Abstract

Language is about symbols and those symbols must be grounded in the physical environment during human development. Most recently, there has been an increased awareness of the essential role of inferences of speakers’ referential intentions in grounding those symbols. Experiments have shown that these inferences as revealed in eye, head and hand movements serve as an important driving force in language learning at a relatively early age. The challenge ahead is to develop formal models of language acquisition that can shed light on the leverage provided by embodiment. We present an implemented computational model of embodied language acquisition that learns words from natural interactions with users. The system can be trained in unsupervised mode in which users perform everyday tasks while providing natural language descriptions of their behaviors. We collect acoustic signals in concert with user-centric multisensory information from non-speech modalities, such as user’s perspective video, gaze positions, head directions and hand movements. A multimodal learning algorithm is developed that firstly spots words from continuous speech and then associates action verbs and object names with their grounded meanings. The central idea is to make use of non-speech contextual information to facilitate word spotting, and utilize user’s attention as deictic reference to discover temporal correlations of data from different modalities to build lexical items. We report the results of a series of experiments that demonstrate the effectiveness of our approach.
## Contents

1 Introduction .................................................. 2

2 How Children Learn the Meanings of Words ............................. 2
   2.1 The Evidence ............................................. 2
   2.2 The Analysis .............................................. 3

3 Related Computational Studies ........................................ 4

4 Embodied versus Symbolic ........................................... 5

5 Theoretical Framework ................................................ 7
   5.1 Learning from Multiple Perceptual Modalities ....................... 8
   5.2 The Role of Mind Reading in Language Acquisition .................. 10
   5.3 An Embodied Language Learning System ............................ 12

6 Method ........................................................... 13
   6.1 Non-linguistic Modalities ..................................... 13
      Attention Detection .......................................... 13
      Attentional Object Spotting .................................. 14
      Action Segmentation and Recognition ............................ 17
      Segmentation of Action Sequences ............................. 17
      Recognition of Human Actions ................................ 18
   6.2 Linguistic modality ............................................ 19
      Phoneme Recognition ......................................... 19
      Phoneme Sequence Comparison ................................ 19
   6.3 Word Learning by Integrating Multimodal Input .................... 23
      Utterance Spotting and Classification .......................... 24
      Word-like Unit Discovery .................................... 24
      Word-like Unit Clustering .................................... 28
      Learning Words .............................................. 28

7 Experiment ....................................................... 30
   7.1 Data Collection .............................................. 31
   7.2 Evaluation Method ........................................... 31
   7.3 Experiment 1: Performing A Natural Task .......................... 31
   7.4 Experiment 2: Reading A Picture Book ............................ 33

8 Conclusion .......................................................... 35
1 Introduction

Children solve many complex learning problems during their first year of life. Among them, one of the most challenging tasks is to gradually develop their speech perception capacities to learn their native language. How an infant’s perceptual system segments continuous speech into lexical units or words? How a young child learns the meanings of words? Those topics have attracted the attention of psychologists and cognitive scientists for several years. A new trend to address those problems is to supplement psycholinguistic studies with a computational account by developing computational models of infant language acquisition. Compared with descriptive models, those computational models can be easily compared and evaluated by processing a set of experimental data. Furthermore, by implementing the descriptions of the theories or claims explicitly in a computer program, insights can be gained in both the nature of the problems needed to be considered and the possible approaches applied by a language learner. In fact, recent computational researches on child-directed corpora (e.g., [Brent, 1999b, Roy and Pentland, 2002]) have shown that relatively simple statistical learning mechanisms could make an important contribution to certain aspects of language acquisition, which shed light on understanding infant language acquisition.

However, most works so far are limited to highly idealized artificial input or constrained speech instead of natural speech. Thus, they can not deal with the acoustic variability of spoken words in different contexts and by various talkers. Furthermore, the current computational studies of language acquisition focus on developing different algorithms for processing purely linguistic information. Non-linguistic sources of information which infant might exploit, including the physical environment infants live in, are ignored. Non-linguistic information undoubtedly plays a major role in language acquisition, but its utility is difficult to computationally evaluate mainly because the child’s representation of the environment is unknown. This provides a methodological reason to focus on speech data and use the statistics of linguistic information to deal with all the problems of language acquisition, from speech segmentation, word discovery to lexicon acquisition. But clearly, linguistic information alone can not be the only basis for language acquisition. Humans develop based on the sensorimotor experiences with the physical environment. Therefore, the problem of language learning should be embodied in the sense that the learner should be situated in an environment to get sensory information and perform actions to interact with other people. Obviously, the approaches based only on purely linguistic statistics have no access to this account.

This paper presents an implemented computational model of word learning. Compared with other computational studies of infant word learning, our system is able to build meaningful semantic representations grounded in perceptual inputs from different modalities. Moreover, our model appreciates the importance of social factors — such as cues regarding the speaker’s focus of attention. The essential structure models the computational role of the body in creating and simplifying brain representations [Ballard et al., 1997]. The movements of different body parts, such as eye movements, head movements and hand movements, provide a substrate for understanding human language development.

The remainder of this paper is organized as follows. Section 2 gives a short overview of the discoveries of infant language acquisition in cognitive studies. Section 3 describes the related works of modeling speech segmentation and lexical development. Section 4 discusses the problems we needed to address in embodied learning. In Section 5, we present the theoretical work of our model. Section 6 focuses on the implementation of the model. The experimental setup and results are provided in Section 7. Section 8 concludes with a discussion of our future work.

2 How Children Learn the Meanings of Words

2.1 The Evidence

Infants seem to have the innate abilities to process acoustic signals and distinguish the phonemes of all languages [Werker and Lalonde, 1999]. Those phonetic knowledges of language seem to be necessary for
the beginning of lexicon acquisition. Based on those knowledges, children learn their first words at some time between 8 and 10 months. After this, most children spend many weeks or months producing single-word utterances. At first, the rate of vocabulary growth is very slow, averaging roughly one word per day, but typically there is a “burst” or an acceleration in the rate of vocabulary growth somewhere between 16 and 20 months. At 18 to 20 months, children begin expressing combinatorial meanings. The related studies of developmental psychology [Tomasello and Bates, 2001] categorize the early grammars into eight patterns, such as “agent + object”, “action+object” and “agent + action”.

Among the first words children learned, a large portion of words can be grounded in the physical senses. Many of them refer to middle-size objects — things that can move and be moved. These include names of specific persons (mama, dada), animals (dog, cat), toys (ball, block), articles of clothing (sock, shirt) and other artifacts (fork, chair). Object names are really special in the early lexical development because they constitute a much larger proportion of children’s early vocabularies compared with the vocabularies of adults. This is true for almost every language that has been studied [Caselli et al., 2000].

During lexical development in English, the proportion of verbs rise rapidly and form the other significant group of infant early words. Statistical analyzes [Caselli et al., 2000] showed that the proportion of verbs that infants comprehension increase from 6% in children with fewer than 20 words to 18% in children with about 200 words. In particular, it has been proposed that children who are learning some languages, such as Chinese and Korean, do not show the same early bias toward nouns, which has been observed in children learning English [Tardif, 1996, Gopnik and Choi, 1995]. Data for children learning Korean or Chinese suggest that verbs may be far more common in these languages, even with the very first stages of lexical learning.

2.2 The Analysis

In acquiring language, children need to solve at least two difficult problems. First, children need to segment continuous speech and identify lexical units. Before children can begin to map words onto objects in the world, they must determine which sound sequences are words. To do so, children must uncover at least some of the units that belong to their native language from a largely continuous stream of sounds. However, spoken language lacks the acoustic analog of blank spaces of written text, which makes the problem of discovering word boundaries in continuous speech quite challenging. In particular, if we assume that children start out with little or no knowledge of the inventory of words, the problem becomes much harder because speech segmentation can not be achieved by identifying known lexical items from connected speech.

Second, young children need to associate those sound patterns with meanings. Learning a word involves mapping a form, such as the sound “car” to a meaning or concept, such as the concept of car. During their development, children sense a multitude of co-occurrences between words and things in the world, and they must determine which co-occurrences are relevant.

Considering the difficulties in early language acquisition, it is an amazing fact that children are able to learn language without any problem. An intuitive explanation of this phenomenon is that the task is simplified. For example, parents may speak clearly and pause between words. Also, they would point to the objects they refer to in speech, and the environments that children live in are relatively not cluttered. However, all the above assumptions are plausible. First, it is less likely that children learn words by attending to single-word utterances. For example, even when instructed to teach their 12-month-olds a noun referring to a body part, some mothers rarely use that word in isolation [Aslin et al., 1996]. Also, in the CHILDES corpus (child-directed speech, [MacWhinney and Snow, 1985]), only around 15% of utterances that children hear are single words. Thus, children do need to segment continuous speech, starting from a state in which they do not know any word. Actually, laboratory evidence has demonstrated that children have substantial segmentation capabilities by the onset of lexical acquisition. For instance, Jusczyk and Aslin [Jusczyk and Aslin, 1995] have shown that at 7.5 months, infants are able to segment speech well enough to identify words in sentential context. Secondly, parents sometimes but not always point to the referents when they utter the corresponding names. Thus, we can not rely on the occasional pointing motions to build general principles of language
development. Thirdly, although the environment of children is simpler than adults, infants who hear a word and know that it might refer to an object in the scene still face with an indefinite number of possible meanings of this word: the target object or some other object in the scene. The ability to determine whether a given word they hear is relevant to something in the context is central to infant word learning.

The topics related to early language learning, such as speech segmentation and lexicon learning, have received much attention in the psycholinguistic community. Recent research has founded a number of relevant cues that are correlated with the presence of word boundaries and could be used to segment the speech signal into words. Cutler and Butterfield [Cutler and butterfield, 1992] argued that English-speaking adults appear to use prosodic cues, such as strong or weak stress, syllable units and subsyllabic units, to parse a continuous acoustic signal into words. Another possible cue for segmentation is to use the distributional statistics of phonemes. Saffran, Aslin, and Newport [Saffran et al., 1996] demonstrated that 8-month-old infants are able to find word boundaries in an artificial language based only on distributional cue. However, the recent experimental studies of Aslin el al. [Aslin et al., 1996] showed that both prosodic and phonological cues in word discovery are not necessary since people are able to segment a continuous speech stream without any prosodic or phonological cues for word boundaries. Perruchet and Vinter [Perruchet and vinter, 1998] proposed an account of the ability to extract words from a continuous flow of speech that lacks prosodic and phonological cues. Their method relies on general principles of memory and associative learning, and requires no computational abilities besides those involved in any memory-based behavior. Other possible cues include phonotactic regularities [Mattys and Jusczyk, 2001] and allophonic variations [Jusczyk et al., 1999].

To summarize, there are several cues that infants might use for speech segmentation, but it remains an open question what strategies are actually utilized by infants.

For the problem of learning word, Lakeoff and colleagues [Lakoff and Johnson, 1999] suggested that a child has reached an adequate level of concept formation prior to the development of language. Children must have some understanding that a dog barks, is furry, and can be played with, even if they don’t know the words “dog”, or “barks” or “furry”. Lakoff argued that children’s sensorimotor experience is continually building up these embodied, pre-linguistic concepts. If we assume that this conceptual machinery is already well established by the time of learning the first words, the language learning problem is simplified by directly associating a linguistic label with a category of sensorimotor experience that has already been built up before [Howell et al., 2001]. As we mentioned before, one interesting finding in developmental psycholinguistics is an overwhelming preponderance of nouns in children’s early speech, not only in English but in most languages. However, Gillette et al. [Gillette et al., 1999] provided strong evidence that learnability is not primarily based on lexical class but upon the word’s imageability or concreteness. The nouns are learned before most of verbs because nouns are more observable than verbs. The imageability of a word is more important than the lexical class and the most observable verbs are learned before the less observable nouns.

