

# A MULTIMODAL LEARNING INTERFACE FOR WORD ACQUISITION

*Dana H. Ballard and Chen Yu*

University of Rochester  
 Department of Computer Science  
 Rochester, NY, 14627, USA  
 {dana,yu}@cs.rochester.edu

## ABSTRACT

We present a multimodal interface 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 temporal correlations of data from different modalities to build hypothesized lexical items. From those items, an EM-based method selects correct word-meaning pairs. Successful learning has been demonstrated in the experiment of the natural task of “stapling papers”.

## 1. INTRODUCTION

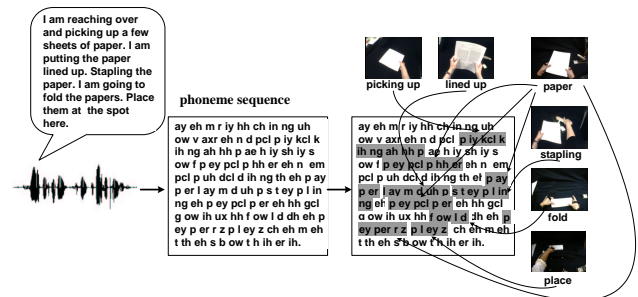
The next generation of computers is expected to interact and communicate with users in a cooperative and natural manner when users carry out everyday activities. Towards this end, a truly intelligent human-machine interface should be able to understand what people are doing and their intentions, and perform helpful speech acts, such as confirming user requests, answering questions and providing related information through speech. In this way, computers will be seamlessly integrated into our everyday lives and work as intelligent observers and human-like assistants.

To progress towards this goal, computers need to know the sound patterns of spoken words and understand their meanings. Most existing speech recognition systems rely on purely acoustics-based statistical models, such as hidden Markov models and hybrid connectionist models. These systems have two inherent disadvantages. First, they require a training phase in which large amounts of spoken utterances paired with manually labeled transcriptions are needed to train the model parameters. This training procedure is time-consuming and needs human expertise to label spoken data. Second, these systems transform acoustic signals to symbolic representations (texts) without regard to their grounded meanings. Humans need to interpret the meanings of these symbols based on their own knowledge. For instance, a speech recognition system can map the sound pattern “car” to the string “car”, but it does not know what this string means.

To overcome the above shortcomings, a few recent studies proposed several unsupervised methods for learning words from lin-

guistic and contextual inputs. Among them, the work of Roy [1] is particularly relevant to our work. He proposed a computational model of infant language acquisition, which utilizes temporal correlation of speech and vision to associate spoken utterances with a corresponding object’s visual features. The model has been implemented to process a corpus of audio-visual data from infant-caregiver interactions. Our work differs from his in that we focus on building a multimodal learning interface that is grounded in naturally-occurring multisensory information in everyday activities. Our learning method incorporates an extensive description of gaze, head and hand movements as well as visual data to provide contextual information when spoken words are uttered.

This paper describes a multimodal learning system that is able to learn perceptually grounded meanings of words from user’s everyday activities. The only requirement is that users need to describe their behaviors verbally while performing those day-to-day tasks. To learn a word (shown in Figure 1), the system needs to discover its sound pattern from continuous speech, recognize its meaning from non-speech context, and associate these two. The range of problems we need to address in this kind of unsupervised word learning is substantial, so to make concrete progress, this paper focuses on how to associate visual representations of objects with their spoken names and map body movements to action verbs. Our work suggests a new trend in developing human-computer interfaces that can automatically learn words by sharing user-centric multisensory information.



**Fig. 1. The problems in word learning.** The raw speech is firstly converted to phoneme sequences. The goal of our method is to discover phoneme substrings that correspond to the sound patterns of words and then infer the meanings of those words from non-speech modalities.

## 2. A MULTIMODAL LEARNING INTERFACE

In typical scenarios, a user performs everyday tasks while describing his/her actions verbally. To learn words from user’s spoken descriptions, three fundamental problems needed to be addressed are:

(1) action recognition and object recognition to provide grounded meanings of words encoded in non-speech contextual information, (2) speech segmentation and word spotting to extract the sound patterns that correspond to words, (3) association between spoken words and their grounded meanings.

