) algorithm [7] is used to estimate both the mixture weights and the underlying Gaussian parameters for each object. K-means clustering [19] is used to provide initial estimates of the parameters.

As shown in Figure 4-1 , input images are split into two parallel processing paths. The first process performs local edge detection in each of the RGB color planes. The outputs from the edge detectors of all three planes are summed to provide an overall estimate of local edges. This results in a binary image in which pixels located at edges are set to 0. In the second process, each 5x5 patch of pixels is classified as belonging either to one of the known set of objects or to the background. This is done by evaluating each color model (described above) for each pixel patch and thresholding the resulting values. Patches that have high matches with any color model are set to 1 and the remainder are classified as 0. The upper middle image in Figure 4-1 shows the result of this stage.

Feature Name	Feature Description
r	The mean R value of the 10x10 pixel region in the center of the object.
g	The mean G value of the 10x10 pixel region in the center of the object.
b	The mean B value of the 10x10 pixel region in the center of the object.
hw-ratio	Height to width ratio of the object contour
area	Contour area
x	X coordinate of the upper left corner of the object bounding box
y	Y coordinate of the upper left corner of the object bounding box
mm-ratio	Ratio of the maximum to the minimum dimension of the object contour

Table 4.1: The Set of Object-related Features Extarcted by the Visual Analysis System

A final step merges the outputs of the two processes by performing a pixel-wise multiplication of the edge and object masks. A contour finding algorithm [46] identifies connected regions in the resulting binary image. The integration deals effectively with partial occlusions of non-similar colored objects.

Figure 4-1, ten objects are identified and assigned random indices.

4.2 Object Properties

Table 4.1 lists the set of object properties related features extracted by the visual analysis module in Fuse. The set of features described in the table form an exhaustive representation of objects with respect to the scene description task. [36] has successfully used these features to develop a spoken language generation system for a similar scene description task.

4.3 Inter-object spatial relations

To enable Fuse to ground the semantics of spatial terms such as "above" and "to the left of", a set of spatial relations similar to [32] is measured between each pair of objects. The first feature measured 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 spatial feature measures the

angle of the line that connects the two most proximal points of the objects. The fourth feature measures the distance between the two most proximal points of the objects.

4.4 Summary

The visual scene analysis module in Fuse captures a video frame during the onset of each spoken language utterance, extracts object properties and inter-object spatial relations. Each object is represented by a ten-dimensional feature vector consisting of three color features, five shape features and two position features. The spatial relations between pairs of objects are represented by four additional spatial features. The set of features are passed to the language model generation and visual attention generation modules.

Chapter 5

Speech Recognition

This chapter describes the speech recognition module of Fuse. Fuse has a conventional speech recognition system that can process a medium size vocabulary of more than 1000 words in real time.

5.1 The Speech Recognition Paradigm

The speech recognition paradigm [29] aims to search for the word string W' such that

$$W' = \arg\max P(A|W)P(W)$$

where $P(A|W)$ is the probability of the observed acoustic input A conditioned on a word string W, and $P(W)$ is the probability of the occurrence of the word string W.

Like most current speech recognition systems, the speech recognition module in Fuse constitutes of

- **Acoustic Front-End**: The acoustic front-end is responsible for spectral feature extraction from acoustic waveforms to obtain sequences of multi-dimensional observation vectors, A'
- **Utterance Segmenter**: The utterance segementer divides a stream of observation vectors into speech segements or silence segments. It passes the speech segments A to the decoder.

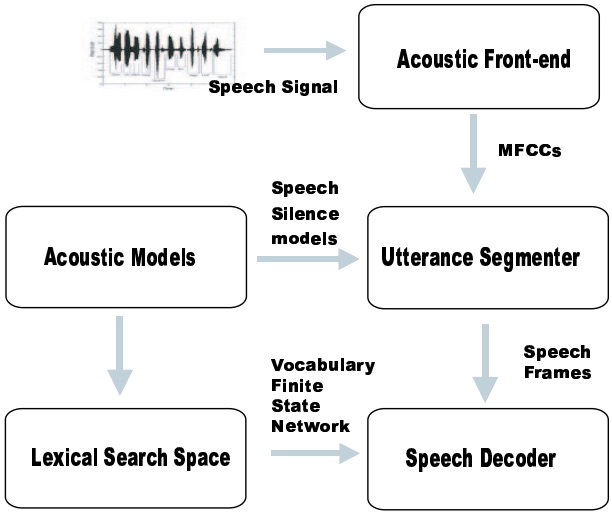

Figure 5-1: The speech recognizer in Fuse

- **Acoustic Models**: Statistical models are trained for all phonemes in the vocabulary. Since word strings are modeled as concatenation of phonemes, the phoneme models are used to estimate $P(A|W)$

- **Language Model**: Conventionally, a statistical language model provides the decoder with estimates of the probability of word strings $P(W)$. In Fuse, the language modeling components of the architecture provide visual-contextual knowledge to the decoder.

- **Lexical Search Space Generator**: The lexical search space generator provides a finite state network comprising all the words related to the task to the decoder.

- **Decoder**: Given an utterance, the decoder uses the acoustic and the language models to search for the best path in the finite state search space that maximizes $P(A|W)P(W)$. The decoder outputs a word lattice corresponding to more probable word strings.

The following subsections describe these components except the language modeling component, that will be described along with the dynamic language modeling components of the architecture in subsequent chapters.

Figure 5-2: The acoustic front-end in Fuse

5.2 Acoustic Front-end

Figure 5-2 gives an overview of the acoustic front-end component in Fuse [47]. The audio from the microphone is pre-emphasized to compensate for the radiation load, which sound undergoes as it exits the mouth. The waveform is then windowed with a Hamming window of length 20 ms, and the window is moved in increments of 10 ms. across the length of the waveform. This creates a vector of features every 10 ms. The samples within this window are passed through a Mel Scale filter bank to determine the amount of each frequency range present in the signal. Following this, the log of each filter bank coefficient is taken. As the average is subtracted off of each component at the end, this allows the feature extraction to concentrate mostly on feature differences. Next, the output of the filter banks is transformed using a discrete Cosine Transform (DCT). 12 filters are used in the bank of filters resulting in 12 transformed coefficients. Liftering is done to attempt to smooth these coefficients. Log energy is included as an extra coefficient. In addition, as the waveform is not stationary, the set of coefficients are augmented to include first and second order derivatives.

24-band Mel scale filters were used in Fuse, and with the data given being sampled at 16 kHz, the front-end generates 10 ms acoustic frames consisting of 39 dimensional feature vectors.

5.3 Utterance Segmenter

The utterance segmenter divides the audio stream into speech and silence segments using a four-state finite state machine to smooth out occasional silence frames in speech segments, and speech frames in silence segments. The model is shown in Figure 5-3. The design, implementation and evaluation details of the utterance segmenter

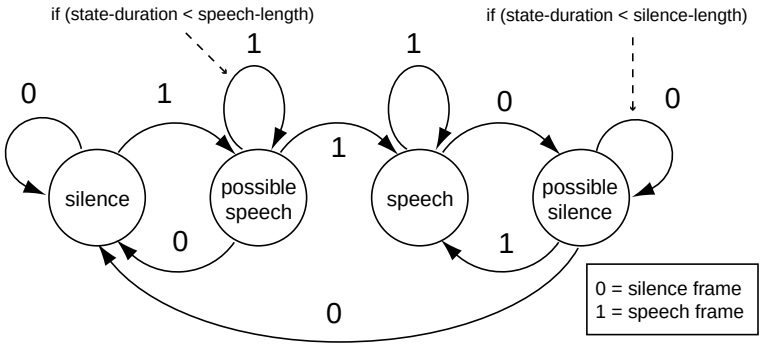

Figure 5-3: Four State Speech Silence Model

component can be found in [48].

The segmenter is initialized in the silence state, and transitions to possible speech, speech, then possible silence based on the input signal. If the state machine is in the speech or silence state, it will stay in that state as long as the speech/silence classifier makes the same classification for an incoming frame. If a differing classification is given for the frame, it switches to the possible-speech or possible-silence state. The possible-speech and possible-silence states insure that a pre-specified number of consecutive frames are necessary in order for a transition between speech and silence segments to occur. The segmenter stays in the possible-speech or possible-silence state until a pre-specified number of consecutive frames with the opposite prediction has occurred. If this specified number of frames do not occur, the state reverts down to the previous speech or silence state. If the minimum number of consecutive frames with the opposite prediction occurs, the finite state machine transitions into the opposite state. In the speech state, feature vectors of the incoming signal are sent on to other parts of the program, whereas in the silence state, information is not sent on to other parts of the system.