3 Related Computational Studies

From a computational perspective, word learning is a complex skill involving the acquisition and integration of information across different modalities. A good survey of the related computational studies of language acquisition can be found in [Brent, 1999b], in which several methods are explained, their performance in computer simulations is summarized, and behavioral evidence bearing on them is discussed.

To explore how young children discover the words embedded in a mostly continuous speech stream, several computational models of speech segmentation have begun to consider the learning problem from the point of view of the statistical properties of language, which might be stored and computed in human brain. Saffran, Newport and Aslin [Saffran et al., 1996] suggested that infants might compute the transitional probabilities between sounds in language and use the relative strengths of these probabilities to hypothesize word boundaries. The method they developed treats syllables rather than phonemes as the fundamental units of input and calculates the probability of each syllable in the language conditioned on its predecessor. They argued that children may segment utterances at low points of the transitional probability between adjacent
syllables. Brent and Cartwright [Brent and Cartwright, 1996] have encoded information of distributional regularity and phonotactic constraints in their computational model. Distributional regularity means that sound sequences occurring frequently and in a variety of contexts are better candidates for the lexicon than those that occur rarely or in few contexts. The phonotactic constraints include both the requirement that every word must have a vowel and the observation that languages impose constraints on word-initial and word-final consonant clusters. More recently, Brent [Brent, 1997, Brent, 1999a, Dahan and Brent, 1999] proposed the model called INCDROP (INCremental Distributional Regularity OPTimization). INCDROP asserts that the process of segmenting utterances and inferring new word-like units is driven by the recognition of familiar units within an utterance. It posits a single mechanism that discovers new units by recognizing familiar units in an utterance, extracting those units, and treating the remaining contiguous stretches of the utterance as novel units. When an utterance contains no familiar units, the whole utterance is treated as a single novel unit, so there is no need to assume a special bootstrapping device that discovers the first units.

Most of these studies, however, deal with phonetic transcriptions of speech instead of raw audio from microphone input. To cope with the variation of pronunciations in different contexts and by various talkers, these systems always encode the knowledge of language, such as language model and word model, to remove the noises and get a clean phonetic transcription. In this way, these systems are not working in entirely unsupervised learning mode because they need human involvement to provide labeled data for training the model parameters. In contrast, infants learn language by processing their perceptual input and no manually labeled transcriptions are needed as “teaching” information. Thus, the model grounded raw speech is more biologically plausible than the transcription-based approaches.

Compared with the studies of early speech segmentation, there are relatively few computational studies of lexical learning. Roy et al. [Roy and Pentland, 1998, Roy, 2001, Roy and Pentland, 2002] use the correlation of speech and vision to associate spoken utterances with a corresponding object’s visual appearance. His work is motivated by modeling the early stages of word acquisition from sensor-grounded speech and visual signals. The learning algorithms based on cross-modal mutual information are utilized to discover words and their visual associations from transcribed speech paired with images of three-dimensional objects. A focus of the work was the discovery and segmentation of word-like acoustic units from spontaneous speech driven by cross-modal analysis. The acquired lexicons of visually grounded words served as the basis for a small vocabulary speech understanding and generation system. Roy’s work is the first implemented model of language acquisition which learns words and their semantics from raw sensory input. However, the audio-visual corpus are collected separately in Roy’s system. Specifically, audio data are gathered from infant-caregiver interactions while visual data are captured by a CCD camera on a robot. Thus, audio and visual input are manually correlated based on the co-occurrence assumption. We will show later that the co-occurrence assumption is not reliable and appropriate for modeling language acquisition.

Luc Steels et al. [Steels and Vogt, 1997, steels, 1997] reported the experiments in which autonomous visually grounded agents bootstrap meanings and language through adaptive language games. He argued that language is an autonomous evolving adaptive system maintained by a group of distributed agents without central control. The computational experiments showed how a coherent lexicon may spontaneously emerge in a group of agents engaged in language games and how a lexicon may adapt to cope with new meanings that arise. The structure of an agent is created by random processes and eliminated based on selectionist principles centering around use and success in use. Agents engage in interactions with the environment or with others to change their internal structures in order to be more successful in the next game.

4 Embodied versus Symbolic

Traditionally, cognition has been considered as a “mental process”. Recently, there is a growing consensus that cognition should be “embodied” in the sense that it merges from physical interactions with the world through a body with given perceptual and motor abilities. Humans develop based on their sensorimotor
experiences with the physical environment. Different levels of abstraction are necessary to efficiently encode those experiences, and one vital role of human brain is to bridge the gap from embodied experience to its expression as abstract symbols. As emphasized by Harnad, to mimic human skills, a challenge in machine intelligence is to specify how to ground symbolic representations from non-symbolic sensorimotor information [Harnad, 1990]. To solve the grounding problem, it has been suggested by various researchers [Brooks, 1990] that a cognitive system has to have a physical body, and be situated in an environment from which it gets sensory information and in which it can perform actions and interact with others. In light of this, we have developed a computational model of embodied language acquisition in which a computer, as a pre-linguistic child, is able to learn word-meaning associations given access to continuous speech signals and signals from other non-speech perceptual inputs. Our approach is quite different from most existing models of language acquisition since they have been evaluated by artificial data of speech and semantics [Brent and Cartwright, 1996, Brent, 1997, Siskind, 1995, Bailey et al., 1998, de marcken, 1996]. In those models, speech is represented by text or phonetic transcriptions and word meanings are usually encoded as symbols or data structures.

Our model appreciates the importance of embodiment for two main reasons. First, the motivation behind this work is that language is grounded in real sensory experiences about the physical world. Unlike dictionary definitions in which words are defined in terms of other words, humans understand basic concepts in terms of sensorimotor experiences. For example, in the Webster’s dictionary, the word “car” is defined as follows: a small vehicle moved on wheels; usually propelled by an internal combustion engine. Humans, however, understand the basic meanings of the concept “car” from their perceptual systems. We perceive the image of a car through visual perception, and understand the functions of car by our own experiences and by observing others people driving cars on the streets. This ability contributes to the conclusion that language is grounded in perception and action. Thus, a fundamental aspect of language acquisition is to associate the body and the environment with words in language. Therefore, instead of considering language learning as a symbol processing problem, it is more reasonable to be treated it as a symbol grounding problem in which spoken words are associated with their grounded referents.

Second, from the perspective of machine learning, our computational model is able to learn grounded words in purely unsupervised mode. Compared with other methods, this unsupervised learning procedure is closer to simulate natural development of infant’s linguistic ability. Infants learn words by sensing the environment from their perceptual system that does not provide the labeled or preprocessed data. Therefore, if we use texts or phonetic transcriptions of speech instead of raw acoustic signals, the mechanism that converts acoustic signals to texts encodes the knowledge of language to deal with the varieties in acoustic signals. To model language learning, however, we can not assume that children utilize learned knowledge of language to solve the problem in the learning procedure.

Obviously, a sensor-grounded model has more advantages than symbol-based models for simulating lexical development. However, it is a significant challenge to produce linguistically meaningful lexicons from raw multimodal data. The followings are the problems we needed to address in our model:

- The first step for language acquisition is to segment connected speech into lexical units. English speech lacks the acoustics analog of blank spaces that people are accustomed to seeing between words in written text. If we assume that children start out with little or no knowledge of the inventory of words, the identification of word boundaries is a significant problem in the domain of child language acquisition. For example, the utterance “my dog” can be represented as a phoneme sequence /m a y d ɹ ɡ/, each of the symbol stands for a phoneme, a perceptual unit of speech. Without any prior knowledge about language, there are 15 possible subsequences that might correspond to a word: /m/, /a y/ /d ɹ/, /l/, /l/, /m a y/ /l/, /m a y d ɹ/, /m a y d ɹ ɡ/, /a y d ɹ/, /a y d ɹ ɡ/, /l/, /l d ɹ ɡ/ /l/ ɹ ɡ/. The problem is how to find the correct word boundaries.
- Furthermore, children have to cope with the acoustic variability of spoken words in different contexts and by various talkers. During speech production sounds blend across word boundaries, and words undergo tremendous phonological and acoustic variation. Thus, before reaching the language learner,
unknown sounds from an unknown number of words drawn from an unknown distribution are smeared across each other and corrupted by various noisy channels. From this, the model must remove the noise from raw signals and extract the representations of the sound patterns of repeated words that are both durable and generalizable across tokens produced by different talkers.

- Based on signals from visual perception and the sense of body movements, the model needs to identify and categorize non-linguistic perceptual input into a set of possible semantic meanings when spoken utterances are produced. Specifically, the system needs to spot and recognize human actions as well as recognize attentional object of talkers’ interest.

- In most context, there are more than one possible meanings that are co-occurring in non-linguistic modalities and can be associated with spoken words uttered in temporal proximity. How to build the correct map? Acquiring lexical semantics involves identifying the meaning of a particular word. Even for concrete nouns, this problem is complicated by the difficulty of detecting which part of the physical environment a speaker is referring to. Even if this can be ascertained, it may still remain unclear whether the term used by the speaker refers to a particular object, a part of that object, or a class of objects. For abstract nouns and other words which have no concrete referents, these difficulties are compounded further.

The following section will provide an overview of theoretical framework of our approach and focus on the novelties of our work. After that, we present the implementation of our model in Section 6.

## 5 Theoretical Framework

Modeling lexical development has received considerable attentions in both experimental psycholinguistics and computational linguistics. Generally, to simplify the problem, the task of learning word meanings is divided into several subtasks. For instance, Siskind [Siskind, 1995, Siskind, 1999] presented a formal version of language acquisition by modeling it into three subtasks shown in Figure 1: (1) identifying the sequence of words in an utterance. (2) identifying a set of likely interpretations of the utterance based on the non-linguistic context when it is produced. (3) inferring the meanings of words given the results of the first two subtasks. Until now, most computational studies of how children learn their native language address one particular subtask. For example, Siskind himself focuses on the subtask of inferring the meanings and the work of Brent et al. [Brent, 1999a] deals with how children discover the sound patterns of words from the continuous spoken language. The studies of object recognition or action recognition in computer vision focus on the second subtask.

![Figure 1](image)

**Figure 1.** Siskind’s framework consists of three subtasks in modeling lexical development.
We take a new look of modeling language acquisition in this paper. Figure 2 illustrates the basic idea of our model. Compared with the three-subtask model of Siskind, the novelty of our model lies in two aspects. First, we treat word learning as a whole task instead of several subtasks. The methodology of dividing a complicated task into several subtasks makes sense in many situations. However, in the context of modeling language acquisition, even after applying this methodology, each of the three subtasks is still very difficult. We notice that two reasons mainly cause the difficulties: (1) Three subtasks are not totally independent so the separation might not simplify the whole problem but make it harder to address. In another word, the sum of three subtasks might increase the complexity of the whole problem. Although three-subtask model is quite clear from the conceptual perspective, this does not mean that the cognitive system really solves the problem in this way. For instance, non-linguistic information, such as visual perception and human body movements, provides a context when spoken words are uttered, so it is quite possible that the cognitive system utilizes this context to infer the associations between spoken utterances and then uses those associations to facilitate the discovery of the sound patterns of words. In this way, the computational cost might be much less than solving three-subtask separately. (2) The temporal co-occurring data from different modalities are correlated. The method based on processing data from individual modality separately can not capture the information encoded in those temporal correlations.

