Dynamic New Vocabulary Enrollment through Handwriting and Speech in a Multimodal Scheduling Application

Edward C. Kaiser

Oregon Health and Science University
OGI School of Science & Engineering
20000 NW Walker Road, Beaverton, OR., 97006, USA
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 better results than achievable by either mode alone.

Introduction

Machines are moving closer to being observant and intelligent assistants for humans (Atkeson, Hale et al. 2000; Bluethmann, Ambrose et al. 2003; Breazeal, Brooks et al. 2004). However, multimodal system interfaces (incorporating speech, gesture, gaze recognition and objection selection mechanisms, e.g. (Kaiser, Olwal et al. 2003)), are typically implemented with static 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. Machines or systems that assist humans in real-time tasks need to be able to learn from being shown — through sketch (Chronis and Skubic 2003; Saund and Mahoney 2004), handwriting (Landay and Myers 2001), teleassistance (Pook and Ballard 1994), speech (Tenenbaum and Xu 2000), or multimodally (as in the work we describe here) through handwriting and speech.

Our aim, as for (Breazeal, Brooks et al. 2004) 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* (Furnas, Landauer et al. 1987). 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-vocabulary (OOV) speech recognition, to leverages two of the richest communicative modes we as humans have available for acquiring new vocabulary. Others have designed OOV speech recognition systems (Asadi 1991; Meliani and O'Shaughnessy 1996; Bazzi and Glass 2000; Galescu 2002; Chung, Seneff et al. 2003), 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 (Roy and Pentland 2002; Gorniak and Roy 2003; Yu and Ballard 2003) 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).

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

Previous Work

Systems that augment speech recognition by visually extracted face and lip movement features (Neti, Potamianos et al. 2001) employ an *early-fusion* approach that discriminately combines both input streams in a single feature space. Previous work in our group (Johnston, Cohen et al. 1997; Kaiser and Cohen 2002; Kaiser, Olwal et al. 2003) as well as our MNVR technique instead employs a *late-fusion* approach to combining speech and handwriting outputs — combining the output of separate modes after recognition has occurred. For our test bed, schedule-chart application *early-fusion* is problematic, because the temporal relation between handwriting and speech associated with it is not yet clear.

Hybrid Fusion for Speech to 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, Seneff et al. 2003) 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 (Bazzi and Glass 2000). These are multi-phone sub-word units extracted from a large corpus with clustering techniques based on a mutual information (MI) metric. (Bazzi 2002) 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 2002) 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 the 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.

Multimodal Semantic Grounding

(Roy 2003) 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 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 and Roy 2003) 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 and Ballard 2003) 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.

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 (Demirdjian, Ko et al. 2003; Kaiser, Demirdjian et al. 2004). 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. 4). The best scoring OPS-tuple, determined through *mutual disambiguation* (MD) (Oviatt 1999), is then enrolled dynamically in the system to become immediately available for future recognition.

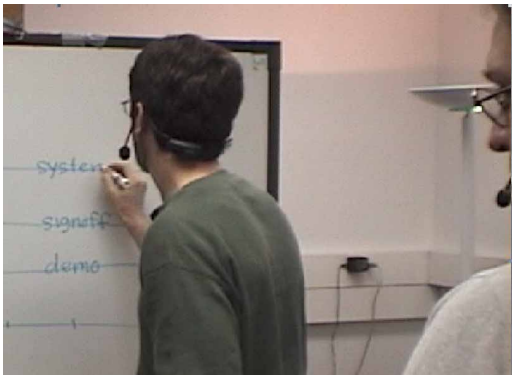

Figure 1: Using handwriting and speech to label task-lines on a Gantt chart in a multimodal, multi-person schedule meeting.

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 (Oviatt 1999; Kaiser, Olwal et al. 2003) 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).

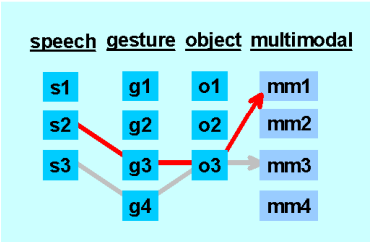

Figure 2: Mutual Disambiguation (MD) over various constraint-related input modes (darker path is correct).