5.4 Acoustic Models

Acoustic models are represented as three state left-to-right Hidden Markov Models (HMMs) [30] for each of the phonemes present in the vocabulary. The HMMs are

initialized through a Segmental K-means algorithm [31]. The output and transition probabilities can be iteratively improved through the Baum-Welch method [1] for Expectation Maximization [7]. However, this method may only reach local maxima, so it is important to initialize these probabilities to reasonable estimates. The transition weights for the HMM arcs are stored as matrices. The probability density function associated with a state is modeled as a continuous density mixture of gaussians.

Ideally the base units chosen for word recognition are independent of context, so that they don't depend on the other units around them. In practice, however, the way phonemes are pronounced varies depending on the phonemes around them. Much of this has to do with co-articulation, where the movements made by the mouth and vocal cords are coordinated to produce smooth speech. This coordination means that adjacent phonemes are produced from one fluent motion, and thus the pronunciation of any phoneme is tied to the other phonemes around it. In order to deal with this effect, models that depend on the context are developed. The speech recognition system uses four models for a phoneme

- "Triphones", phoneme models that depend on the previous and following phonemes,
- "left biphones", phoneme models that depend on the previous context
- "right biphones", phoneme models that depend on the following phonemes
- "uniphones", phoneme models that are context independent

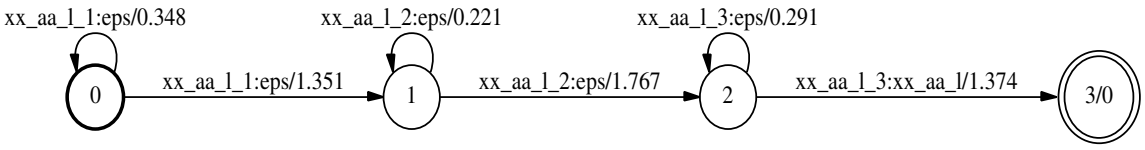

Figure 5-4: Graphical model of a right biphone

Figure 5-4 represnts a right biphone HMM for the phoneme "aa" with phoneme "l" in the right context. Details of implementation of acoustic models can be found in [26].

5.5 Lexical Search Space

Like most state-of-the-art speech recognizers, the search space of Fuse is represented by a tree-structured finite state transducers network constituting the set of the words and the phrases belonging to the vocabulary . In the simplest form, every word in the vocabulary is represented through its pronunciation form using phonemes. Word models are represented as linear sequences of the underlying phoneme HMMs. Figure 5-5 illustrates word models for "red" and "green" using context-dependent phonemes.

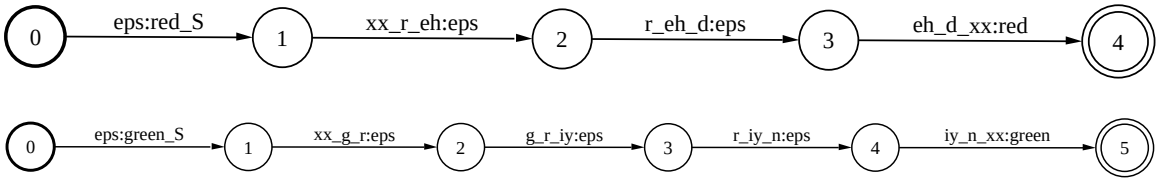

Figure 5-5: Graphical models of the words "red" and "green"

The following algorithm generates word models in terms of the acoustic model parameters.

- Generate an union of all finite state transducers corresponding to the context dependent phonemes that make up the vocabulary.
- Generate a finite state transducer that is equivalent to the Kleene closure [14] of the union of above finite state transducers.
- Compose the above transducer network with each of the word models.

Figure **??** in Appendix **??** illustrates the modified word models for the word "red". The search space is then generated as given by the following algorithm

- Generate union of the entire word model Finite State Transducers comprising the vocabulary.
- Remove epsilon transitions from the generated Finite State Transducer
- Determinize and minimize the above Finite State Transducer

- Apply Kleene Closure to the Finite State Transducer generated in Step 3

Figure ?? in Appendix ?? illustrates the lexical search space for a vocabulary of two words: "red" and "green".

Details of the finite state transducer algorithms such as union, composition, determinization, minimization and Kleene closure can be found in [24, 27]

5.6 Speech Decoder

To start the recognition process, the tree-structured lexicon is loaded into memory. The transitions in the tree-structured lexicon contain information regarding the context of phonemes, the phoneme HMM states distributions and the transition weights. Transitions can take place on several different levels.

- Between states in a phoneme HMM model. This is determined by the transition matrix estimates for the phoneme. The lexicon finite state transducer takes care of the transition values in respective arcs.

- Between phonemes in a word. If the transition matrix allows a transition out of the phoneme model, it transitions to the next phoneme of the word. The lexicon finite state transducer incorporates the information in its within-word transition arcs.

- Between different words. Since a tree-structured lexicon is used, it does not necessarily mean that word-starts follow word-ends. This happens because of the overlap between first few phonemes of different words. However, once final states in the lexicon finite state transducer corresponding to word ends are reached, the lexicon needs to be conceptually expanded to incorporate the probability of transitioning to new words (supplied by the language model estimates). This is made possible by bookkeeping the list of words associated with each phoneme in the transition arcs. This enables the incorporation of language model estimates when the decoder makes transition from a word-end. The cost

of these transitions are determined by the situationally-aware language models generated by Fuse.

The decoder processes observation vectors corresponding to speech segments and outputs the single best word string corresponding to the speech signals as well as partially ordered word graphs [21] containing multiple probable word strings. The decoding algorithm is as follows:

- For all the current hypotheses, make transitions from current active states to new states as allowed in the tree-structured lexicon.
- Add the transition costs as described above.
- Merge two hypotheses when the word-ends in their paths are same, keeping the more probable one.
- Update the state probabilities by adding their probability of creating the current observation vector to their probability.
- Prune states if their probabilities fall below a threshold.
- Mark the remaining states as active hypotheses.
- For all active hypotheses that mark word-ends, append a finite state automaton with a new state corresponding to the word and a transition from the last word-end in its path. Weigh the transition with the posterior probability of the word.
- Repeat the above steps until all the observation vectors in a speech segment are processed.
- At the end of this procedure, a simple back trace starting from the most probable active hypothesis provides the best guess of the most likely word se quence.
- The resulting finite state automaton is determinized and minimized [27] resulting in a reduced search space

Figure **??** shows an example of a compact word lattice.

5.7 Summary

This chapter describes the main components of the speech recognition module in Fuse. Like most speech recognizers, the speech recognition module in Fuse uses acoustic and language models to search for best word strings corresponding to the speech signals. The main components of the speech recognition module are similar to most state-of-the-art systems except the fact that Fuse uses situationally-aware dynamic language models to perform speech decoding.

Chapter 6

Language Modeling Component for Early Integration

This chapter details the language modeling interface in Fuse that performs the early integration of visual contextual knowledge into the lexical search space during the decoding of the incoming utterance. The chapter starts by recapitulating the basics of class-based n-gram language modeling. Algorithms related to the acquisition of visually salient words and phrases and generation of scene-wise visually driven language models are described in the following section. The rest of the chapter deals with the interface of early integration of visual context in the speech decoding process through dynamic update of visual attention, thereby, leading to more likely interpretation of spoken language utterances.

6.1 Class-based n-gram Language Model

This section briefly describes n-gram statistical language models that serve as the basis of the proposed cross-modal integration. Given a word sequence W such that $W = w_1, w_2, w_3, ..., w_n$, the probability of the occurrence of such a word sequence is

given by:

$$P(W) = P(w_1, w_2, w_3, ..., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2, w_1)...P(w_n|w_{n-1}...w_1) \tag{6.1}$$

For n = 2, the above model assumes the Markovian property that the occurrence of any word depends only on the previous word in the word string. Therefore

$$P(W) = P(w_1, w_2, w_3, ..., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_n|w_{n-1}) \tag{6.2}$$

Words may be clustered into equivalence classes leading to class based n-gram models. The set of words may be partitioned into word classes through a function that maps every word w to its corresponding word class $c(w)$. This results in:

$$P(W) = P(c(w_1))P(w_1|c(w_1))P(c(w_2)|c(w_1))P(w_2|c(w_2))...P(w_n|c(w_n)) \tag{6.3}$$

where $P(c(w_i)|c(w_{i-1}))$ denotes class-to-class transition probability while $P(w|c(w))$ denotes class conditional probability. Equation 6.3 can be used to factor $P(w_i|w_j)$ as

$$P(w_i|w_j) = P(w_i|c(w_i))P(c(w_i)|c(w_j)) \tag{6.4}$$

Fuse uses class based language models for the following reasons:

- Class based language models make better use of limited training data to make generalized predictions for word sequences that may not occur in the training data.

