Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application

Edward C. Kaiser
Center for Human Computer Communication
20000 NW Walker Rd.
Beaverton, OR 97006 USA
+1 503 748 1608
kaiser@cse.ogi.edu

ABSTRACT

Our goal is to automatically recognize and enroll new vocabulary in a multimodal interface. To accomplish this our technique aims to leverage the mutually disambiguating aspects of co-referenced, co-temporal handwriting and speech. The co-referenced semantics are spatially and temporally determined by our multimodal interface for schedule chart creation. This paper motivates and describes our technique for recognizing out-of-vocabulary (OOV) terms and enrolling them dynamically in the system. We report results for the detection and segmentation of OOV words within a small multimodal test set. On the same test set we also report utterance, word and pronunciation level error rates both over individual input modes and multimodally. We show that combining information from handwriting and speech yields significantly better results than achievable by either mode alone.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning – *language acquisition, knowledge acquisition*. H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Interaction styles, Input devices and strategies*.

General Terms

Design, Performance, Experimentation.

Keywords

Multimodal interaction, vocabulary learning, mutual disambiguation.

1. INTRODUCTION

Machines are moving closer to being observant and intelligent assistants for humans [1-4]. However, multimodal system interfaces incorporating speech, gesture, gaze recognition and objection selection mechanisms [5], are typically implemented with fixed knowledge spaces, as are unimodal spoken dialogue systems. Automatically acquiring new knowledge as they are running, particularly by a single, natural demonstration would significantly enhance the usability of such systems. Dynamically augmenting vocabularies, pronunciation lexicons and language models is an active area of research in speech and gesture recognition [6-11]. Machines or systems that assist humans in real-time tasks need to be able to learn from being shown — through sketch [12, 13], handwriting [14], teleassistance [15], speech [16], or multimodally through handwriting and speech as in the work we describe here.

Our aim, as for [1] in their work on designing humanoid robots to be cooperative partners for people, is that our system will be able to "acquire new capabilities … as easy and fast as teaching a person." To take some first step in this direction we have focused our efforts on a single, important capability (within the scope of what humans ultimately need to teach a cooperative machine): establishing a common, working vocabulary of spoken words — taught to the machine by natural demonstration as the system is running. We support this capability through our multimodal new-vocabulary recognition (MNVR) technique.

Most computer systems require users to type or speak the right words. However, users — particularly new or intermittent users — often use the wrong words. This is an aspect of the classic *vocabulary problem* [17]. It has been noted in studies of information retrieval searches that users seldom use the same word to refer to a particular concept — even a set of the 15 most common aliases for a concept was shown to cover only 60-80% of the search vocabulary people used for that concept. Our MNVR approach combines handwriting recognition and out-of-

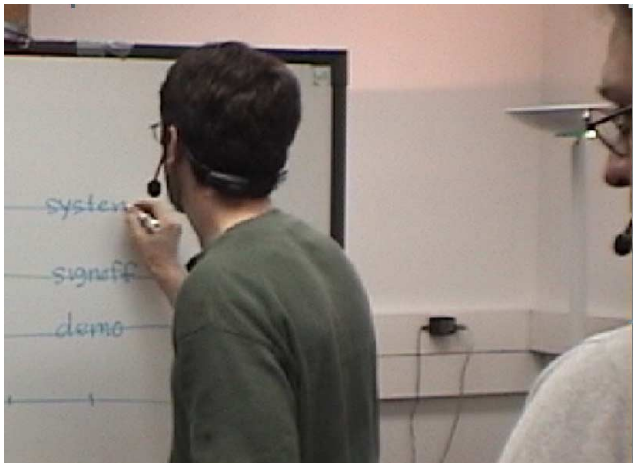

Figure 1: Using handwriting and speech to label task-lines in a multimodal, multi-person schedule meeting.

IUI '05, January 9-12, 2005, San Diego, California, USA.

vocabulary (OOV) speech recognition, to leverage two of the richest communicative modes we as humans have available for acquiring new vocabulary. Others have designed OOV speech recognition systems [7, 18-21], but they are not used in a multimodal context. Related multimodal systems that extract words from statistical associations of object/phone-sequences or action/phone-sequences [22-24] do not leverage the grammatical and linguistic context in the same way we are proposing, nor do they use handwriting as input.

The key components of our approach are (1) highly constrained, real-time out-of-vocabulary (OOV) speech recognition, (2) standard handwriting recognition[1], and (3) a multimodal task domain capable of assigning semantics on the basis of spatial, temporal and in some cases linguistic aspects of the input signals (depicted in Fig. 1). In our task domain the system functions as a real-time, multimodal interface to Microsoft (MS) Project [25]. Recognition of multiple input modes (e.g. speech, 2D pen, handwriting, and 3D gesture) allows the system to dynamically build an MS Project chart as the meeting proceeds. OOV constituent names, like the task-line labels show in Fig. 1, are recognized in real-time and enrolled as part of the MS Project chart.

2. PREVIOUS WORK

Capturing events in which handwriting and speech co-occur and carry redundant information (e.g., as part of labeling a constituent on the whiteboard chart) is integral to our approach. In the human-computer-interaction literature on bi-modal, speech and pen, wizard-of-oz systems for map-based and form-filling tasks speech and handwriting have been found to co-occur redundantly in this way for less than 1% of all interactions [26, 27]. However, in the educational-technology literature on human-human, computer-mediated interactions like the presentation of distance-learning lectures as much as 15% of all pen interactions involve redundant handwriting and speech [28]. One recent study has found that in a tablet-PC-based distance-learning lecture presentation application 100% of the randomly sampled instances of handwritten text were accompanied by semantically redundant speech [29].