Based on the above analyzes, we argue that a good solution of the whole problem could reduce the complexity of the separated subtasks. The advantage of this solution lies in utilizing the information encoded in spatial-temporal correlations between data from multiple modalities. Specifically, we believe that non-linguistic information, such as visual perception and human body movement, provides a context for language development. Children utilize this context as a vital constraint to simply the computation of language learning. For instance, in the subtask of speech segmentation in Siskind’s model, statistic-based methods are widely used on the basis of calculating the probabilities from a large amount of training corpus to get reliable estimates. Those methods need lots of memory to store the data and bring an enormous burden for computation. We will show later that the non-linguistic information can greatly facilitate the speech segmentation problem that is traditionally addressed by solely using linguistic information. The aim of this work is to demonstrate that the method based on both integrating linguistic and non-linguistic information and utilizing temporal information between different modalities can greatly simplify word learning problem. This approach will shed light on the understanding of infant language acquisition.

Second, we believe that social factors, such as inference of speaker’s referential intentions, play a major role in language acquisition by providing an important constraint in delimiting the possible associations between a multitude of co-occurring word-meaning pairs. Moreover, Ballard et al. [Ballard et al., 1997] showed that pointing movements of the body can be used to bind objects in the physical world to cognitive programs. This give rises to the idea that different parts of body movements, especially eye and head movements, can be useful cues to guide the association of data from different modalities.

In the rest of this section, we will present design issues of our model in detail and describe how to utilize the ideas described here in modeling language acquisition.

5.1 Learning from Multiple Perceptual Modalities

A key issue in grounding intrinsic meanings of words with embodiment can be treated as the problem of associating speech with sensory data from other modalities, such as visual perception and proprioception. In most multimodal learning systems, the goal is to overcome the limitations of individual senses for a particular task. When one sense cannot provide all the necessary information, complementary observations may be provided by another sense. For example, haptic sensing complements vision in placing a peg in a hole when the effector occludes the agent’s view. Another type of interplay between modalities is the use of information extracted by one sense to focus the attention of another sense. For instance, audition cuing vision is a common instance of temporal cooperation between modalities. In any case, the fundamental issue is how to combine observations into a coherent percept to improve the overall results[Brunelli and Falavigna, 1995].
Our model treats the problem of word learning as a whole task and our approach utilizes spatial-temporal correlations between data from different modalities.

The integration of data from multiple modalities is a challenging problem. Let us consider the integration of speech and vision as an example. As we all know, interpretation of speech and visual data are two of the most difficult problems in Artificial Intelligence (AI). So the first idea may be that combining two of the most difficult problems in AI results in an even more difficult problem. However, some researchers [Bischoff and Graefe, 1999, Srihari, 1994, Horswill, 1995] argued that a good solution of the combination problem should reduce the complexity of the separated tasks. In practice, the use of multiple modalities is meaningful in many different contexts, so the related research areas are highly diverse and interdisciplinary. In the robotics literature, sensor fusion is an active research area. The typical approach is to use a common representation such as a 3D-model of the environment or a certainty grid to integrate information coming from different modalities. Multimodal processing has also been addressed for the problem of audio-visual lip reading [Durgin, 1995, Stork et al., 1992]. Speech signals and co-occurring visual lip signals are combined to increase accuracy of speech recognition. This approach is very effective in situations with high noise levels.

Different with most of the works of multimodal learning which focus on processing data in the individual modality separately and merging the outputs of multiple streams to make a better decision, we believe that a powerful constraint in multisensory data is coherence in time and space. An unsupervised learning procedure that can capitalize on this constraint may be able to explain much of perceptual self-organization in the humans’ brain. During human development, once the abilities of detecting rudimentary feature have been established, an infant can learn to segment the sensory inputs, and eventually classify them into familiar patterns. These earliest stages of learning seem to be inherently unsupervised. Natural environments do not contain explicit labeling signals, but they contain important information in the form of temporal correlations between sensations to different sensory modalities, and humans are affected by this correlational structure. Infants learn their first words by associating speech patterns with objects, actions, and people. The primitive meanings of words and utterances are inferred by observing the world through multiple senses. Gradually, more complex concepts are formed based on the foundation of early word learning.

In this work, we explore temporally co-occurring multimodal data in three aspects. Firstly, we integrate acoustic signals with visual perception and body movements for learning grounded lexicons. Let us take visual-audio integration as an example. We assume, in general, audio and visual signals are uncorrelated in time. However, when a word is heard by a listener, his eye movement will bring the visual representation of the spoken word in close temporal proximity. The goal of learning in this scenario is to find most reliable audio-visual pairs from multimodal data.

Secondly, our learning algorithm can cluster spoken utterances based on not only their similarities in terms of acoustic features but also their grounded meanings. For example, spoken utterances of the word “car” generated by different people may be various. If we cluster data based on the measurement of their
similarity, we may cluster some instances to wrong groups, such as the cluster for “cow”. In our case, the correlated visual information, however, can provide extra teaching signals for clustering. In more general cases, as we mentioned, it is possible to cluster spoken utterances in different languages, which have the same semantic meaning, into the same group. From the perspective of machine learning, the problem can be viewed as both supervised and unsupervised learning. Viewed as unsupervised learning, input of the system is unlabeled multisensory data. No manually labeled transcriptions are needed as “teaching” information. On the other hand, each stream of data may be treated as noisy labels for the other. For example, spoken utterances are labels for co-occurring visual data, and vice versa. The advantage of the approach is that we do not need to transform signals to symbolic representations but integrate information from multiple modalities directly on the signal level.

Thirdly, non-linguistic context can provide some hints for word discovery. Speech segmentation without prior language knowledge is a challenging problem and attracts more and more attention recently. However, all the works focus on using linguistic information to segment continuous speech. In contrast, our method appreciates the importance of non-linguistic context from different modalities, within which the spoken words are uttered. We believe that sound patterns that frequently appear in a context are more likely to have grounded meanings related to this context. Consider the goal of building grounded lexicons, we are especially interested in spotting the sound patterns that have the meanings related to the context, instead of segmenting the whole utterance. Let us consider a simple example: an utterance “this is a car”. An objective of a speech segmentation algorithm is to obtain the perfect segmentation: /this/is/a/car/. We argue, however, for the purpose of lexical acquisition, it is not mandatory to get a perfect segmentation. As far as we can spot the sound pattern “car” from the utterance, it is enough to build the grounded lexical items. Here we do not acclaim that infants just spot the key words instead of the complete segmentation of speech. Actually many researches have shown that infants do have the ability of speech segmentation. However, from a computational perspective, we argue that at least in the early stage of word learning, it is not necessary to segment every word from utterances. The main advantage of this approach, compared with the methods based on solely linguistic information, is that the non-linguistic context greatly reduces the computational costs of both calculating the statistics from large corpus and storing all the data in memory. Considering those costs, it is plausible that infants would employ statistics-based method to segment the continuous speech into several items among which most of them are useless for the early word learning.

5.2 The Role of Mind Reading in Language Acquisition

A central task of word learning is to figure out which entities specific words refer to. The language children hear and the environment they are in provide a multitude of co-occurrences between words and things in the world, and they must determine which among these co-occurrences are relevant. Thus, children need to know the meanings of words by associating verbal labels with the information from other modalities, such as visual perception. A popular solution to this learning problem is associationism. This is the view that children learn the meaning of “dog” because the word is uttered when they are observing dogs. As a result, the word and the concept become associated, and children could be thought to have learned what the word means. Several computational models [Richards and Goldfarb, 1986, Plunkett, 1997] are derived based on this idea, which focus on establishing connections between a set of stimuli and verbal response. In those models, one of the most important resources is the statistical properties of language that are explored in both computational linguistics and experiment psychology: (1) The studies of natural language processing show that distributional information is a valuable cue for many aspects of language acquisition. (2) Recent experimental evidence in both children and adults shows that the cognitive system is sensitive to features of co-occurrence statistics in language. Therefore, it seems a reasonable working assumption that, given the immense difficulty of the language acquisition problem, the cognitive system is likely to exploit such simple and useful sources of information. But despite the merit of this idea, it sufferes from some problems because it is based on the assumption that words are ordinarily used at the same time that their referrents are perceived. This assumption is plausible from the experimental results of parent-child interactions. For example, within
a supportive family environment, about 30% to 50% of the time that a word is used, young children are
not attending to the object that the adult is talking about [Bloom, 2000]. Also, children in some culture can
learn the meanings of words even if their parents do not carefully name objects for them. Thus, a statistical
covariation between word and percept is useful but not sufficient for word learning.

How children associate words and what they refer to? Bloom [Bloom, 2000] showed the evidence that
children’s word learning actually draws extensively on their understanding of the thoughts of speakers. Thus,
children use their naive psychology to learn the entities to which words refer, and intuit how words related to
one another. He claimed that word learning is a species of intentional inference, called mind reading by Simon
Baron-Cohen [Baron-Cohen, 1995]. It has been shown that an infant at age around nine months will naturally
follow its mother’s gaze and her pointing gesture. These findings give rise to the argument that when a baby
follows the gaze of an adult, he/she might have an implicit assumption that the adult is attending to something
and thinking about or reacting to that object. Thus, children may figure out what adults are intending to refer
to when words are heard based on inference of the intention of others. The experimental results of Tomasello
[Tomasello and Barton, 1994] and Baldwin [Baldwin, 1993] verify the important role of intentional cues in
language learning. They showed that children associate object names with objects only if they believe that
the acts of adults are naming. Specially, Baldwin et al. [Baldwin et al., 1996, Baron-Cohen et al., 1997]
proposed that infants give a special weight to the cues of indexing the speaker’s referential intent when
determining the reference of a novel label. Their experiments showed that infants established a stable link
between the novel label and the target toy only when that label was uttered by a speaker who concurrently
showed his/her attention toward the target, and such stable mapping was not established when labels were
uttered by a speaker who was out of view and hence showing no signs of attention to the target toy. Based
on this fact, they concluded that anytime infants are presented with new words in a context in which cues to
referential intent are impoverished, learning of those words should be compromised. Conversely, enriching
cues to referential intent may be one way in which to enhance vocabulary acquisition.

A complementary picture emerges from studying the computational model of the brain proposed by Bal-
lard et al. [Ballard et al., 1997]. They argue that at time scales of approximately one-third of a second,
orienting movements of the body play a crucial role in cognition and form a useful computational level,
termed the embodiment level. At this level, the constraints of the body determine the nature of cognitive op-
erations, since the natural sequentiality of body movements can be matched to the computational economies
of sequential decision systems. Computation at this level governs the rapid deployment of the body’s sensors
and effectors in order to bind variables in behavioral programs of human brain because of the ability of the
different sensory modalities to quickly direct their foci to localized parts of space. This computation pro-
vides a language that links external sensory data with internal cognitive programs and motor actions. From
a computational perspective, the computation provides a map between lower-level actions and higher-level
symbols. The way it is done is through a system of implicit reference termed deictic, whereby pointing
movements of the body are used to bind objects in the world to cognitive programs.