Our hypothesis is that handwriting and speech are also capable of substantially 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). We believe that this constrained inference of semantic intent both allows for and supports the use of our OOV speech recognition techniques.

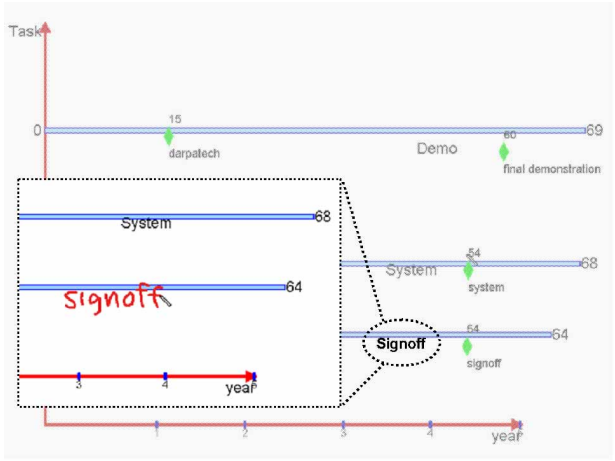

Figure 3: Charter's Gantt schedule-diagram display with before (foreground) and after (background) views of handwriting input.

In our system users layout a schedule grid using our sketch-recognition agent named *Charter* (Fig. 3). 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. 3).

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 language model. 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. 4, 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 (Black and Lenzo 2001) 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. 4). The edit distance is modified to take matching as well as mismatching symbols into account, following (Yu and Ballard 2003). 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".

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 of a single user 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). The 54-instance test set included 18 unique key phrases with 30 unique embedded OOV words.

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 neither desirable to (1) model only

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

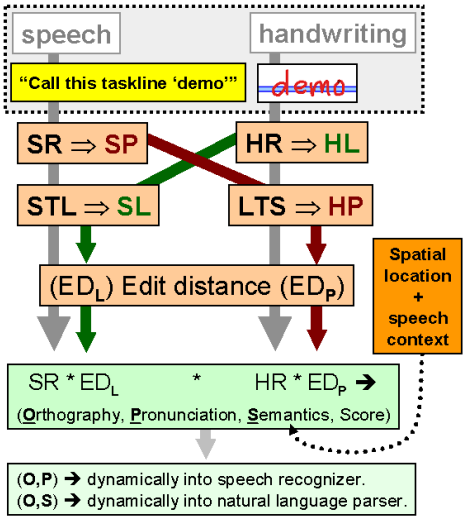

Figure 4: Out-of-Vocabulary (OOV) recognition & system enrollment, via co-temporal, multimodal handwriting and speech.

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 (Bazzi and Glass 2000). 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 2002); 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 (Bazzi and Glass 2000), 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. Since our second-pass search cannot rely on term sequence statistics (from a language model) for pruning, and since 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 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 (Bazzi 2002).

The 54 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 (Kaiser and Cohen 2002). There were an average of 4.5 in-vocabulary (IV) terms in each of the 54 test instances. Of the total 297 word instances 18.2% were

OOV words (Table 1). The OOV recognizer (OR) correctly detected the occurrence of an OOV term in all 54 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 (Gogate, Walker-Andrews et al. 2001). 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 need the large number of correlated occurrences required by the associative statistical categorizers in systems like those of (Roy 2003) or (Yu, Ballard et al. 2003). With a single multimodal demonstration, we not only accomplish OOV detection with a high degree of accuracy, but also achieve accurate segmentation — recognizing 9 out of 10 of the utterances at the IV word level completely correctly (88.89% Utterance correct rate, Table 2, line 1). So we achieve an OOV segmentation error rate (SER) of 10.11%. While our implementation has the ability to learn generally from a single demonstration, it will still 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 process new words in real-time.