Systems that augment speech recognition by visually extracted face and lip movement features [30] employ an *early-fusion* approach that discriminatively combines both input streams in a single feature space. Previous work in our group [5, 31] employs a *late-fusion* approach, which instead combines the output of separate modes after recognition has occurred. This is also true of for the version of our MNVR technique for combining speech and handwriting outputs that we report on this paper. For our test bed, schedule-chart application *early-fusion* is problematic, because the temporal relation between handwriting and the speech associated with it is not yet clearly characterized.

2.1 Hybrid Fusion Phone Recognition

A third possibility, aside from either early or late fusion, is a *hybrid re-recognition* (HRR) approach that takes initial recognition results from all input modes, and then uses information from one input mode to constrain a subsequent re-recognition pass on the input from another mode. We are now actively exploring this approach for MNVR. A variation of this approach has been used by Chung *et al* [7] in their speak and spell technique that allows new users to enroll their names in a spoken dialogue system. User's first speak their name and then spell it, in a single utterance. Thus, there is a single input mode (speech) but separate recognition passes: the first pass employs a letter recognizer with an unknown word model, followed by a second pass OOV recognizer constrained by a sub-word-unit language model and the phonemic mappings of the hypothesized letter sequences from the first pass. On a test set of 219 new name utterances this system achieves a letter-error-rate (LER) of 12.4%, a word-error-rate (WER) of 46.1%, and a pronunciation-error-rate (PER) of 25.5%.

The sub-word-units used by Chung *et al* for modeling OOV words are those of [19]. These are multi-phone sub-word units extracted from a large corpus with clustering techniques based on a mutual information (MI) metric. Bazzi [32] shows that using MI generated sub-word-units outperforms a system that uses only syllabic sub-word units; however, it is interesting to note that 64% of his MI sub-word units are still actual syllables. Chung *et al* extend the space of sub-word units by associating sub-word-unit pronunciations with their accompanying spellings, thereby making a finer grained, grapho-phonemic model of the sub-word-unit space.

Galescu [20] uses an approach similar to Chung *et al*'s in that he chooses *grapheme-to-phoneme correspondences* (GPCs) as his sub-word-units. He uses an MI mechanism like Bazzi's to cluster multi-GPC units (MGUs). His language model (in which MGUs are treated as words) was trained on 135 million words from the HUB4 broadcast news transcriptions, with MGUs first being extracted from the 207,000 unique OOV occurrences in that training data. He tested OOV word modeling on the individual OOV terms occurring in 186 test utterances, yielding between a 22.9% - 29.6% correct transcription rate, and between a 31.2% - 43.2% correct pronunciation rate. Applying the OOV language model to the complete utterances in the 186 instance test sets yielded a false alarm rate of under 1%, a relative reduction in overall WER of between 0.7% - 1.9%, with an OOV detection rate of between 15.4% - 16.8%. For a large vocabulary system these are encouraging results: there is a reduction in WER, whereas other systems report increases in WER.

In designing our algorithm for OOV recognition and multimodal new vocabulary enrollment we have chosen not to use GPCs because they require a large training corpus, whereas our static syllable grammar requires none. Since there is evidence that many if not most MI extracted clusters are actual syllables (64% in Bazzi's work), we feel that the loss in recognition accuracy may be balanced out by the savings in not having to acquire a task-specific corpus.

2.2 Multimodal Semantic Grounding

Roy [33] developed robotic and perceptual systems that can perceive visual scenes, parse utterances spoken to describe the scenes into sequences of phonemes, and then over time and repeated exposure to such combinations extract phonetic representations of words associated with objects in the scene — multimodal semantic grounding. Rather than using string comparison techniques for measuring the similarity between two

[1] Part of the NISSketch™ recognition package from Natural Interaction Systems, LLC: http://www.naturalinteraction.com

speech segments (represented as phone-sequences), he generates an HMM based on a segment's best phone-sequence representation. Then each segment's speech is passed through the other segment's HMM. The normalized outputs are then combined to produce a distance metric. Of the words extracted by this method with audio only input only 7% were lexically correct, while with both visual and audio input (combined through a further Mutual Information measure) 28% of the words extracted were lexically correct, and of those half were correct in their semantic association with the visual object. In related work Gorniak *et al* [22] use these techniques to augment a drawing application with an adaptive speech interface, which learns to associate segmented utterance HMMs with button click commands (rather than associating OOV recognitions with handwriting and contextual semantics as we do).

Yu & Ballard [24] have developed an intelligent perceptual system that can recognize attentional focus through velocity and acceleration-based features extracted from head-direction and eye-gaze sensor measurements, together with some knowledge of objects in the visual scene — based on head-mounted scene cameras. Within that context, measurements of the position and orientation of hand movements (tracked by tethered magnetic sensor) are used to segment spoken utterances describing the actions into phone-sequences associated with the action (e.g. stapling papers, folding papers, etc.), and over time and repeated associations phonetic representations of words describing both the objects and the actions performed on those objects can be statistically extracted.

Rather than using individual HMMs as the basis of measuring distance between phonetic sequences (as Roy does), Yu & Ballard use a modified Levenshtein distance measure based on distinctive phonetic features. In 960 utterances (average six words per utterance) they identify 12% of the words as either action verbs or object names that their system attempts to pair with meanings expressed in the other perceptual modes (gaze, head and hand movement). Their system identifies actions and attentional objects (thus the semantics/meanings of the actions) in non-linguistic modes in 90.2% of the possible cases. Of all possible word-meaning pairs they recall 82.6% of them, and over those recalled pairs achieve an accuracy of 87.9% for correctly pairing words with their associated meanings. The word-like units their method extracts have boundaries that are word-level correct 69.6% of the time. In general the phone-level recognition rate is 75% correct, but their system is offline and as they do not attempt to update the system's vocabulary they don't report phone-error rates.