- Class based language models involve the estimation of class conditional probability mass functions. In context of Fuse, this generation procedure depends on the visual context. The visual context does not affect the underlying grammar. Hence, the class-to-class transition probabilities in Fuse can be estimated statically from the training data.

6.2 Visual Context Driven Language Model

Visual context driven language modeling addresses some of the basic requirements mentioned below:

- The language model should encompass descriptions or expressions that fit all objects in the scene. This is due to the fact that Fuse cannot afford to bias towards individual objects at the onset of an incoming utterance.

- For each object, the language model should consider all possible combinations of words and phrases that speakers may use to describe it. This is due to the fact that speakers may describe individual objects in various ways.

- The language model should guarantee visual generalization while describing individual objects.

We now summarize our approach below. Details of the algorithms and implementation can be found in subsections 6.2.1, 6.2.2 and 6.2.3.

- A vocabulary of visually salient words and pharses and the underlying grammar is acquired from the training dataset. The sematics of the visually salient words are associated with the set of visual features described in chapter 4.

- Descriptions are generated for each object in the current scene using the acquired knowledge base.

- For each object, multiple descriptions of different lengths weighted by their semantic relevance are generated. For example, the word "red" is weighed more than the word "blue" for a red object.

- The resulting descriptions are then weighted by their contextual relevance in the given scene. Ambiguous descriptions are given less weighting than the more salient ones.

- A class-conditional language model is created for each of the objects in the scene. The class-conditionals estimate probability distributions of words and phrases conditioned on the objects in the scene.

- A weighted mixture of all the class-conditional language models is used by the speech decoder. The individual class-conditional language models remain fixed during the processing of an utterance. However, the relative weighting of individual submodels is dynamically updated using partially decoded hypotheses.

To deal with utterances that consist of target as well as landmark objects, the language model also needs to take spatial terms into account. The approach is described in subsection 6.2.3. The following subsections describe the set of algorithms for learning and generating visually grounded language models.

6.2.1 Learning The Description Model

This subsection describes the set of algorithms that have been developed for acquiring the structures necessary to produce object related descriptions. Details of the learning methods may be found in [36]. All parameters of the description model are learned from examples of objects embedded in scenes that are labeled with descriptive phrases. A set of 60 training examples were collected from each participant, resulting in a total of 476 examples in the training dataset described in section 3.1. The examples were chosen such that the accompanying spoken language descriptions are simple expressions referring to individual objects.

Detection of Visually Salient Words / Phrases by Feature Selection

Feature selection proceeds by assigning features to each word on an individual basis. Features for an individual word are selected to maximize the symmetrized KL distance [36] between the word conditional distribution and the background unconditioned

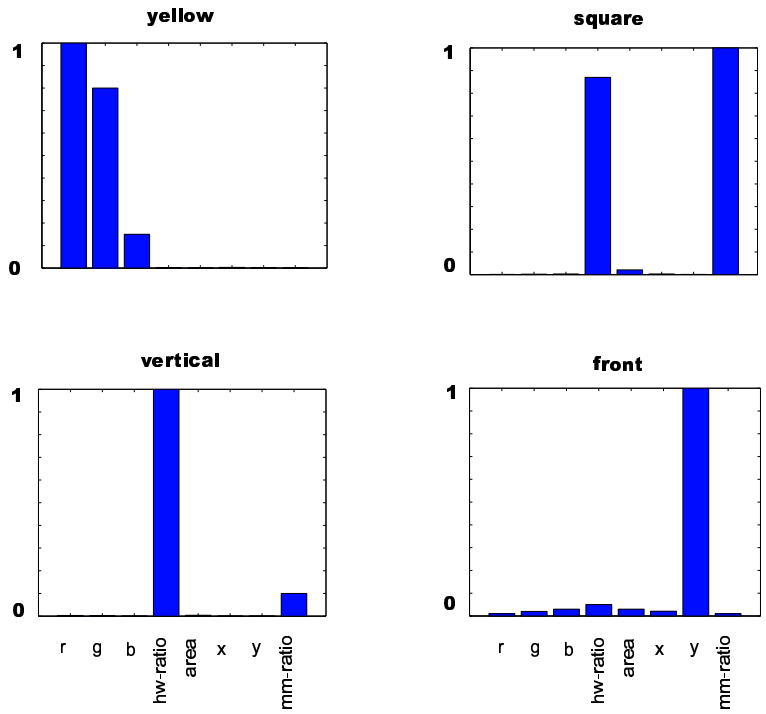

Figure 6-1: Feature associations with visually salient words.

distribution,

$$KL(p_1(x)||p_2(x)) = \frac{1}{2}tr(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I) + \frac{1}{2}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) \tag{6.5}$$

where Σ is the full covariance matrix and μ is a mean vector.

Feature selection is achieved using a greedy algorithm. It starts by selecting the single feature which leads to the highest symmetrized KL distance between conditioned and unconditioned distributions. The feature selection procedure iteratively finds the next best feature which maximally increases equation 6.5. Each KL distance is normalized by the number of selected features (i.e., the number of dimensions). After each feature is added, the increase in normalized KL distance is computed. No further iterations are made when no increase is obtainable.

Visually salient words are retrieved from this feature search algorithm. If the selected feature set for a word is a subset of the entire visual feature set described in Table 4.1, the word is denoted as a visually salient one. Words that are not able to pull up discriminative feature subsets are marked as ungrounded or visually irrelevant.

Word Class	Class Members
C0	large, big, small, little, bigger, smaller
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 6.1: Word Classes and their Members

Figure 6-1 illustrates a couple of visually salient words pulling out subsets of features from the training data.

Word Class Formation

Once visually relevant words are retrieved from the training examples, they are clustered into word classes. Visually ungrounded words serve as singleton equivalence sets.

1. Clustering based on syntactic similarities: A bigram-pair occurence indicator variable is defined as follows:

$$I(w_i, w_j) = 1, \quad \text{if } w_i \text{ follows } w_j \text{ in the corpus}$$

$$I(w_i, w_j) = 0, \quad \text{otherwise}$$

A distortion metric between two word classes is defined as

$$d_1(C_i, C_j) = \Sigma_{k=1}^{N_i} \Sigma_{l=1}^{N_j} (\Sigma_{m=1, m \neq C_i(k), m \neq C_j(l)}^{V} (I(C_i(k), w_m) - I(C_j(l), w_m))^2) \tag{6.6}$$

where V is the vocabulary size.

A greedy algorithm is used to merge clusters as follows

- Begin with K = V clusters.

- Find C_i and C_j such that $d_1(C_i, C_j)$ falls below a chosen threshold, $1 \leq i, j \geq K$
- If there are no such clusters, then stop
- Merge the elements of clusters that do not contain visually ungrounded words.
- Goto Step 2

On termination, the visually relevant words get clustered into equivalence classes while the ungrounded words remain as singleton word classes.

2. Clustering based on semantic similarities: This method clusters words that co occur in similar visual contexts. Details of the clustering algorithm can be obtained from [36].

A weighted combination of the above clustering algorithms is used to generate a set of visually salient word classes. Table 6.1 enumerates the list of word classes acquired by Fuse.

