MULTIMODAL HUMAN-COMPUTER INTERACTION

Minh Tue Vo
tue@cs.cmu.edu
(412) 268-3076

Alex Waibel
ahw@cs.cmu.edu
(412) 268-7676

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890, U.S.A.

Keywords: Multiple modalities, multimodal interface, speech recognition, lip-reading, eye-tracking, gesture recognition, handwriting recognition

MULTIMODAL HUMAN-COMPUTER INTERACTION

Minh Tue Vo and Alex Waibel

School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213-3890, U.S.A.
E-Mail: tue@cs.cmu.edu, ahw@cs.cmu.edu

ABSTRACT

While human-to-human communication takes advantage of an abundance of information and cues, human-computer interaction is limited to only a few input modalities (usually only keyboard and mouse) and provides little flexibility as to choice of communication modality. In this paper, we present an overview of a family of research projects we are undertaking at Carnegie Mellon and Karlsruhe University to overcome some of these human-computer communication barriers. Multimodal interfaces are to include not only typing, but speech, lip-reading, eye-tracking, face recognition and tracking, and gesture and handwriting recognition. Initial experiments aimed at exploiting the complementary nature of these alternate modalities in interpreting user intent in a user interface are discussed.

KEYWORDS: Multiple modalities, multimodal interface, speech recognition, lip-reading, eye-tracking, gesture recognition, handwriting recognition.

1. INTRODUCTION

With multimedia workstations and high-speed data-links coming of age, we expect delivery and transmission of information to improve dramatically over the coming years. What is sorely lacking is the capture of information or the analysis and interpretation of human communicative signals. Human interaction is characterized by a multiplicity of signals, which generate redundant and complementary information that makes human communication robust, flexible and natural. To endow computer interfaces with similar flexibility, robustness and naturalness, we have begun to develop multimodal human-machine interfaces that incorporate human gesture, lip and face recognition, as well as hand modeling, character recognition and eye-tracking to understand the underlying intent and goals of the human user. Such multimodal interfaces are expected to be useful in human-to-human communication (e.g., video conferencing, speech translation), but also in human-computer interaction, such as database access, appointment scheduling, document production, CAD-design, operating machinery, etc.

In the following paper, we describe these efforts. We will begin by briefly describing the processing and recognition algorithms developed to recognize and understand each modality individually, e.g., speech recognition systems, lip-readers, eye-trackers, character and gesture recognizers. We will then describe first experiments aimed at combining some of these sources of information. and demonstrate that even in limited cases and domains, improvements can be obtained from combining complementary communication modalities.

2. PROCESSING OF DIFFERENT INPUT MODALITIES

2.1 Speech Recognition

Foremost among human communication modalities, speech and language undoubtedly carries a significant part if not most of the information in human communication. A multimodal human-computer interface should certainly take advantage of state-of-the-art speech understanding front ends. At Carnegie Mellon several approaches toward robust high performance speech recognition is under way. They include Hidden Markov Models (HMM) and several hybrid connectionist and statistical techniques. Several applications of speech processing is also under investigation, amongst them large vocabulary speech recognition, special vocabulary recognition and word spotting. Some of these will be described in the following.

2.1.1 Large Vocabulary Continuous Speech Recognition

Amongst the pure HMM systems, the Sphinx system is available as a large vocabulary speaker independent speech recognition server [9]. Some experiments aimed at speech interface design are carried out using this server. Several additional experiments exploring Learning Vector Quantization (LVQ-2) and Multi-State Time Delay Neural Network (MS-TDNN) and some other connectionist models are under way to attempt to provide highest acoustic phonetic recognition performance for large vocabulary continuous speech recognition [18][22].

Following acoustic phonetic modeling, an efficient search algorithms must find the most likely word sequence in real time. The search module of the recognizer builds a sorted list of sentence hypotheses using the word-dependent N-best algorithm [1]. The resulting N-best list is resorted using trigrams to further improve results. Resorting improves the word accuracy for the best scoring hypothesis (created using smoothed bigrams) from 91.5% to 98.8% [22] on a conference registration task [18]. Speed and memory requirements have been dramatically improved from earlier versions by using information collected in the first-best search for aggressive pruning in the N-best search; by dynamically adapting

the beam width to keep the number of active states constant; and by carefully avoiding the evaluation of states in large inactive regions of words. Although the number of computed hypotheses was increased from 6 to 100, the time required for their computation was reduced from typically 3 minutes to 3 seconds. Beyond improving acoustic recognition, the N-best search can also be used effectively for disambiguation and rescoring using multimodal information.

