SHACER: a Speech and Handwriting Recognizer

Edward C. Kaiser[1]
Natural Interaction Systems, LLC
10260 SW Greenburg Road
Portland, OR 97223 USA
+1 503 748 1608
ed@naturalinteraction.com

ABSTRACT

Within the task domain of a multi-party, multimodal meeting focused on the creation of a whiteboard schedule chart, we have designed and implemented a general method of aligning handwriting and speech for capturing out-of-vocabulary terms, dynamically enrolling them in the system's recognition modules, and then using them to improve subsequent tracking and recognition. Our approach involves the use of an ensemble of syllable and phoneme recognizers for speech whose output is integrated with redundantly delivered handwriting recognition. We refer to our conceptual framework as Multimodal Out-Of-Vocabulary Recognition (MOOVR — pronounced *mover*). Within that framework this paper describes our Speech and HAndwriting reCognizER module (SHACER — pronounced *shaker*), which observes human-to-human spoken and handwritten interactions, analyzes them off-line and contributes improved recognitions to a record of the meeting in the form of a project schedule. We examine an example meeting and show how our technique corrects four of five label recognition errors including implicitly discovering the semantics of a handwritten abbreviation.

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

The goal of our MOOVR framework is to automatically recognize and enroll new vocabulary in a multimodal interface. Dynamically augmenting vocabularies, pronunciation lexicons and language models is an active area of research in speech and gesture recognition [1-6]. Computer systems that track or assist in human-human real-time interactions need to be able to learn from observation — of sketch [7, 8], of handwriting [9], of speech [10], or of related modes like handwriting and speech as in the work we describe here. To accomplish this our technique leverages the mutually disambiguating aspects of redundantly delivered handwriting and speech (Fig. 1).

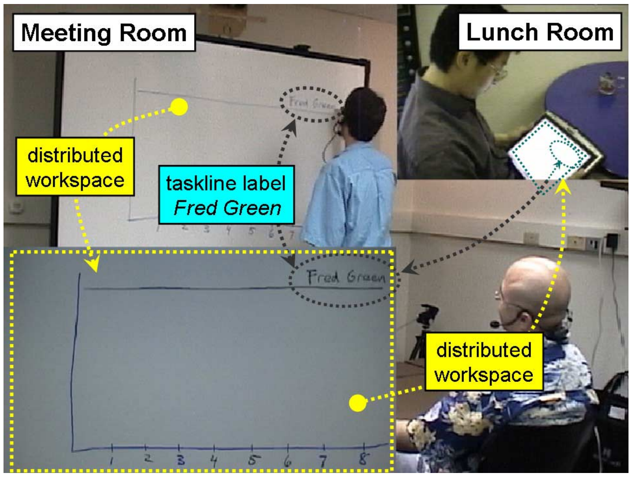


Figure 1: SHACER being used in a distributed multiparty meeting to recognize out-of-vocabulary terms like the taskline label, *Fred Green*, here being redundantly introduced through speech and handwriting.

In our earlier work on Multimodal New Vocabulary Recognition (MNVR) [11] we showed how combining speech-recognizer-generated phonetic pronunciations with letter-to-sound generated phonetic pronunciations from handwriting recognitions significantly improves the quality of enrolled pronunciations for OOV terms. For example, when a user creating a schedule chart at the whiteboard says, "Call this task-line *handoff*," where h*andoff* is an out-of-vocabulary (OOV) term, while also 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 letter-to-sound (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

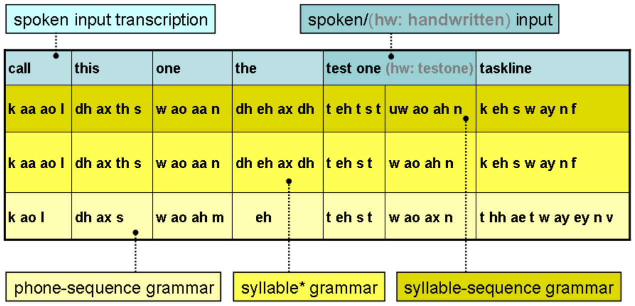
ICMI '05, October 4-6, 2005, Trento, Italy.
Workshop on Multimodal, Multiparty Meeting Processing

[1] Work on this paper was largely done at OHSU's OGI School of Science & Engineering, in the Center for Human Computer Communication (CHCC).

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.

MNVR, however, constrains users to only utter OOV terms in certain grammatically specified positions within a larger *carrier phrase*. For instance in the example above the *carrier phrase* is, "Call this task-line <*oov_term*>", and the <*oov_term*> can only be recognized in the specified position. The advantage of this approach is accuracy and tractability: it is a real-time method, and the *carrier phrase* aids in accurate segmentation of the OOV term within the larger utterance. For some applications with fixed vocabularies (e.g. certain classes of military applications) this may be a viable approach; but, in general requiring the use of *carrier phrases* is too restrictive, and a more general approach is called for.

Our Speech and Handwriting reCognizER (SHACER) is a general, unconstrained method for capturing speech via an ensemble of syllable/phone recognizers and aligning it with handwriting recognition results by means of an articulatory-feature based metric. Along with the ensemble of syllable/phone recognizers we also employ a dedicated Word/Phrase Spotting Recognizer (WPSR) into which newly recognized terms are enrolled and then made available for subsequent recognition, and a large vocabulary continuous speech recognizer, Carnegie Mellon Univerisity's Sphinx 3.5 recognizer implemented as a *Speechalyzer* agent within our systems distributed Open Agent Architecutre [12].


spoken input transcription
spoken/(hw: handwritten) input
phone-sequence grammar
syllable* grammar
syllable-sequence grammar

call	this	one	the	test one (hw: testone)		taskline
k aa ao l	dh ax th s	w ao aa n	dh eh ax dh	t eh t s t	uw ao ah n	k eh s w ay n f
k aa ao l	dh ax th s	w ao aa n	dh eh ax dh	t eh s t	w ao ah n	k eh s w ay n f
k ao l	dh ax s	w ao ah m	eh	t eh s t	w ao ax n	t hh ae t w ay ey n v

Figure 2: Phone sequence outputs for different ensemble recognizers: (bottom) unconstrained phone-sequence, (middle) unconstrained syllable sequence grammar, (top) constrained syllable sequence grammar.