### 2.1. Recognition of Actions and Objects

The non-speech inputs of the system consist of visual data from a head-mounted camera, head and hand positions in concert with gaze-in-head data. Those data provide a context in which spoken utterances are produced. Thus, the possible meanings of spoken words that users utter are encoded in this context, and we need to extract those meanings from raw sensory inputs. Specifically, the system should spot and recognize actions from user’s body movements, and discover the objects of user interest.

We observe that in accomplishing well-learned tasks, the user’s focus of attention is linked with body movements. In light of this, our method firstly utilizes eye and head movements as cues to estimate the user’s focus of attention. Attention, as represented by gaze fixation, is then utilized for spotting the target object of user interest. Attention switches are calculated and used to segment a sequence of hand movements into action units which are then recognized by Hidden Markov Models(HMMs). The results are two temporal sequences of grounded meanings as depicted by the box labeled “contextual information” in Figure 2. Further information about attentional object spotting and action recognition can be obtained from [2, 3].

### 2.2. Speech Processing

We describe our methods of phoneme recognition and phoneme string comparison in this subsection, which provide a basis for further processing. Detailed technical descriptions of algorithms can be obtained from [4].

#### 2.2.1. Phoneme Recognition

We have implemented an endpoint detection algorithm to segment the speech stream into several spoken utterances. Then the speaker-independent phoneme recognition system developed by Robinson [5] is employed to convert spoken utterances into phoneme sequences. The method is based on Recurrent Neural Networks (RNN) that perform the mapping from a sequence of the acoustic features extracted from raw speech to a sequence of phonemes. The training data of RNN 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 networks, each sentence is presented to the recurrent back-propagation procedure. The target outputs are set using the phoneme transcriptions provided in the TIMIT database. Once trained, a dynamic programming match is made to find the most probable phoneme sequence of a spoken utterance, for example, the boxes labeled “phoneme strings” in Figure 2.

#### 2.2.2. Comparing Phoneme Sequences

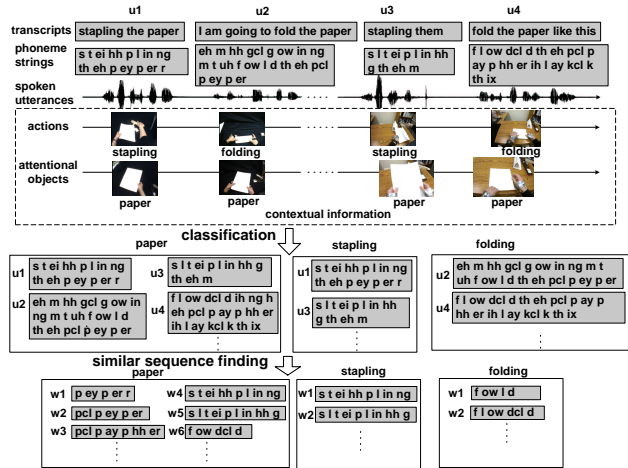
In our system, the comparison of phoneme sequences has two purposes: one is to find the longest similar substrings of two phonetic sequences (word-like units spotting described in Subsection 2.3.1), and the other is to cluster segmented utterances represented by phoneme sequences into groups (word-like units clustering presented in Subsection 2.3.2). In both cases, an algorithm of the alignment of phoneme sequences is a necessary step. Given raw speech input, the specific requirement here is to cope with the acoustic variability of spoken words in different contexts and by

various talkers. Due to this variation, the outputs of the phoneme recognizer previously described are noisy phoneme strings that are different from phonetic transcriptions of text. In this context, the goal of phonetic string matching is to identify sequences that might be different actual strings, but have similar pronunciations.

To align phonetic sequences, we first need a metric for measuring distances between phonemes. We represent a phoneme by a 15-dimensional binary vector in which every entry stands for a single articulatory feature called a distinctive feature. Those distinctive features are indispensable attributes of a phoneme that are required to differentiate one phoneme from another in English. We compute the distance between two individual phonemes as the Hamming distance. Based on this metric, a modified dynamic programming algorithm is developed to compare two phoneme strings by measuring their similarity. 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 the operations of insertion and deletion to convert one sequence to another. The optimal alignment is the one that maximizes the accumulated score. See [4] for further information about our method of phoneme sequence comparison.

