MULTIMODAL LEARNING INTERFACES

Minh Tue Vo[1], Ricky Houghton[1], Jie Yang[1], Udo Bub[1],
Uwe Meier[2], Alex Waibel[1,2], Paul Duchnowski[2]

[1] Carnegie Mellon University, Pittsburgh, PA, USA

[2] University of Karlsruhe, Karlsruhe, Germany

1. INTRODUCTION

While significant advances have been made in recent years to continuously expand and improve speech recognition performance, speech recognition systems have still not found broad acceptance in everyday life. In searching to eliminate their shortcomings, we have begun to focus our efforts on producing a sensible and useful *user interface*, rather than a better recognizer alone. Such useful speech interfaces should not only recognize speech but also

- recognize other communication modalities such as gesture, handwriting, and pointing,
- provide freedom from headsets and push-buttons,
- allow for graceful recovery from errors and miscommunications, and
- know what they don't know, and what the user does or doesn't know.

At our laboratories we have embarked in an effort aimed at solving some of these problems by designing multimodal interfaces that

- combine speech, pen-based gesture, and handwriting,
- combine lip-reading and speech for robust recognition in the presence of noise,
- combine visual (face tracking) and acoustic processing (microphone arrays) to extract better speech signals in the presence of jamming noise and to identify focus of attention and address of a given speaker, and
- combine alternate modalities for speech recognition error repair.

In this paper we review these activities and our performance results to date.

2. COMBINATION OF SPEECH, GESTURE, AND HANDWRITING

We have developed a multimodal interface for an appointment scheduling task on a computerized calendar. Figure 1 shows a block diagram of the system.

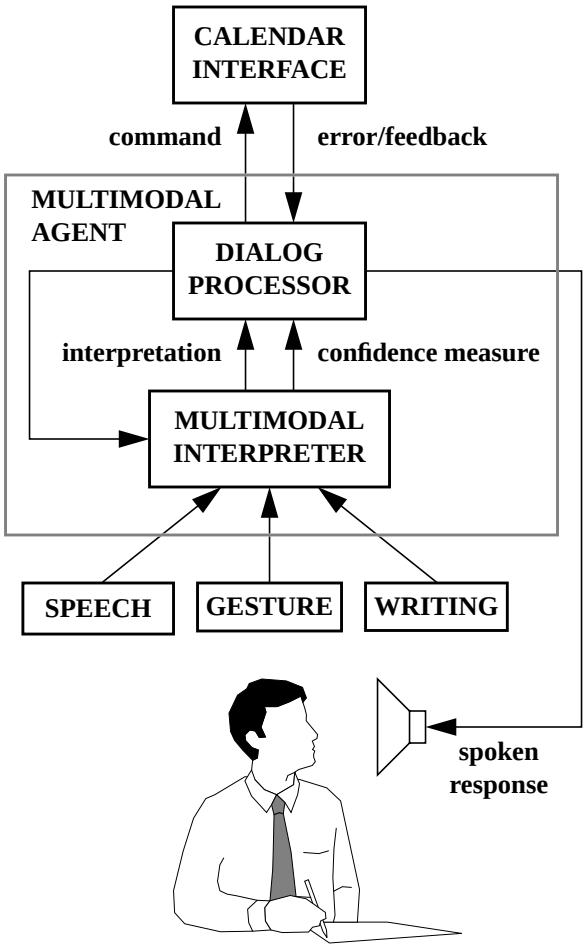

Figure 1: Block diagram of the multimodal interface

The user can use any combination of spoken input, gesturing with a pen on a touch-sensitive screen, or handwritten words to interact with the system. In a typical scenario, the user might say "Schedule a meeting on Monday," while at the same time drawing a line on the calendar to indicate when the meeting should start and how long it should last; 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," etc.

2.1. Separate Modality Recognition

Our speech recognition subsystem is based on the recognition frontend of the JANUS speech translation system [1][2][3] and the SPHINX continuous speech recognition system [4]. The JANUS recognizer is capable of processing speaker-independent, spontaneous speech and was trained on human-human dialogs in the appointment scheduling domain. We are currently collecting data on human-computer interaction to retrain the recognizer for our multimodal interface.

The gesture recognition module [5] is a TDNN classifier [6] trained to recognize 8 editing gestures drawn with a stylus on a touch-sensitive screen or a digitizing tablet. The input to the network is a sequence of coordinates tracking the strokes made with the stylus, preprocessed to extract local geometric features [7]. With training data of 80 samples per gesture, we have achieved "gesturer"-dependent recognition rate of over 98% on an independent test set. In addition, we have done some experiments on providing the gesture recognizer with the capability of learning new gesture variants incrementally during actual use. We achieve this by adding extra template-matching units to the TDNN [8].