Grounding Word Semantics

For each visually salient word, a multivariate Gaussian model is estimated using the observations which co-occur with that word. Each visually relevant word class inherits the conjunction of features assigned to its members. The word-conditional model specifies a probability density function (pdf) over the subset of visual features which have been inherited by the corresponding word class. Table 6.2 enumerates the set of features associated with individual word classes

Word Class	Feature Subset
C8	area, mm-ratio
C10	area, mm-ratio
C11	x,y
C12	x,y
C13	r, g, b
C25	hw-ratio,

Table 6.2: Feature Subsets Associated with Individual Word Classes

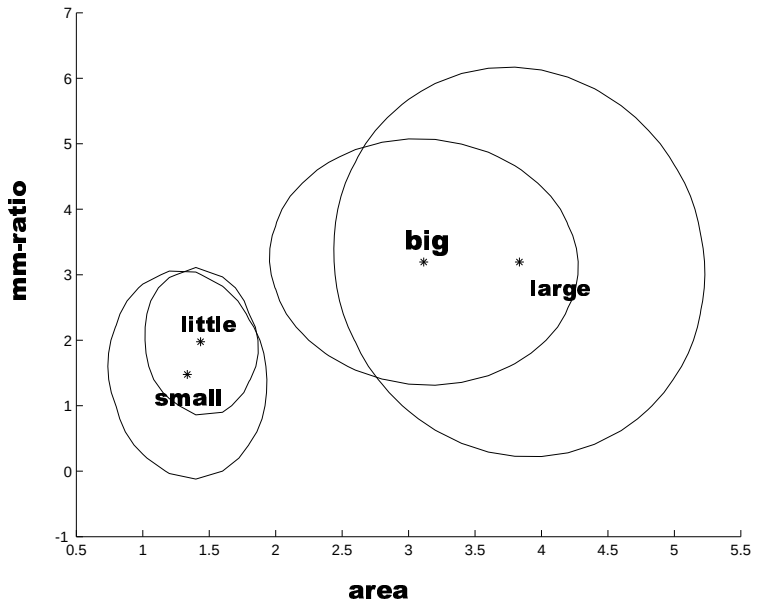

Figure 6-2: Example of a word class with four members.

Figure 6-2 shows the visual models associated with some of the members of the word class "C8" in Fuse.

Class-based Bi-gram Language Model

A class-based bigram statistical language model [3] is estimated (based on class frequencies) to model the syntax of referring expressions. Visually "ungrounded" words form singleton word classes (classes with only one member). The bigram statistical language model representing the word class syntax is learned from the training data in a leave-one-speaker-out form. Due to the fact that the underlying grammar of referring expressions does not change with visual context, the class based bi-gram model remains static irrespective of changes in the visual scene.

Figure 6-3 shows a subset of phrase level bigrams in the form of a transition network. Each arc is labeled with the transition probability between the connected words / word classes. Any path through this network constitutes a possible description of an object.

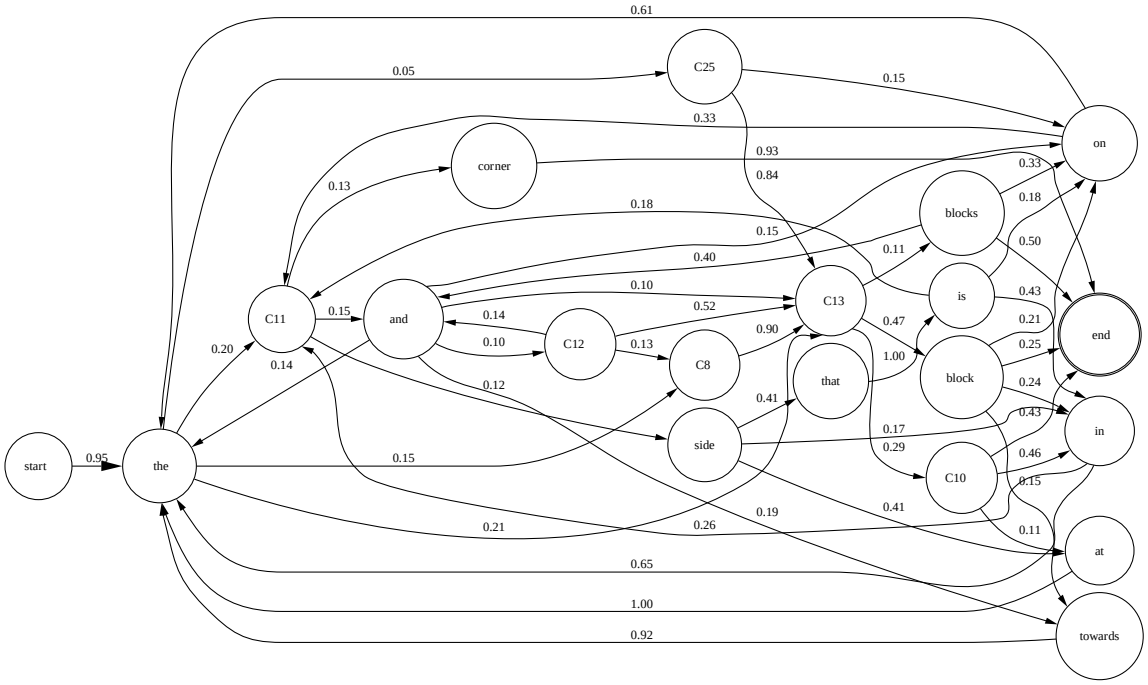

Figure 6-3: The bi-gram finite state network used to generate descriptions of objects. To allow legibility, the full grammar has been pruned for the figure (18 of 55 nodes are shown).

6.2.2 Generation of Dynamic Language Model

In our implementation, the speech recognizer requires a language model consisting of a set of word bigram transition probabilities (refer to chapter 5). As Equation 6.3 shows, the word bigram can be obtained from the product of word class transition probabilities $P(c(w_i)|c_(w_{i-1}))$ and class conditional word probabilities $P(w_i|c(w_i))$. The word class transition probabilities are fully determined from training data (Figure 6-3) and remain static during speech processing. The transition probabilities between classes remain static with respect to the immediate visual context. However, the probabilities of words within each word class do depend on visual context. There are two aspects in generating such probabilities.

- Given a scene, the language model generates word probabilities conditioned on the objects present in the scene. The model is reestimated when a new scene is presented to the system.

- During the onset of an utterance all objects are given equal priors. The priors are updated based on the partially decoded hypotheses obtained from the speech decoder.

Therefore, class conditional word probabilities are dynamically estimated as a function of the scene and visual attention using a six step process:

1. A graph-theoretic path search algorithm enumerates all left-to-right paths connecting the *start* and *end* states of the finite state transition network. The process results in a set of N variable length sequences $\{C_1, C_2, \ldots, C_N\}$, where each sequence C_i consists of a ordered set of T_i word classes:

$$C_i = c_i^1, c_i^2, \ldots, c_i^{T_i} \tag{6.7}$$

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

2. Given a sequence C_i and an object O_j the word classes $C_i = c_i^1, c_i^2, \ldots, c_i^{T_i}$ are mapped to a sequence of words $W_i^j = w_{ij}^1, w_{ij}^2, \ldots, w_{ij}^{T_i}$. There are two cases:

   - A word w_{ij} belongs to a visually irrelevant word class. Since these classes are singleton sets, the single word member is chosen.
   - A word w_{ij} belongs to a grounded word class. In this case, w_{ij}^k is selected according to:

$$w_{ij}^k = \overset{\text{argmax}}{m}\ p(O_j|w_m),\ \ c(w_m) = c_i^k,\ \ m = 1 \ldots n \tag{6.8}$$

     where $p(O_j|w_m)$ is derived from the visually grounded model associated with the word w_m.

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

3. Similar to the procedure described in [36], each description is evaluated by computing the product of the word conditional probabilities of the observed object properties, which is equivalently expressed as a sum of log probabilities:

$$fit(W_i^j, O_j) = \frac{\Sigma_{t=1}^{T_i} \log p(O_j|w_{ij}^t)}{G(C_i)} \tag{6.9}$$

where $G(C_i)$ is the number of visually grounded word classes in the sequence C_i, and $p(O_j|w_{ij}^t))$ evaluated 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 a constant. The denominator term normalizes effects due to the length of the description.

The fitness function measures how well a descriptive phrase matches the properties of the target object.

4. The fitness function does not take into account contextual effects due to other objects in the scene. For example, a description may be ambiguous over several objects if two or more objects are visually similar. To capture contextual effects, [36] defines a context-sensitive fitness, which is assigned to the source word class sequence:

$$\psi(C_i, O_j) = fit(W_i^j, O_j) - max_{k \neq j} fit(W_i^j, O_k) \tag{6.10}$$

5. For a given object and word class sequence, object conditional probabilities are assigned to each visually grounded word:

$$P(w|O_i, c(w)) = \frac{p(O_i|w) \sum_{C_j, C(w) \epsilon C_j} \psi(C_j, O_i)}{\sum_{k=1}^{M} p(O_k|w) \sum_{C_j, C(w) \epsilon C_j} \psi(C_j, O_k)} \tag{6.11}$$

where $c(w)$ is the word class to which w belongs.