In the remainder of the paper, we first describe the syllable/phone recognition ensemble and articulatory-feature based alignment mechanism. Then we discuss related work, and examine in detail two example meetings in which our approach yields substantial improvements in recognition. Among those improvements we describe the system's ability to accumulate knowledge about learned words and learned pronunciation variations both within individual meetings and persistently across a series of meetings. We end our description of the system by looking at how the persistent enrollment of learnt new words allows us to dynamically acquire the semantics of handwritten abbreviations. Finally we conclude and discuss our future work.

2. SHACER

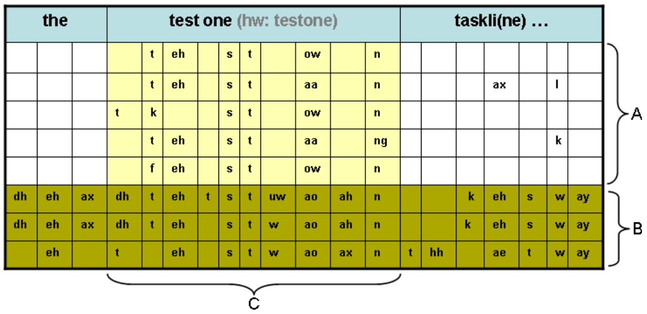
SHACER is a general approach for capturing unconstrained speech, which may contain OOV terms, via an ensemble of syllable and phoneme recognizers. The ensemble of phone sequence representations of the input speech are aligned with an articulatory-feature based alignment mechanism. Figure 2 illustrates some of the various phone sequence recognitions and their alignment with respect to each other.

2.1 Phonetic Interpretation and Alignment

letters	HW likelihood	LTS phones
testone	0.752	t eh s t ow n
testonl	0.598	t eh s t aa n ax l
tcstone	0.480	t k s t ow n
testonc	0.371	t eh s t aa ng k
festone	0.299	f eh s t ow n

Figure 3: List of handwriting recognitions — spelling, score, letter-to-sound (LTS) phones — for the handwritten phrase, *test one*, from **Figure 2**.

We employ an ensemble approach to phone recognition because phone recognizers have high error rates and our speech recognizer (an augmented version of Carnegie Mellon University's Sphinx 2) is not optimized for phone recognition. Each of our grammar-based recognizers [11] is tuned to yield somewhat different phone sequence interpretations as shown in Figure 2. It is possible for a single grammar-based recognizer to yield multiple phone-level interpretations from a second pass lattice search, but given that we use no stochastic model of English phone sequences such a lattice search in our approach is intractable. If we use threshold pruning to ensure tractability, then variations in resulting interpretations tend to be bunched toward the end. Using an ensemble of Viterbi, first-pass recognizers allows full variation across each interpretation, rather than just at the sequence ends.



the			test one (hw: testone)										taskli(ne) …						
				t	eh		s	t		ow		n							
				t	eh		s	t		aa		n				ax		l	
			t	k			s	t		ow		n							
				t	eh		s	t		aa		ng						k	
				f	eh		s	t		ow		n							
dh	eh	ax	dh	t	eh	t	s	t	uw	ao	ah	n			k	eh	s	w	ay
dh	eh	ax	dh	t	eh		s	t	w	ao	ah	n			k	eh	s	w	ay
	eh		t		eh		s	t	w	ao	ax	n	t	hh		ae	t	w	ay

Figure 4: Phonetic alignment matrix based on articulatory-feature distance. (A) LTS phone sequences from HW recognition – note that the handwriting is recognized as one word, "testone," thus the consistently incorrect pronunciation sequence is generated by the LTS engine. (B) Ensemble speech recognition phone sequence outputs. (C) HW LTS aligned segment accurately bounded within the larger utterance – over this segment the dynamic phone bigram sequence model is built that constrains second pass speech recognition.

Although our phone recognition rates are low (less than 70%) we do know that each interpretation is of the same utterance. Thus, they can be reliably aligned using our phonetic articulatory-feature-based aligner (not described here). This aligner allows us to compare phone hypotheses by feature sets rather then by phone name: so instead of assigning the phone match between *d* and *t* an absolute score of 0 because they are not the same phone we can instead assign them a metric that takes into account the fact that they are identical in all articulatory features except voicing. This results in phone hypotheses matrices against which letter-to-sound (LTS) interpretations of the handwriting letter-string hypotheses (Figure 3) can also be aligned to discover their segmental boundaries in the spoken utterance with which they redundantly occur (as shown in Figure 3).

2.2 Meeting Recording and Playback

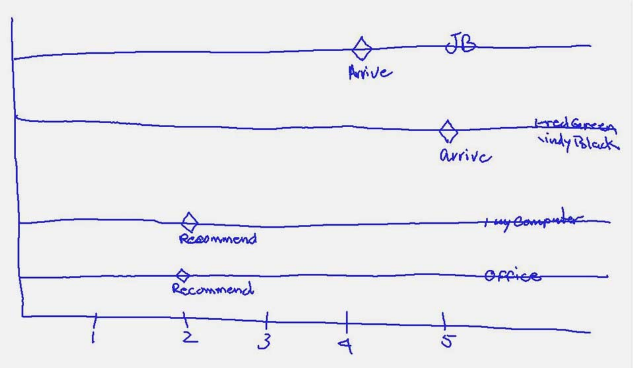


Figure 5: The ink for the *G2* meeting: second in a series of meetings referred to as the *G* meeting series.

The example meetings examined here were recorded at the level of speech, ink and 3D gesture. Speech was recorded at 11.025 KHz with head-worn, close-talking microphones (both wired and wireless). Ink was captured on an Interwrite™ interactive whiteboard. 3D gesture was captured with vision-based body-tracking using a whiteboard-mounted stereo camera [13]**.** The whiteboard ink for the example meeting used as a basis for discussion below is shown in Figure 5.

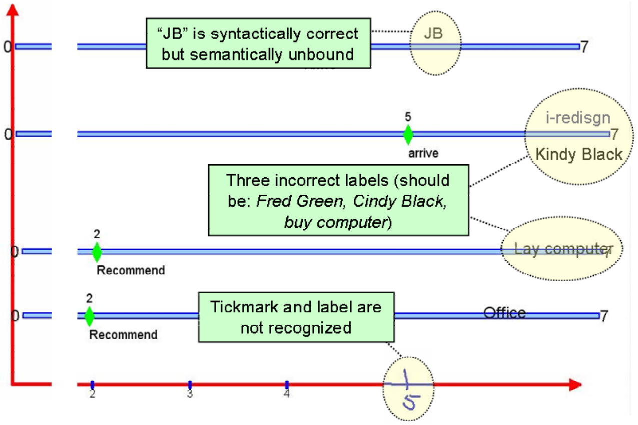

"JB" is syntactically correct but semantically unbound
JB
0
7
5
arrive
i-redisgn
Kindy Black
Three incorrect labels (should be: *Fred Green, Cindy Black, buy computer*)
2
Recommend
Lay computer
Tickmark and label are not recognized
Office
2
3
4
5

Figure 6: G2 meeting analysis errors when SHACER is **not** used. *Top*: missing semantics for abbreviation. *Middle*: three misspelled constituent labels due to incorrect handwriting (HW) recognition. *Bottom*: unrecognized tick mark and label due to sketch mis-segmentation.