The handwriting recognizer developed by Stefan Manke at University of Karlsruhe based on the MS-TDNN [9][10] is capable of processing writer-independent, continuous (cursive) handwriting [11][12]. The MS-TDNN integrates recognition with automatic segmentation by combining the high accuracy character recognition capabilities of a TDNN with a non-linear time alignment procedure (Dynamic Time Warping) for finding an optimal alignment between strokes and characters. In the most recent experiments, we achieved 93% writer-independent word recognition rates on a database of 400 handwritten words. Recognition experiments on a 20,000-word vocabulary task are in progress.

2.2. Multimodal Interpretation

The multimodal interpreter module uses the mutual information between all the input sources and output actions to derive a joint interpretation of user intent. In this manner it can incrementally learn input/output associations during actual use instead of having to go through lengthy training by examples. This capability could potentially be very valuable because it allows the system to adapt to a particular user over time.

The dialog processor employs simple discourse modeling to permit interactive, multimodal error correction. By querying the application (calendar) interface, which embodies domain knowledge, the dialog processor can detect missing or conflicting information and provide specific feedback responses much more useful than the typical "I don't understand, please repeat." For instance, if the user says "Schedule a meeting on Monday" without specifying starting time and duration, the system's response will indicate that those specific pieces of information are missing, and the user has the choice of using speech, gesture, or a combination of modalities to supply them. A dialog with the system can stretch over several individual multimodal interactions to culminate in the determination of an action to carry out. The user's acceptance or rejection of this interpretation provides cues to the system to update its input/output associations.

3. FACE AND ATTENTION TRACKING

To understand a user's intention better and to allow for freedom of movement, freedom from headsets and push-buttons, better modeling of a user is required. Such modeling includes identifying where a user is as well as what he/she is currently doing, saying, and looking at.

3.1. Face Tracker

We are developing a system capable of identifying where one or more people are in a room. Locating and tracking human faces is a way to achieve this goal.

The face tracker under development is a system that can locate and track people moving freely within a room. The task of the face tracking system, described in detail elsewhere [13], is to supply other recognition/understanding systems with the coordinates and a stable image of the speaker's face. While tracking a face, the position of the camera and the zoom lens are automatically adjusted to maintain a centered position of the face at a desired size within the camera image.

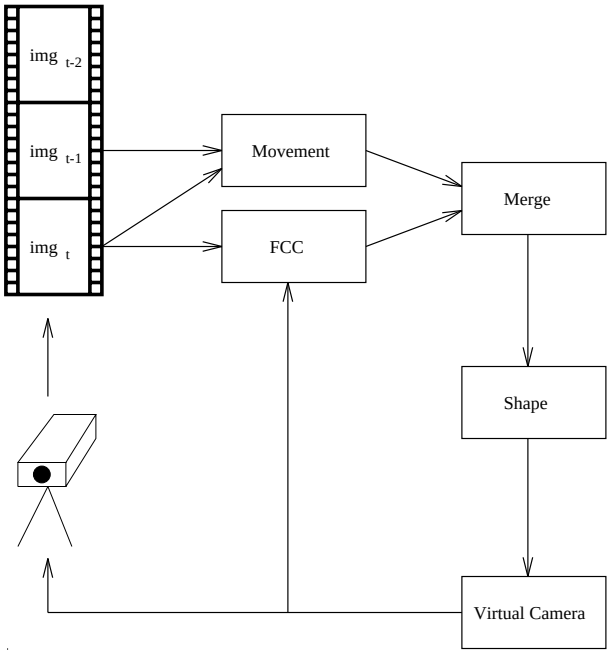

Figure 2: Face tracker system architecture

The system consists of a Canon pan-tilt-zoom camera (VC-C1) controlled by a workstation via a serial port. Color information is extracted by the Face Color Classifier (FCC) which maps each pixel into a two-dimensional brightness-normalized color space divided into colors belonging to faces and all others. As few as five sample images of faces with various skin colors have been found sufficient to establish this color distribution. Movement is computed from successive frames and merged with the color

information. The resulting candidate face objects are fed into a neural network which considers object shapes in producing the coordinates of the *virtual camera*, indicating the region actually containing the face. Appropriate pan-tilt-zoom commands are issued to the camera if the face moves out of a predefined area in the center of the physical camera.