### 2.3. Word Learning

In this subsection, we describe our approach to integrating multi-modal data for word acquisition. We divide this problem into two basic steps: speech segmentation shown in Figure 2 and lexical acquisition illustrated in Figure 4.

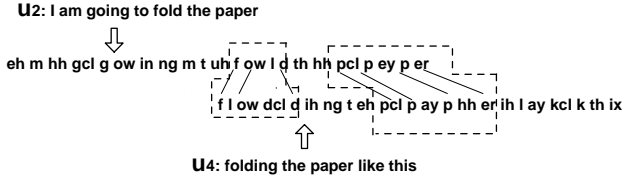


**Fig. 2. Word-like unit segmentation.** Spoken utterances are categorized into several bins that correspond to temporally co-occurring actions and attentional objects. Then we compare any pair of spoken utterances in each bin to find the similar subsequences that are treated as word-like units.

#### 2.3.1. Word-like Unit Spotting

Figure 2 illustrates our approach to spotting word-like units in which the central idea is to utilize non-speech contextual information to facilitate word spotting. The reason we use the term “word-like units” is that some actions are verbally described by verb phrases (e.g. “line up”) but not single action verbs. The inputs are phoneme sequences ( $u_1, u_2, u_3, u_4$ ) and possible meanings of words (objects and actions) extracted from non-speech perceptual inputs. Those phoneme utterances are categorized into several bins based on their possible associated meanings. For each meaning, we find the corresponding phoneme sequences uttered in temporal proximity, and then categorize them into the same bin labeled by that meaning. For instance,  $u_1$  and  $u_3$  are temporally correlated

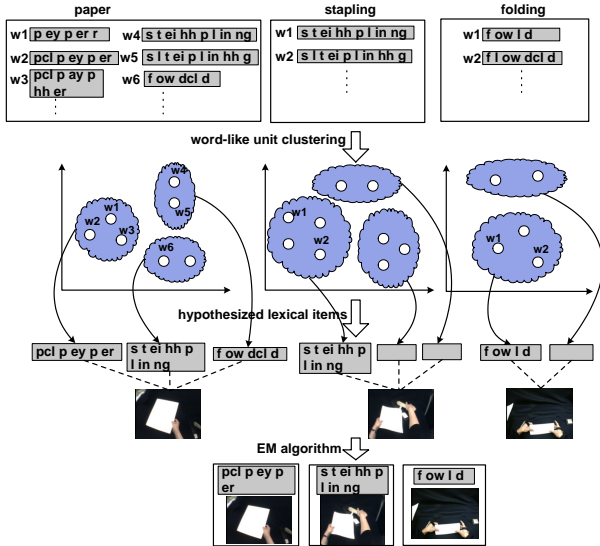
with the action “stapling”, so they are grouped in the same bin labeled by the action “stapling”. We need to point out here that, since one utterance could be temporally correlated with multiple meanings grounded in different modalities, it is possible that an utterance is selected and classified in different bins. For example, the utterance “stapling a few sheets of paper” is produced when a user performs the action of “stapling” and looks toward the object “paper”. In this case, the utterance is put into two bins: one corresponding to the object “paper” and the other labeled by the action “stapling”. Next, based on the method described in Subsection 2.2.2, we compute the similar substrings between any two phoneme sequences in each bin to obtain word-like units. Figure 3 shows an example of extracting word-like units from the utterance  $u_2$  and  $u_4$  that are in the bin of the action “folding”.



**Fig. 3. An example of word-like unit spotting.** The similar substrings of two sequences are /f ow l d/ (fold), /f l ow dcl d/ (fold), /pcl p ey p er/ (paper) and /pcl p ay p hh er/ (paper).