Ink and speech can be played back in appropriate order within our MultiPlayer Suite (not described here) for off-line analysis and integration. All examples discussed result from this off-line playback analysis, which simulates real-time input by playing back logged messages from multiple input streams in a time-synchronous, lock-step mode. The errors that occur in this Multiplayer analysis for meeting G2 are depicted in Figure 6.

Note that stroke skipping visible in the ink shown in Figure 5 causes incorrect handwriting recognition for three of the constituent labels in Figure 6. With no other evidence these mis-recognition errors are not recoverable. However in these cases the label names were spoken redundantly as they were handwritten, so both the abbreviation semantics and the three incorrect labels (along with their pronunciations and semantics) are recoverable using SHACER's second-pass recognition (as shown in Fig. 7).

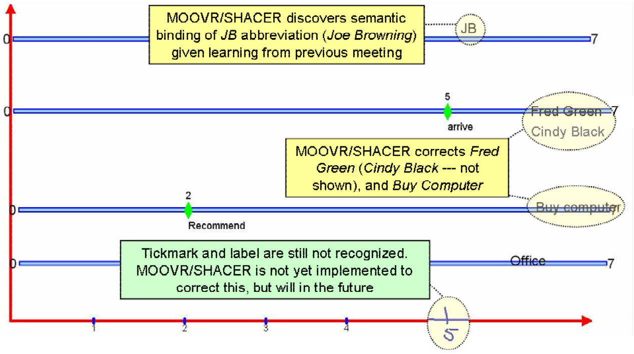
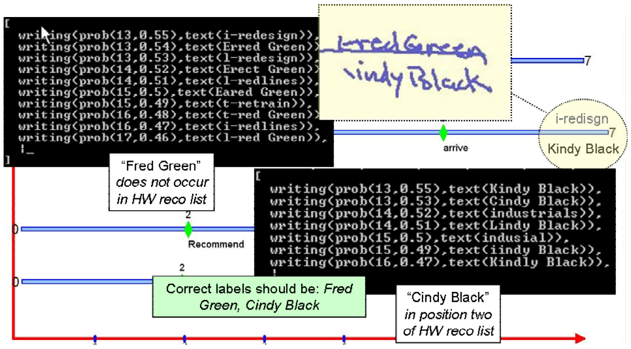

MOOVR/SHACER discovers semantic binding of *JB* abbreviation (*Joe Browning*) given learning from previous meeting
JB
0
7
5
arrive
Fred Green
Cindy Black
MOOVR/SHACER corrects *Fred Green* (*Cindy Black* --- not shown), and *Buy Computer*
2
Recommend
Buy computer
Tickmark and label are still not recognized. MOOVR/SHACER is not yet implemented to correct this, but will in the future
Office
1
2
3
4
5

Figure 7: G2 meeting analysis corrections when SHACER **is** used. *Top*: abbreviation semantics discovered. *Middle*: 3 constituent labels dynamically enrolled in WPSR with correct spelling, semantics and pronunciation tuples. *Bottom*: Unrecognized tickmark and label not processed by SHACER at this time, still incorrect.

2.3 Caching and Second-Pass Recognition

All spoken inputs to SHACER are first decomposed into Mel Cepstral feature vectors and then cached within the system in a sliding window that acts like a short term memory over the most recent speech events. Caching the speech as feature vector arrays saves space and processing time later when the multimodal integration agent – after having received some handwriting interpretations – calls for a second pass recognition over the integrated speech and handwriting phone sequences. Also saved in sliding window caches are both the time-segmented transcripts and lattices from the parallel Speechalyzer recognition and the time-bounded term sequences from the Word/Phrase-Spotting Recognizer (WPSR).

During the *G2* meeting each of the handwriting events is accompanied by redundant speech. In the human-computer-interaction (HCI) 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 [25,27]. However, in the more recent educational-technology literature on human-human,



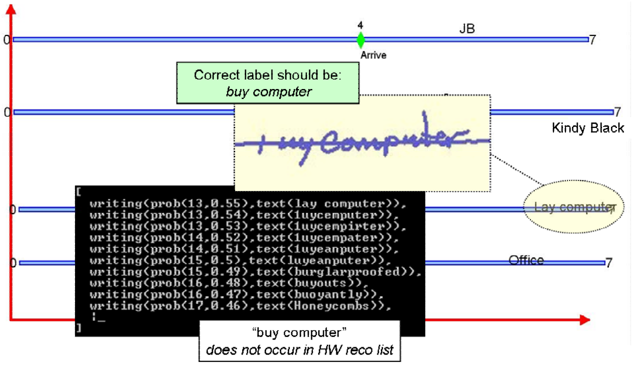
```
[
  writing(prob(13,0.55),text(i-redesign)),
  writing(prob(13,0.54),text(Erred Green)),
  writing(prob(13,0.53),text(l-redesign)),
  writing(prob(14,0.52),text(Erect Green)),
  writing(prob(14,0.51),text(l-redlines)),
  writing(prob(15,0.5),text(Eared Green)),
  writing(prob(15,0.49),text(t-retrain)),
  writing(prob(16,0.48),text(t-red Green)),
  writing(prob(16,0.47),text(i-redlines)),
  writing(prob(17,0.46),text(l-red Green)),
  !_
]
```

```
[
  writing(prob(13,0.55),text(Kindy Black)),
  writing(prob(13,0.53),text(Cindy Black)),
  writing(prob(14,0.52),text(industrials)),
  writing(prob(14,0.51),text(Lindy Black)),
  writing(prob(15,0.5),text(indusial)),
  writing(prob(15,0.49),text(iindy Black)),
  writing(prob(16,0.47),text(Kindly Black)),
  !
```

