CONNECTIONIST MODELS IN MULTIMODAL HUMAN-COMPUTER INTERACTION

Alex Waibel *Paul Duchnowski*

Carnegie Mellon University, Pittsburgh PA, USA
University of Karlsruhe, Karlsruhe, Germany

Abstract

We present an overview of our laboratories' research on Multimodal Human-Computer Interfaces. By exploiting all available channels of human communication we aim to increase flexibility, robustness, and naturalness of human-computer interaction. The information sources we process include Speech-, Character-, and Gesture Recognition, Face- and Eye Tracking, Lipreading, and Sound Source Localization. Connectionist and hybrid techniques are used throughout.

Introduction

Recent developments in the computer and communication industries are rapidly increasing the amount and variety of information available to a wide and diverse audience. The multi-media nature of this data explosion, heralded by the concept of the "Information Superhighway", offers images, sound, text, etc. as the output presented to the information consumer. This is in stark contrast to the impoverished set of input options which are still largely limited to the keyboard and mouse. Attempts at the use of alternate modalities have mostly focused on single alternatives and are finding limited acceptance.

In an effort to improve this situation, we have begun to develop ways to process a multiplicity of signals that are believed to all carry meaning in human communication. These include: Speech Understanding, Written Character- and Gesture Recognition, Lipreading, Face-Tracking, Eye-Tracking, and Sound Source Localization. In combination, these different sources of information are known to provide humans with sometimes crucial information for effective face-to-face communication. They allow for greater robustness by taking advantage of redundant information and their availability provides flexibility and freedom to choose a suitable/convenient communication channel. Such multimodal interfaces are expected to be useful in human-to-human communication (e.g., video conferencing, speech translation), as well as human-computer interaction such as database access, document production, CAD use, machinery control, etc.

To create multimodal interfaces, we are developing technology that improves processing and interpretation of each modality, while at the same time pursuing the integration of the information sources in a single framework. Connectionist models are used throughout because of their superior performance as pattern classifiers as well as for the ease with which they can integrate heterogeneous signals and features (sensor fusion).

Separate Modality Recognition

This section concentrates on the recognition challenges and solutions specific to single modalities.

Speech Recognition

Foremost among human communication modalities, speech and language arguably carry most of the information in human communication. Automatic Speech Recognition (ASR) naturally constitutes an integral part of an advanced human-computer interface. In our laboratories several approaches toward robust high performance speech recognition are under way.

We continue to experiment with several connectionist, stochastic and hybrid approaches for Large Vocabulary Continuous Speech Recognition and spontaneous speech recognition. These include Multi-Layer Perceptrons (MLP), Time Delay Neural Networks (TDNN), Learning Vector Quantization-2 (LVQ-2) and Hidden Markov Model (HMM) techniques and combinations of these. Detailed descriptions of these systems and performance measures are reported elsewhere.[3,5,9,10]

Our modality integration experiments (see below) have employed our word-spotting system for continuous spontaneous speech.[11] Because of their small vocabulary and size, word spotters offer a practical and efficient solution for many speech recognition problems that depend on the accurate recognition of a few important keywords. The word spotting system architecture is based upon the TDNN and more recently the Multi-State TDNN (MS-TDNN).[3] The network consists of a common input layer and hidden layer, connected to a state layer and output layer for each keyword. In the state layer, each keyword is represented by a sequence of sub-word states over time. A dynamic time warping algorithm is used to find the best state sequence, from which we can hypothesize the presence or absence of a keyword when its score reaches a threshold.

Training and testing of the system was performed on two separate databases, the Roadrally corpus, and the Switchboard credit card corpus.[11] Each of these databases contains a set of 20 keywords to be spotted (including variants), embedded in extraneous speech. 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. These figures compare favorably to those of other keyword spotting systems in its class evaluated by ARPA.

More extensive word-spotting as well as topic-spotting is under development. We are making use of our continuous speech recognition and translation system JANUS.[10]

Gesture Recognition

We have been investigating pen-based gestures drawn using a stylus on a digitizing tablet. This kind of gesture is simpler to handle than hand gestures captured with a camera but still allows for rich and powerful expressions. The initial multimodal editor we developed currently uses 8 editing gestures. Some of these were inspired by standard markup symbols used by human editors. Others, such as the "delete" symbols, are what most people would automatically use when correcting written text on paper.