Such a framework presents a balance between the visual fitness criterion as well as the contextual fitness criterion. It can handle situations where

- If two words are visually salient to a similar extent for a given object , they are given higher priorities by the fitness criterion, irrespective of whether they tend to increase ambiguity in the scene.
- If two words describe the object equally, the contextual fitness criterion ensures less weightage of the word that tends to increase ambiguity in the scene

6. The final step is to weigh the influences of all objects in the scene:

$$P(w|c(w)) = \sum_{i=1}^{M} P(w|O_i, c(w))P(O_i) \tag{6.12}$$

Relative emphasis of objects is controlled by Fuse's visual attention state, $P(O_i)$, described in section 6.4.

The above procedure generates a set of class conditional word probabilities 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. The class conditionals generated can be used to compute bi-grams as given by equation 6.3. As Fuse updates its visual attention based on partial linguistic information, the words that describe objects that have a larger attention spread become more probable and thus bias the speech recognizer search towards them. Since visual context in this domain does not affect order of words, the grammar remains static throughout the processing

6.2.3 Generation of Spatial Language Model

- **Learning of Spatial Terms**: To learn spatial relations between a target and a landmark object, a user interface was created to enable the user to load an image on the screen and type in spatial phrases such as "above", "below" and "left_of". Participants were instructed to select two objects from the scene that he or she found suitable to serve as the target and the landmark objects for the spatial phrase. Spatial relations between the target and the landmark objects are extracted by the visual analysis system and a multidimensional Gaussian

model is computed for each of the spatial lexical items using all observations that co-occur with the lexical item.

- **Spatial Clauses Grammar**: Phrase bigrams are used to model the use of relative spatial clauses. For example, "The red block beneath the small green one" contains references two objects, the target and a *landmark* ("the small green one"). The spatial relation "beneath" describes the relation between target and landmark. Figure 6-4 illustrates the spatial clauses grammar. The states "TARGET OBJECT" and "LANDMARK OBJECT" correspond to the class bigram network shown in figure 6-3.

- **Grounding of Spatial Terms**: Spatial connective terms may consist of multiple words (e.g., "to the left of") but are tokenized and treated as a single acoustic unit during speech decoding. Each spatial term is grounded in a Guassian model that is defined over the spatial features described in Chapter 4.

- **Generation of Spatial Language Model**: All distinct paths connecting the *start* and *end* nodes of the transition network are enumerated. 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 the form:

$$C_i = \text{(target object, spatial\_term, landmark object)}$$

For all combinations of target and landmark object pairs, each of the sequences are evaluated as follows:

$$\psi(O_i, w, O_j) = logP(w|O_i, O_j)T(w|target)T(w|spatial), \;\; i \neq j \qquad (6.13)$$

where $P(w|O_i, O_j)$ is derived from the grounded models while the following terms are derived from the spatial grammar.

The spatial language model is estimated through the following equations:

$$P(w|O_i, c(w)) = \Sigma_{j=1, j\neq i}^{N} \psi(O_i, w, O_j), \;\; i \neq j, w \in spatial\ terms \qquad (6.14)$$

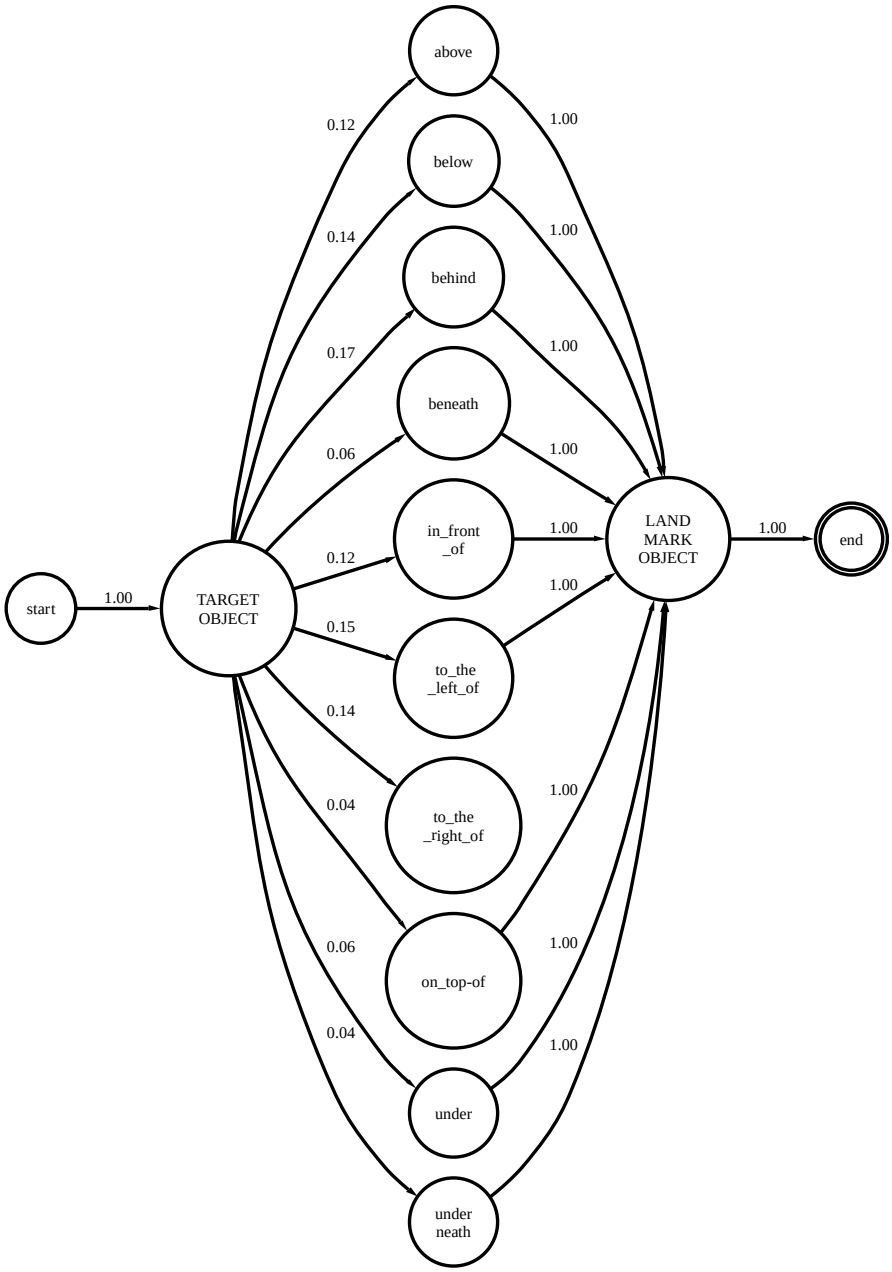

Figure 6-4: The probabilistic grammar used to generate descriptions with relative spatial clauses. A pruned grammar has been shown in the figure.

$$P(O_j|w, O_i) = \frac{\psi(O_i, w, O_j)}{\Sigma_{j=1, j\neq i}^{N} \psi(O_i, w, O_j)}, \quad i \neq j, w \in \textit{spatial terms} \tag{6.15}$$

6.3 Integration of Visually Steered Models in Speech Recognition

As we have mentioned in chapter 5, the speech recognition paradigm [29] aims to search for the word string W' such that

$$W' = \arg\max P(A|W)P(W)$$

where $P(A|W)$ is the probability of the observed acoustic input A conditioned on a word string W, and $P(W)$ is the probability of the occurrence of the word string W. The effect of visual contextual information does not play any role to infer on $P(A|W)$ since they are estimated f rom the acoustic signal alone. The visually steered linguistic models generated by the methods described above are used to evaluate $P(W)$.

A class-based language model approximates $P(w_i|w_{i-1})$ by

$$P(w_i|w_{i-1}) = P(c(w_i)|c(w_{i-1}))P(w_i|c(w_i)) \tag{6.16}$$

where $P(c(w_i)|c(w_{i-1}))$ is the estimate of the probability of the wordclass $c(w_i)$ following $c(w_{i-1})$ and $P(w_i|c(w_i))$ is the class conditional distribution of w_i in the wordclass $c(w_i)$.