3. OUR APPROACH

Our technique enrolls new words into the vocabulary of a system that tracks a collaborative, multi-person scheduling meeting (Fig. 1): one person standing at a touch sensitive whiteboard creating a Gantt chart, while another person looks on in view of a calibrated stereo camera, for vision-based body-tracking [25, 34]. When a user at the whiteboard speaks an OOV label name for a chart constituent, while also writing that label name on a task-line of the Gantt chart, the OOV speech is combined with letter sequences hypothesized by the handwriting recognizer to yield an orthography, pronunciation and semantics (OPS-tuple) for the new label (Fig. 5). The best scoring OPS-tuple, determined through *mutual disambiguation* (MD) [35], is then enrolled dynamically in the system to become immediately available for future recognition.

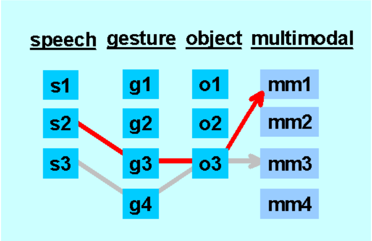

Figure 2: Mutual Disambiguation (MD) over ranked, constituent recognition lists (darker path is correct).

Because the handwriting, speech and application modules are imperfect recognizers uncertainty is a major concern. In our previous work on handling uncertainty in multimodal interfaces [5, 35] we have illustrated the importance of *mutual disambiguation* (MD). MD derives the best joint interpretation by unification of meaning fragments across the ranked inputs of the various modes (Fig. 2). In MNVR MD plays a role both in (*a*) identifying the occurrence of a labeling event, and then for that event we use (*b*) a variation of MD to identify the set of best pronunciations for the new word. The MD rate over (*a*) is:

$$MD = \frac{1}{N}\sum_{i=1}^{N} Sign\left(\frac{\sum_{c=1}^{C} R_i^c}{C} - R_i^{\mathrm{MM}}\right)$$

This equates MD rate to the average over *N* commands of those for which the average rank (*Ri*) of the constituent recognitions (*C*) that contribute to the multimodal interpretation is higher than the rank of the correct multimodal integration on the *n*-best list (Ri^{MM}), minus those in which that average is less than Ri^{MM}. In Fig. 2 MD occurs whenever the correct path is not drawn straight across the top. In situation (*b*) we again have constituent lists of handwriting-derived and speech-derived phone sequences, but instead of using grammatical constraints to highlight allowed combinations within the cross-product of the two lists (since they are all allowed) we use a simple edit-distance measure as the basis for re-scoring and re-ordering that cross-product list. Thus our hypothesis is that handwriting and speech are also capable of significantly disambiguating each other, particularly in a constrained task domain like the creation of a Gantt scheduling chart, where the temporal/spatial ontology of the task itself offers clear indications of the user's semantic intent for a given set of handwriting and speech inputs (e.g., creation of a schedule grid must precede the creation of task-lines, which in turn must precede the creation of task milestones).

In our system users layout a schedule grid using our sketch-recognition agent named *Charter* (Fig. 4). It employs a 2D sketch recognizer for the necessary constituents of the scheduling chart (dot, line, axis-grid, diamond, area, etc.), and has an associated handwriting recognizer (Calligrapher 5). Charter also displays the beautified Gantt chart produced by the multimodal integration of observed, interpreted speech, sketch and handwriting (Fig. 4).

To implement OOV speech recognition (**SR**) we have augmented CMU's Sphinx2 speech recognizer to use an embedded Recursive Transition Network (RTN) grammar in place of a standard *n*-gram

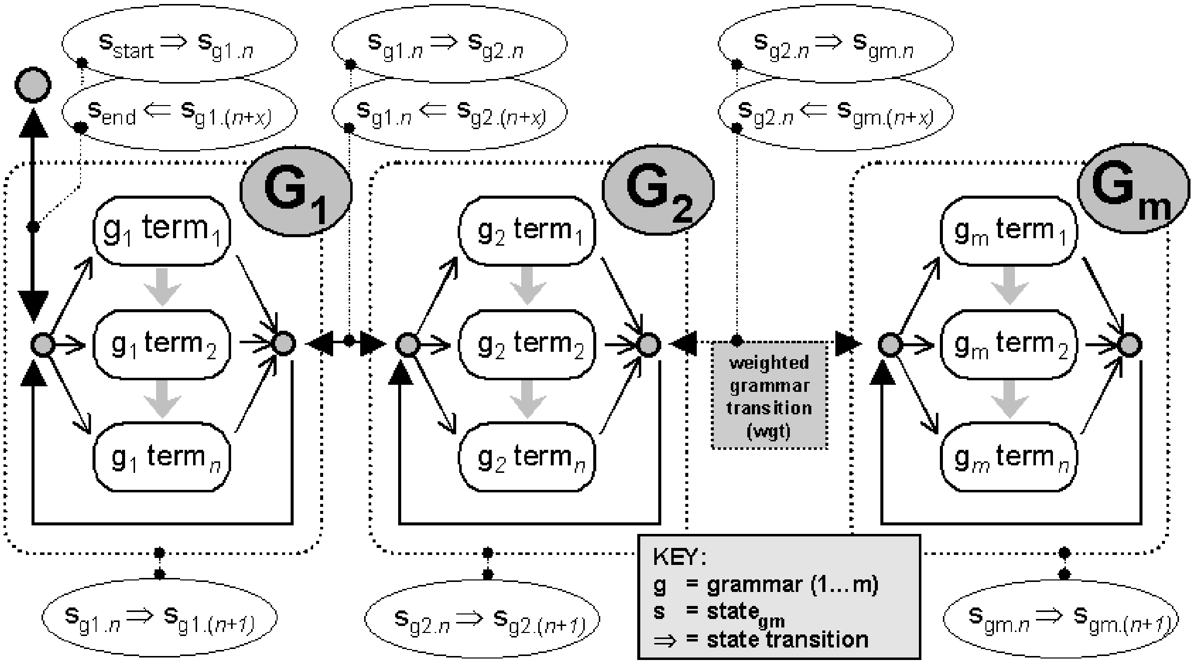
$s_{start} \Rightarrow s_{g1.n}$
$s_{end} \Leftarrow s_{g1.(n+x)}$
$s_{g1.n} \Rightarrow s_{g2.n}$
$s_{g1.n} \Leftarrow s_{g2.(n+x)}$
$s_{g2.n} \Rightarrow s_{gm.n}$
$s_{g2.n} \Leftarrow s_{gm.(n+x)}$
G_1
G_2
G_m
g_1 term$_1$
g_1 term$_2$
g_1 term$_n$
g_2 term$_1$
g_2 term$_2$
g_2 term$_n$
g_m term$_1$
g_m term$_2$
g_m term$_n$
weighted grammar transition (wgt)
$s_{g1.n} \Rightarrow s_{g1.(n+1)}$
$s_{g2.n} \Rightarrow s_{g2.(n+1)}$
$s_{gm.n} \Rightarrow s_{gm.(n+1)}$
KEY:
g = grammar (1…m)
s = state$_{gm}$
⇒ = state transition