Using a temporal representation, a gesture is captured as a sequence of coordinates tracking the stylus as it moves over the tablet's surface. This dynamic representation was motivated by its successful use in handwritten character recognition[7] and is preferred to a static, bitmapped representation of gesture's shape. The coordinates are normalized and resampled at regular intervals to eliminate differences in size and drawing speed; from these resampled coordinates we extract local geometric information at each point, such as the direction of pen movement and the curvature of the trajectory.

Each coordinate is represented in the classifying TDNN by eight such low–level features. Their temporal sequence constitutes the input layer. Ten units in the first hidden layer extract patterns from the input, eight units in the second hidden layer spot patterns typical of a given gesture. Output units (one per gesture) integrate over time the evidence from the corresponding unit in the second hidden layer. The output unit with the highest activation level determines the classification. The network is trained on a set of manually classified gestures using a modified backpropagation algorithm. With training data of 80 samples/gesture, we have achieved "gesturer"-dependent recognition rate of 98.8% on an independent test set.

Our gesture recognizer also incorporates a method for acquiring new gestures "on the fly", i.e., while the system is in use. When a recognition error occurs, the system queries the user for the correct output and creates new template-matching hidden units that project onto the output units. If a subsequent input pattern is similar to the template used to create an extra unit, it is turned on and can influence the corresponding output unit. This technique is called an Incremental TDNN.[8]

Handwriting Recognition

The recognition of continuous handwriting, on a touch screen or digitizing tablet, has not only scientific but also considerable practical value, such as for notepad computers or for providing redundant or alternative input options in a multimodal system. The main advantage of on-line handwriting recognition is the availability of temporal information much as in gesture recognition as presented above. Handwritten words can be represented as a time-ordered sequence of coordinates with varying speed and pressure in each coordinate. As 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 MS-TDNN has been applied successfully to overcome the problem of recognizing continuous (cursive) handwriting.[7] 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) for finding an optimal alignment between strokes and characters in handwritten continuous words (see Figure 1). In the most recent experiments, we achieved 98.7%/82.0% writer-dependent/independent word recognition rates on a database of 400 handwritten words. Recognition experiments on a 20,000-word vocabulary task are in progress.

Lip-reading

It is well known that hearing-impaired listeners and those listening in adverse acoustic environments (noise, reverberation, multiple speakers) rely heavily on the visual input to disambiguate among acoustically confusable speech elements. The usefulness of lip movement information stems in large part from its rough complementariness to the acoustic signal: the former is most reliable for distinguishing the place of articulation, the latter conveys most robustly manner and voicing information. ASR systems' performance is, if anything, even more sensitive to degradation of the acoustic input. Therefore, it is only natural to try to supplement the acoustic data with lip movement information.

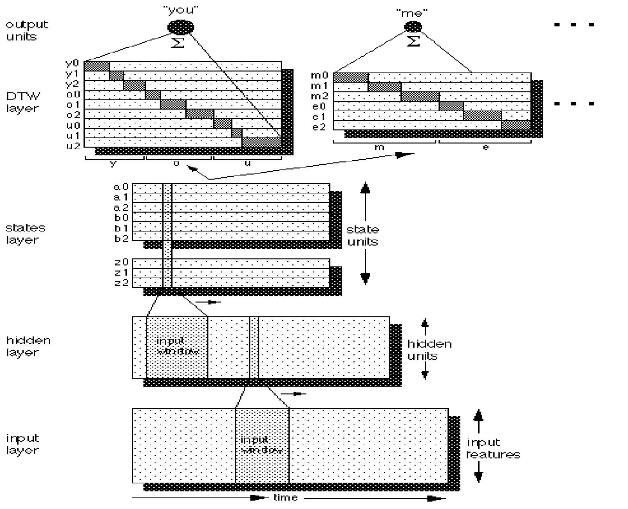

Figure 1. MS-TDNN architecture as used for handwriting recognition.

The visual evidence is obtained by "frame-grabbing" the output of a conventional camcorder camera at 30 frames/sec, with 8-bit gray level resolution. Currently, speakers are asked to position themselves such that their lips appear within an 144x80 pixel frame that is simultaneously shown on the screen of a workstation. However, no special markers, restraints or position indicators are used. The image within the frame is normalized for lighting variations and a data vector to be used by the recognition algorithm is extracted from it. Best results have been achieved with Linear Discriminant Analysis coefficients of downsampled (to 24x16) frame image.

