BUILDING AN APPLICATION FRAMEWORK FOR SPEECH AND PEN INPUT INTEGRATION IN MULTIMODAL LEARNING INTERFACES

Minh Tue Vo and Cindy Wood

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15213-3890, U.S.A.
Email: {tue, caw}@cs.cmu.edu

ABSTRACT

While significant advances have been made in recent years to improve speech recognition performance, and more recently, gesture and handwriting recognition as well, speech- and pen-based systems have still not found broad acceptance in everyday life. One reason for this is the inflexibility of each input modality when used alone. Human communication is very natural and flexible because we can take advantage of a multiplicity of communication signals working in concert to supply complementary information or increase robustness with redundancy. In this paper we present a multimodal interface capable of jointly interpreting speech, pen-based gestures, and handwriting in the context of an appointment scheduling application. The interpretation engine based on semantic frame merging correctly interprets 80% of a multimodal data set assuming perfect speech and gesture/handwriting recognition; in the presence of recognition errors the interpretation performance is in the range of 35-62%. A dialog processing scheme uses task domain knowledge to guide the user in supplying information and permits human-computer interactions to span several related multimodal input events.

1. INTRODUCTION

Speech recognition is coming of age and is being incorporated into modern computer systems. Although pen input processing is still in its infancy compared to speech, pen-based systems have begun to appear in the form of personal digital assistants (PDAs) and pen laptops. However, current speech and pen systems are still not very popular and did not live up to the promise of bringing the power of computing to the population at large. Their failure is partly due to users' frustrations generated by inadequate recognition performance, especially for handwriting recognition. However, this failure is also an indication that current systems are too inflexible for the important task of facilitating the use of computers for everyone, especially people who may not be computer-literate. PDA users frustrated by gesture and handwriting recognition errors would naturally wish they could talk to their machines to correct the errors or simply to avoid having to write everything. Likewise users of speech-enabled systems quickly find out that there are tasks that cannot be conveniently expressed by spoken commands but would be enormously simplified by the ability to point to or circle objects on the screen in addition to speaking commands. These intuitive assertions were confirmed by a user study [1] conducted at Carnegie Mellon University, in which people interacting with a computer much preferred a combination of both speech and gestures over speech or gestures alone. Our own experiments with a calendar interface also show that when given a choice, people tend to use the communication method (or methods) most suitable for the task at hand and freely switch among available methods.

Research efforts at our Interactive Systems Laboratories (Carnegie Mellon University and University of Karlsruhe) are focused on producing a sensible and useful user interface by integrating multiple input modalities, rather than building better speech and gesture recognizers alone. An interface supporting the kind of highly flexible interaction we envision must be capable of integrating information from both speech and non-verbal input sources to arrive at a correct understanding of complete multimodal events. Some of our initial works along this line have been reported in previous publications [2][3][4][5]. This paper presents an approach to integrating speech, pen-based gestures, and handwriting, in the context of an appointment scheduling application. A multimodal interpretation engine jointly interprets information from all input sources by merging semantic frames. A domain-independent dialog processor maintains context information across input events. We evaluated the system with data collected in a user study conducted using the Wizard-of-Oz paradigm [6].

2. JEANIE: A MULTIMODAL CALENDAR

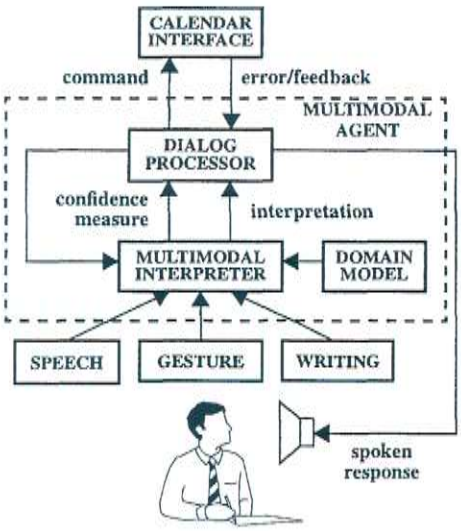
We have developed a prototype of a multimodal interface for an appointment scheduling program. A person using our Jeanie multimodal calendar can employ any combination of spoken input, gesturing with a pen on a touch-sensitive screen, or handwritten words to interact with the system. In typical scenarios, the user might say "Schedule a meeting on Monday" while at the same time drawing a box on the calendar to indicate where the new meeting should be inserted; write words on the newly scheduled meeting to annotate it; draw a cross on another meeting to cancel it; or point to a meeting and say "Reschedule this on Tuesday" or simply draw an arrow from that meeting to the new time slot on Tuesday.