Probability mass functions in the form of $P(c_i|c_{i-1})$ are estimated from the static class based bigram language model. The dynamic language models described in the previous sections are used to evaluate $P(w_i|c(w_i))$.

Three independent cases are considered:

- w_i is a visually grounded word depicting an intra-object property such as 'color', 'shape' and 'size': We linearly interpolate the generated class conditionals over the number of objects in the scene to evaluate $P(w_i|c(w_i))$

$$P(w_i|c_i) = \Sigma_{j=1}^{n} \lambda_j P(w_i|c(w_i), O_j) \quad \Sigma_{j=1}^{n} \lambda_j = 1 \tag{6.17}$$

where n is the number of objects in the scene.

- w_i is a visually grounded word depicting a spatial relation such as 'above', 'below', etc.: We perform a linear interpolation over the spatial language model. The interpolation is in the form of

$$P(w_i|c_i) = \Sigma_{j=1}^{n} \lambda_j P(w_i|c(w_i), O_j) \quad \Sigma_{j=1}^{n} \lambda_j = 1 \tag{6.18}$$

  where O_j is the target object and n is the number of objects in the scene. $P(w_i|c(w_i), O_j)$ is derived from the generated spatial language model using Equation 6.14

- w_i is a visually ungrounded or visually irrelevant word such as 'the', 'by': $P(w_i|c(w_i))$ is estimated from the static leave-one-speaker-out corpus.

The priors provide a measure of visual attention in context of the current scene. Their values are determined dynamically from active partial utterance hypotheses during speech decoding.

6.4 Using Incremental Speech Processing to Drive Visual Attention

The visual attention generator module enables the early integration of visual context to provide dynamic re-estimation of the priors associated with the interpolated class conditional probabilities. In other words, the model uses the visual context to immediately determine the attention distribution spread over the objects in the current scene. Given a partial utterance hypothesis, the model rank-orders and scores each object in the current scene based on the visual semantic fit over the partially decoded utterance. These scores are used as the interpolation weights to calculate the class conditional in the form of $P(w_i|c(w_i))$.

The priors $\lambda_j, \quad j = 1, ..., n$ get dynamically updated when the decoder search algorithm leaves a state that marks the end of a word w_m. The partial hypothesis

is of length m at this point. From the nature of the collected utterances, there exist three cases that are described below:

- w_m is a visually relevant word depicting intra-object property, for example "large", "vertical", etc. Here, the update rule is as follows

$$P(O_j)_{w_m} = P(O_j|w_m)P(O_j)_{w_{m-1}},\ j = 1, 2, ..., n \tag{6.19}$$

where $P(O_j|w_m)$ is derived from the visully grounded models. Therefore

$$\lambda_j = \frac{P(O_j)_{w_m}}{\Sigma_{i=1}^{n} P(O_i)_{w_m}}, \quad j = 1, 2, ..., n \tag{6.20}$$

- w_m is a visually relevant word / phrase depicting a spatial relation, for example "above", "beneath" and so on. Here, the update rule is in the form

$$P(O_j)_{w_m} = \Sigma_{i=1, i\neq j}^{n} P(O_j|w_m, O_i)P(O_i)_{w_{m-1}} \tag{6.21}$$

$$j = 1, 2, ..., n$$

where $P(O_j|w_m, O_i)$ is derived from the dynamic spatial language model using Equation 6.15.

Again,

$$\lambda_j = \frac{P(O_j)_{w_m}}{\Sigma_{i=1}^{n} P(O_i)_{w_m}}, \quad j = 1, 2, ..., n \tag{6.22}$$

- w_m is a visually irrelevant or ungrounded word such as "the", "by", etc. In this case, we have the following update rule:

$$P(O_j)_{w_m} = \gamma P(O_j)w_{m-1},\ j = 1, 2, ..., n \tag{6.23}$$

where γ is a constant score given to the likelihood of ungrounded words. The priors are updated by the same rules described above.

6.5 Spoken Language Understanding for Visual Focus of Attention

Fuse performs spoken language understanding to identify target objects referred to by the spoken language descriptions besides visually grounded speech recognition. Given a spoken language input referring to individual objects in a scene, Fuse receives output from the speech decoder in the form of word strings, $W = w_1 \ldots w_N$. Fuse is able to handle two classes of expressions listed below

- **Simple expressions**: In the case of a simple referring expression, Fuse selects the object with greatest visual attention:

$$object' = \underset{i}{\text{argmax}} \; P(O_i)_{w_N} \tag{6.24}$$

- **Complex expressions**: 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:

$$object' = \underset{i}{\text{argmax}} \; P(O_i)_{w_{m-1}} \sum_{j=1, j \neq i}^{M} p(O_j | w_m, O_i) P(O_j)_{w_N} \tag{6.25}$$

  where $p(O_j | w_m, O_i)$ is derived from equation 6.15

6.6 A Detailed Example of Visually-Steered Speech Processing

To make the interaction between visual attention and speech processing more concrete, we take a closer look at an example. Shown below is the transcription of a sample utterance, the output of the speech decoder using standard bigrams without use of the visual attention model, and the de coder's output using visual attention:

Transcript	The large green block in the back above the yellow block.
No visual context	[The] lower green block in the back of [the] yellow block
Visual context	The *large* green block in the back *above* the yellow block

Table 6.3: Speech Recognition Output with and without Visual Context

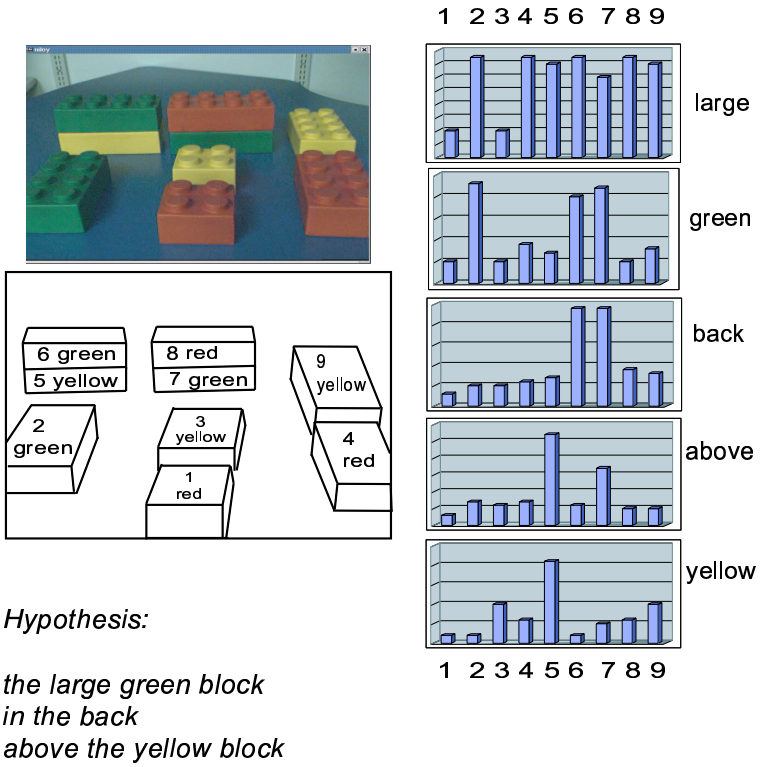

Figure 6-5: Evolution of attention during processing of the utterance, "The large green block in the back above the yellow block".

The evolution of visual attention is illustrated for this example in Figure 6-5. Each plot shows the spread of attention across the nine objects in the scene after integrating the words shown to the left of that plot. This occurs for all active partially decoded hypotheses but only the hypothesis containing the most likely word sequence is shown in the diagram. Words that have no visual salience have no effect on visual attention and are not shown.

6.7 Summary

Fuse acquires visually grounded vocabulary from a "show-and-tell" procedure in which spoken language descriptions accompany visual scenes from camera images. Given a new scene, Fuse generates a language model as a mixture of language sub models for each of the objects in the scene. Fuse also generates a spatial language model to handle relative spatial clauses in complex referring expressions. A dynamic visual attention model, driven by incremental partially decoded hypotheses, generates mixture priors at each time step in the speech decoding process. The visual attention module has been designed as a simple computational model of belief propagation for the purposes of the thesis. Future directions may include the design of a more complex computational model for visual attention.