### 2.3.2. Word-like Unit Clustering

As shown in Figure 4, the extracted phoneme substrings of word-like units are clustered by a hierarchical agglomerative clustering algorithm that is implemented based on the method described in Subsection 2.2.2. The centroid of each cluster is then found and adopted as a prototype to represent this cluster. Those prototype strings are associated with their possible grounded meanings to build hypothesized lexical items. Among them, some are correct ones, such as /s t ei hh p l in ng/ (stapling) associated the action of “stapling”, and some are incorrect, such as /s t ei hh p l in ng/ (stapling) paired with the object “paper”. Now that we have hypothesized word-meaning pairs, the next step is to select reliable and correct lexical items.



**Fig. 4. Word learning.** The word-like units in each bin are clustered based on the similarities of their phoneme strings. The EM-algorithm is applied to find lexical items from hypothesized word-meaning pairs.

### 2.3.3. Multimodal Integration

Next, we utilize the co-occurrence of multimodal data to select meaningful semantics 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 with action verbs can be viewed as the problem of identifying word correspondences between English and “body language”. In light of this, we apply a technique from machine translation to address this problem. We model the probability of each word as a mixture model that consists of the conditional probabilities of each word given its possible meanings. In this way, the Expectation-Maximization(EM) algorithm is employed to find the reliable associations of spoken words and their grounded meanings that will maximize the probabilities.

We assume that every meaning  $m$  can be associated with a word-like phoneme string  $w$ . We can find the word  $\hat{w}$  that is associated with the meaning  $m$  by choosing the one that maximizes  $P(w|m)$ . Let  $N$  be the number of meanings,  $W_n$  be the number of words in the  $n$ -th meaning, and let  $a_n$  represent a set of the possible assignments:  $(a_{n1}, a_{n2}, \dots, a_{nW_n})$ , 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 similar to that of Duygulu et al. [6]:

$$p(w|m) = \prod_{n=1}^N \prod_{j=1}^{W_n} p(a_{nj})p(w_{nj}|m_n) \quad (1)$$

We can estimate  $p(w_{nj}|m_n)$  from data directly and the only incomplete data is  $p(a_{nj})$ . The remaining problem is to find the maximum likelihood parameter:

$$\tilde{p}(a) = \arg \max_{p(a)} p(w|m, p(a)) = \arg \max_{p(a)} \sum_a p(a, w|m, p(a)) \quad (2)$$

The EM algorithm can be expressed in two steps. Let  $p(a)^{[k]}$  be our estimate of the parameters at the  $k$ th iteration. In **E-step**: we compute the expectation of the log-likelihood function:

$$\begin{aligned} Q(p(a)|p(a)^{[k]}) &= E[\log p(a, w|m, p(a)^{[k]})] \\ &= \sum_{n=1}^N \sum_{j=1}^{W_n} p(a_{nj}|w_{nj}, m_n, p(a)^{[k]}) \times \\ &\quad \log [p(a_{nj})p(w_{nj}|m_n)] \end{aligned} \quad (3)$$

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

$$p(a)^{[k+1]} = \arg \max_{p(a)} Q(p(a)|p(a)^{[k]}) \quad (4)$$

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

$$\sum_{j=1}^{W_n} p(a_{nj}) = 1 \quad (5)$$

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

$$h(p(a), \lambda) = Q(p(a)|p(a)^{[k]}) + \sum_{n=1}^N \lambda_n (1 - \sum_{j=1}^{W_n} p(a_{nj})) \quad (6)$$

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

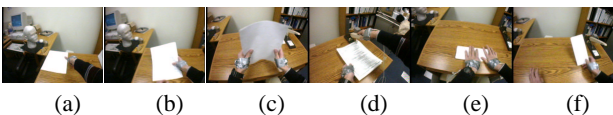
$$p(a_{n_j}) = \frac{p(a_{n_j}|w_{n_j}, m_n, p(a)^{[k]})p(w_{n_j}|m_n)}{\sum_{j=1}^{W_n} p(a_{n_j}|w_{n_j}, m_n, p(a)^{[k]})p(w_{n_j}|m_n)} \quad (7)$$

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  $m_n$ , the system selects all the words with the probability  $p(a_{n_j})$  greater than a pre-defined threshold. In this way, one meaning can be associated with multiple words. This is because people may use different names to refer to the same object and the spoken form of an action verb can be expressed differently. For instance, the phoneme strings of both “staple” and “stapling” correspond to the action of stapling. Therefore, the system is developed to learn all the spoken words that have high probabilities in association with a meaning.

