

CONNECTIONIST MODELS IN MULTIMODAL HUMAN-COMPUTER INTERACTION

Alex Waibel

Paul Duchnowski

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

Abstract

Speech recognition are un-
h several connectionist
Large Vocabulary Con-
eous speech recog-
neurons (MP),
g Vector
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 hu-
man-computer interaction. The information sources we pro-
cess include Speech-, Character-, and Gesture Recognition,
Face- and Eye Tracking, Lipreading, and Sound Source Lo-
calization. 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
upt of the "Information Superhighway", offers images,
text, etc. as the output presented to the informa-
consumer. This is in stark contrast to the impover-
of input options which are still largely limited to
keyboard and mouse. Attempts at the use of alternate
have mostly focused on single alternatives and
limited acceptance.

to improve this situation, we have begun to
process a multiplicity of signals that are
carry meaning in human communication.

Understanding, Written Character-

Lipreading, Face-Tracking, Eye-

ocalization. In combination,

rnation are known to pro-

l information for effec-

they allow for greater

undant information

ty and freedom to

n channel. Such

ful in human-

g, speech

such as

ery

ch-

n

...e a system of network. The network consists of... shape of the objects in pro-
 and enhanced using the... of the... camera, indicating the
 ... individual region... of the face. Appropriate commands
 ... to carry the... and... issued if the face moves out of
 ... etc. On the... and... of the physical camera. Fig-
 ... shows... extracted image and the area classified
 ... by the... system
 ... approach to providing clean
 ... source.

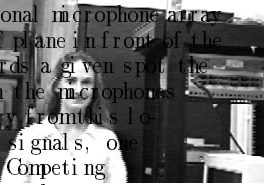


Figure 2. Camera image and extracted largest skin-colored object.

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 be used to determine the user's focus of attention in a graphical interface; for instance, knowing whether the user is looking at the screen or somewhere else while talking is important in deciding whether automated speech should be activated.

Now we have developed a neural-network-based gaze tracker based on camera input. In an advanced gaze tracking, the user is not required to use any special equipment, nor to keep the system close to the camera. The system comes from a camera monitor. An infrared light source is directed at the eye. The gaze direction is determined by the relative positions of the eye and the light source. The system extracts the eye position from the gray-scale image and sends the position as input to a neural network. The output units for the neural network are performed by...

...y we

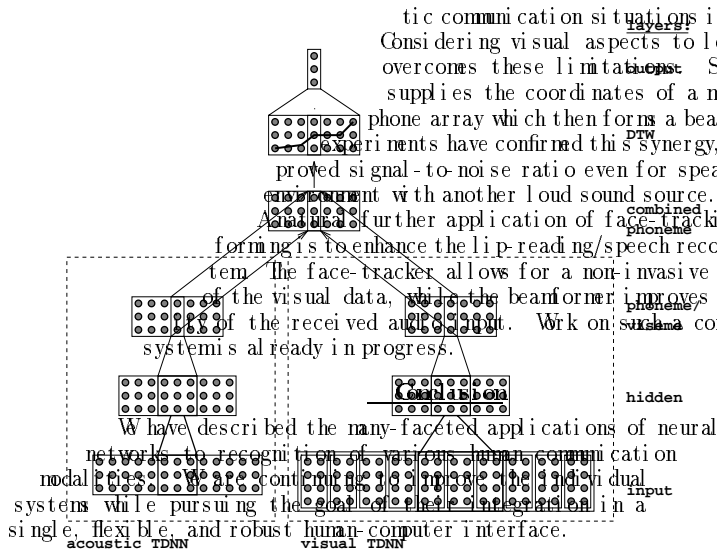


Figure 3. Basic recognition network architecture (integration of the acoustic and visual input).

The authors gratefully acknowledge the support of the Advanced Research Projects Agency, Microtechnology Office and the Defense Research Agency, Microtechnology Office. This research was made possible by the support of the Defense Research Agency, Microtechnology Office. 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-visual error rate by 33.7%.

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-visual error rate by 33.7%.

Speech and Gesture Recognition

References

Bel (1993). Non-Intrusive Gaze Detection Using Neural Networks, in *Neural Networks in Signal Processing* (NIPS-6), Morgan Kaufmann, San Mateo, CA.

Bel (1994). See *Speech Recognition: A Practical Approach*, P.94. Integ.

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 are 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 required.

Tracking and Beamforming

The beamformer described earlier picks its target as the speaker in its vicinity. It, therefore, encounters problems when trying 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.

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, which the beamformer improves the quality of the received audio input. Work on such a complete system is already in progress.

We have described the many-faceted applications of neural networks to recognition of acoustic and visual communication modalities.

We have described the many-faceted applications of neural networks to recognition of acoustic and visual communication modalities while pursuing the goal of developing a single, flexible, and robust human-computer interface.