7
i-redisgn
7
Kindy Black
arrive
"Fred Green" *does not occur in HW reco list*
2
0
Recommend
2
0
Correct labels should be: *Fred Green, Cindy Black*
"Cindy Black" *in position two of HW reco list*

Figure 8: Handwriting recognition (HWR) for *Fred Green* and *Cindy Black*, a label-list for a chart taskline. Due to ink-skipping *Fred Green* is not found in its HWR hypothesis list, and *Cindy Black* is not the first hypothesis on its list.

computer-mediated interactions like the presentation of distance-learning lectures as much as 15% of all pen interactions were found to be handwriting [28], and of such handwriting events a follow-on study found that 100% of the randomly sampled instances of handwritten text were accompanied by semantically redundant speech [26]. The *G2* handwriting events for the *Fred Green, Cindy Black* and *buy computer* taskline labels are shown in Figure 8 and Figure 9. By handwriting recognition alone none are correct.



```
[
  writing(prob(13,0.55),text(lay computer)),
  writing(prob(13,0.54),text(1uycemputer)),
  writing(prob(13,0.53),text(1uycempirter)),
  writing(prob(14,0.52),text(1uycempater)),
  writing(prob(14,0.51),text(1uyeanputer)),
  writing(prob(15,0.5),text(luyeanputer)),
  writing(prob(15,0.49),text(burglarproofed)),
  writing(prob(16,0.48),text(buyouts)),
  writing(prob(16,0.47),text(buoyantly)),
  writing(prob(17,0.46),text(Honeycombs)),
  !_
]
```

4
JB
0
7
Arrive
Correct label should be: *buy computer*
0
7
Kindy Black
0
~~Lay computer~~
0
~~Office~~
7
"buy computer" *does not occur in HW reco list*

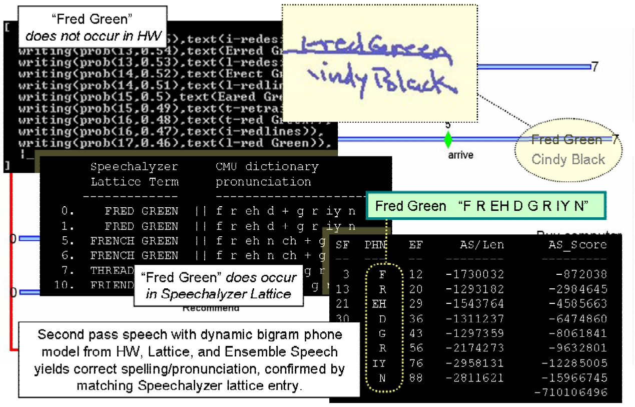
Figure 9: Handwriting recognition (HWR) for *buy computer*, a chart taskline label. Due to ink-skipping the correct interpretation does not occur in the HWR hypothesis list.

When a speech and handwriting integration is hypothesized, a second-pass recognition is requested to evaluate it. Information from the following sources is examined to propose an integrated hypothesis:

1. Aligned HandWriting/Speech Matrix (AHWSM, e.g. Fig. 4) — exposes HW segment location within accompanying speech on the basis of articulatory-feature alignment.
2. Term recognition in WPSR — a match of WPSR terms to an HW hypothesis is strong evidence for fusing speech/handwriting segment information; otherwise, the WPSR term segment can be used to expose HW abbreviations.
3. Use AHWSM segmentation bounds to extract terms from Speechalyzer transcript — if HW match exists this is very strong evidence for fusing speech/handwriting segment information.
4. Use AHWSM segmentation bounds to extract terms from Speechalyzer lattice — if HW match or near match exists this is very strong evidence for fusing speech/handwriting segment information.

2.3.1 *Lattice Alignment Fusion: 'Fred Green'*

In finding the correct spelling and pronunciation for *Fred Green* (as illustrated in Figure 10), list point 4 is the critical piece of information. Although *Fred Green* does not exist in the HW hypothesis list (see Fig. 9), there is nonetheless enough phonetic information in the HW hypothesis list to make a correct phonetic segmentation (by alignment to the ensemble speech output) and thus discover the correct HW segmentation bounds within the utterance. Given these bounds, term sequences can be extracted from the Speechalyzer lattice (see the middle block of Fig. 10). Among the term sequences within those bounds is *Fred Green,* which is discovered to be the closest match to the bounded segment from the ensemble speech matrix, and thus provides the strongest stochastic bias for the second pass evaluation of the cached acoustic feature vectors. The bounded matrix segment of speech and handwriting phone sequences is used to dynamically build a positional bigram model, which provides the stochastic constraint for both the second-pass recognition's Viterbi search and lattice search. The second-pass search yields an *alternates* list of pronunciation interpretations of the handwritten term. In this case the output of the second-pass recognition (lower block Fig. 10), when passed through a sound-to-letter engine, exactly confirms the Speechalyzer lattice hypothesis, *Fred Green*. All speech and handwriting combinations are scored by combining their speech, handwriting and alignment scores, and an *alternates* list of the top scoring combinations is returned. The label semantics (e.g. taskline versus milestone label) are determined by the spatial location of ink and the current state of the chart.



```
[
          5),text(i-redes
  writing(prob(13,0.54),text(Erred Gr
  writing(prob(13,0.53),text(l-redes
  writing(prob(14,0.52),text(Erect Gr
  writing(prob(14,0.51),text(l-redlin
  writing(prob(15,0.5),text(Eared Gre
  writing(prob(15,0.49),text(t-retra
  writing(prob(16,0.48),text(t-red Gr
  writing(prob(16,0.47),text(i-redlines)),
  writing(prob(17,0.46),text(l-red Green)),
  !_
]
```

```
      Speechalyzer      CMU dictionary
      Lattice Term      pronunciation
     -------------      ---------------------
 0.     FRED GREEN   || f r eh d + g r iy n
 1.     FRED GREEN   || f r eh d + g r iy n
 5.   FRENCH GREEN   || f r eh n ch + g
 6.   FRENCH GREEN   || f r eh n ch + g
 7.   THREAD                           g r
10.   FRIEND                           g
```

```
SF   PHN   EF     AS/Len       AS_Score
--   ---   --   --------      ---------
 3    F    12   -1730032        -872038
13    R    20   -1293182       -2984645
21   EH    29   -1543764       -4585663
30    D    36   -1311237       -6474860
      G    43   -1297359       -8061841
      R    56   -2174273       -9632801
     IY    76   -2958131      -12285005
      N    88   -2811621      -15966745
                              -710106496
```