Figure 3: Schematic diagram of Recursive Transition Network (RTN) architecture embedded within Sphinx2 Speech Recognizer

language model. This architecture is schematically illustrated in Fig. 3. Symbolic grammar-based speech recognition is often implemented on top of a flat lexical tree. We have instead used a separate re-entrant lexical prefix tree for each sub-grammar within the RTN (shown as terms *1-n* in each grammar in Fig. 3). Thus when we dynamically add new words they are added only to the appropriate grammar's lexical prefix tree. During the first pass Viterbi search we search all sub-grammars in parallel, constrained only by the *a priori* state transitions specified in the compiled grammars.

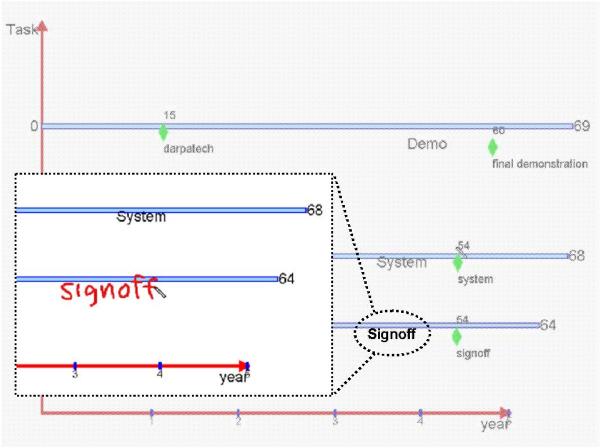

Figure 4: Charter's Gantt chart display with before (foreground) and after (background) handwriting input.

In our MNVR approach the RTN is only two levels deep: (1) a task grammar, and (2) a syllabic sequence grammar to cover OOV words. Therefore our actual implementation is very similar to Bazzi's approach [32]. Where he uses a class-based *n*-gram model — with OOV terms being modeled as a single word-class — we use an RTN with an OOV sub-grammar. The key point of penalizing the transition into the OOV component (e.g. the **wgt** in Fig. 3 above) is the same. The basic formula for speech recognition, P(W|A) = argmax (W∈£) P(A|W)*P(W), (where A=acoustic features, £=language, W=word(s), P(A|W)=acoustic model, P(W)=language model) is unchanged except that for MNVR the language model value is either 1 or 0 depending on whether the hypothesized state transition is licensed in the RTN.

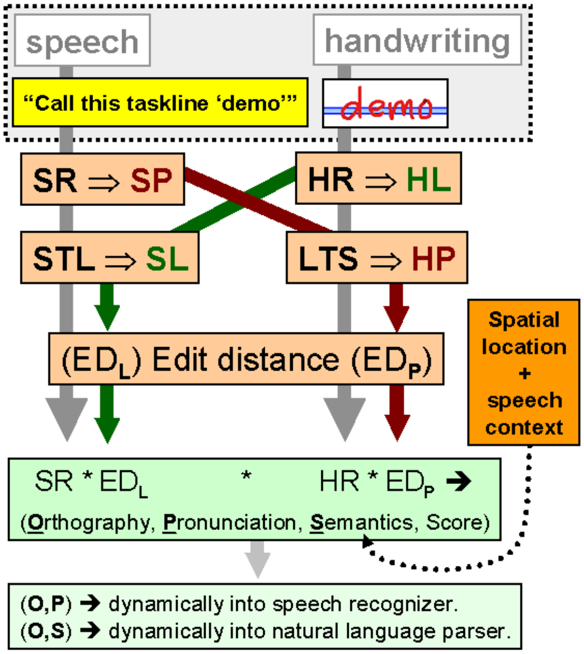
speech
handwriting
"Call this taskline 'demo'"
demo
SR ⇒ SP
HR ⇒ HL
STL ⇒ SL
LTS ⇒ HP
(ED_L) Edit distance (ED_P)
Spatial location + speech context
SR * ED_L * HR * ED_P →
(Orthography, Pronunciation, Semantics, Score)
(O,P) → dynamically into speech recognizer.
(O,S) → dynamically into natural language parser.

Figure 5: Out-of-Vocabulary (OOV) recognition & system enrollment, via multimodal handwriting and speech.