2.1.2 Word Spotting

Word spotting systems for continuous, speaker independent speech recognition are becoming more and more popular [13][21] because of the many advantages they afford over more conventional large scale speech recognition systems. Because of their small vocabulary and size, they offer a practical and efficient solution for many speech recognition problems that depend on the accurate recognition of a few important keywords. At Carnegie Mellon we have implemented and tested such a system on two spontaneous continuous speech databases (Zeppenfeld et al. [23][24]).

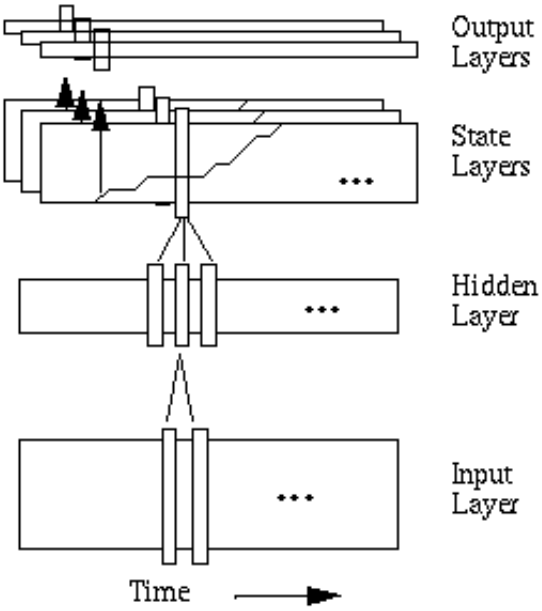

Figure 1: Word-Spotter System Architecture

The word spotting system architecture is based upon the Time Delay Neural Network (TDNN) [17], and more recently the Multi-State Time Delay Neural Network (MS-TDNN) [6][23]. A diagram of the basic network architecture is shown in Figure 1. [24] presents several recent improvements such as training with noise, average spectrum removal, equal occurrence keyword training, word duration modelling, state duration modelling, enforced minimum state durations, training with context frames, and keyword variant modeling.

Training and testing of the system was performed on two separate databases, the Roadrally corpus, and the new Switchboard corpus [24]. The system's performance is measured by plotting the keyword detection rate for several false alarm rates per keyword per hour (fa/(kw*hr)). By changing the thresholds of the word-output units, the detection rate can be improved at the expense of increasing the number of false alarms. The Figure of Merit (FOM) for the system is the averaged keyword detection rate over the false alarms from 0 to 10 fa/(kw*hr). Our system achieves an FOM = 72.2% for the Roadrally corpus and 50.9% on the much more difficult Switchboard corpus [24]. These figures compare favorably to those of other keyword spotting systems in its class evaluated by DARPA.

Our word spotting system has proved to be a viable alternative to the much larger full vocabulary speech recognition systems. With relatively few parameters we are able to achieve good performance and speed in recognizing select keywords on noisy, telephone quality spontaneous recordings.

2.2 Lip-reading

Most approaches to automated speech perception are very sensitive to background noise or fail totally when more than one speaker talk simultaneously (cocktail-party effect), as it often happens in offices, conference rooms, outdoors, and other real-world environments. Humans deal with these distortions by considering additional sources such as context information and visual information, such as lip movements. This latter source is subconsciously involved in the recognition process and is even more important for hearing-impaired people, but also contributes significantly to normal hearing recognition.

In order to exploit lip-reading as a source of information complementary to speech, we developed a lip-reading system based on the MS-TDNN and testing it on a letter spelling task for the German alphabet (Bregler et al. [4]). The recognition performance is understandably poor (31% using lip-reading only) because some phonemes cannot be distinguished using pure visual information; however, the thrust of this work is to show how a state-of-the-art speech recognition system can be significantly improved by considering additional visual information in the recognition process. This section presents only the lip-reading component; its combination with speech recognition is described in a later section.