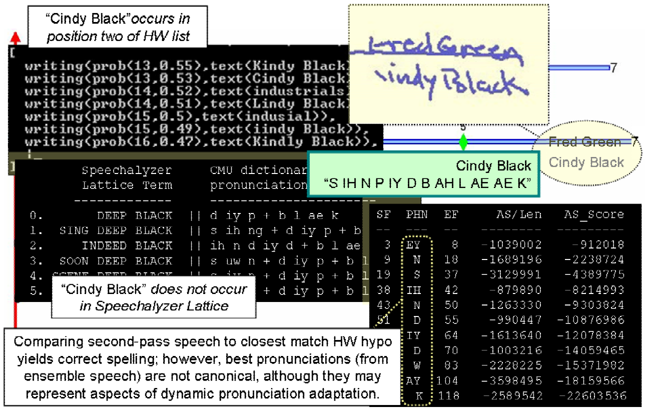
"Fred Green" *does not occur in HW*
7
5
arrive
~~Fred Green~~ 7
Cindy Black
Fred Green "F R EH D G R IY N"
0
0
"Fred Green" *does occur in Speechalyzer Lattice*
Recommend
Second pass speech with dynamic bigram phone model from HW, Lattice, and Ensemble Speech yields correct spelling/pronunciation, confirmed by matching Speechalyzer lattice entry.

Figure 10: Discovering the correct spelling, semantics and pronunciation of *Fred Green* by aligning and fusing speech and handwriting information sources.

Once a label has been recovered, as for *Fred Green* above, it is enrolled in the Word/Phrase-Spotting Recogizer (WPSR), which is optimized for word and phrase-spotting. This grammar is capable of recognizing the enrolled words or phrases when they are subsequently spoken, for instance while participants are pointing to a diagram element and speaking about it (Fig. 18).

2.3.2 *Speech/Handwriting Fusion: 'Cindy Black'*

In finding the correct spelling of *Cindy Black* (as illustrated in Figure 11), list point 1 is of primary importance, because in this case the correct term sequence does not occur in either the Speechalyzer transcript or lattice. However since *Cindy Black* does occur as the second hypothesis on the HW *alternates* list, and its LTS translation is closest to the phone matrix resulting from the second-pass search over the cached speech features there is enough evidence to choose it as the best spelling. Although the canonical pronunciation is not among the results of either the ensemble speech or of the second pass search, the pronunciations returned, like that highlighted in Figure 11, do show intriguing evidence of phonetic adaptation (e.g., the common tendency to say words like *black* as two syllables, *bah-lack*, instead of one), which we will examine more closely as part of our future work.



"Cindy Black"*occurs in position two of HW list*

```
writing(prob(13,0.55),text(Kindy Black
writing(prob(13,0.53),text(Cindy Black
writing(prob(14,0.52),text(industrials
writing(prob(14,0.51),text(Lindy Black
writing(prob(15,0.5),text(indusial)),
writing(prob(15,0.49),text(iindy Black)),
writing(prob(16,0.47),text(Kindly Black)),
```

```
       Speechalyzer      CMU dictionar
       Lattice Term      pronunciation
       ------------      --------------------
 0.       DEEP BLACK  || d iy p + b l ae k
 1.  SING DEEP BLACK  || s ih ng + d iy p + b
 2.    INDEED BLACK   || ih n d iy d + b l ae
 3.  SOON DEEP BLACK  || s uw n + d iy p + b l
 4.                                + d iy p + b l
 5.                                + d iy p + b l
```

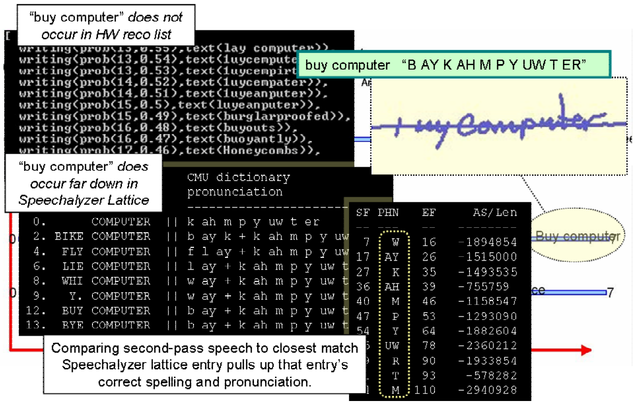
"Cindy Black" *does not occur in Speechalyzer Lattice*

```
SF  PHN   EF      AS/Len    AS_Score
--  ---   --    --------   ---------
 3  EY     8    -1039002     -912018
 9   N    18    -1689196    -2238724
19   S    37    -3129991    -4389775
38  IH    42     -879890    -8214993
43   N    50    -1263330    -9303824
51   D    55     -990447   -10876986
    IY    64    -1613640   -12078384
     D    70    -1003216   -14059465
     W    83    -2228225   -15371982
    AY   104    -3598495   -18159566
     K   118    -2589542   -22603536
```

Cindy Black
"S IH N P IY D B AH L AE AE K"

Comparing second-pass speech to closest match HW hypo yields correct spelling; however, best pronunciations (from ensemble speech) are not canonical, although they may represent aspects of dynamic pronunciation adaptation.

Figure 11: Discovering the correct spelling for *Cindy Black* and introducing the possibility of pronunciation adaptation.



"buy computer" *does not occur in HW reco list*

```
writing(prob(13,0.55),text(lay computer)),
writing(prob(13,0.54),text(1uycompute
writing(prob(13,0.53),text(1uycempirt
writing(prob(14,0.52),text(1uycempater)),
writing(prob(14,0.51),text(1uyeanputer)),
writing(prob(15,0.5),text(luyeanputer)),
writing(prob(15,0.49),text(burglarproofed)),
writing(prob(16,0.48),text(buyouts)),
writing(prob(16,0.47),text(buoyantly)),
writing(prob(17,0.46),text(Honeycombs)),
```

"buy computer" *does occur far down in Speechalyzer Lattice*

```
                        CMU dictionary
                        pronunciation
                        ---------------------
 0.        COMPUTER  || k ah m p y uw t er
 2.   BIKE COMPUTER  || b ay k + k ah m p y uw
 4.    FLY COMPUTER  || f l ay + k ah m p y uw
 6.    LIE COMPUTER  || l ay + k ah m p y uw t
 8.    WHI COMPUTER  || w ay + k ah m p y uw t
 9.     Y. COMPUTER  || w ay + k ah m p y uw t
12.    BUY COMPUTER  || b ay + k ah m p y uw t
13.    BYE COMPUTER  || b ay + k ah m p y uw t
```

```
SF  PHN   EF     AS/Len
--  ---   --   --------
 7   W    16   -1894854
17  AY    26   -1515000
27   K    35   -1493535
36  AH    39    -755759
40   M    46   -1158547
47   P    53   -1293090
54   Y    64   -1882604
    UW    78   -2360212
     R    90   -1933854
     T    93    -578282
     M   110   -2940928
```

buy computer "B AY K AH M P Y UW T ER"

Comparing second-pass speech to closest match Speechalyzer lattice entry pulls up that entry's correct spelling and pronunciation.