Details of the recognition algorithm are given below in the section on combined lip-reading and speech recognition. The lip-reader alone gives 40–50% letter accuracy scores on a spelling task, performance perhaps not useful by itself but helpful in combination with ASR.

Face Tracking

The task of the face tracking system, described in detail elsewhere,[6] 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.

A conventional camcorder, mounted on a pan/tilt unit (PTU), supplies roughly 10 images per second. Color information is extracted by the Face Color Classifier (FCC). The FCC maps each pixel into a two-dimensional brightness–normalized color space and divides it 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. The network considers shape of the objects in producing the coordinates of the *virtual camera*, indicating the region actually containing the face. Appropriate commands to the PTU and zoom lens are issued if the face moves out of a pre-defined area in the center of the physical camera. Figure 2 shows an example of an image and the area classified as a face by the tracking system.

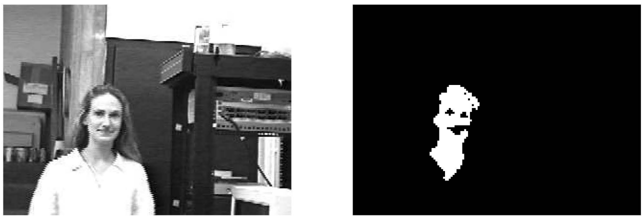

Figure 2. Camera image and extracted largest skin–colored object.

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 with 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) performing arbitrary movements 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 were less than 10% of the size of the head.

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 and as an analysis tool for human-computer interaction studies. 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.

At Carnegie Mellon we have developed a neural-network-based non-intrusive gaze tracker based on camera input only.[1] Unlike in most advanced gaze tracking, the user is required neither to wear any special equipment, nor 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 extracts a 15x30 window surrounding the reflection. The gray-scale values of the window's pixels become the input to a neural network comprising 4 hidden units and 50 output units for each of the coordinates (X and Y). Training is performed by backpropagation.

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 or head sensors typical of other approaches.

Acoustic Localization and Beamforming

For applications such as video conferencing it is desirable to allow several partipiciants of either party to move freely in a room while a system of sensors keeps track of the person of interest and enhances speech and other information modalities of this individual. This person should not be encumbered by having to carry sensors such as a close-talking microphone, etc. On the other hand, the communication/recognition systems should not be distracted by background noises or other speakers. Beamforming with a multi-microphone array is one approach to providing clean acoustic input from a single sound source.

We have constructed a one-dimensional microphone array consisting of 8 sensors spanning the half plane in front of the array. 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, uncorrelated with the signal and coming from other locations are added out of phase and attenuated. This procedure is well known as *delay and sum* beamforming. The characteristic delays for a point are determined mathematically, assuming a spherical form of speech radiation.

We conducted experiments with the JANUS[10] ASR system in a noisy environment to assess the effectiveness of the array. With a close-talking microphone, word accuracy of 85.6% was obtained, while a single microphone placed away from the speaker resulted in only 15.5%. By using the microphone array we improved this score to 79.1%.

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.

Automatic Speech Recognition and Lip-reading

Our audio-visual speech recognizer has been developed for the German spelling task mainly in the 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 the basic set-up, we record, in parallel, the acoustic speech and the corresponding series of mouth images of the speaker. Conventional pre-processing of the acoustic input gives 16 Melscale Fourier coefficients at a 10 ms frame rate. Data extraction from the visual input was described above.

A modular MS-TDNN, drawing on a pure acoustic spelling recognizer,[4] performs the recognition. Figure 3 is a schematic of the 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 decodes the recognized letter sequence. The weights in the parallel networks are trained by backpropagation. There are 15 hidden units in both sub-nets. 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. Initial results suggesting an advantage of hidden layer combination as well as a more complete description of the system can be found elsewhere.[2]