Chapter 7

Evaluation

This chapter describes the set of experiments performed to evaluate Fuse. Section 7.1 details the data collection procedure. The following section evaluates the speech recognition / understanding performance of Fuse. The rest of the chapter presents an analysis of possible sources of errors in the system performance.

7.1 Data Collection

A corpus of 990 utterances paired with corresponding visual camera images was collected from eight speakers. The speakers were instructed to talk naturally and describe target objects such that a listener could later select the same target object from the identical scene . A data collection program was written that displays images and records spoken responses. A set of 60 images was captured from a camera embodied on a two-degree freedom robot in preparation of data collection. Each of the speakers wore a headset microphone. A user interface was designed to display a random image, highlight a random object in the image and record the spoken language response. This was repeated 100 times on an average for each of the speakers. The speech segmenter in Fuse was used to segment the captured audio into separate spoken utterances. Each utterance was saved as a separate file tagged with the identity of the corresponding image and the target object highlighted at the time.

Each spoken utterance was manually transcribed at the word level. The utter-

Type	Utterance
Simple	the red square
Simple	the large yellow rectangle
Simple	the large yellow block in front
Simple	the vertical green block
Simple	the frontmost and rightmost yellow block
Complex	the red block above the green block
Complex	the yellow block in the front beneath the red block
Complex	the green block to the left of the yellow block in the back
Complex	the vertical large green block in front of the small blue block
Complex	the large yellow block in the far right beneath a pile of yellow blocks
Complex	the large green block above the red and the yellow blocks

Table 7.1: Typical utterances in the scene description task

ances were divided into two types: simple referring expressions and complex referring expressions as have been described in the previous chapter. Table 7.1 lists some representative utterances from each of the types. Out of the 990 utterances collected, 476 simple examples were chosen to build the visually grounded word models.

A separate user interface was developed to collect examples corresponding to spatial clauses. The interface enabled the participants to display an image on the screen and select a spatial phrase such as "above", "below", "left_of" from a list of phrases. Participants were instructed to select two objects from the scene that he or she found suitable to serve as the target and the landmark objects for the selected spatial phrase. Each selected spatial phrase was saved separately tagged with the image identity as well as the target and landmark object identities.

7.2 Experimental Evaluation

To study the effect of visual context on speech processing, the speech recognition performance accuracy of the system was evaluated with and without the accompanying visual information. During the evaluation without visual context, the class conditionals were distributed equally among words occurring in the same word class for all visually relevant word classes. Table 7.2 reports speech recognition errors over the set of speakers. The speech recognition errors are measured using the standard

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 7.2: Speech recognition word error rates (%). 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 7.3: Speech understanding accuracy results (%). Averaged across all eight speakers, the early integration of visual context reduced the language understanding error rate by 41%.

NIST measurement package [23] that compares speech decoder with ground truth penalizing insertions, deletions and substitutions . The introduction of visual context led to a 31.3% reduction in word error rates averaged over eight speakers, a significant improvement over the baseline system.

To study the overall performance of the system components, namely, the visual system, the dynamic language model acquisition/generation system and the visual attention model, the spoken language understanding accuracy of the system was evaluated. The evaluation was performed in a leave-one-speaker-out fashion by comparing the target object identity chosen by the system to the identity of the object that was highlighted during the generation of the training dataset, given the visual

scene. Table 7.2 reports the speech understanding errors over the set of speakers with and without early integration of visual context in the speech recognition search process.

7.3 Analysis of Errors

Although the overall performance of the system improves with integration of visual context in speech recognition/understanding, there are shortcomings in the system that cause it to fail for a certain percentage of utterances. Mistakes in spoken language understanding occur due to several causes:

- Automatic Speech Segmentation Errors: The speech segmentation module in our real time speech recognition system sometimes merges utterances into one that should have been separate utterances. This attributes the combination of two descriptions to a single object identity.

- Descriptions with more than one landmark object: The system assumes that a complex referring expression consists of a target object description, a landmark object description and a relative spatial phrase connecting the two descriptions. The system can not handle cases where the referring expressions contain descriptions of more than one landmark objects in conjunction or groups of landmark objects.

- Error Propagation: Some errors are due to the fact that the attention distribution gets propagated frame by frame over the search space. Due to the feed-forward behavior of the algorithms, errors that creep in during the initial stages get propagated throughout the entire utterance.

- Visual Segmentation Errors: Some errors in understanding occur due to imperfect image segmentations performed by the visual system. Such segmentations may merge more than one objects or divide an object into two or more parts. These cause mismatch among descriptions and the corresponding objects

- Visual-Semantics Acquisition: A few errors are due to the fact that the visually grounded word models are not generalizable enough to capture variances in object related properties due to the absence of sufficient training data. These result in ambiguities in the visual attention model.

Chapter 8

Conclusion

This chapter presents a brief summary of the thesis along with suggestions for future work on the research. It also enumerates possible deliverables of the thesis and contributions in specific application scenarios.

8.1 Thesis Summary

This thesis demonstrates a visual context-aware speech processing (recognition understanding) system, Fuse, which supports natural human-computer conversational interaction in a simple situated non-mobile world consisting of configurations of objects placed between a robot and its human partner. The thesis explores a novel method of predicting occurrences of words and phrases (statistical language modeling) using visual contextual knowledge from a physical environment. Fuse receives visual input from a colored video camera and audio input from a microphone headworn by the user. During the onset of an utterance, Fuse performs dynamic language modeling based on the visual information contained in the immediate environment. Fuse uses these models to drive a medium vocabulary speech recognizer search to interpret more likely sequences of word strings corresponding to an input utterance. For the purposes of the thesis, it is assumed that the spoken language utterances refer to individual target objects in the environment. Fuse also performs spoken language understanding to select target objects that are referred to in the user speech.

The word and phrase predictions are dynamically updated in real-time as the human participant changes the configuration of the environment by adding, removing or substituting new objects in the environment.

The main components in Fuse are listed below:

- **Visual Scene Analysis:** This component is responsible for extracting object properties (color, shape, size) and spatial relations from an input camera image. The extracted information is passed to the language modeling components.

- **Speech Recognizer:** The speech recognition component performs searches for the most likely word strings embedded in speech segments using acoustic knowledge as well dynamic language models derived from the visual environment.

- **Visually Driven Language Model Generator:** This component generates probability distributions of occurrences of object-level words and phrases as well as relative spatial clauses based on the visual information.

- **Visual Attention Generator:** This component provides probability distributions of attention spread over all the objects in the scene. As the speech recognizer provides partially decoded hypotheses, the distributions are dynamically updated to prioritize objects that match the partial descriptions more than the others.

- **Spoken Language Understanding:** This component selects an object from the visual scene based on the speech recognition output.

8.2 Future Directions

The system presented in the thesis has several limitations. One of the primary limitations is that Fuse is able to handle only two classes of utterances referring to individual objects in a visual scene, simple expressions describing single objects and complex expressions containing descriptions of target objects along with individual landmark objects. Human participants can use various different strategies to describe objects in

a visual scene. Another drawback of the system is that the visual processing handles objects of simple shapes and uniform colors. Future work may assist in increasing visual complexity that will result in an increase in spoken language complexity. The visual attention generation module uses simple feed-forward propagation of attention based on partial linguistic evidence. More complex models of attention need to be derived as the task complexity increases. We are performing initial experiments to study human eye movement behavior patterns during on-line spoken language comprehension. Detailed study of human eye movements at various levels of sub word granularity may lead to develop more sophisticated visual attention models in Fuse.

8.3 Deliverables

- A novel language modeling interface that
  - generates dynamic predictions of words and phrases from visual information,
  - biases speech recognition search towards more likely interpretations
- A dynamic computational model of visual attention that dynamically propagates object priorities based on linguistic input.
- A medium vocabulary speech recognition system that can deal with more than 1000 words.

8.4 Application Scenarios

The research presented in this thesis may find applications in various areas. Some of them are described below:

- **Human-robot interaction:** The proposed approach or framework is primarily meant to assist human-robot interaction through spoken language conversations. The approach will be particularly useful in a manipulator robot that

performs spoken language interaction with a human partner in a shared visual environment.

- **Context-sensitive assistive aids:** Besides supporting visual context, the framework allows support for other sources of contextual cues such as geographical information and commonsense information. These sources of information may be integrated with the main framework in order to develop context-aware assistive aids. Context-sensitive medical transcription is a well-defined task that may use a visual context-aware spoken language understanding approach to assist flawless interpretations of health-related queries and solutions.