Two neural networks are used for centering and size estimation, respectively. They were trained by backpropagation on 5000 artificially scaled and shifted example images generated from a database containing 72 images of 24 faces of different sex, age, hair style, skin color, etc. Performance was evaluated on test sequences of over 2000 images of 7 persons (with different skin types) moving arbitrarily about in front of different backgrounds. Depending on the sequence, the face was located in 96% to 100% of all images in the sequence. The average difference of the actual position of the face and the output of the system was less than 10% of the size of the head.

3.2. Attention Finder

Many Human-Computer-Interaction applications need to know where a person is looking, and what he/she is paying attention to. This information can provide valuable communication cues to a multimodal interface. Such information can be obtained from tracking the orientation of a human head, or gaze. While current approaches to gaze tracking tend to be highly intrusive (the subject must either be perfectly still, or wear a special device), we aim to develop a more flexible system using computer vision technology. We have developed a system, Attentionfinder, that can identify a person's focus of attention based on face orientation tracking. Our system allows a person to move freely in a room while finding his/her face orientation. A person's face image is captured by the face tracker described above. The orientation of the face is then computed by non-linear mapping between input image and angular coordinates [14]. The system can provide both binary output and angular information from -90^o to 90^o.

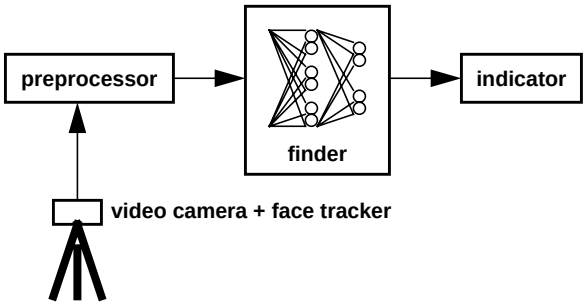

Figure 3: Block diagram of the Attentionfinder

We used four sets of 15 images of 7 different people for training. The 15 images correspond to various face orientations from -70^o to 70^o. The images are artificially shifted to create a total of 50,820 training images. An independent test set consists of 14,520 images. If only 3 outputs (right, straight, left) are required, the system classifies the images correctly for 99.7% of the training data and 95.7% of the test set. Another setup allows the determination of 19 face orientations in 10^o steps from -90^o to 90^o. In this case the classification performance is 30% for the training data and 24% for the test set, but the average error is only 10^o on training data and 12^o on test data. These results are still encouraging because the training set itself contained some inaccuracy due to the data collection method and the artificial generation of extra data.

4. COMBINATION OF FACE TRACKING AND MICROPHONE ARRAY

Another subsystem coupled to our face tracker is a microphone array intended to replace the close-talking microphone and provide freedom from intrusive headsets. The microphone array, consisting of 8 to 16 sensors arranged in a horizontal formation to span the half plane in front of the array, locates a sound source by phase delay measurements and enhances it using a beamforming technique. In order to steer the array towards a given spot, the differences of sound arrival time between the microphones are compensated for waves originating exactly from this location. By summing these aligned (in phase) signals, one achieves an enhancement of the desired signal. Competing sounds, not correlated with the signal and coming from other locations, are added out of phase and attenuated.

We conducted experiments with a context-independent version of JANUS [2] in a noisy environment to assess the effectiveness of the array. With a close-talking microphone, an initial word accuracy of 88% over sentences rerecorded (by a non-native speaker in a noisy room) from Resource Management text material was observed. This result deteriorated rapidly to 58.3% when a single microphone was placed away from the speaker, even with channel adaptation. By using the microphone array this score improves to 79.8%.

By replacing the acoustic localization procedure, which is limited to finding the loudest sound source in a room, with visual localization based on face tracking, our system allows the speaker of interest to move freely in a room in the presence of noise [15]. We investigated two noise situations: *background noise* consisting of low-level signals such as humming fans (in this case the main speaker is clearly dominant), and *competing noise* in the form of music from a radio at high output level.

	Backg. noise	**Comp. noise**
Single microphone	59.8	14.5
Acoustically guided beam	69.5	43.4
Visually guided beam	68.9	54.6
Close-talking microphone	88.1	88.4

Table 1: Word accuracy for microphone array with moving speakers in background- and competing-noise environments

Table 1 shows that visually and acoustically guided beamforming lead to comparable recognition rates in the case of background noise, but the presence of competing noise loud enough to distract the array's focus renders acoustic guidance less effective than visual localization. The recognition rates with beamforming are lower than the 79.8% reported earlier since in this experiment the speaker is moving instead of standing still in front of the microphone array.