We arranged to record acoustic and visual data in parallel (see Figure 7 in section 3.1). The video images covering the full face of the speaker are recorded in real-time (30 frames/sec) and saved as 256x256 pixel images with 8-bit grey-level information per pixel. We applied two alternative preprocessing techniques: histogram normalized grey-value coding and 2-dimensional Fourier transformation [4], performed on an area of interest (AOI) centered around the lips.

As recognition system we use a modular MS-TDNN (see Figure 6 in section 3.1). The visual information is processed by one of two front-end TDNN's. The classification is based on "visemes", or the smallest set of visually distinguishable units in speech, which is a subset of phonemes.

Simulation results are presented in section 3.1. Related papers are listed in [4].

2.3 Eye-tracking

The goal of gaze tracking is to determine where a person is looking from the appearance of his eye. Two potential uses of a gaze tracker are as an alternative to the mouse as an input modality [20] and as an analysis tool for human-computer interaction studies [11]. The direction of eye fixation can also be used to determine the user's focus of attention in a multimodal interface; for instance, knowing whether the user is looking at the screen or somewhere else while talking may be important in deciding whether automated speech recognition should be activated.

The most accurate gaze tracking so far has come from intrusive systems which either require the subject to keep their head stable, through chin rests etc., or systems which require the user to wear cumbersome equipment, ranging from special contact lenses to a camera placed on the user's head to monitor the eye. At Carnegie Mellon we have developed a neural-network-based non-intrusive gaze tracker based on camera input only (Baluja and Pomerleau [2]); the user is neither required to wear any special equipment, nor required to keep his head still.

Input to the system comes from a camera mounted on top of the computer monitor. An infrared light source creates a specular reflection on the eye. The gaze direction can be computed from the relative positions of the reflection and the pupil's center; the system performs this computation using a neural network. One of the primary benefits of the NN based gaze tracker is that it is non-intrusive; the user is allowed to move his head freely. In order to account for the shifts in the relative positions of the camera and the eye, the system searches for the specular reflection in the eye image and extracts a 15x30 window surrounding the reflection. The 15x30 window containing the image of the eye is used as the input to the neural network (see Figure 2). The output units are organized with 50 output units for specifying the X coordinate, and 50 units for the Y coordinate. The gaussian output representation used is similar to that used in ALVINN [12]. Training is performed by backpropagation [2].

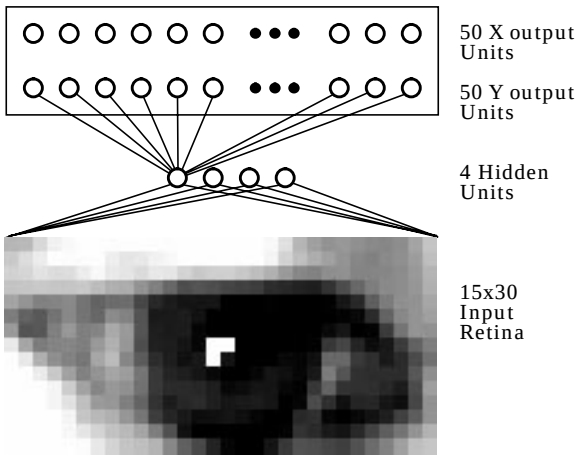

Figure 2: Network Architecture for Gaze Tracking

The current system works at 10 Hz. The best accuracy we have achieved is 1.5 degrees with the freedom of head movement up to 30 cm. Although we have not yet matched the best gaze tracking systems, which have achieved approximately 0.75 degree accuracy, our system is non-intrusive, and does not require the expensive hardware which many other systems require.

2.4 Gesture Recognition

We have been investigating pen-based gestures drawn using a stylus on a digitizing tablet. We developed a multimodal text editor capable of recognizing speech and gesture commands (Vo and Waibel [16]). The gesture component of the editor currently supports 8 gestures (see Table 1).

	Select		Begin selection
	Delete		End selection
	Delete		Transpose
	Paste		Split line

Table 1: Text-Editing Gestures

2.4.1 Input Representation and Preprocessing

We use a temporal representation of gestures, i.e. a sequence of coordinates tracking the stylus as it moves over the tablet's surface, as opposed to a static bitmapped representation of the shape of the gesture. This dynamic representation was motivated by its successful use in handwritten character recognition [5]. Results of experiments described in that work suggest that the time-sequential signal contains more information relevant to classification than the static image, leading to better performance.