Based on the above analysis, we have built a model to explore the computational role of mind reading in
learning language. To achieve this goal, we utilize the direction of gaze to detect the speaker’s attention. It
has been shown that, compared with the movement of other parts of body, eye movements are particularly
instructive: visual fixations provide an explicit indicator that information is being represented in cognitive
programs [Ballard et al., 1992]. Therefore, it is straightforward to utilize eye movements to estimate the
intention of people. Moreover, Cooper [Cooper, 1974] found that people have a strong tendency to look
toward objects referred to in conversations. He showed that the response system of eye movements in the
presence of an on-going conversation is always characterized by a high degree of linguistic sensitivity. These
findings provide the possibility to develop a computational model that associates the spoken words with other
perceptual information by employing gaze to find referential intent.
5.3 An Embodied Language Learning System

In light of infant language acquisition, we are developing a computational model that can learn meaningful semantic representations grounded in the physical environment around us. In order to ground language, we have to know what the body is doing from moment to moment. To do this we attach different kinds of sensors to a real person as shown in Figure 3. Those sensors include a head-mounted CCD camera to capture a first-person point of view, a microphone to sense acoustic signals, an eye tracker to track the course of eye movements that indicates the agent’s attention, and position sensors attached to the head and hands of the agent to simulate proprioception in sense of motion. The functions of those sensors are similar to human sensory system and they allow the intelligent system to collect multisensory data to simulate the development of human-like capabilities. In the learning phase, the real agent performs some everyday tasks, such as making a sandwich, pouring some drinks or stapling a letter, while describing his/her actions verbally. The system builds the grounded semantics by associating object names and action verbs with visual perception and body movements. The advantages of our approach are as follows:

- The system is embodied by sharing sensory experiences about the physical world with a real agent. Thus, an intelligent system sees as the agent sees, hears as the agent hears, and experiences the life and the environment of the agent in a first-person sense.
- The system can utilize the correlations and constraints of data from different modalities, temporally and spatially, to improve its intelligence. Many learning problems, such as language acquisition, are difficult because no single feature of the input correlates with the relevant aspects of language structure. Although it is a natural way to study cues in isolation, the problem of acquisition might be easier when multiple cues are taken into account. As the number of cues that learner considers increases, the difficulty of the learning problem may decrease. This suggests that the cognitive system may aim to exploit as many sources of information as possible in language acquisition. For example, the objects of user interest in time can help action recognition since those objects are closely correlated with some specific actions, such as a knife related to the action of cutting and a stapler related to the action of stapling, etc. Such object contexts would suggest specific actions for action recognition.
- With the ability to track the course of eye movements, we can obtain the center of eye in time that explicitly indicates the agent’s attention which can be used for spotting the objects involved in actions. Also, attention switches can be calculated and used to segment the agent’s action sequences.
- In this work we are specifically concerned with the unsupervised word acquisition given no prior language-specific knowledge. In contrast to several prior proposals, our algorithm makes no assumptions about the presence of facilitative side information, or of cleanly spoken and segmented speech, or about the distribution of sounds within words. Furthermore, no language-specific knowledge means
that our model is independent of a specific language. It could be applied to shed light on general mechanisms of language acquisition.

- Different from other approaches that purely focus on statistical learning from speech corpus, we appreciate the importance of attention and utilize signals from non-linguistic modalities to help solve the problems, such as speech segmentation, word discovery and lexicon learning, which are traditionally addressed using purely linguistic signals. Also, our method does not need to acquire the specific words and phonological structure of a language, which requires exposure to a significant corpus of language input.

- Compared with traditional video processing based on fixed camera observations, the dynamic properties of the agent-centered view captured by a head-mounted camera provide image sequences that are more informative because the agent uses the same data for visual processing.

6 Method

As described in Section 4, the range of problems we need to address in embodied language learning is substantial, so to make concrete progress, we need to narrow the scope of our effort. Specifically, we have implemented the model in the domain of learning the specific groups of words: object names and action verbs. The system implements all components of the model and is grounded in perceptual input. This section provides the implementation details of processing data from multiple modalities.

6.1 Non-linguistic Modalities

The non-linguistic input of the system consists of visual information from a head-mounted CCD camera, head and hand positions in concert with gaze-in-head data. The goal of non-linguistic processing system is to:

- detect user’s focus of attention and attention switch.
- extract a representation of the attentional object that the subject is focusing on.
- segment the continuous action streams into several action primitives and acquire the motion types which serve as possible meanings of spoken words.

Attention Detection

In the context of our application of eye gaze, the primary objective of eye data analysis is to determine where and when the user looks at the objects in the visual scene. Although there are several different modes of eye movement, the two most important modes for directing cognitive works are saccades and fixations. Saccades are rapid eye movements that allow the fovea to view a different portion of the visual scene. Often a saccade is followed by one or more fixations when objects in a scene are viewed. Our goal is to find the fixations from continuous data stream of eye movement. The existing fixation finding methods [Salvucci and Goldberg, 2000] can be categorized into three groups: velocity-based, dispersion-based and region-based. Velocity-based methods find fixations according to the velocities between consecutive data points. Dispersion-based methods identify fixation points as the points that are grouped closely together with the assumption that the fixation points generally occur near one another. Region-based methods identify fixation points as points that fall within a fixed region called areas of interest (AOIs).

We developed a velocity-based method to model eye movements using a hidden Markov model (HMM) representation that has been widely used in speech recognition with great success [Rabiner and Juang, 1989].
A hidden Markov model consists of a set of $N$ states $S = \{s_1, s_2, s_3, ..., s_N\}$, the transition probability matrix $A = a_{ij}$, where $a_{ij}$ is the transition probability of taking the transition from state $i$ to state $j$, prior probabilities for the initial state $\pi_i$, and output probabilities of each state $b_i(O(t)) = P\{o(t)|s(t) = s_i\}$. Salvucci et al. [Salvucci and Anderson, 1998] first proposed a HMM-based fixation identification method that uses probabilistic analysis to determine the most likely identifications for a given protocol. Our approach is different from his in two ways. First, we use training data to estimate the transition probabilities instead of setting pre-determined values. Secondly, we notice that head movements provide valuable cues to model focus of attention. This is because when users look towards an object, they always orient their heads towards the object of interest so as to make it in the center of their visual fields. As a result of the above analysis, head positions are integrated with eye positions as the observations of HMMs.

**Figure 4.** The HMM of eye movement

Figure 4 shows a 2-state HMM that is used in our system for eye fixation finding. One state corresponds to saccade and the other represents fixation. The observations of HMM are 2-dimensional vectors consisting of the magnitudes of the velocities of head rotations in three dimensions and the magnitudes of velocities of eye movements. We model the probability densities of the observations using a two-dimensional Gaussian:

$$b_j(O_t) = \frac{1}{\sqrt{(2\pi)^2 | \sigma_j^2}} e^{-\frac{1}{2} (O_t - \mu_j)^T \sigma_j^{-1} (O_t - \mu_j)} \quad (1)$$

The parameters of HMMs needed to be estimated comprise the observation and transition probabilities. Specifically, we need to compute the means($\mu_j$) and variances($\sigma_j$) of two-dimensional Gaussian (four parameters) for each state and the transition probabilities (2 parameters) between two states. Thus, a total of 10 parameters need to be estimated in the HMM. The estimation problem concerns how to adjust the model $\lambda$ to maximize $P(O | \lambda)$ given an observation sequence $O$. We can initialize the model with flat probabilities, then the forward-backward algorithm [Rabiner and Juang, 1989] allows us to evaluate this probability. Using the actual evidence from the training data, a new estimate for the respective output probability can be assigned:

$$\bar{\mu}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)O_t}{\sum_{t=1}^{T} \gamma_t(j)} \quad (2)$$

and

$$\bar{\sigma}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)(O_t - \bar{\mu}_j)(O_t - \bar{\mu}_j)^T}{\sum_{t=1}^{T} \gamma_t(j)} \quad (3)$$

where $\gamma_t(j)$ is defined as the posterior probability of being in state $j$ at time $t$ given the observation sequence and the model.

As a result of learning, the saccade state contains an observation distribution centered around high velocities and the fixation state represents the data whose distribution is centered around low velocities. The transition probabilities for each state represent the likelihood of remaining in that state or making a transition to another state. An example of the results of eye data analysis is shown in Figure 5.

### Attentional Object Spotting

This section describes the method of automatic object spotting by integrating visual information with eye gaze data. For an eye fixation, the object of user interest is extracted from the snapshot of the scene. Figure 6 shows the overview of our approach composed of three steps: image segmentation, object representation and object recognition.
Figure 5. Eye Fixation finding. The top plot: Point-to-point velocities of eye positions. The middle plot: The velocity profile of head. The bottom plot: A temporal state sequence of HMM ("1" indicates the fixation state and "0" represents the saccade state).

Figure 6. The overview of Object Spotting

Image Segmentation The image segmentation in our system consists of two steps. First, we apply seeded region growing (SRG) algorithms [Adams and Bischof, 1994] to segment objects from the background. SRG is based on the conventional region growing postulate of similarity of pixels with regions, but whose mechanism is closer to that of the watershed. The method starts from an initial, incomplete segmentation and tries to aggregate the unlabeled pixels to one of the given regions. The decision whether a pixel should join a region or not is based on the fitness function that reflects the similarity between the region and the candidate pixel. The order in which the pixels are processed is determined by a global priority queue which sorts all candidate pixels by their fitness values. Second, eye gaze is utilized as a cue to extract the object of user interest from all the objects detected. This is implemented by coordinating the eye gaze position into \((x, y)\) position in the image and choosing the region that contains this position. Figure 7 shows a scene snapshot and the segmentation result when the user is grasping a peanut butter jar.

Object Representation The extracted object is represented by a model that contains color, shape and texture features. Based on the works of [Mel, 1997, Schiele and Crowley, 2000, Swain and Ballard, 1991], we
constructed the visual features of objects that are large in number, invariant to different viewpoint, and are driven by multiple visual cues. Specifically, 64-dimensional color features are extracted by color indexing method [Swain and Ballard, 1991], and 48-dimensional shape features are represented by calculating histograms of local shape properties [Schiele and Crowley, 2000]. The Gabor filters with three scales and five orientations are applied to the segmented image. It is assumed that the local texture regions are spatially homogeneous, and the mean and the standard deviation of the magnitude of the transform coefficients are used to represent the object in a 48-dimensional texture feature vector. The feature representations consisting of a total of 160 dimensions are formed by combining color, shape and texture features, which provide fundamental advantages for fast, inexpensive recognition. Most classification algorithms, however, do not work efficiently in higher dimensional spaces because of the inherent sparsity of the data. This problem has been traditionally referred to as the dimensionality curse. In our system, we deduced the 160-dimensional feature vectors into the vectors with the dimensionality of 30 by principle component analysis (PCA) [Aggarwal and Yu, 2000], which represents the data in a lower dimensional subspace by pruning away those dimensions that result in the least loss of information.