The grammar writer can semantically label specific contextual locations in the grammar where out-of-vocabulary (OOV) words are licensed to occur. At run-time, when these grammatical contexts occur in the speech input, OOV words are recognized as sequences of phones (speech-phones, **SP**), as illustrated in Fig. 5, using a syllabic sub-grammar[2]. These phone sequences are then mapped to orthographies using a sound-to-letter (**STL**) module (speech-letters, **SL**). If semantically interpretable handwriting recognition (**HR**) occurs co-temporally then the letter string hypotheses from the handwriting recognizer (handwriting-letters, **HL**) are mapped to corresponding phone strings (handwriting-phones, **HP**) by an embedded letter-to-sound (**LTS**) module [36] and paired with the OOV-based OPS-tuples using a combined edit distance measure: $\mathbf{ED_L}$ = edit-distance between letter strings, $\mathbf{ED_P}$ = edit-distance between phone strings (Fig. 5). The edit distance is modified to take matching as well as mismatching symbols into account, following [37]. The best scoring OPS-tuple (score = SR x EDL x HR x EDP) is then dynamically enrolled in the system at points pre-specified during creation of the grammar. For example, task-line labels may be specified to act as modifiers for spoken references to milestones occurring on that task-line, like "move that 'signoff' milestone to year two," where the modifier has been enrolled simultaneously along with the new task-line's label, 'signoff'.

3.1 Baseline Performance Test

To provide baseline performance test results we ran our test bed system — with a scenario of scheduling the tasks and milestones involved in collaboratively designing, creating and presenting a demonstration system — and collected 54 instances from each of three users labeling task-lines on a Gantt chart. The labeling events involved both speaking key phrases like, "Let's call this task-line *concur*," or "Label this one the *trial* task-line," (where *concur* and *trial* are examples of OOV words) and co-temporally writing the OOV label names (in this example, *concur* and *trial* respectively) on the task-line (Fig. 1).

User's read their spoken input from a script, and audio was recorded using a Samson AH-1 QV wireless, close-talking microphone. Gain control was difficult to set for this microphone, so many recordings were over-modulated and unusable for speech recognition. Thus we excluded input combinations in which the speech was over-modulated from our test set. Some OOV terms were combinations of two short words, like *dry run* or *hand shake*. In some instances the handwriting recognizer interpreted such written inputs as two separate inputs, especially when the user paused between writing each word. Our system is not yet equipped to handle two handwriting recognitions spanning a single OOV term, so such input combinations were also excluded from the current test set. After exclusions we were left with 100 combinations of co-temporal speech and handwriting for labeling Gantt chart task-lines from the three users. The 100-instance data set included 18 unique key phrases with 51 unique embedded OOV words. One user was female, while two were male. One user was familiar with the system, while two were not at all familiar with it. The data used in this experiment (speech recordings and handwriting recognition lists) is available upon request from the author.

The OOV recognizer's syllabic sub-grammar has 19006 unique syllable entries spread across four categories (first-last-syllable, first-syllable, last-syllable, middle-syllable). Since we have no large corpus of task-specific speech in this domain on which to build a plausible *n*-gram model over sub-word units, we instead rely on a symbolic grammar. Thus we have no probabilities on either syllable sequences or rule occurrences over the non-terminal categories (as would be the case with either an *n*-gram model or a stochastic context free grammar model). We view this as an advantage of our approach, because in modeling OOV terms it is neither desirable to (1) model only the OOV labeled words in a corpus, nor to (2) model cross-word occurrences for OOV words only at the boundaries occurring in the corpus. Both can result in over-training [19]. We argue that for task-independence, it is better to use a large dictionary (we use CMU Dictionary, version 6) to model a more general representation of the possible sub-word unit combinations of which OOV terms may be comprised.

Our choice of non-terminal categories is very similar to those used by Galescu [20]; however, we restrict sub-word unit combinations to a 3-syllable length limit. This is somewhat longer than Bazzi's length limit of 3-5 phones [19], while both Chung *et al*'s and Galescu's systems have built in language-model-based length biases determined by the types of OOV terms occurring in their respective corpora. Our systems' current 3-syllable length limit is partly due to tractability issues that arise from not having a stochastic language model. Our second-pass speech recognition search cannot rely on term sequence statistics (from a language model) for pruning. Given this and the fact that our syllabic vocabulary is relatively large, we cannot tractably perform a complete backward-forward A* search. So, we instead rely on a depth-first beam search with a one term look-ahead (over normalized acoustic scores) that attempts to heuristically guess the best partial paths to keep in the search beam. If the search dead-ends then it back tracks to the closest previous point where a new group of partial paths outside the previous beam limit can be found and moves forward again until either the specified number of alternatives has been found or the search space is exhausted. Transitions into the syllabic sub-grammar are weighted, similar to the approach used by [32].

The 100 test instances of multimodal speech and handwriting for labeling a Gantt chart task-line were fed into the system via the regression testing mechanism described in [31]. There was an average of 4.5 in-vocabulary (IV) terms in each of the 100 test instances. Of the total 548 word instances 18.2% were OOV words (Table 1). The OOV recognizer (OR) correctly detected the occurrence of an OOV term in all 100 instances (100% detection as shown in Table 1).

Our approach uses syntactic fragments in a grammar-based speech recognizer to frame and constrain OOV recognition to a small set of licensed linguistic contexts. These framed syntactic fragments are designed with the fact in mind that human caregivers naturally use intuitively simple syntax in addressing infants [38]. Our intuition is that the use of linguistic constructions used for teaching language to human infants may also come naturally to people for instructing a computer system. Certainly the 100% OOV detection rate we see in these test results bears witness to the effectiveness of leveraging sentence final position of new words (a characteristic of the prosodic delivery typical of infant caregivers) to more effectively segment the phone sequences to be learned. With this approach we don't

[2] Based on syllabifications of the CMU Dictionary, version 6.

need the large number of correlated occurrences required by the associative statistical categorizers in systems like those of [33] or [39]. With a single multimodal demonstration, we not only accomplish OOV detection with a high degree of accuracy, but also achieve accurate segmentation — recognizing 8.4 out of 10 of the utterances at the IV word level completely correctly (84% *Utterance correct* rate, Table 2, line 1). So we achieve an OOV segmentation error rate (SER) of 16%. While our implementation has the ability to learn generally from a single demonstration, it will also in the future be able to benefit from multiple presentations over time to refine pattern recognition accuracy.