An earlier version of this system was described in [4]. The interpretation engine in that system was based on an information-theoretic connectionist network [7] capable of incremental learning during use. Although that approach worked well for the preliminary system designed to explore multimodal interpretation, we found that the connectionist network was difficult to scale up when the task domain representation was significantly expanded in the new version. A semantic frame merging scheme was found to work much better on the larger and more complex action space, although we had to sacrifice the incremental learning capability.

Figure 1 shows a block diagram of the Jeanie system. The individual modality components (speech, gesture, and handwriting recognizers) are separate modules that can be replaced with ease. The calendar interface embodies domain knowledge and serves to isolate task-specific components. The heart of the system is the multimodal agent which interprets user input and performs

0-7803-3192-3/96 $5.00©1996 IEEE

FIGURE 1: The Jeanie multimodal calendar



requested operations through the calendar interface. This section describes the calendar interface and the modality processors. The next section presents the multimodal agent and the multimodal interpretation algorithm.

2.1. Calendar Interface

The object-oriented calendar interface is based on ICAL, a public domain calendar program developed by Sanjay Ghemawat at MIT. We extended ICAL to support a weekly view and client/server operations. The multimodal agent acting as a client of the calendar interface can query it for information such as gesture contexts (see Section 2.2) and ask the interface to carry out scheduling operations which include adding, removing, and changing appointments and notes.

2.2. Modality Processors

Inputs captured by a speech recorder and a touch-sensitive screen are processed by separate recognizers for each modality. Semantic frames extracted from recognizer outputs are merged by the multimodal interpreter to arrive at a unified interpretation as described in Section 3.

Speech. Our speech recognition subsystem is based on the recognition front-end of the JANUS speech translation system [8][9] which is capable of processing speaker-independent, spontaneous speech and was trained on human-human dialogs in the appointment scheduling domain. On a small data set of 128 utterances collected in user study experiments, the word recognition rate was 82%; if we take into account insertion, substitution, and deletion errors, the word accuracy is 76%.

Gesture. In earlier implementations we employed a gesture recognition module [2] based on a TDNN classifier [10]. The present implementation increases flexibility by decomposing gestures into sequences of strokes recognized as basic shapes such as lines, arcs, arrows, circles, crosses... Each gesture component is augmented by *gesture contexts* indicating spatial relationships between the gesture and nearby objects in the calendar interface. The current implementation of the gesture recognizer uses simple template matching and does not perform well on noisy input. A version based on the handwriting recognition algorithm described below is being developed and should greatly improve recognition performance.

Handwriting. Our handwriting recognizer developed by Stefan Manke at University of Karlsruhe based on the MS-TDNN [11] is capable of processing writer-independent, continuous (cursive) handwriting [12] at a recognition rate of over 90% on a 20,000-word vocabulary. Handwriting recognition is performed only when the gesture recognizer cannot identify the input strokes as basic shapes. This simple heuristic works fine for true cursive handwriting but can mistake writing strokes for gestures when it encounters printed block letters. This was unfortunately the case for some of the data collected in our user study; in addition most of the collected handwriting included uppercase letters for which the recognizer was not trained. These facts combined with gestures that were not adequately covered by templates led to a combined gesture/handwriting recognition rate of only 32%.

3. MULTIMODAL INTERPRETATION BY SEMANTIC FRAME MERGING

The multimodal interpreter is responsible for producing an interpretation of user intent (i.e., a command to send to the calendar interface) from the output of the modality processors. We represent this interpretation as a frame consisting of slots specifying pieces of information such as the action to carry out or the date and time of a meeting. Recognition output from the modality processors are parsed into partially filled frames that are merged together to produce the combined interpretation. The underlying algorithm is domain-independent although the encoding of information in frames must necessarily depend on the task domain.