In our current implementation, the stream of data from the digitizing tablet is preprocessed [5] by normalizing and resampling the coordinates to eliminate differences in size and drawing speed, and extracting local geometric information such as the direction of pen movement and the curvature of the trajectory. These features are believed to hold discriminatory information that could help in the recognition process.

2.4.2 Gesture Classification Using Neural Networks

We use a TDNN [17] (see Figure 3) to classify each preprocessed time-sequential signal as a gesture among the predefined set of 8 gestures. Each gesture in the set is represented by an output neuron. The network is trained on a set of manually-classified gestures using a modified backpropagation algorithm [17]. The output neuron with the highest activation level determines the recognized gesture.

Our gesture recognizer achieves 98.9% recognition rate on the training data set (640 samples) and 98.8% on an independent test set (160 samples).

2.5 Handwriting Recognition

The recognition of continuous handwriting, as it is being writ-

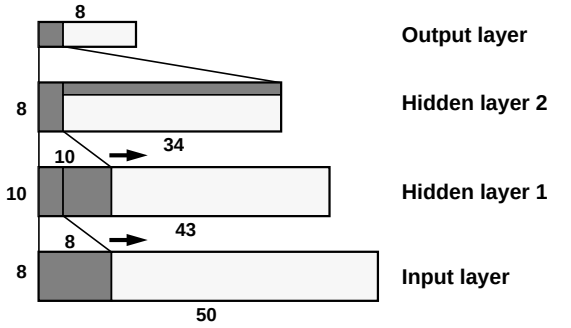

Figure 3: TDNN Architecture for Gesture Recognition

ten on a touch screen or digitizing tablet, has not only scientific but also enormous practical value, such as for note pad computers or for integration into multimodal systems. The main advantage of on-line handwriting recognition is that temporal information of writing can be recorded and used for recognition, much like for gestures as presented above. Handwritten words can be represented as a time-ordered sequence of coordinates with varying speed and pressure in each coordinate. Like in speech recognition the main problem of recognizing continuous words is that character or stroke boundaries are not known (in particular if no pen lifts or white space indicate these boundaries) and an optimal time alignment has to be found.

The TDNN [17] has been applied successfully for on-line single character recognition [5]. With single character, no automatic segmentation is necessary; however, some conflicts may arise that are unresolvable without context information. For instance, it is impossible to distinguish among “o”, “O”, and “0” by looking at a character in isolation.

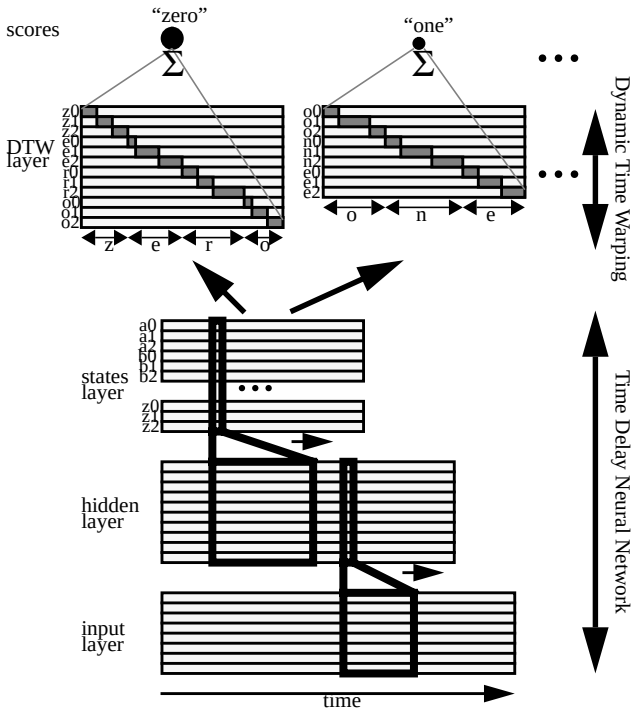

Figure 4: MS-TDNN for Handwriting Recognition

The MS-TDNN [6] was applied successfully to overcome the problem of recognizing continuous (cursive) handwriting (Bodenhausen and Manke [3]). This problem is much more difficult than the single character problem because of the need for automatic segmentation; however, it is possible to resolve the type of conflicts presented above using context. The MS-TDNN integrates the recognition and segmentation processes by combining the high accuracy character recognition capabilities of a TDNN with a non-linear time alignment procedure (Dynamic Time Warping) [10] for finding an optimal alignment between strokes and characters in handwritten continuous words (see Figure 4). In the most recent experiments, we achieved up to 94.7% word recognition rate on a database of 400 handwritten words.