We reduce the scope of the language acquisition problem to that of recognizing out-of-vocabulary (OOV) words in grammatically specified positions. Thus, instead of posing the problem as that of language acquisition we modify the problem to be *additional language acquisition* for an established language syntax. By using both the temporal/spatial coherence constraints of the scheduling task itself, and the contextual grammatical constraints to isolate the system's efforts at OOV recognition, we are able to process new words in real-time.

Table 1: OOV Speech Recognition test set statistics (scored on best-of-5 output)

Utterances	100
Words	548
OOV words	100
OOV rate	18.20%
OOV detection	**100.00%**

Table 2: Unimodal OOV Speech Recognition (scored on best-of-5 output)

IV Utterance correct	84.00%
IV substitutions	1.79%
IV insertions	3.57%
IV deletions	1.34%
IV accuracy	93.30%
IV Word Error Rate (**WER**)	**6.70%**
Phone-correct OOV words	13.00%
Phone substitutions	23.03%
Phone insertions	16.10%
Phone deletions	9.50%
Phone accuracy	51.37%
Phone Error Rate (**PER**)	**48.63%**

Note that the IV statistics given in Table 1 are computed over the best five transcript alternatives produced by the recognizer. In multimodal systems it is not necessary that the best recognizer transcript be correct. Mutual disambiguation from other input modes can "pull-up" the correct transcripts [5], so we take that into account by scoring over the top five alternative transcripts. For this test set there are only seven instances in which the best word-level transcript is not the recognizer's highest ranked alternative. For scoring phoneme recognition we also score over the five best alternatives from the speech recognizer, because each alternative represents a different pronunciation and only one of them has to be correct for the word to be recognized the next time it is uttered by a user. For phonetic pronunciations, the recognizer's highest ranked alternative is the best match only 32% of the time.

For in-vocabulary (IV) recognition, taking into account the number of substitution, insertion, and deletion errors, we achieve word-level recognition accuracy of 93.3%, and thus an IV word error rate (WER) of 6.7% (Table 1). The unimodal speech recognition of phonetic pronunciations is much less accurate. We achieve an accuracy of 51.37% (Table 2) for a phone error rate (PER) of 48.63%. Recall that Chung *et al*'s Speak and Spell system on a test set of 219 utterances a pronunciation-error-rate (PER) of 25.5% (much lower than our unimodal rate), and a letter-error-rate (LER) of 12.4%. Currently our system's word spelling (and thus LER) depends solely on the best alternative from the handwriting recognizer, because although there can be alternative pronunciations for the same lexical item we must still choose one single lexical representation for an item. In future versions we intend to use orthographies generated via sound-to-letter (STL) rules from the speech generated phone-sequences to help in mutually disambiguating the best lexical representation, but here we have not done that. Thus, we achieved a letter-level accuracy of 88.65% (Table 3) for an 11.35% LER (somewhat lower than Chung's above).

Table 3: Unimodal Handwriting (HW) letter recognition statistics. (Scored on first-best handwriting alternative)

HW OOV Term letter correct	46.00%
HW OOV Term letter substitutions	7.84%
HW OOV Term letter insertions	0.81%
HW OOV Term letter deletions	2.70%
HW OOV Term letter accuracy	88.65%
HW OOV Term Letter Error Rate	11.35%

Table 4: Phone recognition via unimodal (UM) Handwriting (HW) using Letter-to-Sound (LTS) rules over handwriting letters. (Scored on top 5 alternatives)

UM HW Phone-correct OOV words	25.00%
UM HW Phone substitutions	14.10%
UM HW Phone insertions	1.62%
UM HW Phone deletions	6.32%
UM HW Phone accuracy	77.96%
UM HW Phone Error Rate	**22.14%**

Table 5: Phone recognition via multimodal (MM) Speech + Handwriting (SHW) using Letter-to-Sound (LTS) rules over handwriting, and Sound-to-Letter (STL) rules over speech phone sequences. (Scored on top 5 combinations)

MM SHW Phone-correct OOV words	28.00%
MM SHW Phone substitutions	13.94%
MM SHW Phone insertions	2.43%
MM SHW Phone deletions	4.21%
MM SHW Phone accuracy	79.42%
MM SHW Phone Error Rate	**20.58%**

Our unimodal PER of 48.63% is closer to that of [20] which was 31.2% - 43.2%; however, when we use LTS to generate phone sequences from the handwriting alternatives and then use these to disambiguate the speech phone sequences we improve our PER to 20.58% (Table 5) This surpasses the accuracy of Chung *et al*'s system (25.5%), and represents a 57.5% relative error reduction between unimodal speech pronunciations and multimodal speech plus handwriting pronunciations.

Of course, given such a large improvement in pronunciation recognition from unimodal speech to multimodal speech plus handwriting, we must ask how much of this improvement we could achieve solely by deriving pronunciations from the handwritten spellings transformed via LTS rules. It may be the case that speech-only information is simply not accurate enough, and we would be better off extracting pronunciations just from the handwriting. This certainly seems plausible when we recall that for this test set the letter-level accuracy of handwriting recognition is 88.65% (Table 3). Table 4 shows that using handwriting alone (with LTS transformations) we could achieve an accuracy of 77.96% in predicting the phonemic pronunciations — for a PER of 22.14%. However, when we again look at the results of combining speech and handwriting streams to arrive at pronunciations, where the PER is 20.58% (Table 5), we find that mutual disambiguation across multiple input modes still yields 7.04% relative error reduction compared to extracting pronunciations unimodally from handwriting alone. This phone-level recognition improvement due to mutual disambiguation across combined speech and handwriting inputs compared to the phone-level pronunciations generated from unimodal handwriting alone is significant by a McNemar test, which yields a probability of this difference in recognition results having occurred due to chance as only 3.1e-8.