Figure 12: Discovering the correct spelling and pronunciation of the taskline label, *buy computer*.

2.3.3 *Speech/Handwriting Fusion: 'buy computer'*

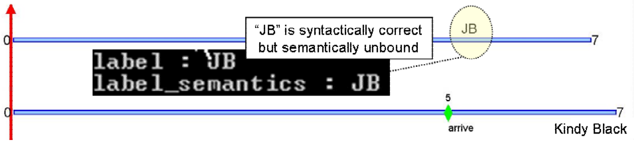
In finding the correct spelling of *buy computer* (as illustrated in Figure 12), we are able to leverage the refined phone matrix produced by the second-pass recognition over the cached speech features as an anchor for comparison. Comparing the term sequences extracted from the Speechalyzer lattice to the second-pass phone matrix yields a closest match for the 13th alternative (very low on the alternates list by virtue of its Speechalyzer score). This strong comparative match boosts it to have the best combined score, and thus allows SHACER to recover the correct spelling and pronunciation in this instance.

2.3.4 *Speech/Handwriting Fusion: Summary*

In summary it seems clear that the array of evidence (e.g., ensemble speech, handwriting, Speechalyzer transcripts and lattices, WPSR recognition) that we have at our disposal is very rich, and provides a basis for making many reasonable recognition choices in context. We are just beginning to explore the types of features available in this space and the ways in which we can take advantage of this rich information across redundant modal inputs.

2.3.5 *Discovering HW Abbreviation Semantics*

In Figure 6 the handwritten abbreviation, *JB*, is syntactically correct but semantically unbound. The system only knows that the symbol *JB* is a handwritten taskline label (see Fig. 13).

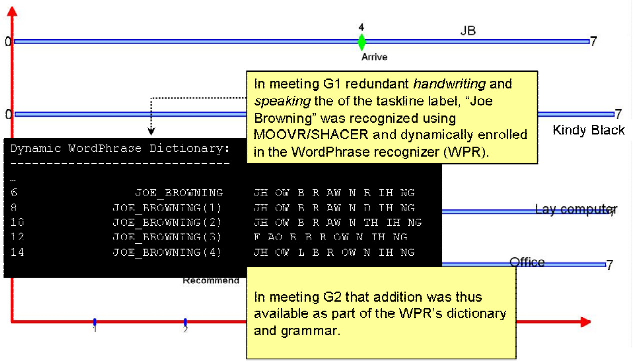


"JB" is syntactically correct but semantically unbound

```
label : JB
label_semantics : JB
```

Figure 13: Unbound semantics of taskline label, *JB*.

Without SHACER the system does not know that *JB* has a broader sphere of reference, and indeed shares the same meaning as the spoken and handwritten term, *Joe Browning*. SHACER has the capability to make this discovery and to do it dynamically based on WPSR enrollments from the previous meeting, *G1* (as shown in Figure 14). WPSR acts a persistent store of enrolled spelling/pronunciation combinations that is cumulative either within a single meeting or across a series of meetings, thus supporting boot-strapped recognition improvements the more often the system is used.



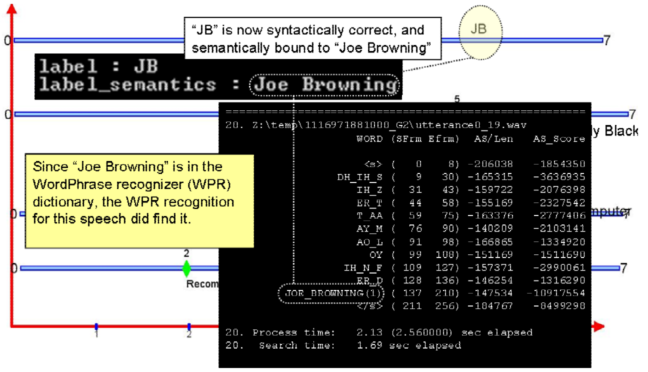
In meeting G1 redundant *handwriting* and *speaking* the of the taskline label, "Joe Browning" was recognized using MOOVR/SHACER and dynamically enrolled in the WordPhrase recognizer (WPR).

```
Dynamic WordPhrase Dictionary:
------------------------------
…
6              JOE_BROWNING     JH OW B R AW N R IH NG
8           JOE_BROWNING(1)     JH OW B R AW N D IH NG
10          JOE_BROWNING(2)     JH OW B R AW N TH IH NG
12          JOE_BROWNING(3)     F AO R B R OW N IH NG
14          JOE_BROWNING(4)     JH OW L B R OW N IH NG
```

In meeting G2 that addition was thus available as part of the WPR's dictionary and grammar.

Figure 14: Word/Phrase-Spotting Recognizer (WPSR) acting as a persistent store of spelling and pronunciations across a series of meetings (in this case from meeting *G1* to meeting *G2*).

In meeting *G2* as the user wrote *JB* he also said, *"This is our timeline for 'Joe Browning.'"* The Speechalyzer recognition for


"JB" is now syntactically correct, and semantically bound to "Joe Browning"
JB
Since "Joe Browning" is in the WordPhrase recognizer (WPR) dictionary, the WPR recognition for this speech did find it.
0
7
5
2
Recom
y Black
mputer
1
2

```
label : JB
label_semantics : Joe Browning
```