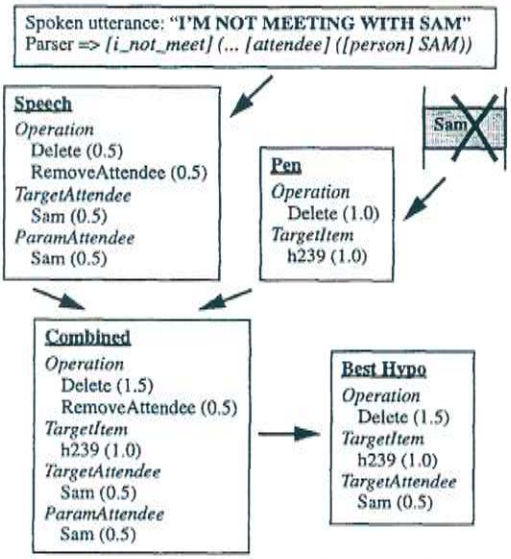
This frame merging technique represents a much extended and improved implementation of the approach sketched in [2]. It leads to uniform handling of high-level information from all input sources, which is very important for modularity and extensibility. To add another input modality we need only provide a module to convert low-level recognizer output to a partially filled frame to be merged with others. In addition, context information can be retained across input events by merging with previous interpretation frames as implemented in our dialog processor (Section 3.3).

3.1. Parsing Inputs From Individual Modalities

The text string output from the speech recognizer is processed by the Phoenix semantic parser developed by Ward [13]. The parser compiles a grammar specifying semantically meaningful fragments of text into an efficient recurrent transition network that identifies these fragments in the recognized utterance. It is capable of skipping unknown words and unmatched fragments and can therefore deal gracefully with ungrammatical sentences. The concepts identified by the parser serve to determine which slots of the speech frame should be filled and what scores should be assigned to the slot values.

The gesture/handwriting recognizers produce a sequence of gesture/handwriting components. The same gesture shape may mean different things depending on the calendar context surrounding it; for instance, a line across an appointment may indicate deletion but the same line falling on an empty time slot may visually describe a new meeting to be scheduled. In the present implementation of Jeanie, gestures and handwritten words are encoded into frames according to simple hard-coded heuristics. The development of a grammar to drive this process may become necessary for more complex gestures and operations.

FIGURE 2: Example of interpretation using frames



3.2. Merging Frames and Extracting Hypotheses

Frames are merged by taking the union of the sets of values filling each slots and adding corresponding scores. Embedded frames are merged recursively. This produces an aggregate frame encoding all alternative interpretation hypotheses. The multimodal interpreter extracts these hypotheses in descending order of aggregate scores and sends the best ones to the dialog processor. Figure 2 shows an example taken from our actual test data.

3.3. Dialog Processing

A dialog between a user and the system can span several related multimodal events, where each event may contain one or more spoken utterances and/or pen gestures. The dialog processor maintains a single frame per dialog to hold all information obtained since the start of the dialog; new information from each additional input event is merged into this frame and the scores are weighted appropriately so that frames produced by different input events in the same dialog contribute equally to the end result.

The domain-independent dialog processor also consults the domain interface (i.e., the calendar in this case) to filter out hypotheses that would cause domain errors such as missing or conflicting parameters. If it fails to find an acceptable hypothesis with a high enough score, it uses the error information returned by the domain interface to construct an appropriate feedback message to guide the user (only this process depends on the task domain). As soon as a good hypothesis is found, the dialog processor sends it to the domain interface as the operation to carry out and starts a new dialog. The user has the option of undoing the operation just performed if it is incorrect.

4. USER STUDY AND DATA COLLECTION

In order to create a useful and robust multimodal system, we need to find out how people would use a system with such capabilities. We have begun a series of experiments following the Wizard-of-Oz paradigm [6], in which the test subjects were presented with the Jeanie calendar interface and a hidden operator takes place of the real system in interpreting and carrying out multimodal commands. In parallel with the design of Jeanie, we conducted pilot experiments that contributed to the design of the frame representation for our task domain and yielded preliminary test data to evaluate the system.