To see how using the speech-generated pronunciations helps us to improve on the handwriting generated pronunciations, we will analyze an example. The user says, “Call this task-line *handoff*,” (in which *handoff* is OOV) while writing *handoff* on the whiteboard chart to label a task-line (similar to the labeling event depicted in Figure 1). The correct spelling (as the user wrote it) is *handoff*, but the handwriting recognizer reports the spelling to be *handifi*. Using LTS rules on *handifi* yields the pronunciation string, “hh ae n d iy f iy,” which is one substitution and one insertion away from the correct pronunciation of, “hh ae n d ao f.” In this case the best pronunciation alternative from the speech recognizer is, “hh ae n d ao f,” which is the correct pronunciation. So by using the phone string generated by the speech recognizer we are able to enroll the correct pronunciation despite errors in the handwriting recognition, thus demonstrating the effectiveness of using multimodal speech and handwriting to achieve a level of pronunciation modeling accuracy for new (OOV) words not achievable by either mode alone.

4. CONCLUSION

We have described a system capable of multimodal speech and handwriting recognition (along with other recognition modes such as 2D and 3D gesture recognition which are not within the scope of this paper). We have described a test environment where speech and handwriting in combination are used to label elements of a whiteboard chart (e.g. task-lines, as depicted in Figure 1). Over a small test set of 100 speech and handwriting events collected from three users we have shown that combining speech and handwriting information multimodally results in significantly greater accuracy than that achievable in either mode alone. For example, the phone-error-rate (PER) over phone sequence pronunciations generated by speech alone was 48.63%, by handwriting alone it was 22.14%, while by multimodal combination of speech plus handwriting it was 20.58%. That represents a 57.5% relative error reduction compared to speech-only pronunciations, and a 7.04% relative error reduction compared to handwriting-only pronunciations (generated by LTS rules). This supports our hypothesis that handwriting and speech are capable of significantly disambiguating each other in a constrained task domain like that of labeling whiteboard Gantt chart constituents.

We have implemented a system that demonstrates the base-line capability of using multimodal speech and handwriting for new (OOV) word recognition. This capability allows users to teach our system their chosen vocabulary, thus shifting the burden of learning off the user and onto the system. We believe this is an important step towards making pen-based interaction more intelligent and natural.

5. ACKNOWLEDGEMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

Our thanks to Alan Black of CMU for providing CMU’s syllabified dictionary, FLITE’s letter-to-sound (LTS) module [36], and a sound-to-letter (STL) module within FLITE to support this work. The author would also like to thank Xiaoguang Li and Matt Wesson for designing and implementing the sketch/handwriting recognition module on which our experiment depended, and Phil Cohen (CHCC Co-Director) for his advise, and many valuable suggestions and discussions.

6. REFERENCES

[1] Breazeal, C., A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda, *Humanoid Robots as Cooperative Partners for People.* International Journal of Humanoid Robots (Forthcoming), 2004. **1**(2).

[2] Atkeson, C.G., J.G. Hale, F. Pollick, M. Riley, S. Kotosaka, S. Schaal, T. Shibata, G. Tevatia, A. Ude, S. Vijayakumar, and M. Kawato, *Using Humanoid Robots to Study Human Behavior.* IEEE Intelligent Systems, 2000. **16**(4): p. 46-56.

[3] Bluethmann, W., R.O. Ambrose, M. Diftler, S. Askew, E. Huber, M. Goza, F. Rehnmark, C. Lovchik, and D. Magruder, *Robonaut: A Robot Designed to Work with Humans in Space.* Autonomous Robots, 2003. **14**(2-3): p. 179-197.

[4] Franklin, D. and K. Hammond. *The Intelligent Classroom: Providing Competent Assistance*. in *In Proceedings of International Comference on Autonomous Agents (Agents-2001)*. 2001.

[5] Kaiser, E., A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. *Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality*. in *International Conference on Mutimodal Interfaces (ICMI '03)*. 2003.

[6] Chung, G., C. Wang, S. Seneff, E. FIlisko, and M. Tang. *Combining Linguistic Knowledge and Acoustic Information in Automatic Pronunciation Lexicon Generation*. in *Interspeech '04*. 2004. Jeju Island, Korea.

[7] Chung, G., S. Seneff, and C. Wang. *Automatic Acquistion of Names Using Speak and Spell Mode in Spoken Dialogue Systems*. in *Proceedings of HLT-NAACL 2003*. 2003. Edmonton, Canada.

[8] Chung, G., S. Seneff, C. Wang, and L. Hetherington. *A Dynamic Vocabulary Spoken Dialogue Interface*. in *Interspeech '04*. 2004. Jeju Island, Korea.

[9] Roy, D. and N. Mukherjee, *Visual Context Driven Semantic Priming of Speech Recognition and Understanding.* Computer Speech and Language (In press).

[10] Kara, L.B. and T.F. Stahovich. *An Image-Based Trainable Symbol Recognizer for Sketch-Based Interfaces*. in *AAAI Fall Symposium Series 2004: Making Pen-Based Interaction Intelligent and Natural*. 2004. Arlington, Virginia.

[11] Porzel, R. and M. Strube, *Towards Context-adaptive Natural Language Processing Systems*, in *Computational Linguistics for the New Millenium: Divergence or Synergy*, M. Klenner and H. Visser, Editors. 2002: Lang, Frankfurt am Main.