**Object Recognition** We employ an appearance-based object recognition method that makes our system more general and more easily trainable from visual data. The system essentially operates by comparing a feature representation of object appearance against many prototype representations stored in the memory to find the closest match. Three-dimensional objects are represented by using a view-based approach in which multiple two-dimensional images of an object are captured from multiple perspectives and grouped to collectively form a model of the object. In the training phase, the feature vectors are used to train the classifier whose rule is to divide the feature space into regions that correspond to different objects. Kohonen’s Learning Vector Quantization (LVQ) algorithm [Kohonen, 1990] has been applied that allows us to build a classifier from labeled data samples. Instead of modeling the class densities, LVQ models the discrimination function defined by the set of labeled codebook vectors and the nearest neighborhood search between codebooks and data. The training algorithm involves an iterative gradient update of the winner codebook. The direction of the gradient update depends on the correctness of the classification using a nearest-neighborhood rule in Euclidean space. If a data sample is correctly classified (the labels of the winner unit and the data sample are the same), the codebook closest to the data sample is attracted toward the sample; if incorrectly classified, the data sample has a repulsive effect on the codebook. The update equation for the winner unit \( m^c \) defined by the nearest-neighbor rule and a data sample \( x(t) \) is

\[
m^c(t + 1) = m^c(t) \pm \alpha(t)[x(t) - m^c(t)]
\]

where the sign depends on whether the data sample is correctly classified (+) or misclassified (-). The learning rate \( \alpha(t) \in [0, 1] \) must decrease monotonically in time and the training procedure is repeated iteratively until convergence. In the recognition phase, a data point \( x_i \) is assigned to a class according to the class labels of the k-closest codebooks.
**Action Segmentation and Recognition**

We make extensive use of the ability to track the course of eye movements in a task and introduces the usefulness of head movements. Particularly in hand-eye coordination tasks, head movements provide valuable cues for the segmentation of such tasks. Gaze, head and hand movements provide a language for interpreting actions in tasks. In this section, we present a method which can utilize the movements of different body parts to parse typical tasks into actions. The ability to do such parsing is extremely important, as it is a precursor for representing the task linguistically.

**Segmentation of Action Sequences**

The segmentation of human action sequences has been a topic of considerable interest in computer vision. For example, Kuniyoshi et al.[Kuniyoshi and Inoue, 1993] focus on detecting meaningful changes in the environment to check for action switches. Ogawara et al.[Ogawara et al., 2000] attempt to recognize and extract meaningful segments by gesture spotting. The method of Rui et al.[Rui and Anandan, 2000] is based on detecting temporal discontinuities of the spatial pattern of image motion that captures the action.

The novel approach of this paper is to segment the continuous actions in natural tasks by detecting agent-centered switches of attention. Based on the fact that eye and head movements are closely linked to attention[Land et al., 1999], we develop a method to detect attention by integrating eye gaze and head position information. In our experiments, we noticed that the temporal relations between eye fixating and hand manipulating are quite tight and predictable, with vision leading action. For example, in the simple action of “picking up an object”, the performer rotates his head toward the object first, followed by fixating the eye on the target and moving the arm to approach and grasp it. At the end of the grasping action, the head and eye always move toward another location, which indicates the beginning of the next action.

We observed that actions can occur in two situations: during head fixations and during eye fixations. For example, in a “picking up” action, the performer focuses on the object first then his motor system moves the hand to approach it. During the procedure of approaching and grasping, the head moves toward to the object as the result of the upper body movements but eye gaze remains stationary on the target object. The second case includes such actions as “folding a piece of paper” where the head fixates on the object involved in the action. During the head fixation, eye-movement recordings show that there can be a number of eye fixations ranging from 1 to 6. For example, when the performer folds a piece of paper, he spends 5 fixations on the different part of the paper and 1 look-ahead fixation to the location where he will place it after folding. In this situation, the head fixation is a better cue than eye fixations to segment the actions.

Based on the above analysis, we developed an an algorithm for action segmentation, which consists of the following three steps:

1. **Head fixation finding** is based on the positions and orientations of the head. We use \((x, y)\) position on the table plane and 3D orientations to calculate the velocity profile of the head, as shown in the first two rows of Figure 8.

2. **Eye fixation finding** is accomplished by a velocity-threshold-based algorithm. The algorithm significantly reduces the size and complexity of eye data by removing raw saccade(rapid eye movement) points and collapsing raw fixation points into a single representative tuple. A sample of the results of eye data analysis is shown in the third and fourth rows of Figure 8.

3. **Action Segmentation** is achieved by analyzing head and eye fixations, and partitioning the sequence of hand positions into the action segments(shown in the bottom row of Figure 8) based on the following three cases:

   - Within the head fixation, there are one or more than one eye fixations. This corresponds to actions, such as “folding”. “Action 3” in the bottom row of Figure 8 represents this kind of action.
During the head movement, the performer fixates on the specific object. This situation corresponds to actions, such as “picking up.” “Action 1” and “Action 2” in the bottom row of Figure 8 represent this class of actions.

During the head movement, eyes are also moving. It is most probable that the performer is switching attention after the completion of the current action.

Figure 8. Segmenting actions based on head and eye fixations. The first two rows: Point-to-point velocities of head data and the corresponding fixation groups(1–fixating, 0–moving). The third and fourth rows: Eye velocity data and the eye fixation groups(1–fixating, 0–moving) after removing saccade points. The bottom row: The results of action segmentation by integrating eye and head fixations.

Recognition of Human Actions

We collect the raw position $\mathbf{x} = (x, y, z)$ and the rotation $\mathbf{R} = (h, p, r)$ data of the hands from which feature vectors are extracted for recognition. In our system, we want to recognize the types of motion not the accurate trajectory of the hand. The same action performed by different people varies. Even in different instances of a simple action of “picking up” performed by the same person, the hand goes roughly in a different trajectory. This indicates that we cannot directly use the raw position data to be the features of the actions. As pointed out by Campbell et al.[Campbell et al., 1996], features designed to be invariant to shift and rotation perform better in the presence of shifted and rotated input. The feature vectors should be chosen such that large changes in the action trajectory produce relatively small excursions in the feature space, while the different types of motion produce relatively large excursions. In the context of our experiment, we calculated three element feature vectors consisting of the hand’s velocity on the table plane($d \sqrt{x^2 + y^2}$), the position in the z-axis, and the velocity of rotation in the 3 dimensions($d \sqrt{h^2 + p^2 + r^2}$).

Hidden Markov Models(HMMs) have been widely used in speech recognition with great success. Recently, HMMs have been applied within the computer vision community to address problems where time variation is significant, such as action recognition[Brand et al., 1997, Bobick, 1997] and gesture recognition [Starner and Pentland, 1996]. In our experiments, the six actions we sought to detect are: “picking up”, “placing”, “lining up”, “stapling”, “folding”, “unscrewing”. We model each action as a forward-chaining continuous HMM plus a HMM for any other motions. Each HMM consists of 6 states, each of which can jump to itself and the next two forward-chaining states(Figure 9). Given a sequence of feature vectors extracted from hand positions, we determine which HMM most likely generates those observations by calculating the log-probability of each HMM and picking the maximum. Further information can be obtained from [Yu and Ballard, 2002].
6.2 Linguistic modality

The inputs of linguistic processing system are acoustic signals collected from microphone. The linguistic studies always assume that a string of phonemes represented as a set of phonetic features is an appropriate representation of speech stream. Inherited from this assumption, our hypothesis is that the learner has knowledge of the phonetic structure of English prior to lexical development. Thus, we firstly convert acoustic signals into phonetic strings by a recurrent neural network, then we employ a modified algorithm of dynamic time-warping to compare and cluster phonetic representations.

Phoneme Recognition

We assume that some phonetic abilities are learned or pre-borned before lexical development. In light of this, The speaker-independent phoneme recognition system developed by Robinson [Robinson, 1994] is employed to distinguish the phonemes based on the recurrent error propagation network recognizer. The input of the system is the acoustic signal and a preprocessor generates a range of types of output including bark scaled spectrum, energy and estimates of formant positions. The Recurrent Neural Network (RNN) is an extension of a standard feed-forward neural network and it performs the mapping from a sequence of the acoustic features to a sequence of phone labels associated with those frames of parameterized speech. The network is trained with a variation on the stochastic gradient descent procedure which updates the weights by an adaptive step in the direction given by the sign of the gradient. The training data are from the TIMIT database — phonetically transcribed American English speech — which consists of read sentences spoken by 630 speakers from eight dialect regions of the United States. To train the network, each sentence is presented to the back propagation in time during the training procedure. The target outputs are set using the transcriptions provided in the TIMIT database. Once trained, a dynamic programming match is made to find the most probable symbol string of phonetic segments. Figure 10 shows output of the RNN that extracts phonetic strings from microphone input.

Phoneme Sequence Comparison

In our model, the processing of phonetic sequence is for two purposes: one is to find the longest common substring of two phonetic sequences, and the other is to cluster segmented utterances represented by phoneme subsequences into groups. In both cases, an algorithm of the alignment of phonetic sequences is necessary. In our case, the specific requirement of speech processing is to cope with the acoustic variability of spoken words in different contexts and by various talkers. Due to this variation, the output of phoneme recognizer are noised phoneme strings that are different from phonetic transcriptions of texts. In this context, the goal
of phonetic string matching is to identify sequences that might be different actual strings, but have similar pronunciations.

Several researchers have addressed the problem of aligning phoneme sequences for different purposes. Covington [Covington, 1996] developed an algorithm for the alignment of cognates based on depth-first search and a special distance function to align words. The work of Somers [Somers, 1998, Somers, 1999] concerns the implementation and testing of similarity metrics for the alignment of phonetic segments in transcriptions of children’s articulations with the adult model. He implemented three versions of the algorithm using different methods to compute the cost of substitution in sequence comparison. Kondrak [Kondrak, 1999, Kondrak, 2000, Kondrak, 2001] critically evaluated several approaches to phonetic alignment that have been reported within the last few years. He also presented an algorithm that combines a number of techniques developed for sequence comparison with a scoring scheme to compute phonetic similarity on the basis of multivalue features of individual phoneme.

**Comparing individual phonemes** To align phonetic sequences, we first need a metric for measuring distance between phonemes. We represent a phoneme by a binary vector in which every entry stands for a single articulatory feature called a distinctive feature. Distinctive features are those indispensable attributes of a phoneme that are required to differentiate one phoneme from another in a language. For example, the phonemes /b/ and /p/ in English are differentiated by the feature “voicing”, which in turn is an indispensable attribute in differentiating phonemes in English. When two phonemes differ by only one distinctive feature, they are known as being minimally distinct from each other; e.g., phonemes /p/ and /b/ are minimally distinct because the only feature that distinguishes them is “voicing”. Similarly, /p/ and /t/ are minimally distinct, separated by the feature “liability”. On the basis of the articulatory descriptions and their acoustic correlates, a metric can be constructed to describe phonemes in terms of distinctive features. Table 1 presents a distinctive feature system of English consonants and vowels. In this table, the columns represent the phonemes and the rows represent the features; the number one represents the presence of a feature in a phoneme and zero represents the absence of that feature. The numerical values, ones and zeros, are used here instead of plus and minus (mostly used in phonological literatures) in order to calculate numerically the differences between phonemes. We compute the distance between two individual phonemes by summing of all feature value differences for each of the 15 features in the vector: $\sigma(X, Y) = \sum_{i=1}^{n} |X_i - Y_i|$. Such metrics not only have the ability to distinguish a large number of different phonemes, but also stand on a much firmer theoretical base. The underlying assumption is that the number of binary features by which two given sounds differ is a good indication of their proximity. Moreover, phonological rules can often be expressed as a modification of a limited number of feature values. Therefore, sounds that differ in a small number of features are more likely to be related.