Before the pilot experiment we conducted some small preliminary experiments to find out the types of gestures and speech people would use to make changes to a schedule, and to see if one modality was preferred over the other for certain tasks. That information was used to design pilot experiments that would illicit a mixture of speech and gestures from the test subjects. To get a variety of gestures that people might use the test subjects were first asked to make changes to a schedule using gestures only in any way they like. The same experiment was then repeated asking subjects to use speech only.

For the pilot experiments subjects were asked to make the changes by either giving voice commands or drawing on the touch screen or any combination of the two. Based on observations from the preliminary experiments we settled on instructing subjects by giving them a printed calendar with handwritten changes using as few words as possible and varying the wording to avoid speech contamination.

8 test subjects were used and a total of 13 trials were done for the preliminary and first set of pilot tests. The second pilot experiment consisted of 8 trials for 4 subjects using two completely different calendars to include a greater variety of tasks. In the pilot tests it was observed that some subjects preferred to use gestures more often, some preferred speech, others mixed the two. It was also observed that some tasks proved to be much easier to do by one method or the other; in those cases subjects chose the easier method regardless of general preference.

The pilot experiments showed, among other things, that test subjects often referred to a calendar event by using the name of a person or event, not just by a date and time; in addition, some subjects gave implied commands such as "On Tuesday I'm meeting with Bill, Tony, and Melanie" instead of "Add Tony and Melanie to the meeting with Bill on Tuesday". Test subjects with little computer experience tended to use implied commands most often, although we still do not have enough data to confirm such trends. The above observations significantly influenced the design of the frame representation.

5. PRELIMINARY PERFORMANCE EVALUATION

From out pilot Wizard-of-Oz experiments we obtained a small data set consisting of 185 multimodal events (77 speech-alone events, 57 pen-alone events, and 51 combination events). We transcribed each spoken utterance and pen gesture/handwriting as well as passing them through the modality recognizers. Results for all-transcribed inputs give us an idea of the performance of the multimodal interpreter, while results for recognized inputs offer some indication of performance degradation due to recognition errors.

The top-scored hypothesis produced by the interpreter for each input event is classified as *perfect*, *ok* (producing the intended result despite not being strictly correct), *ambiguous* (score too close to the next best hypothesis), *partially bad* (mostly correct except for one parameter), or *bad*. Table 1 shows the performance results for the pilot data. The pen data from the first pilot experiment could be transcribed but was unusable for recognition because of a flaw in the recording module which was cor-

TABLE 1: Multimodal interpretation performance

(TS/TP: transcribed speech/pen, RS/RP: recognized speech/pen)

	number of inputs	perfect (%)	ok (%)	ambi. (%)	partially bad (%)	bad (%)
TS/TP	185	78	2	7	3	10
RS/TP	185	57	5	16	6	16
TS/RP	100	49	2	26	9	14
RS/RP	100	28	7	29	12	24

rected for the second pilot experiment; because of this the bottom half of the table shows results for only the data set collected in this second experiment.

In the absence of recognition errors the multimodal interpreter will do the right thing 80% of the time (total of the *perfect* and *ok* columns), which is adequate for a usable system. At the other end of the spectrum the interpretation rate for all-recognized inputs is only 35%, a direct result of the poor 32% gesture/handwriting recognition rate.

The second row of the table shows that with a 76% word accuracy in speech recognition the interpretation rate drops to 62%. A closer analysis reveals that of the 18% that went from good to bad, about 15% occurred on speech-only events and only 3% occurred on combined speech/pen events. To put it another way, 27 of 77 (=35%) speech-only events caused interpretation errors due to speech recognition errors, compared to only 6 of 51 (=12%) combined events. On the pen-input side (only for the second data set of 100 input events), 11 of 18 (=61%) pen-only events caused interpretation errors due to gesture/handwriting recognition errors, compared to 19 of 39 (=49%) combined events. The sample is too small to assert any statistical significance, but this may be an indication of cross-modal redundancy at work.