[12] Chronis, G. and M. Skubic. *Sketched-Based Navigation for Mobile Robots*. in *In Proceedings of the 2003 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2003)*. 2003. St. Louis, MO.

[13] Saund, E. and J. Mahoney. *Perceptual Support of Diagram Creation and Editing*. in *Diagrams 2004 - International Conference on the Theory and Applications of Diagrams*. 2004. Cambridge, England.

[14] Landay, J.A. and B.A. Myers, *Sketching Interfaces: Toward More Human Interface Design.* IEEE Computer, 2001. **34**(3): p. 56-64.

[15] Pook, P.K. and D.H. Ballard. *Deictic Teleassistance*. in *Proc. IEEE/RSJ/GI Int'l Conf. on Intelligent Robots and Systems*. 1994. Muenchen, Germany.

[16] Tenenbaum, J.B. and F. Xu. *Word learning as Bayesian inference*. in *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. 2000.

[17] Furnas, G.W., T.K. Landauer, L.M. Gomez, and S.T. Dumais, *The vocabulary problem in human-system communication.* Communications of the Association for Computing Machinery, 1987. **30**(11): p. 964-971.

[18] Asadi, A.O., *Automatic Detection and Modeling of New Words in a Large Vocabulary Continuous Speech Recognition System*, in *Department of Electrical and Computer Engineering*. 1991, Northeastern University: Boston.

[19] Bazzi, I. and J.R. Glass. *Modeling Out-of-Vocabulary Words for Robust Speech Recognition*. in *Proceedings of the 6th International Conference on Spoken Language Processing*. 2000. Beijing, China.

[20] Galescu, L., *Sub-lexical language models for unlimited vocabulary speech recognition*. 2002, ATR: Kyoto, Japan.

[21] Meliani, R.E. and D. O'Shaughnessy. *New efficient fillers for unlimited word recognition and keyword spotting*. in *ICSLP '96*. 1996. Philadelphia, Pennsylvania, USA.

[22] Gorniak, P. and D.K. Roy. *Augmenting User Interfaces with Adaptive Speech Commands*. in *In Proceedings of the International Conference for Multimodal Interfaces*. 2003. Vancouver, B.C., Canada.

[23] Roy, D. and A. Pentland, *Learning Words from Sights and Sounds: A Computational Model.* Cognitive Science, 2002. **26**(1): p. 113-146.

[24] Yu, C. and D.H. Ballard. *A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions*. in *International Conference on Multimodal Interfaces (ICMI '03)*. 2003. Vancouver, B.C., Canada: ACM Press.

[25] Kaiser, E., D. Demirdjian, A. Gruenstein, X. Li, J. Niekrasz, M. Wesson, and S. Kumar. *Demo: A Multimodal Learning Interface for Sketch, Speak and Point Creation of a Schedule Chart*. in *International Conference on Multimodal Interfaces (ICMI '04)*. 2004. State College, PA.

[26] Oviatt, S. and E. Olsen. *Integration Themes in Multimodal Human-Computer Interaction*. in *International ConferenceonSpoken Language Processing (ICSLP '94)*. 1994.

[27] Oviatt, S.L., A. DeAngeli, and K. Kuhn. *Integration and synchronization of input modes during multimodal human-computer interaction*. in *Proceedings of Conference on Human Factors in Computing Systems: CHI '97*. 1997. New York:: ACM Press.

[28] Anderson, R.J., R. Anderson, C. Hoyer, and S.A. Wolfman. *A Study of Digital Ink in Lecture Presentation*. in *CHI 2004: The 2004 Conference on Human Factors in Computing Systems*. 2004. Vienna, Austria.

[29] Anderson, R., C. Hoyer, C. Prince, J. Su, F. Videon, and S. Wolfman. *Speech, Ink and Slides: The Interaction of Content Channels*. in *ACM Multimedia*. 2004.

[30] Neti, C., G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri. *Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop*. in *Proc. IEEE Workshop on Multimedia Signal Processing*. 2001. Cannes.

[31] Kaiser, E.C. and P.R. Cohen. *Implementation Testing of a Hybrid Symbolic/Statistical Multimodal Architecture*. in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*. 2002. Denver.

[32] Bazzi, I., *Modelling Out-of-Vocabulary Words for Robust Speech Recognition*, in *Electrical Engineering and Computer Science*. 2002, Massachusetts Institute of Technology. p. 153.

[33] Roy, D., *Grounded Spoken Language Acquisition: Experiments in Word Learning.* IEEE Transactions on Multimedia., 2003. **5**(2): p. 197-209.

[34] Demirdjian, D., T. Ko, and T. Darrell. *Constraining Human Body Tracking*. in *Proceedings of the International Conference on Computer Vision*. 2003. Nice, France.

[35] Oviatt, S.L. *Mutual Disambiguation of Recognition Errors in a Multimodal Architecture*. in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 1999.

[36] Black, A.W. and K.A. Lenzo. *Flite: a small fast run-time synthesis engine*. in *The 4th ISCA Worskop on Speech Synthesis*. 2001. Perthshire, Scotland.

[37] Yu, C. and D.H. Ballard, *A Computational Model of Embodied Language Learning*. 2003, Computer Science Deptartment, University of Rochester: Rochester, New York.

[38] Gogate, L.J., A.S. Walker-Andrews, and L.E. Bahrick, *The Intersensory Origins of Word Comprehension: an Ecological-Dynamic Systems View.* Development Science, 2001. **4**(1): p. 1-37.

[39] Yu, C., D.H. Ballard, and R.N. Aslin. *The Role of Embodied Intention in Early Lexical Acquisition*. in *25th Annual Meeting of Cognitive Science Society (CogSci 2003)*. 2003. Boston, MA.