**Similarity score** Next we consider how to compare two phonetic strings. Basic to the literature on sequence comparison is the concept of distance from one sequence to another and algorithms for calculating the distance. Levenshtein distance [Kruskal, 1999] is the most common measurement of string distance which may be understood as the cost of converting one sequence into the other using substitutions and indels (insertions and deletions). In our system, we argue that distance-based comparison does not satisfy our requirement for two reasons. First, distance-based comparison does not consider the similarity of strings but only reflects difference of strings. For example, the cost (distance) of the long strings /aaabbbccc/ and /aaabbbcc/ is the cost of two substitutions. In the other case, the cost of two short strings /ab/ and /aa/ is only one substitution. For the clustering purpose, if we apply the cost (or distance) to measure the similarity of two strings, the short string pair seems to be more similar than the long string pair. Thus, the basic Levenshtein distance measurement will bias the results so that the shorter strings will always look similar to each other while the longer words would tend to need more editing cost for matching changes. This is because the Levenshtein distance measurement can only reflect difference but not the similarity of two strings. Secondly, in finding the similar portions of two sequences, if our approach is based on minimum distance and in the absence of any constraint
### Table 1. Distinctive features of phonemes

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>b</th>
<th>t</th>
<th>d</th>
<th>k</th>
<th>g</th>
<th>ch</th>
<th>jh</th>
<th>f</th>
<th>v</th>
<th>th</th>
<th>dh</th>
<th>s</th>
<th>z</th>
<th>sh</th>
<th>zh</th>
<th>m</th>
<th>n</th>
<th>ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonantal</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Vocalic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Continuant</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nasal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Anterior</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Coronal</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Back</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Voice</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Strident</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sonorant</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>l</td>
<td>hh</td>
<td>y</td>
<td>w</td>
<td>iy</td>
<td>ih</td>
<td>ey</td>
<td>eh</td>
<td>ae</td>
<td>uh</td>
<td>uw</td>
<td>ow</td>
<td>ao</td>
<td>ah</td>
<td>aa</td>
<td>oy</td>
<td>aw</td>
<td>er</td>
</tr>
<tr>
<td>Consonantal</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vocalic</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Continuant</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Nasal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anterior</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Coronal</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>High 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Back</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Voice</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Strident</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sonorant</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
| Tense | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21
to keep the portions long, the minimum distance of zero will always be possible with a meaningless align-
ment of very short portions, such as a single element from one sequence with a single element from the other.
Again, this is because the distance measurement only consider penalties (for indels and substitutions) and the
matching portions of strings do not get rewards (positive scores). This problem can be automatically avoid in
our approach based on the scores that are both positive (rewards for string matching) and negative (penalties
for mismatching).

We apply the concept of similarity to compare phonetic strings. A similarity scoring scheme assigns large
positive scores to pairs of matching segments; large negative scores to pairs of dissimilar segments, and small
negative scores to indels. The optimal alignment is the one that maximizes the overall score. The additional
advantage of the similarity approach is that it implicitly includes the length information in comparing the
segments. In our system, the phoneme recognizer encodes 39 phonemes. We compute the similarity matrix
which consists of $39 \times 39$ elements. Each element is assigned to a score which represents the similarity
of two phonemes. The diagonal elements are set to a positive value as the rewards of matching (with the
same phoneme). We set this value to be the half of the dimension of distinctive features ($\frac{1}{2} \times 16 = 8$). The
other elements in the matrix are assigned to negative values which correspond to the Hamming distance of
distinctive features between two phonemes.

![Figure 11. Subsampling the phonetic strings.](image)

**Comparison algorithm** The algorithm is based on the Dynamic Programming (DP) principle [Kruskal, 1999].
Fundamental to the idea of dynamic programming is the notion of string-changing operations. To determine
the extent to which two phonetic strings differ from each other, we need to define a set of primitive string
operations. By applying several string operations, one phonetic string is aligned with the other. Also, the
cost of each operation can be assigned so that we could measure the similarity of two phonetic strings by
summing both the cost of individual string operation in alignment and the reward of matching symbols. To
identify the phonetic strings that may be of similar pronunciation, the method needs to consider both the
duration and the similarity of phonemes. Thus, each phonetic string is subject not only to alternation by the
usual additive random error but also to variations in speed (the duration of the phoneme being uttered). Such
variations can be considered as compression and expansion of phoneme with respect to the time axis. Our
method include two additional steps to compare temporal sequences of phoneme. First, the outputs of the
phoneme recognizer are phonetic strings with timestamps of the beginning and the end of each phoneme.
We subsample the phonetic strings so that symbols in the resulting string have the same duration, which is
shown in Figure 11. Second, compared with the basic dynamic programming or the Dynamic Time-Warping
(DTP), our method performs the comparison in a way that allows for deletion-insertion-substitution as well
as compression-expansion. The followings are string operations allowed which are illustrated in Figure 12:

1. (a) Insertion: insert one phoneme symbol to make two strings similar.
2. (b) Deletion: delete extra phoneme in one string.
3. (c) Substitution: replace one phone symbol with the other one.
• (d) Compression: compress several phonemes in one phonetic string into one phoneme in the other phonetic string.

• (e) Expansion: expand one phoneme in one phonetic string into several phonemes in the other string.

Figure 12. The five string operations

Assume that we have two sequences $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_n$. Let $S_{ij}$ be the similarity score of the shortest possible time-warping between the initial subsequences of $a$ and $b$ containing $i$ and $j$ elements, respectively. The similarity score can be calculated recursively for successively larger subsequences as follows:

$$S_{ij} = \min \begin{cases} S_{i-1,j} + \frac{1}{2}w[a_i, b_j] \\ S_{i-1,j} + w_{del}[a_i] \\ S_{i-1,j-1} + w[a_i, b_j] \\ S_{i,j-1} + w_{ins}[b_j] \\ S_{i,j-1} + \frac{1}{2}w[a_i, b_j] \end{cases} \tag{5}$$

where $w$ is the metric of the similarity score and $w_{del}[a_i] = \min(w[a_i, a_{i-1}], w[a_i, a_{i+1}])$ and $w_{ins}[b_j] = \min(w[b_j, b_{j-1}], w[b_j, b_{j+1}])$. Figure 13 contains the modified DP algorithm to compute the similarity score of two phonetic strings.

6.3 Word Learning by Integrating Multimodal Input

Based on signal processing algorithms previously described, we have developed a method to learn the grounded lexicons of object names and action verbs. In this section, we present an approach to integrating multimodal data for speech segmentation shown in Figure 14 and lexical acquisition illustrated in Figure 15. We first provide an overview of our approach, then each module of the system is described in detail.

The inputs are phoneme sequences $(u_1, u_2, u_3, u_4)$ and possible meanings (objects and actions) of word extracted from non-speech perceptual inputs. First, for each meaning, we find the corresponding phoneme sequences in temporal proximity and save them in the short-term memory. Those phoneme sequences that have same possible meanings are categorized in the same bin. For instance, $u_1$ and $u_2$ are in the same bin labeled by the meaning “stapling”. We need to point out here that, since one utterance could consist of several words that have different meanings grounded in different modalities, it is possible that the utterance is selected and classified in different bins. For example, the utterance “stapling papers” is produced when the subject performs the action of “stapling” and looks toward the object “paper”. In this case, the utterance will be put in two bins, in which one corresponds to the object “paper” and the other is the group of the action “stapling”. Next, for each bin, we compute the similar substrings between any two phoneme sequences. For instance, if there are four phoneme strings in the bin of “paper”, we will get $C_4^2 = 6$ pairs of substrings. Figure 16 shows the result of comparing the utterance $u_2$ with the utterance $u_4$. After applying the algorithm of similar sequence spotting, four similar sequences extracted are /fl ow dcl/d/, /fl ow l/d/, /pcl p ey p er/ and /pcl p ay p hh er/.

We then treat those substrings as word-like units that could have grounded meanings.
Algorithm: phonetic string comparison

For $i = 0$ to $m$ do
  $s_{i0} = 0$
end for

For $j = 0$ to $n$ do
  $s_{0j} = 0$
end for

For $i = 0$ to $m$ do
  For $j = 0$ to $n$ do
    $S_{ij} = \min (\frac{1}{2} w[a_i, b_j],\frac{1}{2} w[a_i, b_j],\frac{1}{2} w[a_i, b_j])$
  end for
end for

Figure 13. The algorithm for computing the similarity of two phonetic strings

The next step is lexical acquisition shown in Figure 15. The word-like units in each bin are clustered based on the similarity on their phonetic strings. We select a prototype of each cluster and form the hypothesized lexical items, such as (/s t ei hh pl in ng/,”stapling”), (/p ey p er r/,”paper”) and (/f ow l d/,”fold”). Now we have hypothesized word-meaning associations and want to pick the most reliable lexical items. We apply the technique in machine translation to address this problem. We can model the probability of each word as a mixture model which consists of the conditional probabilities of each word given their possible meanings. In this way, EM algorithm is employed to find the best associations that will maximize the probabilities. The learning results are grounded lexicons with confidence values, which are saved in the long-term memory. As a developmental model, the confidence value of individual lexicon may change with new training data. If the confidence value is less than a threshold, the lexicon will be removed from the long-term memory. Also, there are new lexical items added in the memory from new training data. The following sections will describe each module of our approach in detail.

Utterance Spotting and Classification

In the previous section, we have described how to process the information from non-linguistic modalities to extract the possible grounded meanings of spoken words. In our system, the grounded meanings consist of objects and actions which will be mapped to object names and action verbs separately. Spoken utterances are categorized into several bins based on their possible associated meanings. The argument here is that the grounded meaning and the utterance in temporal proximity are very likely to be related. Therefore, the utterances that are produced in temporal proximity to the same grounded meanings(object or action) are categorized in the same bin labeled by that meaning. We need to point out here that one utterance could be classified into several bins in case there are several possible meanings co-occurring on the non-linguistic channels when the utterance is produced. Table 2 shows the collection of utterances in the bin of the action “stapling”. Those utterance are generated when the subjects perform the action of stapling papers.

Word-like Unit Discovery

One utterance may contain one or several words. Thus, we need to segment spoken utterances into discrete words. Different from traditional approaches that are purely based on linguistic information, we argue that
word-like unit clustering

hypothesized lexical items

EM algorithm

Figure 15. Spoken word learning
Table 2. The spoken utterances and the phonetic sequences when the subjects perform the action of stapling papers

<table>
<thead>
<tr>
<th>u1: stapling the paper</th>
<th>/n s t ey p l ng ih p ey p er/</th>
</tr>
</thead>
<tbody>
<tr>
<td>u2: staple them</td>
<td>/s t ey p l ng th ae m/</td>
</tr>
<tr>
<td>u3: now I will staple them</td>
<td>/n ow ah ay s t ey p l ih th ae m p/</td>
</tr>
<tr>
<td>u4: now I am stapling it</td>
<td>/m ow ay m s t ey p l ih ng ih t/</td>
</tr>
<tr>
<td>u5: I am stapling the paper</td>
<td>/a y ih n s t ey p l iy ng eh n p ey ih p hh er ch/</td>
</tr>
<tr>
<td>u6: stapling the paper</td>
<td>/sh s t ey p l ng ih p ey hh er/</td>
</tr>
</tbody>
</table>

non-linguistic contexts provide useful cues for speech segmentation. Specifically, since spoken utterances have been grouped into several categories based on their potential meanings as shown in the previous section, we can extract common substrings of the utterances in each group and treat them as word-like units. In the examples shown in Table 2, the common parts of results of u1 and u2 are /s t ey p l ng/ and /p ey p er/. Those word-like units are common patterns of sound in the specific context and serve as candidates for building the grounded lexical items. Assume that we have n utterances in a group, we needs to compare any pairs, so the total number of comparison is $C_n^2 = \frac{n \times (n-1)}{2}$.