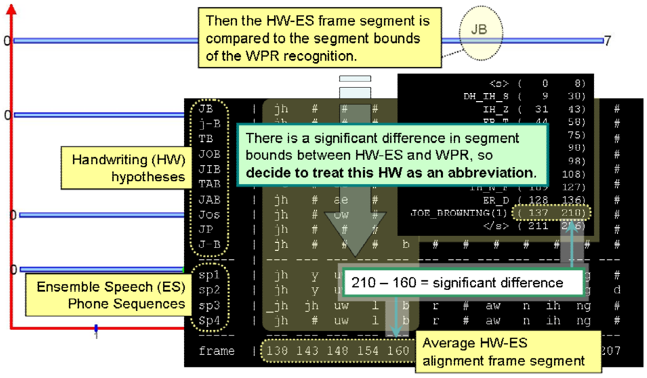
```
=====================================================
20. Z:\temp\1116971881000_G2\utterance0_19.wav
                    WORD  (SFrm Efrm)  AS/Len    AS_Score

                     <s> (    0    8) -206038    -1854350
                DH_IH_S (    9   30) -165315    -3636935
                   IH_Z (   31   43) -159722    -2076398
                   ER_T (   44   58) -155169    -2327542
                   T_AA (   59   75) -163376    -2777406
                   AY_M (   76   90) -140209    -2103141
                   AO_L (   91   98) -166865    -1334920
                     OY (   99  108) -151169    -1511690
                 IH_N_F (  109  127) -157371    -2990061
                   ER_D (  128  136) -146254    -1316290
        JOE_BROWNING(1) (  137  210) -147534   -10917554
                   </s> (  211  256) -184767    -8499298

20. Process time:    2.13 (2.560000) sec elapsed
20.  Search time:    1.69 sec elapsed
```

Figure 15: Word/Phrase-Spotting Recognition of *Joe Browning* as the basis of binding the semantics of HW abbreviation *JB*.

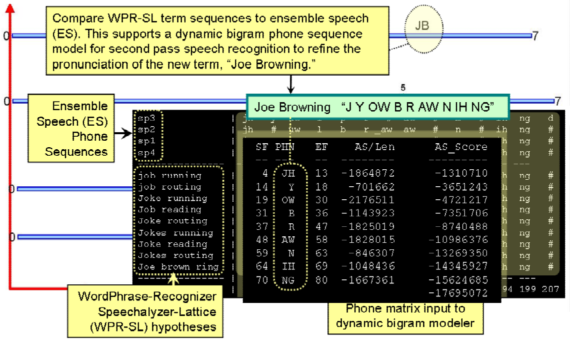
this utterance was, *"This is our timeline for **job running**,"* because *Browning* is not in the Speechalyzer dictionary. However, since *Joe Bronwing* was enrolled in the WPSR by SHACER during meeting *G1*, it is recognized by WPSR now in meeting G2 for this user utterance, and this recognition provides the basis for binding *JB* to *Joe Browning* as depicted in Figure 15.


Then the HW-ES frame segment is compared to the segment bounds of the WPR recognition.
JB
Handwriting (HW) hypotheses
There is a significant difference in segment bounds between HW-ES and WPR, so **decide to treat this HW as an abbreviation.**
210 – 160 = significant difference
Ensemble Speech (ES) Phone Sequences
Average HW-ES alignment frame segment
0
7
1

```
                                          <s> (    0    8)
                                      DH_IH_S (    9   30)
                                         IH_Z (   31   43)
                                                        58)
                                                        75)
                                                        90)
                                                        98)
                                                       108)
                                                       127)
                                         ER_D (  128  136)
                              JOE_BROWNING(1) (  137  210)
                                         </s> (  211  256)

JB     | jh  #   #   #                                    #
j-B    |                                                  #
TB     |                                                  #
JOB    |                                                  #
JIB    |                                                  #
TAB    |                                                  #
JAB    | jh  #   ae  #                                    #
Jos    | jh  #   ow  #                                    #
JP     | jh  #   #   #                                    #
J-B    | jh  #   #   #   b   #   #   #   #   #   #        #
-----  | --- --- --- --- --- --- --- --- --- --- --- ---
sp1    | jh  y                                            #
sp2    | jh  y                                            d
sp3    | _jh jh  uw  l   b   r   #   aw  n   ih  ng       #
Sp4    | jh  #   uw  l   b   r   #   aw  n   ih  ng       #
-----  | --- --- --- --- ---                          ---
frame  | 138 143 148 154 160                          207
```

Figure 16: Phonetic alignment matrix across handwriting hypotheses (for the abbreviation *JB*) and the ensemble speech phone sequences for *Joe Browning* section of, "This is our timeline for *Joe Browning*."


Compare WPR-SL term sequences to ensemble speech (ES). This supports a dynamic bigram phone sequence model for second pass speech recognition to refine the pronunciation of the new term, "Joe Browning."
JB
Ensemble Speech (ES) Phone Sequences
Joe Browning "J Y OW B R AW N IH NG"
WordPhrase-Recognizer Speechalyzer-Lattice (WPR-SL) hypotheses
Phone matrix input to dynamic bigram modeler
0
5
7

```
sp3
sp2   | jh  #  uw  l  b  r _aw  aw  #  n  #  ih  ng  #
sp1
sp4
---------------
job running
job routing
Joke running
Job reading
Joke routing
Jokes running
Joke reading
Jokes routing
Joe brown ring
---------------
```

```
SF PHN   EF     AS/Len       AS_Score
-- ---   --    --------     ---------
 4  JH   13   -1864872      -1310710
14   Y   18    -701662      -3651243
19  OW   30   -2176511      -4721217
31   B   36   -1143923      -7351706
37   R   47   -1825019      -8740488
48  AW   58   -1828015     -10986376
59   N   63    -846307     -13269350
64  IH   69   -1048436     -14345927
70  NG   80   -1667361     -15624685
                           -17695072
```

Figure 17: Spelling and pronunciation of the proper name semantically attached to the HW abbreviation *JB*.

As mentioned in list item 2 above, term recognition in WPSR is first used to match to an HW hypothesis. If the bounds of the HW hypothesis are significantly different, then the WPSR term segment can be used to expose the existence of a handwritten abbreviation. This situation is shown in Figure 16. The HW abbreviation phone sequence hypotheses cover a segment across the ensemble speech much shorter then the bounds of the WPSR term's end boundary. This significant difference triggers a decision to explore this WPSR event as the semantics of an HW abbreviation.