### 3. EXPERIMENT

We collected data from multiple sensors with timestamps. A Polhemus 3D tracker was utilized to acquire 6-DOF hand and head positions at 40Hz. A user wore a head-mounted eye tracker from Applied Science Laboratories(ASL). The headband of the ASL holds a miniature “scene-camera” to the left of the user’s head, which 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 15Hz. The gaze positions on the image plane were reported at the frequency of 60Hz. The acoustic signals were recorded using a headset microphone at a rate of 16 kHz with 16-bit resolution.

Six users participated in the experiments. They were asked to sit at a table and performed the task of “stapling papers” while describing their actions verbally. Each user performed the task six times. Figure 5 shows the snapshots captured from the head-mounted camera when a user performed the task.



**Fig. 5.** The snapshots of an action sequence when a user performed the task of stapling several sheets of paper: (a) picking up papers (b) placing them to the position close to the body (c) lining up (d) stapling (e) folding (f) placing them to the target location.

To evaluate experimental results, we define the following three measures: (1) **Semantic accuracy** is to measure the recognition accuracy of processing non-linguistic information, which consists of recognizing both human actions and visual attentional objects. (2) **Speech segmentation accuracy** is to measure whether the beginning and the end of phoneme strings of word-like units are correct word boundaries. (3) **Word learning accuracy** is to measure the percentage of successfully segmented words that are correctly associated with their meanings.

Table 1 shows the results of three measures. 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 71.6%. Naturally, an improved phoneme recognizer based on a language model would improve the overall results, but the intent here is to study the model-independent learning interface. The error in word learning is mainly caused by a few words (such as “several” and “here”)

**Table 1.** Results of word acquisition

	semantics	speech segmentation	word learning
overall	92.9%	71.6%	90.2%
picking up	96.3%	73.2%	89.6%
placing	93.6%	68.9%	92.3%
lining up	73.2%	73.6%	88.9%
stapling	86.2%	72.9%	86.3%
folding	83.6%	71.5%	86.9%
paper	96.7%	70.8%	92.1%

that frequently occur in some contexts but do not have grounded meanings. Considering that the system processes natural speech and our method works in unsupervised mode without manually encoding any linguistic information, the accuracies for both speech segmentation and word learning are impressive.

### 4. CONCLUSIONS

This paper presents a multimodal learning interface for word acquisition. The system is able to learn the sound patterns of words and their semantics while users perform everyday tasks and provide spoken descriptions of their behaviors. From the perspective of multimodal integration, we believe that a powerful constraint in multisensory data is coherence in time and space. Our method of learning words exhibits how to capitalize on this constraint for word acquisition without manual transcriptions and human involvement. From an engineering perspective, our system demonstrates a new approach to developing human-computer interfaces, in which computers seamlessly integrate in our everyday lives and are able to learn lexical items by sharing user-centric multisensory information.

### 5. REFERENCES

- [1] Deb Roy, “Integration of speech and vision using mutual information,” in *Proceedings of Int. Conf. Acoustics, Speech and Signal Processing(ICASSP)*, Istanbul, Turkey, June 2000.
- [2] Chen Yu, Dana H. Ballard, and Shenghuo Zhu, “Attentional object spotting by integrating multisensory input,” in *IEEE Proceedings of the 4th International Conference on Multimodal Interface*, Pittsburg, PA, U.S.A., Oct 2002.
- [3] Chen Yu and Dana H. Ballard, “Learning to recognize human action sequences,” in *IEEE Proceedings of the 2nd International Conference on Development and Learning*, Boston, U.S., June 2002, pp. 28–34.
- [4] Chen Yu and Dana H. Ballard, “A computational model of embodied language learning,” Tech. Rep. 791, Department of Computer Science, University of Rochester, 2002.
- [5] Tony Robinson, “An application of recurrent nets to phone probability estimation,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, 1994.
- [6] P. Duygulu, K. Barnad, J.F.G. de Freitas, and D.A. Forsyth, “Object recognition as machine translation: learning a lexicon for a fixed image vocabulary,” in *Proceedings of European Conference on Computer Vision*, Copenhagen, 2002.