From a perspective of phonetic sequence comparison, the method of local alignment is needed to compare two utterances that contain a word or phrase in common, even though they differ elsewhere. The approach presented here is the modified form of the method proposed by Kruskal and Sankoff [Kruskal and Sankoff, 1999]. To find the local alignment between phonetic strings $a = a_1a_2......a_m$ and $b = b_1b_2......b_n$, we first define the substrings $a^i = a_1a_2......a_i$ and $b^j = b_1b_2......b_j$ that are the portions of $a$ and $b$ ending at $a_i$ and $b_j$. Let $q_{ij}$ to be the maximum value of any local alignment between $a^i$ and $b^j$ that ends at $a_i$ and $b_j$. The best local alignment corresponds to $q_{max} = \max_{1 \leq i \leq m, 1 \leq j \leq n} q_{ij}$. $q_{ij}$ is calculated recursively as follow using the algorithms shown in Figure 13:

$$q_{ij} = \max \begin{cases} q_{i-1,j} + \frac{1}{2}w[a_i, b_j], \\ q_{i-1,j} + w_{del}[a_i], \\ q_{i,j-1} + w[a_i, b_j], \\ q_{i,j-1} + w_{ins}[b_j], \\ 0 \ (if \ all \ alignments \ have \ negative \ values) \end{cases}$$

The only difference between equation 5 and equation 6 is that $q_{ij}$ is set to be 0 if all possible alignments have negative values. This is because in this situation, it is better to set the current symbol to be the starting of the new string and form new local homology compared with including some dissimilar portions. When we find the maximal $q_{ij}$, we need to backtrack the optimal path to obtain the similar portions of phonetic strings. In practice, there might be several portions in a sequence which are similar to the corresponding portions in the other, and we will find all of the possible portions as candidates. Figure 16 shows the extraction of word-like units from two utterances.

**U2: I am going to fold the paper**

```
  ey m hh g cl g ow in ng m t uh f ow l d th hh pcl p ey p er
```

**U4: folding the paper like this**

```
  ey m hh g cl g ow in ng m t uh f ow l d th hh pcl p ey p er
```

Figure 16. An example of word-like unit spotting.
Word-like Unit Clustering

In this step, the substrings of word-like units extracted from the previous step are clustered based on their distances. The centroid of each cluster is then found and adopted as a prototype to represent this cluster. The distance between two substrings is computed by using the DP procedure shown in Figure 13. Given the distance matrix calculated above, we sought clusters of phonetically close word-like units by applying hierarchical agglomerative clustering. At each step of the procedure, we select the shortest distance in the matrix, and then group two data points which give rise to it. Since we need to iterate the procedure, we have to assign a distance from the newly formed cluster to all the remaining points. For this purpose we take a weighted average of the distances from each of the points in the cluster. The weighting is determined by the size of the elements being clustered. If the distance between Cluster $i$ and Cluster $j$ is minimal, then we form a cluster $C(i, j)$ and calculate the distance from each $k$ to the new cluster:

$$D(k, C(i, j)) = \frac{|i|}{|i| + |j|} \sum_i D(i, k) + \frac{|j|}{|i| + |j|} \sum_j D(j, k)$$

where $|k|$ is the number of elements in Cluster $k$. Note that the right-hand side is the sum of two terms, each of which is the product of a weighting and an average distance. Based on the clustering results, we select the prototype strings for each reliable clusters. Those prototype strings will be associated with their possible grounded meanings to build the hypothesized lexical items. For instance, two prototypes of word-like units in the context of performing the action of stapling are /s t ey p l ih ng/ and /s t ey p hh l/, which correspond to the words “staple” and “stapling” separately.

Learning Words

A central issue here is the correspondence problem. Thus, we need to utilize the co-occurrence of multimodal data to extract meaningful lexical items that associate visual representations of objects and body movements with spoken words. We take a novel view of this problem as the word correspondence problem in machine translation. For example, body movements can be looked as a “body language”. Thus, associating body movements and action verbs can be viewed as the problem of identifying word correspondences between English and “body language”. In light of this, we develop two approaches to building associations of meaning-word pairs. One is based on statistical measurement of co-occurring data, and the other applies EM algorithm to a mixture model of word-by-word alignment.

The method based on statistical measurement We have clustered the word-like units into clusters that can be expressed by symbols. The problem we need to address is how to identify symbol correspondence. Given any pair of symbols $(u, v)$ in which $u$ represents a visual cluster and $v$ represents an audio cluster, we
can measure the association between \(u\) and \(v\) by making use of mutual information [Gale and Church, 1991]. \(\phi^2\), a \(\chi^2\)-like statistic, seems to be a good measurement of correlation:

\[
\phi^2 = \frac{(p(u, v)p(u, v) - p(u)p(v))^2}{(p(u, v) + p(u))(p(u, v) + p(v))(p(u) + p(u))(p(v) + p(v))}
\]

(8)

where \(p(u, v)\) is the probability of \(u\) and \(v\) co-occurrence, \(p(u)\) is the probability of \(u\) occurrence but \(v\) is not in close temporal proximity; \(p(v)\) is the probability of \(v\) occurrence but \(u\) is not in close temporal proximity. \(\bar{p}(u, v)\) is the probability that neither \(u\) nor \(v\) occurs.

We compute the variance of \(\phi^2\) as follows:

\[
\text{var}(\phi^2) \approx \frac{\partial(\phi^2)}{\partial p(u, v)} \text{var}(p(u, v)) + \frac{\partial(\phi^2)}{\partial p(u)} \text{var}(p(u)) + \frac{\partial(\phi^2)}{\partial p(v)} \text{var}(p(v)) + \frac{\partial(\phi^2)}{\partial \bar{p}(u, v)} \text{var}(\bar{p}(u, v))
\]

(9)

\[
\text{var}(\phi^2) \approx \phi^2 (4 \frac{p(u, v)^2 \text{var}(\bar{p}(u, v)) + \bar{p}(u, v)^2 + p(u)^2 p(v) + (p(u) + p(v))^2 p(u)}{(p(u, v) + p(u))(p(u, v) + p(v))(p(u) + \bar{p}(u, v))(p(v) + \bar{p}(u, v))}) + \frac{1}{p(u, v) + p(u)} + \frac{1}{p(u, v) + p(v)} + \frac{p(v) + \text{var}(\bar{p}(u, v))}{(p(u) + p(v))^2}
\]

To determine whether \((u, v)\) is correlated, we apply the following rules based on the measurement of \(t\)-score:

1. If there is only one hypothesized pair that \(u\) and \(v\) co-occur, then we measure the confidence by \(t = \phi^2 / \sqrt{\text{var}(\phi^2)}\).

2. If there are multiple token \(v_1, v_2, \ldots, v_N\) that could be associated with \(u\), we can select the best pair \((u, v_i)\) as follows:

\[
\max_i \left( \frac{2 * \phi^2(u, v_i) - \sum_{j=1}^{N} \phi^2(u, v_j)}{\sqrt{\sum_{j=1}^{N} \text{var}^2(\phi^2(u, v_j))}} \right)
\]

(10)

We will apply this method to build grounded lexical items in the experiments depicted in the following sections.

**EM-based Method** The other way is to treated the problem as statistical translation. We take the view that every meaning \(m\), can be associated with a word-like phonetic string \(w\). \(P(m|w)\) is the probability of the meaning \(m\) given the word \(w\). We want to find the word \(w\) that is associated with the meaning \(m\). We can choose the word \(\tilde{w}\) which maximize \(P(w|m)\). Let \(N\) be the number of meaning, \(W_n\) be the number of words in the \(n\)th meaning, and \(a_n\) denotes a set of the possible assignments: \(a_{n1}, a_{n2}, \ldots, a_{nN_n}\), such that \(a_{nj}\) assigns the word \(w_{nj}\) to the meaning \(m_n\). \(P(a_{nj})\) is the probability that the meaning \(m_n\) is associated with a specific word \(w_{nj}\) and \(P(w_{nj}|m_n)\) is the probability of obtaining the word \(w_{nj}\) given the meaning \(m_n\). We use the model proposed by Brown et al. [Brown et al., 1990, Brown et al., 1993]:

\[
P(w|m) = \prod_{n=1}^{N} \prod_{j=1}^{W_n} P(a_{nj}) P(w = w_{nj}|b = m_n)
\]

(11)
We can estimate $P(w = w_{nj}|b = m_n)$ from data directly and the only incomplete data is $P(a_{nj})$. The problem can be stated as to find the maximum likelihood parameter:

$$\tilde{P}(a) = \arg \max_{P(a)} p(w|m, P(a)) = \arg \max_{P(a)} \sum_a p(a, w|m, P(a)) \quad (12)$$

The EM algorithm can be expressed in two steps. Let $P(a)^{[k]}$ be our estimate of the parameters at the $k$th iteration.

1. **E-step:** we compute the expectation of the log-likelihood function:

$$Q(P(a)|P(a)^{[k]}) = E[\log p(a, w|m, P(a)^{[k]})] \quad (13)$$

2. **M-step:** let $P(a)^{[k+1]}$ be that value of $P(a)$ which maximizes $Q(P(a)|P(a)^{[k]})$:

$$P(a)^{[k+1]} = \arg \max_{P(a)} Q(P(a)|P(a)^{[k]}) \quad (14)$$

In our case, the $Q(P(a)|P(a)^{[k]})$ function is given by

$$Q(P(a)|P(a)^{[k]}) = \sum_{n=1}^{N} \sum_{j=1}^{W_n} p(a_{nj}|w_{nj}, b_n, P(a)^{[k]}) \log[P(a_{nj})P(w = w_{nj}|b = m_n)] \quad (15)$$

We wish to find the assignment probabilities so as to maximize $Q(P(a)|P(a)^{[k]})$ subject to the constraints that for each $m_n$,

$$\sum_{j=1}^{W_n} p(a_{nj}) = 1 \quad (16)$$

Therefore, we introduce Lagrange multipliers $\lambda_n$ and seek an unconstrained maximization:

$$h(P(a), \lambda) = \sum_{n=1}^{N} \sum_{j=1}^{W_n} p(a_{nj}|w_{nj}, b_n, P(a)^{[k]}) \log[P(a_{nj})P(w = w_{nj}|b = m_n)] + \sum_{n=1}^{N} \lambda_n (1-\sum_{j=1}^{W_n} p(a_{nj})) \quad (17)$$

We compute derivatives with respect to the multipliers $\lambda$ and the parameters $P(a)$ to estimate $P(a_{nj})$:

$$P(a_{nj}) = \frac{p(a_{nj}|w_{nj}, b_n, P(a)^{[k]})P(w = w_{nj}|b = m_n)}{\sum_{j=1}^{W_n} p(a_{nj}|w_{nj}, b_n, P(a)^{[k]})P(w = w_{nj}|b = m_n)} \quad (18)$$

The algorithm sets an initial $P(a)^0$ to be flat distribution and performs the E-step and the M-step successively until convergence. Then for each meaning $m_n$, we select the word $j$ with the highest probability $P(a_{nj})$.