2.6 Incremental Learning in Gesture/Handwriting Recognition

The usefulness of gesture and handwriting recognition depends largely on the ability to adapt to new users because of the great range of variability in the way individuals write or make gestures. No matter how many tokens we put in the training database to cover different gestures that mean “delete text”, for example, there may always be totally different gestures that are not yet part of the gesture vocabulary. This is particularly troublesome for neural network-based systems because usually the network has to be retrained using all the old training data mixed with a large number of new examples, in order to be able to recognize new patterns without catastrophically forgetting previously learned patterns. Because of the large number of examples needed and the long retraining time, this clearly cannot be done on-line in a way that would enable the user to continue to work productively. A good system should be able to query the user for correction and remember this particular input pattern in order to make intelligent guesses when similar inputs occur; during the subsequent work sessions new data can be quietly collected for off-line training of a regular network that will do a better job later on..

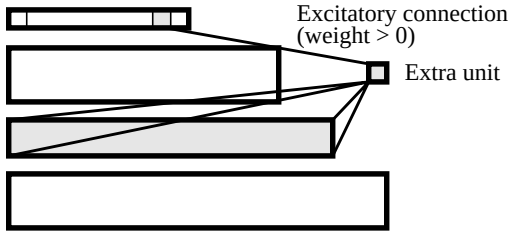

Figure 5: The Incremental TDNN Architecture

We have developed a method to accomplish this using an Incremental TDNN (ITDNN) architecture. A regular TDNN is trained using the available data as a base network. When a recognition error occurs during use, the system queries the user for the correct output and creates template-matching hidden units that influence the output units via excitatory or inhibitory connections (see Figure 5). In experiments involving a simple handwritten digit recognition task, we presented the base network with examples of an input pattern it had

never seen before. The recognition rate on this data set was near zero. With a single additional hidden unit, the network was able to recognize 97% of these new patterns with a performance drop of only 0.4% on the old training data set

These experiments show that the ITDNN is capable of quickly adding coverage for a new input variation without forgetting previously learned information and thus is a good candidate for systems requiring on-line, immediate recognition improvements during use.

3. COMBINATION OF MODALITIES

Beyond better recognizing and understanding each human communication event individually, we are mostly interested in combining multiple modalities to improve robustness and flexibility by offering complementary information. Several experiments aimed at such multimodal synergies have been undertaken.

3.1 Speech Recognition and Lip-reading

The lip-reading system described in section 2.2 is actually one half of a speech perception system using both acoustic and visual signals. The system was constructed for the connected German letter spelling task which features a small but highly ambiguous vocabulary, with no grammar or other high-level information. The system is described in detail in [4].

The system shown in Figure 6 is based on a modular MS-TDNN architecture [8]. The preprocessed acoustic and visual data are fed into two front-end TDNN's [17], respectively. Backpropagation is used to train the networks separately in a bootstrapping phase, to fit phoneme targets. The last layers (phone-state layers) of the TDNN's are combined using "entropy weights" [4], and the DTW algorithm [10] is applied to find the optimal path of phone-hypotheses for the word models. A second training phase backpropagates the error

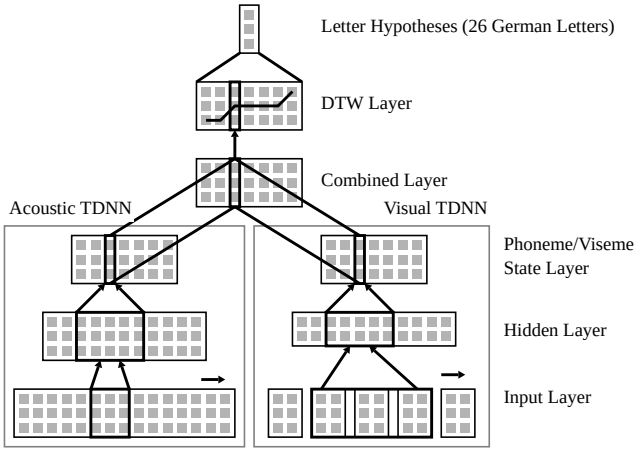

Figure 6: Speech Recognition/Lip-reading System

derivatives from the letter units through the best DTW path down to the front-end TDNN's, in order to optimize the overall network for the actual evaluation task, which is letter and not phoneme recognition.