5. COMBINATION OF SPEECH AND LIP-READING

Most approaches to automated speech recognition (ASR) that consider solely acoustic information are very sensitive to background noise or fail totally when two or more voices are presented simultaneously (cocktail party effect). Humans deal with these distortions by considering additional sources such as directional, contextual, and visual information, primarily lip movements. We are interested in emulating some of these capabilities by combining speech recognition with lipreading to improve robustness and flexibility by offering complementary information.

Our audio-visual speech recognizer has been developed for a German spelling task mainly in speaker-dependent mode. Letter sequences of arbitrary length and content are spelled without pauses. The task is thus equivalent to continuous recognition with small but highly confusable vocabulary.

In order to give the speaker reasonable freedom of movement within a room, the speakers face is automatically acquired and followed by the face tracker subsystem described above, which delivers constant-size, centered images of the face in real time. The image of the lips is automatically extracted from the camera picture of the speakers's face by the lip locator module consisting of two neural networks. The first network estimates the initial position of the mouth from two directional edgemaps. The second network detects the corners of the mouth, from which a window showing only the mouth area is extracted from the image. The sequence of extracted lip images is the input to the lip-reader module.

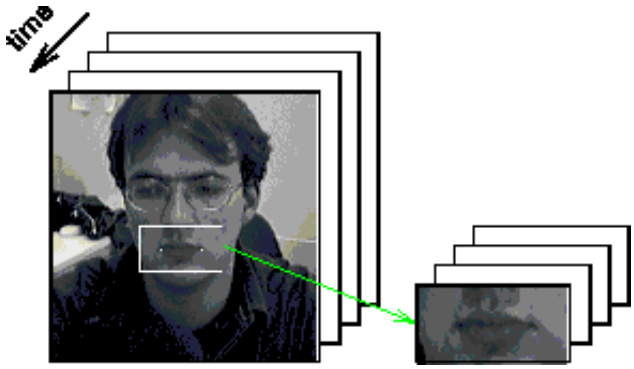

Figure 4: Extraction of lip image sequence

We record acoustic and visual data in parallel. Conventional preprocessing of the acoustic input gives 16 Melscale Fourier coefficients at a 10-ms frame rate. For visual data preprocessing, we have been investigating several alternate visual data representations: direct gray-level values, Fourier magnitude coefficients (averaged in rings in the frequency domain), principal components, and discriminant analysis coefficients.

A modular MS-TDNN, drawing on a pure acoustic spelling recognizer [16], performs the recognition. Figure 5 shows the network architecture. Through the first three layers (input-hidden-phoneme/viseme) the acoustic and visual inputs are processed separately. The third layer produces activations for 62 phoneme or 42 viseme (the rough visual correlate of a phoneme) states for acoustic and visual data, respectively. Weighted sums of the phoneme and corresponding viseme activations are entered in the combined layer and a one-stage DTW algorithm finds the optimal path through the combined states that decode the recognized letter sequence. The weights in the parallel networks are trained by backpropagation. There are 15 hidden units in both subnets. The combination weights are computed dynamically during recognition to reflect the estimated reliability of each modality. We have also investigated alternative methods of combining the audio and visual information at the input and hidden layer levels of the network [17].

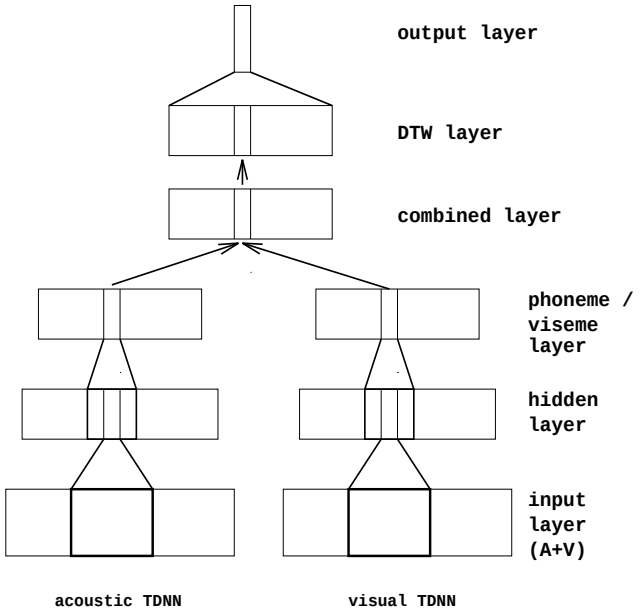

Figure 5: Audio-visual speech recognizer architecture

We have tested the recognizer on data sets of 200 letters sequences from single speakers. The performance measured by word accuracy is reported in Table 2.