The recognition rate over IV utterance words was 88.89% (Table 2), with 63% of the IV recognition errors being due to deletions. For example, in the utterance, "Let's label this the *handover* task-line," (in which *handover* is OOV) the word 'task-line' is deleted because the OOV recognizer doesn't find the correct boundary for stepping out of the syllable-based OOV sub-grammar in the weighted recursive-transition-network (RTN) parser embedded in the speech recognizer. Instances similar to this example account for four out of the five of the utterance level deletion errors. Adjusting the weights on the transitions between the task grammar and its embedded syllabic sub-grammar (within the RTN language model) can ameliorate this error; however, we currently have no mechanism in place for dynamically adjusting this weight. This is a topic for future research.

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 (Kaiser, Olwal et al. 2003), so we take that into account by scoring over the top five alternative transcripts. For this test set there are only two 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 48.15% of the time.

Utterances	54
Words	297
OOV words	54
OOV rate	18.20%
OOV detection	**100.00%**

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

IV Utterance correct	88.89%
IV substitutions	0.41%
IV insertions	0.82%
IV deletions	2.06%
IV accuracy	96.71%
IV Word Error Rate (**WER**)	**3.29%**
Phone-correct OOV words	9.26%
Phone substitutions	18.33%
Phone insertions	21.67%
Phone deletions	7.33%
Phone accuracy	52.67%
Phone Error Rate (**PER**)	**47.33%**

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

For IV recognition, taking into account the number of substitution, insertion, and deletion errors, we achieve word-level recognition accuracy of 96.71%, and thus an IV word error rate (WER) of 3.29% (Table 1). The unimodal speech recognition of phonetic pronunciations is much less accurate. We achieve an accuracy of 52.67% (Table 2) for a phone error rate (PER) of 47.33%. Recall that Chung *et al*'s Speak and Spell system on a test set of 219 utterances achieved a word-error-rate (WER) of 46.1% (much higher than ours), 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 95.13% (Table 3) for a 4.87% LER (much lower than Chung's above, indicating the accuracy of handwriting as opposed to spoken spelling for lexical identification).

Our unimodal PER of 47.33% is closer to that of (Galescu 2002) 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 16.33% (Table 5) This surpasses the accuracy of Chung *et al*'s system (25.5%), and represents a 65.5% relative error reduction between unimodal speech pronunciations and multimodal speech plus handwriting pronunciations.

HW OOV Term letter correct	75.93%
HW OOV Term letter substitutions	1.43%
HW OOV Term letter insertions	0.57%
HW OOV Term letter deletions	2.87%
HW OOV Term letter accuracy	95.13%
HW OOV Term Letter Error Rate	**4.87%**

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

UM HW Phone-correct OOV words	35.19%
UM HW Phone substitutions	13.67%
UM HW Phone insertions	1.00%
UM HW Phone deletions	4.67%
UM HW Phone accuracy	80.67%
UM HW Phone Error Rate	**19.33%**

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

MM SHW Phone-correct OOV words	38.89%
MM SHW Phone substitutions	11.33%
MM SHW Phone insertions	1.33%
MM SHW Phone deletions	3.67%
MM SHW Phone accuracy	83.67%
MM SHW Phone Error Rate	**16.33%**

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 speech and handwriting alternatives)

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 95.13% (Table 3). Table 4 shows that using handwriting alone (with LTS transformations) we could achieve an accuracy of 80.67% in predicting the phonemic pronunciations — for a PER of 19.33%. However, when we again look at the results of combining speech and handwriting streams to arrive at pronunciations, where the PER is 16.33% (Table 5), we find that mutual disambiguation across multiple input modes still yields 15.5% relative error reduction compared to extracting pronunciations unimodally from handwriting alone.

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. This improvement due to speech occurs altogether seven times across this small test set of utterance/handwriting events, 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.

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 54 speech and handwriting events we have shown that combining speech and handwriting information multimodally results in 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 47.33%, by handwriting alone it was 19.33%, while by multimodal combination of speech plus handwriting it was 16.33%. That represents a 65.5% relative error reduction compared to speech-only pronunciations, and a 15.5% relative error reduction compared to handwriting-only pronunciations (generated by LTS rules). This supports our hypothesis that handwriting and speech are capable of substantially disambiguating each other in a constrained task domain like the labeling of 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.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under Contract No. NBCHD030010.