We use two pieces of evidence to make the final decision on binding the HW abbreviation. First we measure the distance of each HW hypothesis from the WPSR output. Currently we only consider the HW hypotheses as first letter abbreviatons (but in the future we will be expanding to consider other varieties of abbreviation). This measurement gives us the best first letter abbreviation interpretation (Figure 17). Second we examine the Speechalyzer lattice for term sequences across the boundaries found in the WPSR recognition, and compare them. Then after these two comparisons, if the first letter abbreviation distance is close enough and there is a sufficient match between WPSR output and the Speechalyzer lattice we decide to treat the HW as an abbreviation and bind its semantics to the proper name represented by the spelling associated with the WPSR speech recognition. We also use the Speechalyzer lattice terms, WPSR output and ensemble speech phone sequences to constrain a second-pass recognition on cached speech features. Figure 18 shows the result of the second-pass recognition, a plausible pronunciation adaptation for the term Joe Browning, which in turn is added back into the WPSR as another pronunciation alternative. In the future we will use such additions to refine the WPSR pronunciation alternatives (using clustering and centroid pronunciations, along the lines of what Roy and Yu & Ballard

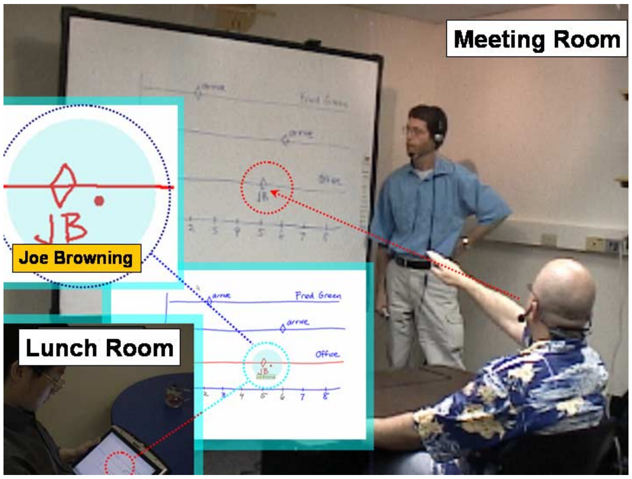


Figure 18: A distributed, multimodal, multiparty meeting during which MOOVR/SHACER has dynamically discovered the semantics of the abbreviaton, *JB*. As the seated user points toward *JB* his pointing gesture is distributed via the blue circle representing the gestural area of confidence. As he says, "There is a problem with **his** office space," the remote user also sees a hover label below *JB* which contains the semantics of the abbreviation, *Joe Browning*.

have outlined in their works – see Section 3.2), but for now we just expand the number of alternative pronunciations.

2.3.6 Using HW Abbreviation Semantics

Given MOOVR/SHACER's ability to dynamically discover the semantics of handwritten abbreviations, we have also augmented our display system so that in a distributed, multimodal, multiparty meeting setting (e.g. Fig. 1 and Fig. 18) a remote user (the *Lunch Room* user shown in Figure 18) can see a hover label floating below the abbreviation, *JB*, as another user (the seated *Meeting Room* user in Fig. 18) is pointing at it and referring to it. In Fig. 18 the seated user, as he is pointing toward the *JB* milestone on the whiteboard, is actually saying, "There is a problem with **his** office space." Since he does not refer to Joe Browning by name the hover semantics is an important means of contextualization for the remote user listening and watching the shared display on his tablet PC in the *Lunch Room*.

3. RELATED WORK

Early fusion systems like those that augment speech recognition by visually extracted face and lip movement features [14] employ an approach that discriminatively combines both input streams in a single feature space. Previous work in our group [15, 16] employs a *late-fusion* approach, which instead combines the output of separate modes after recognition has occurred. This is true as well for both our earlier work with MNVR and this work with MOOVR/SHACER for combining speech and handwriting outputs. For now *early-fusion* of speech and handwriting remains problematic, because of the temporal distance between handwriting and the speech that sometimes can occur.

3.1 Hybrid Fusion Phone Recognition

A third possibility, aside from either early or late fusion, is a *hybrid re-recognition* (HRR) approach. A variation of this approach has been used by Chung *et al* [1] in their speak and spell technique that allows new users to enroll their names in a spoken dialogue system.

The sub-word-units used by Chung *et al* for modeling OOV words are those of [17]. These are multi-phone sub-word units extracted from a large corpus with clustering techniques based on a mutual information (MI) metric. Bazzi [18] shows that using such 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 [19] 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). Applying his OOV language model to the complete utterances in a 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 algorithms for OOV recognition and multimodal new vocabulary enrollment (MNVR and MOOVR/SHACER) 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.

3.2 Multimodal Semantic Grounding

Roy [20] 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. In related work Gorniak *et al* [21] 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 [22] 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.

4. CONCLUSION AND FUTURE WORK

We have described a system capable of multimodal speech and handwriting recognition. MOOVR/SHACER is capable of leaning new terms dynamically from single instance observations of natural human-human interactions during multiparty meetings.

Knowledge of the learnt new terms persists across either a single meeting or across a series of related meetings. This capability supports both improved recognition of label names attached to chart constituents on the schedule chart created in our testbed domain of a multiparty, multimodal meeting, and also allows us to determine the semantics of a handwritten abbreviation for a term introduced in an earlier meeting. We have shown an example meeting series in which the use of multimodal integration of redundant speech and handwriting corrects three out of four chart constituent labeling errors, and persistent information about a learnt term from an earlier meeting is used to bind and display the semantics of a handwritten abbreviation on the schedule chart.

In the future we will attempt to move beyond chart constituent name enrollment to general grammar induction. In this way we believe it may be possible to recover the computational advantages demonstrated by our earlier Multimodal New Vocabulary Recognizer (MNVR) [11], which contextualized out-of-vocabulary words in specific grammar defined locations within a *carrier phrase*. In effect, we believe it may be possible to start dynamically learning the phrases that people in conversation actually use as carriers for the important content words.

Much more research needs to be done to understand how and why people deliver multimodal input redundantly in some settings. We will be studying various corpora to hone our understanding of these issues and also to begin forming better heuristic and statistical characterizations of the temporal, spatial and referential aspects of semantically redundant spoken and handwritten expression that can aid in building the next generation of robust multimodal recognizers.

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).

6. REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

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

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

[11] Kaiser, E.C. *Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application*. in *Proceedings of the International Conference on Intelligent User Interfaces*. 2005. San Diego, CA.

[12] Singh, R., *The Sphinx Speech Recognition Systems*, in *Encyclopaedia of Human Computer Interaction*, W. Bainbridge, Editor. 2004, Berkshire Publishing Group.

[13] Ko, T., D. Demirdjian, and T. Darrell. *Untethered Gesture Acquistion and Recognition for a Multimodal Conversational System*. in *Fifth International Conference on Multimodal Interfaces, ICMI '03*. 2003. Vancouver, B.C., Canada.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

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

[21] 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.

[22] 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.