We also observe that one significant effect of recognition errors is to increase the number of ambiguous outputs, more than the number of incorrect outputs. Recognition errors obviously reduce the amount of information available for interpretation, thereby causing more confusion. In interactive mode, the dialog processor can prompt the user with judicious guidance messages to supply clarifying information that could resolve ambiguities.

6. CONCLUSION

We have presented an architecture capable of joint interpretation of multimodal inputs and an example of a speech- and pen-enabled application using that architecture. The interpretation engine combines multiple input sources in a uniform and flexible way. The dialog processor maintains context information across input events and produces context-sensitive feedback to the user. The interpretation error rate is not unduly high but recognition errors may degrade performance significantly; however, we have seen indications that cross-modal redundancy can partly compensate for this degradation.

The pilot experiments have shown that the type of tasks we used in our Wizard setup will provide data for gesture, speech and a mixture of the two and will therefore be useful for the development and training of the Jeanie system. We are working on collecting more data to evaluate the system more fully, improve recognition rates, and refine parsing grammars. Experiments with the actual system in the loop rather than a Wizard operator are also being planned. Future studies should include experiments to investigate further the effect of cross-modal synergy in the presence of recognition errors to prove conclusively our intuitive assertions about the benefits of multimodal integration.

7. ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Office of Naval Research under Grant number N00014-93-1-0806.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Navy or the U.S. Government.

The authors would like to thank Dr. Alex Waibel for his guidance in this research and for his help in preparing this paper.

8. REFERENCES

[1] Hauptmann, A.G., "Speech and Gestures for Graphic Image Manipulation," *Proc. CHI'89* (Austin, Texas).

[2] Vo, M.T. and Waibel, A., "A Multimodal Human-Computer Interface: Combination of Speech and Gesture Recognition," *Adjunct Proc. InterCHI'93* (Amsterdam, The Netherlands).

[3] Vo, M.T. and Waibel, A., "Multimodal Human-Computer Interaction," *Proc. ISSD'93* (Waseda, Japan).

[4] Vo, M.T., Houghton, R., Yang, J., Bub, U., Meier, U., Waibel, A., and Duchnowski, P., "Multimodal Learning Interfaces," *Proc. ARPA SLT Workshop 95* (Austin, Texas).

[5] Waibel, A., Vo, M.T., Duchnowski, P., and Manke, S., "Multimodal Interfaces," *Artificial Intelligence Review, Special Volume on Integration of Natural Language and Vision Processing*, McKevitt, P. (Ed.), Vol. 10, Nos. 3-4, 1995.

[6] Salber, D. and Coutaz, J., "Applying the Wizard of Oz Technique to the Study of Multimodal Systems," *Proc. EWHCI'93* (Moscow, Russia).

[7] Gorin, A.L., Levinson, S., Gertner, A., and Goldman, E., "Adaptive Acquisition of Language," *Computer, Speech and Language*, Vol. 5, No. 2, pp. 101-132, April 1991.

[8] Woszczyna, M., Aoki-Waibel, N., Buø, F.D., Coccaro, N., Horiguchi, K., Kemp, T., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C.P., Schultz, T., Suhm, B., Tomita, M., and Waibel, A., "JANUS 93: Towards Spontaneous Speech Translation," *Proc. ICASSP'94* (Adelaide, Australia).

[9] Suhm, B., Geutner, P., Kemp, T., Lavie, A., Mayfield, L., McNair, A., Rogina, I., Schultz, T., Sloboda, T., Ward, W., Woszczyna, M., and Waibel, A., "JANUS: Towards Multilingual Spoken Language Translation," *Proc. ARPA SLT Workshop 95* (Austin, Texas).

[10] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K., "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 3, 1989, pp. 328-339.

[11] Haffner, P. and Waibel, A., "Multi-State Time Delay Neural Networks for Continuous Speech Recognition," *Advances in Neural Network Information Processing Systems 4*, Morgan Kaufmann Publishers, 1992, pp. 135-142.

[12] Manke, S., Finke, M., and Waibel, A., "The Use of Dynamic Writing Information in a Connectionist On-Line Cursive Handwriting Recognition System," *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann, 1994.

[13] Ward, W., "Understanding Spontaneous Speech: the Phoenix System," *Proc. ICASSP'91* (Toronto, Canada).