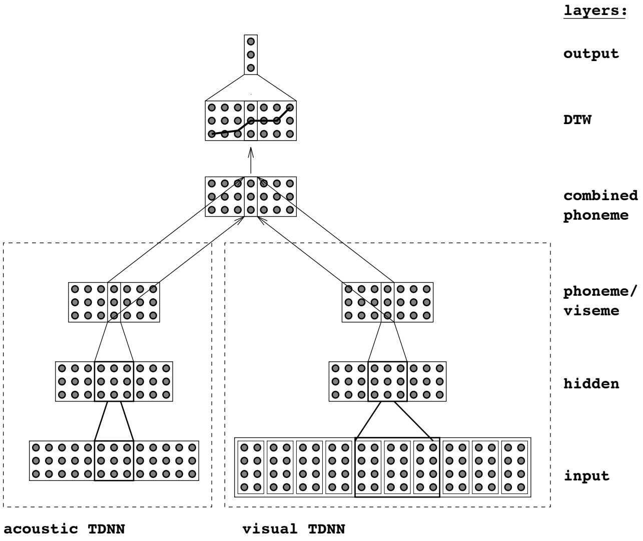

Figure 3. Basic recognition network architecture (integration at the phoneme/viseme level).

We have tested the recognizer on data sets of 200 letter sequences from single speakers. On the average, LDA-preprocessed visual input produces best results, reducing the audio-alone error rate by 33.7 %.

Speech and Gesture Recognition

We have developed a speech- and gesture-based text editor as another step towards modality integration. The word spotter (see above) 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 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. For example, 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.

Face Tracking and Beamforming

The beamformer described earlier picks its target as the loudest source in its vicinity. It, therefore, encounters problems while attempting to track a moving talker in realistic communication situations including competing speakers. Considering visual aspects to locate the speaker's position overcomes these limitations. Specifically, the face-tracker supplies the coordinates of a moving speaker to the microphone array which then forms a beam to that location. Our experiments have confirmed this synergy, demonstrating improved signal-to-noise ratio even for speakers moving in an environment with another loud sound source.

A natural further application of face-tracking and beamforming is to enhance the lip-reading/speech recognition system. The face-tracker allows for a non-invasive acquisition of the visual data, while the beamformer improves the quality of the received audio input. Work on such a complete system is already in progress.

Conclusion

We have described the many-faceted applications of neural networks to recognition of various human communication modalities. We are continuing to improve the individual systems while pursuing the goal of their integration in a single, flexible, and robust human-computer interface.

Acknowledgements

The authors gratefully acknowledge the support of the Advanced Research Projects Agency, Microtechnology Office that made this research possible.

Thanks to Udo Bub, Martin Hunke, Stefan Manke, Uwe Meier, Dean Pomerleau, Ivica Rogina, Minh Tue Vo, and Torsten Zeppenfeld for their work on each of the modules described here and for various assistance in preparation of this manuscript.

References

[1] S. Baluja and D. Pomerleau (1993). Non-Intrusive Gaze Tracking Using Artificial Neural Networks, in *Neural Information Processing Systems* (NIPS-6), Morgan Kaufmann.

[2] P. Duchnowski, U. Meier, and A. Waibel (1994). See Me, Hear Me: Integrating Automatic Speech Recognition and Lipreading. to appear in *Proc. ICSLP 94.*

[3] P. Haffner, M. Franzini, and A. Waibel (1991). Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition, *Proc. ICASSP'91.*

[4] H. Hild and A. Waibel (1993). Connected Letter Recognition with a Multi-State Time Delay Neural Network. in *Neural Information Processing Systems* (NIPS-5), Morgan Kaufmann.

[5] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee, and R. Rosenfeld (1993). The SPHINX-II Speech Recognition System: An Overview, *Computer Speech and Language,* 7, pp.137–148.

[6] H.M. Hunke (1994). Locating and Tracking of Human Faces with Neural Networks. Technical Report CMU–CS–94–155, Carnegie Mellon Univ.

[7] S. Manke and U. Bodenhausen (1994). A Connectionist Recognizer for On-Line Cursive Handwriting Recognition, *Proc. ICASSP'94.*

[8] M.T. Vo (1994). Incremental Learning Using the Time Delay Neural Network, *Proc. ICASSP'94.*

[9] A. Waibel, A. Jain, A. McNair, H. Saito, A. Hauptmann, and J. Tebelskis (1991). JANUS: a Speech-to-speech Translation System Using Connectionist and Symbolic Processing Strategies, *Proc. ICASSP'91.*

[10] M. Woszczyna et al (1994). JANUS 93: Towards Spontaneous Speech Translation, *Proc. ICASSP'94.*

[11] T. Zeppenfeld, R. Houghton, and A. Waibel (1993). Improving the MS-TDNN for Word Spotting, *Proc. ICASSP'93.*