Preproc.	Processing	Clean	16 dB	8 dB
Gray-level images	Visual only	30.0	30.0	30.0
	Acoustic only	93.9	64.2	41.2
	AV combined	97.6	75.2	49.1
LDA	Visual only	53.0	53.0	53.0
	Acoustic only	93.9	64.2	41.2
	AV combined	97.6	77.0	63.6

Table 2: Lipreading performance for various signal/noise ratios

6. WORK IN PROGRESS AND FUTURE DIRECTIONS

We are currently working towards integrating all the above subsystems into useful and flexible interfaces that allow the user to interact with the computer using either voice or pen-based input or any combination thereof, and interact with other people in a computer-assisted video conference. We envision systems that allow for freedom of movement in a possibly noisy room without the bother of intrusive devices such as headsets and close-talking microphones.

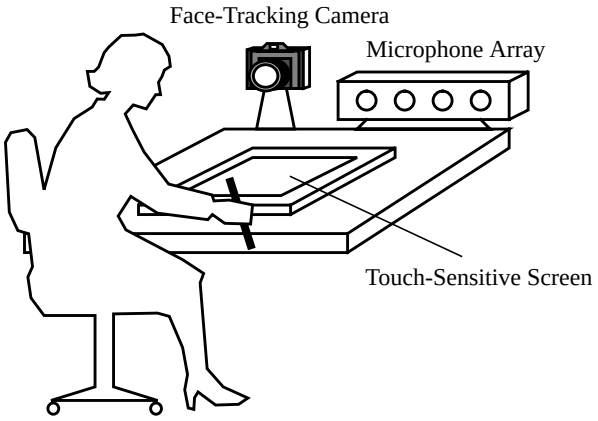

Figure 6: Multimodal human-computer interface setup

REFERENCES

1. Woszczyna. M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C.P., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W., "Recent Advances in JANUS, a Speech Translation System," *Proc. EUROSPEECH'93* (Berlin, Germany, Sept. 1993), Vol. 2, pp. 1295-1298.

2. 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, April 1994), Vol. 1, pp. 345-348.

3. 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," published in these proceedings.

4. Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K., and Rosenfeld, R., "The SPHINX-II Speech Recognition System: An Overview," *Computer Speech and Language*, Vol. 7, No. 2, 1993, pp. 137-148.

5. Vo, M.T. and Waibel, A., "A Multimodal Human-Computer Interface: Combination of Speech and Gesture Recognition," *Adjunct Proc. InterCHI'93* (Amsterdam, The Netherlands, April 1993).

6. 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.

7. Guyon, I., Albrecht, P., LeCun Y., Denker, J., and Hubbard, W., "Design of a Neural Network Character Recognizer for a Touch Terminal," *Pattern Recognition*, Vol. 24, No. 2, 1991, pp. 105-119.

8. Vo, M.T., "Incremental Learning Using the Time Delay Neural Network," *Proc. ICASSP'94* (Adelaide. Australia, April 1994), Vol. 2, pp. 629-632.

9. Haffner, P., Franzini, M., and Waibel, A., "Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition," *Proc. ICASSP'91* (Toronto, Canada, May 1991), Vol. 1, pp. 105-108.

10. 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.

11. Manke, S. and Bodenhausen, U., "A Connectionist Recognizer for On-Line Cursive Handwriting Recognition," *Proc. ICASSP'94* (Adelaide, Australia, April 1994), Vol. 2, pp. 633-636.

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 Publishers, 1994.

13. Hunke, M., "Locating and Tracking of Human Faces with Neural Networks," Technical Report CMU-CS-94-155, Carnegie Mellon University.

14. Schiele, B. and Waibel, A., "A Connectionist Attentionfinder Based on a Face-Color-Intensifier," to be published.

15. Bub, U., Hunke, M., and Waibel, A., "Knowing Who to Listen to in Speech Recognition: Visually Guided Beamforming," to appear in *Proc. ICASSP'95*.

16. Hild, H. and Waibel, A., "Connected Letter Recognition with a Multi-State Time Delay Neural Network," *Advances in Neural Information Processing Systems 5*, Morgan Kaufmann Publishers, 1993, pp. 712-719.

17. Duchnowski, P., Meier, U., and Waibel, A., "See Me, Hear Me: Integrating Automatic Speech Recognition and Lipreading," *Proc. ICSLP'94* (Yokohama, Japan, Sept. 1994).