The author would like to thank Alan Black of CMU for providing CMU's syllabified dictionary, FLITE's letter-to-sound (LTS) module (Black and Lenzo 2001), and for creating a sound-to-letter (STL) module within FLITE to support this work. The author would also like to thank Xiaoguang Li formerly of OHSU's Center for Human Computer Communication (CHCC) and Matt Wesson of CHCC and Natural Interaction Systems 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.

References

Asadi, A. O. (1991). Automatic Detection and Modeling of New Words in a Large Vocabulary Continuous Speech Recognition System. Ph.D. diss., *Dept. of Electrical and Computer Engineering*. Boston, Northeastern University.

Atkeson, C. G., J. G. Hale, et al. (2000). "Using Humanoid Robots to Study Human Behavior." *IEEE Intelligent Systems* **16**(4): 46-56.

Bazzi, I. (2002). Modelling Out-of-Vocabulary Words for Robust Speech Recognition. Ph.D. diss., *Electrical Engineering and Computer Science*, Cambridge, Mass. Institute of Technology.

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

Black, A. W. and K. A. Lenzo (2001). *Flite: a small fast run-time synthesis engine*. The 4th ISCA Workshop on Speech Synthesis, Perthshire, Scotland.

Bluethmann, W., R. O. Ambrose, et al. (2003). "Robonaut: A Robot Designed to Work with Humans in Space." *Autonomous Robots* **14**(2-3): 179-197.

Breazeal, C., A. Brooks, et al. (2004). "Humanoid Robots as Cooperative Partners for People." *International Journal of Humanoid Robots (Forthcoming)* **1**(2).

Chronis, G. and M. Skubic (2003). *Sketched-Based Navigation for Mobile Robots*. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2003), St. Louis, MO.

Chung, G., S. Seneff, et al. (2003). *Automatic Acquisition of Names Using Speak and Spell Mode in Spoken Dialogue Systems*. Proceedings of HLT-NAACL 2003, Edmonton, Canada.

Demirdjian, D., T. Ko, et al. (2003). *Constraining Human Body Tracking*. Proceedings of the International Conference on Computer Vision, Nice, France.

Furnas, G. W., T. K. Landauer, et al. (1987). "The vocabulary problem in human-system communication." *Communications of the Association for Computing Machinery* **30**(11): 964-971.

Galescu, L. (2002). Sub-lexical language models for unlimited vocabulary speech recognition. *Technical Report of IEICE*, **102**(108), SP2002-30, May 2002, pp. 37-42.

Gogate, L. J., A. S. Walker-Andrews, et al. (2001). "The Intersensory Origins of Word Comprehension: an Ecological-Dynamic Systems View." *Development Science* **4**(1): 1-37.

Gorniak, P. and D. K. Roy (2003). *Augmenting User Interfaces with Adaptive Speech Commands*. International Conference for Multimodal Interfaces, Vancouver, B.C., Canada.

Johnston, M., P. R. Cohen, et al. (1997). *Unification-based Multimodal Integration*. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics.

Kaiser, E., D. Demirdjian, et al. (2004). *Demo Proposal: A Multimodal Learning Interface for Sketch, Speak and Point Creation of a Schedule Chart*. International Conference on Multimodal Interfaces (ICMI '04), State College, PA.

Kaiser, E., A. Olwal, et al. (2003). *Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality*. International Conference on Mutimodal Interfaces (ICMI '03).

Kaiser, E. C. and P. R. Cohen (2002). *Implementation Testing of a Hybrid Symbolic/Statistical Multimodal Architecture*. ICSLP '02, Denver, CO., USA.

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

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

Neti, C., G. Potamianos, et al. (2001). *Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop*. Proc. IEEE Workshop on Multimedia Signal Processing, Cannes, France.

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

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

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

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

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

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

Yu, C. and D. H. Ballard (2003). A Computational Model of Embodied Language Learning. Technical Report 791, Computer Science Dept., University of Rochester.

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

Yu, C., D. H. Ballard, et al. (2003). *The Role of Embodied Intention in Early Lexical Acquisition*. 25th Annual Meeting of Cognitive Science Society (CogSci 2003), Boston, MA.