### 7 Experiment

We have designed a series of experiments to evaluate our model. This section first describes the experimental setup and the collection of multimodal data, then we present results and discussions.
7.1 Data Collection

We collected data from multiple sensors with timestamps. Monocular (left) eye position was monitored with an Applied Science Laboratories (ASL) Model 502 eye tracker, which is a head-mounted, video-based, IR reflection eye tracker [Pelz et al., 1999]. The eye position signals were sampled at 60 Hz and had a real time delay of 50 msec. The accuracy of the eye-in-head signal is approximately $1^\circ$ over a central $40^\circ$ field. Both pupil and first Purkinje image centroids are recorded, and horizontal and vertical eye-in-head position is calculated based on the vector difference between the two centroids. A Polhemus 3D tracker was utilized to acquire 6-DOF head and hand positions at $40 Hz$. The headband of the ASL holds a miniature “scene-camera” to the left of the user’s head that provides the video of the scene from the “first person” perspective. The video signals were sampled at the resolution of 320 columns by 240 rows of pixels at the frequency of 15 Hz. The acoustic signals are recorded using a headset microphones at a rate of 16 kHz with 16-bit resolution.

7.2 Evaluation Method

To evaluate the results of the experiments, we define the following five measures on the word-like unit and the grounded lexical items.

- **Semantic accuracy** is to measure the recognition accuracy of processing non-linguistic information, which consists of recognizing both human actions and visually attentional objects.

- **Segmentation accuracy** is to measure whether the beginning and the end of phonetic string of word-like units are word boundaries. For example, the string /k aa r/ is a positive instance while the string /k aa r l/ is negative. The phrases with correct boundaries are also treated as position instances for two reasons. One is that those phrases do not break word boundaries but only combine some words together. The other reason is that some phrases correspond to concrete grounded meanings, which are exactly spoken units we want to extract. For instance, the phrases, such as “pick up” or “line up”, specify some human actions.

- **Word spotting accuracy** is to measure the percentage of extracted word-like units that have grounded meanings. This reflects the semantic accuracy of word-like units that the model spots. The examples of good instances are /k eh t/(cat), /d uh k/(duck) and /b oy/ and the phonetic string /th ih s/ will be counted as an error.

- **Lexical finding accuracy (recall)** is to measure the percentage of word-meaning pairs that are spotted by the model.

- **Word learning accuracy (precision)** is a measure of the percentage of lexical items that are correctly associated with their meanings. For some phrases, such as “pick the paper up”, are treated as good instances if they are associated with the corresponding actions.

Segmentation accuracy and word spotting accuracy focus on evaluating two different aspects of extracted word-like units. Lexical finding accuracy demonstrates the effectiveness of correspondence finding based on machine translation techniques. Word learning accuracy shows the overall performance.

7.3 Experiment 1: Performing A Natural Task

Six subjects participated in the study. They were asked to be seated on the table and performed the everyday task of “stapling papers” while describing their actions verbally. Each subject was instructed to perform the task six times. Figure 18 shows the snapshots captured from the head-mounted camera when a subject
performed the task. We collected video, eye positions, head and hand positions in concert with spoken descriptions. Figure 19 demonstrates that the task in this experiment is to find the phonetic strings that correspond to action verbs from continuous spoken descriptions.

Figure 18. The snapshots of the action sequences when the subject performed the task of stapling several sheets of paper

Figure 19. Spoken word learning

For the evaluation purpose, we manually annotated speech data and calculate the frequencies of words. The action sequences in the experiments consist of several motion types: “pick up”, “line up”, “staple”, “fold” and “place”. The only object mentioned in the speech is “paper”. Several statistics of speech data obtained in the experiments are shown in Table 3. On average, spoken utterances contain six words, which illustrates the necessity of word segmentation from a continuous speech. Among all these words, approximately 30% of them are action verbs and object names that we want to spot and associate with their grounded meanings. This statistics further demonstrates the difficulty of learn lexical items from naturally co-occurring data. These annotations were only used for the evaluation purpose and our model did not use them as extra information.

Table 4 shows the results of the experiments. The recognition rate of the phoneme recognizer we used is 75% because it does not encode any language model and word model. Based on this result, the overall accuracy of speech segmentation is 69%. The error in word learning is mainly caused by a few words, such as “several” and “here”) that frequently occur in some contexts but do not have grounded meanings. Considering that the system processes natural speech and our method purely works in unsupervised mode without manually encoding and training any language model and word model, the accuracies for both speech segmentation and word learning are impressive.
Table 3. Speech data The second column shows the number of spoken utterances produced by the subjects. The third column shows the number of words in those utterances. The number of object names uttered is showed in the fourth column and the fifth one shows the percentage of object names in a spoken description.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Utterances</th>
<th>Words</th>
<th>Keywords</th>
<th>Percent keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>34</td>
<td>230</td>
<td>81</td>
<td>35%</td>
</tr>
<tr>
<td>MH</td>
<td>43</td>
<td>350</td>
<td>103</td>
<td>29%</td>
</tr>
<tr>
<td>SU</td>
<td>39</td>
<td>266</td>
<td>92</td>
<td>35%</td>
</tr>
<tr>
<td>JJ</td>
<td>49</td>
<td>263</td>
<td>52</td>
<td>20%</td>
</tr>
<tr>
<td>BS</td>
<td>43</td>
<td>256</td>
<td>69</td>
<td>27%</td>
</tr>
<tr>
<td>DB</td>
<td>36</td>
<td>233</td>
<td>66</td>
<td>28%</td>
</tr>
</tbody>
</table>

Table 4. Experimental results

<table>
<thead>
<tr>
<th>Subject</th>
<th>Semantic accuracy</th>
<th>Segmentation accuracy</th>
<th>Word spotting accuracy</th>
<th>Lexical finding accuracy</th>
<th>Word learning accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>89%</td>
<td>67%</td>
<td>86%</td>
<td>89%</td>
<td>86%</td>
</tr>
<tr>
<td>MH</td>
<td>91%</td>
<td>68%</td>
<td>87%</td>
<td>92%</td>
<td>91%</td>
</tr>
<tr>
<td>SU</td>
<td>93%</td>
<td>71%</td>
<td>92%</td>
<td>91%</td>
<td>92%</td>
</tr>
<tr>
<td>JJ</td>
<td>92%</td>
<td>68%</td>
<td>83%</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td>BS</td>
<td>89%</td>
<td>69%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
</tr>
<tr>
<td>DB</td>
<td>89%</td>
<td>72%</td>
<td>89%</td>
<td>91%</td>
<td>93%</td>
</tr>
</tbody>
</table>

7.4 Experiment 2: Reading A Picture Book

The experiment involves three subjects that were asked to tell a story based on the picture book “I went walking” [Williams and Vivas, 1989]. Several images collected from the head-mounted camera from the first-person perspective are in Figure 20. We focused on learning the names of objects that appear in the picture book since there are no actions involved in the experiment except for the action of “turning to the next page”. The task of learning object names is illustrated in Figure 21. We noticed that eye gaze is especially useful and can be utilized as the pointing movement of human body to bind objects in the picture book with speech. The temporal proximity between spoken utterances produced by subjects and visual attentional objects targeted by subjects’ eye gaze allows us to associate the visual representations of objects with their spoken names.

For the evaluation purpose, we manually annotated speech data and calculate the frequencies of words. The object names that appear in the transcription are: “boy”, “cat”, “dog”, “cow”, “pig”, “duck” and “horse”. Several statistics of speech data obtained in the experiments are shown in Table 5. On average, spoken utterances contain six words, which illustrates the necessity of word segmentation from a continuous speech. Among all these words, approximately 10% of them are object names that we want to spot and associate with their grounded meanings. This statistics further demonstrates the difficulty of learn lexical items from...
Table 5. **Speech data** The second column shows the number of spoken utterances produced by the subjects. The third column shows the number of words in those utterances. The number of object names uttered is showed in the fourth column and the fifth one shows the percentage of object names in a spoken description.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Utterances</th>
<th>Words</th>
<th>Object names</th>
<th>Percent object names</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>57</td>
<td>256</td>
<td>39</td>
<td>15%</td>
</tr>
<tr>
<td>MH</td>
<td>42</td>
<td>196</td>
<td>46</td>
<td>23%</td>
</tr>
<tr>
<td>DB</td>
<td>50</td>
<td>266</td>
<td>49</td>
<td>18%</td>
</tr>
</tbody>
</table>

Now the whole group of them, the boy, the duck, the cow, the cat, the horse are walking up to a big pig. Phonetic strings:

<table>
<thead>
<tr>
<th>phonetic strings:</th>
</tr>
</thead>
<tbody>
<tr>
<td>n ow th eh ah hh ow gcl g r uw</td>
</tr>
<tr>
<td>pcl p aw f hh ih m hh eh b oy</td>
</tr>
<tr>
<td>th eh d uh kcl k th ih k ow aw r</td>
</tr>
<tr>
<td>th eh k eh th eh h a s z a aa ar</td>
</tr>
<tr>
<td>w aw l k in ng ah p t ux aa ar</td>
</tr>
<tr>
<td>bcl b ih k p ly g.</td>
</tr>
</tbody>
</table>

**Figure 21. Spoken word learning**

naturally co-occurring data. These annotations were only used for the evaluation purpose and our model did not use them as extra information.

Table 6. **Experimental results**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Semantic accuracy</th>
<th>Segmentation accuracy</th>
<th>Word spotting accuracy</th>
<th>Lexical finding accuracy</th>
<th>Word learning accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>83%</td>
<td>69%</td>
<td>86%</td>
<td>89%</td>
<td>88%</td>
</tr>
<tr>
<td>MH</td>
<td>86%</td>
<td>68%</td>
<td>90%</td>
<td>91%</td>
<td>86%</td>
</tr>
<tr>
<td>DB</td>
<td>90%</td>
<td>70%</td>
<td>92%</td>
<td>92%</td>
<td>89%</td>
</tr>
</tbody>
</table>

Table 6 shows the results of the experiments. Segmentation accuracy is the most challenge measure because the system process raw acoustic data and the phoneme recognizer we employed can only obtain 75% accuracy without encoding any language model and word model. Word spotting accuracy is quite impressive that shows the effectiveness of our method. Lexical finding accuracy shows that eye movements as deictic reference to estimate the speaker’s referential intents, play a key role in spotting the objects in the picture book. Word learning accuracy evaluates the overall performance of the model.
8 Conclusion

We proposed a computational model of embodied language acquisition that can simulate the development of human-like capabilities. The intelligent system presented in this paper implements the model and represents a step toward machines that are grounded in the physical world. Our model is unique in that we encode both the social factors and temporal-spatial correlations between data from different modalities with our model.

In our current implementation, we associate the spoken word “car” with car images based on the statement that an infant is told “car” when shown a toy car. This assumption ignores the point that at the early stage, the classification abilities of infants in both object recognition and speech recognition are not well developed. Thus, in order to more fully simulate infant development, it is necessary to develop a multimodal learning algorithm without assuming that the abilities to classify acoustic and visual data have been already developed. In light of this, the immediate goal is to integrate this work with the cross-modal unsupervised classification algorithm [de Sa and Ballard, 1998], which uses data from one modality to teach clustering in the other modality based on minimizing the disagreements of the clustering outputs between two modalities. In this way, the system could develop single modality classification abilities as well as cross-modal integration and association abilities, which is a more biologically plausible model of human development.
References