We obtained data to train and test the system using a bimodal data acquisition procedure (Figure 7) in which acoustic and visual data are recorded in parallel. Visual data acquisition and preprocessing was described in section 2.2. The acoustic data is sampled at a 16-kHz rate and 12-bit resolution. For acoustic preprocessing we follow the established approach of applying the Fast Fourier Transform (FFT) on the Hamming-windowed speech data in order to get 16 Melscale Fourier coefficients at a 10-ms frame rate. Our database consists of 2 sets of 114 and 350 letter sequences (names and random sequences) respectively, spelled by two male speakers. The first data set was split into 75 training and 39 test sequences; the second set was split into 200 training and 150 test sequences. Table 2 summarizes the recognition performance results on the sentence level. It can be seen that the additional visual information can help reduce the error rate by 40-50%,

Figure 7: Bimodal Data Acquisition for Speech Recognition and Lip-reading

especially in the presence of noise.

	Acoustic	Visual	Combined
msm/clean	88.8	31.6	93.2
msm/noisy	47.2	31.6	75.6
mcb/clean	97.0	46.9	97.2
mcb/noisy	59.0	46.9	69.6

Table 2: Word Accuracy of Speech/Lip System

3.2 Speech and Gesture Recognition

Joint interpretation of multimodal events was successfully demonstrated in our speech- and gesture-based text editor [16]. Figure 8 shows a block diagram of the multimodal interpreter module.

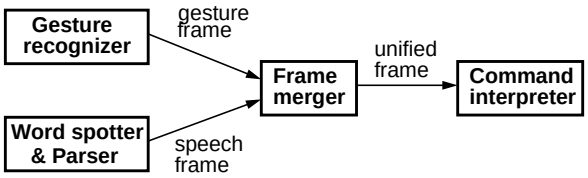

Figure 8: Joint Interpretation of Gesture and Speech

The TDNN-based gesture recognizer was described in section 2.4. For the speech component we use the word spotter (section 2.1.2) coupled with a semantic-fragment parser [19]. The word spotter was trained to spot 11 keywords representing editing commands such as *move, delete,...* and textual units such as *character, word,...* The effect is to let the user speak naturally without having to worry about grammar and vocabulary, as long as the utterance contains the relevant keywords. For example, an utterance such as "Please delete this word for me" is completely equivalent to "Delete word".

We based the interpretation of multimodal inputs on frames consisting of slots representing parts of an interpretation. The speech and gesture recognizers produce partial hypotheses in the form of partially filled frames. The output of the interpreter is obtained by unifying the information contained in the partial frames.

Consider an example in which a user draws a circle and says "Please delete this word". The gesture-processing subsystem recognizes the circle and fills in the command scope (what to operate on) specified by the circle in the gesture frame. The word spotter produces "delete word", from which the parser fills in the *action* and *textual unit* slot in the speech frame. The frame merger then outputs a unified frame indicating that the operation *delete* is to be carried out on the word specified by the scope of the circle.

One important advantage of this frame-based approach is its flexibility, which will facilitate the integration of more than two modalities. All we have to do is define a general frame for interpretation and specify the ways in which slots can be filled by each input modality. In a general implementation, it is possible that the slots may be filled in different ways, and performing a search to find the best merge would be superior.

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we have reported on a number of experiments aimed at integrating multiple sources of sensory information into joint multimodal human computer interfaces. Combining modalities could be seen to

- improve recognition performance significantly by exploiting redundancy (e.g. speech & lip-reading)
- provide greater expressiveness and flexibility by exploiting complementary information in different modalities (e.g. speech & gesture)
- improve understanding in allowing for complementary modalities to take effect.