- **Behavioral studies:** The eye movement data processing interface can serve as a computational tool to automatically record and process eye movements of human participants. The tool may assist cognitive science researchers in studying human behaviors.

Bibliography

[1] L. E. Baum. *An Inequality and Associated Maximization Technique Occuring in Statistical Ananlysis of Probabilistic Functions of Markov Chains.* Inequalities, 1972.

[2] P. J. Brown, J. D. Bovey, and X. Chen. *Context-Aware Applications: from the Laboratory to the Marketplace.* Number 5. IEEE Pernonal Communications, 1997.

[3] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza andJennifer C. Lai, and Robert L. Mercer. *Class based n-gram Models of Natural Language*, pages 467–479. Number 4. Computational Linguistics, 1992.

[4] Eugene Charniak. *Statistical Parsing with a Context-free Grammar and Word Statistics.* Proceedings of the Fourteenth National Conference on Artificial Intelligence, 1997.

[5] C. Chelba and F. Jelinek. *Recognition Performance of a Structured Language Model*, pages 1567–1570. Proceedings of the Eurospeech, 1999.

[6] N. Chompsky. *Aspects of the Theory of Syntax.* MIT Press, 1965.

[7] A. Dempster, N. Laird, and D. Rubin. *Maximum Likelihood of Incomplete Data via the EM Algorithm*, pages 1–38. Number 1. Journal of Royal Statistical Society, 1993.

[8] A. K. Dey and G.D. Abowd. *Toward a better understanding of context and context-awareness.* GVU Technical Report GIT-GVU-99-23, 1999.

[9] K. M. Eberhard, M. J. Spivey, J. C. Sedivy, and M. K. Tanehaus. *Eye movements as a window into real-time spoken language comprehension in natural contexts.* Journal of Psycholinguistic Research, 1995.

[10] J. A. Fodor. *Modularity of Mind.* MIT Press, Cambridge, MA, 1983.

[11] A. L. Gorin, G. Riccardi, and J. H. Wright. *How May I Help You*, volume 23, pages 113–127. Speech Communication, 1997.

[12] A.L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright. *Automated Natural Spoken Dialog.* IEEE Computer Magazine, 2002.

[13] D. Hindle, A. Ljolie, and M. Riley. *Recent Improvements to the AT&T Speech-To-Text (STT) System.* Proceedings of ARPA Speech Recognition Workshop, February 1996.

[14] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley Publishing, 2000.

[15] Frederick Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, MA, 1997.

[16] Davy Temperley John D. Lafferty, Danny Sleator. *Grammartical trigrams: A Probabilistic Model of Link Grammar.* Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, 1992.

[17] Daniel Jurafsky, Chuck Wooters, Gary Tajchman, Jonathan Segal, Andreas Stolcke, Erfic Fosler, and Nelson Morgan. *The Berkeley Restaurant Project.* Proceedings of the Intl. Conference on Acoustics, Speech, and Signal Processing, 1995.

[18] C. Kamm, M. Walker, and L. Rabiner. *The Role of Speech Processing in Human Computer Intelligent Communication*, volume 23, pages 263–278. Speech Communications, 1986.

[19] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. *An Efficient K-means Clustering Algorithm: Analysis and Implementation*, pages 881–892. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.

[20] Reinhard Kneser and Hermann Ney. *Improved Clustering Techniques for Class-based Statistical Language Modeling.* Proceedings of the European Conference on Speech Communication and Technology, 1993.

[21] L. Mangu, E. Brill, and A. Stolcke. *Finding Consensus Among Words: Lattice-based Word Error Minimization*, pages 495–498. Proceedings of Eurospeech, 1999.

[22] M. Marcus, B. Santorini, and M. Marcinkiewiez. *Building a Large Annotated Corpus of English: the Penn Treebank.* Computational Linguistics, 1993.

[23] Alvin Martin and Mark Pfzybocki. *The NIST 1999 Speaker Recognition Evaluation - An Overview.* Number 1. Digital Signal Processing, 2000.

[24] Mehryar Mohri, Michael Riley, and Fernando C. N. Pereira. *Weighted Finite-State Transducers in Speech Recognition.* Proceedings of ASR2000, International Workshop on Automatic Speech Recognition: Challenges for the Next Millennium, 2000.

[25] R. Moore, D. Appelt, J. Dowding, J.M. Gawron, and D. Moran. *Combining Linguistic and Statistical Knowledge Sources in Natural Language Processing for ATIS*, pages 261–264. Spoken Language Systems Technology Workshop, February 1995.

[26] Niloy Mukherjee and Deb Roy. *Spontaneous Speech Recognition.* The Media Laboratory Internal Technical Document, 2002.

[27] Fernando C. N. Pereira and Michael Riley. *Speech Recognition by Composition of Weighted Finite Automata.* MIT Press, Cambridge, MA, 1997.

[28] Patti J. Price. *Evaluation of Spoken Language Systems: The ATIS Domain.* Proceedings of the DARPA Speech and Natural Language, 1990.

[29] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Prentice-Hall, Eaglewood Cliffs, NJ, 1993.

[30] L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, pages 257–286. Number 2. Proceedings of the IEEE, 1989.

[31] L. R. Rabiner. *The segmental K-means algorithm for estimating parameters of hidden Markov models.* Number 9. IEEE Transactions on Signal Processing, 1990.

[32] Terry Regier. *The Human Semantic Potential.* MIT Press, 1996.

[33] Ronald Rosenfeld. *A Maximum Entropy Approach to Adaptive Statistical Language Modeling*, pages 187–228. Computer Speech and Language, 1996.

[34] Ronald Rosenfeld. *Two Decades of Statistical Language Modeling: Where do we go from here?*, volume 88. Proceedings of the IEEE, 2000.

[35] Deb Roy. *Learning Visually Grounded Words and Syntax of Natural Spoken Language.* Number 1. Evolution of Communication, 2000.

[36] Deb Roy. *A Trainable Visually-Grounded Spoken Language Generation System.* Proceedings of ICSLP, 2002.

[37] Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. *A Trainable Spoken Language Understanding System for Visual Object Selection.* Proceedings of ICSLP, 2002.

[38] Deb Roy and Alex Pentland. *Learning Words from Sights and Sounds: A Computational Model*, pages 113–146. Number 1. Cognitive Science, 2002.

[39] Deb Roy, Kai yuh Hsiao, Peter Gorniak, and Niloy Mukherjee. *Human Robot Interaction.* AAAI Fall Symposium Report, 2002.

[40] Deb Roy, Kai yuh Hsiao, and Nikolaus Mavridis. *Conversational Robots: Building Blocks for Grounding Word Meanings.* HLT-NAACL03 Workshop on Learning Word Meaning from Non-Linguistic Data, 2003.

[41] S. Seneff, E. Hurley, C. Pao, P. Schmid, and V. Zue. *Galaxy-II: A Reference Architecture for Conversational System Development.* Proceedings of ICSLP, 1998.

[42] Danny Sleator and Davy Temperley. *Parsing English with a Link Grammar.* Technical Report CMU-CS-91-196, 1991.

[43] M. J. Spivey, M. J. Tyler, K. M. Ebehard, and M. K. Tanehaus. *Linguistically Mediated Visual Search*, pages 282–286. Number 4. Psychological Science, 2001.

[44] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. E. Sedivy. *Integration of Visual and Linguistic Information in Spoken Language Comprehension*, pages 1632–1634. Science, 1995.

[45] Wayne Ward. *The CMU Air Travel Information Service: Understanding Continuous Speech*, pages 127–129. Proceedings of the DARPA Speech and Natural Language, 1990.

[46] T. Westman, D. Harwood, T. Laitinen, and M. Peitikinen. *Color Segmentation by Hierarchical Connected Components Analysis with Image Enhancement by Symmetric Neighbourhood Filters*, pages 796–802. Proceedings of ICPR, 1990.

[47] Benjamin Yoder. *Spontaneous Speech Recognition Using Hidden Markov Models.* M.S. Thesis, Massachusetts Institute of Technology, 2001.

[48] Norimasa Yoshida. *Automatic Utterance Segmentation in Spontaneous Speech.* M.S. Thesis, Massachusetts Institute of Technology, 2002.