While our experiments are encouraging, much remains to be done. In addition to continuing improvements on the underlying pattern processing methods, we are beginning to scale our experiments to larger, less constrained human computer interfaces. To assess the relative effectiveness of individual modalities in different tasks, we are beginning to perform user studies interactively, while designing systems that support promising multimodal enhancements. Wizard-of-Oz-style simulations of potential interfaces may be used to investigate user preferences, short of having a finished system. Further research on capturing sensory information more flexibly and combining it more robustly is also under way. Amongst them are face-tracking and recognition algorithms, that derive visual information on eye and lip motion, and acoustic speech information, even when more than one speaker move about the room.

5. REFERENCES

[1] S. Austin and R. Schwartz. A Comparison of Several Approximate Algorithms for Finding N-best Hypotheses. In *Proc. ICASSP'91*, pp. 701-704.

[2] S. Baluja and D. Pomerleau. Non-Intrusive Gaze Tracking Using Artificial Neural Networks. To appear in *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann, 1993.

[3] U. Bodenhausen and S. Manke. Connectionist Architectural Learning for High Performance Character and Speech Recognition. In *Proc. ICASSP'93*.

[4] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving Connected Letter Recognition by Lipreading. In *Proc. ICASSP'93*.

[5] I. Guyon, P. Albrecht, Y. LeCun, J. Denker, and W. Hubbard. Design of a Neural Network Character Recognizer for a Touch Terminal. *Pattern Recognition*, 1990.

[6] P. Haffner, M. Franzini, and A. Waibel. Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition. In *Proc. ICASSP'91*.

[7] A. Hauptmann. Speech and Gestures for Graphic Image Manipulation. In *Proc. CHI'89*, ACM Press, pp. 241-245.

[8] H. Hild and A. Waibel. Connected Letter Recognition with a Multi-State Time Delay Neural Network. *Neural Information Processing Systems* (NIPS 5).

[9] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee, and R. Rosenfeld. The SPHINX-II Speech Recognition System: An Overview. *Computer Speech and Language* (in press), 1993.

[10] H. Ney. The Use of a One-Stage Dynamic Programming Algorihtm for Connected Word Recognition. In *Proc. ICASSP'84*.

[11] C. Nodine, H. Kundel, L. Toto, and E. Krupinski. Recording and Analyzing Eye-position Data Using a Microcomputer Workstation. *Behavior Research Methods, Instruments & Computers* 24 (3), 1992, pp. 475-584

[12] D. Pomerleau. Neural Network Perception for Mobile Robot Guidance. Ph.D. Thesis, Carnegie Mellon University, CMU-CS-92-115.

[13] R. Rose and D. Paul. A Hidden Markov Model Based Keyword Recognition Systems. In *Proc. ICASSP'90*.

[14] O. Schmidbauer and J. Tebelskis. An LVQ-based Reference Model for Speaker-Adaptive Speech Recognition. In Proc. ICASSP'92, pp. 441-444.

[15] J. Tebelskis and A. Waibel. Performance Through Consistency: MS-TDNNs for Large Vocabulary Continuous Speech Recognition. *Advances in Neural Information Processing Systems*, Morgan Kaufmann.

[16] M.T. Vo and A. Waibel. A Multimodal Human-Computer Interface: Combination of Speech and Gesture Recognition. In *Adjunct Proc. InterCHI'93*.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1989.

[18] A. Waibel, A. Jain, A. McNair, H. Saito, A. Hauptmann, J. Tebelskis. JANUS: a Speech-to-speech Translation System Using Connectionist and Symbolic Processing Strategies. In *Proc. ICASSP'91*.

[19] W. Ward. Understanding Spontaneous Speech: the Phoenix System. In *Proc. ICASSP'91*, pp. 365-367.

[20] C. Ware and H. Mikaelian. An Evaluation of an Eye Tracker as a Device for Computer Input. In *Human Factors in Computing Systems IV,* 1987.

[21] J. Wilpon, L. Miller, and P. Modi. Improvements and Applications for Keyword Recognition Using Hidden Markov Modeling Techniques. In *Proc. ICASSP'91*.

[22] M. Woszczyna et al. Recent Advances in JANUS:A Speech Translation System. In *Proc. EUROSPEECH'93*.

[23] T. Zeppenfeld and A. Waibel. A Hybrid Neural Network, Dynamic Programming Word Spotter. In *Proc. ICASSP'92*.

[24] T. Zeppenfeld, R. Houghton, and A. Waibel. Improving the MS-TSNN for Word Spotting. In *Proc. ICASSP'93*.