TOWARD MOVEMENT-INVARIANT AUTOMATIC LIP-READING AND SPEECH RECOGNITION

Paul Duchnowski[1] *Martin Hunke*[1,2] *Dietrich Büsching*[1] *Uwe Meier*[1] *Alex Waibel*[1,2]

Interactive Systems Laboratories
[1]University of Karlsruhe, Karlsruhe, Germany
[2]Carnegie Mellon University, Pittsburgh PA, USA

ABSTRACT

We present the development of a modular system for flexible human–computer interaction via speech. The speech recognition component integrates acoustic and visual information (automatic lip-reading) improving overall recognition, especially in noisy environments. The image of the lips, constituting the visual input, is automatically extracted from the camera picture of the speaker's face by the lip locator module. Finally, the speaker's face is automatically acquired and followed by the face tracker sub-system. Integration of the three functions results in the first bi-modal speech recognizer allowing the speaker reasonable freedom of movement within a possibly noisy room while continuing to communicate with the computer via voice. Compared to audio-alone recognition, the combined system achieves a 20 to 50 percent error rate reduction for various signal/noise conditions.

1. INTRODUCTION

An obvious, if non-trivial, way to enhance, simplify, and popularize human-computer interaction is by taking advantage of all the modalities normally used by people in everyday interactions. Departure from the keyboard as the primary input modality should encompass integration of such information sources as speech, lip movement, handwriting, gaze direction, gesture, facial expression, etc. A survey of multiple projects in these areas undertaken in our labs at the University of Karlsruhe and Carnegie Mellon University can be found in [11].

This paper focuses on the development of a speech recognition system incorporating automatic lip-reading while allowing the user reasonable freedom of movement within a room. Lip-reading plays an important role in communication by the hearing-impaired and by individuals listening in difficult acoustic environments [10]. Several studies have demonstrated the utility of augmenting automatic speech recognition (ASR) with visual information (eg. [6, 7, 9]). Our own work in this area has been previously reported in [1, 3]. However, a major limitation of virtually all the systems was the method by which visual data was acquired. This included such invasive techniques as head-mounted cameras, reflective markers placed on the speaker's lips, and manual extraction of relevant face image sections, effectively precluding practical applications. In our system as described in [3] the process was continuous, automatic, and without special markers but required the speaker to position himself such that his lips appeared within a window shown on a workstation screen.

The goal of present research is to free the user from all such interference. A face-tracking algorithm automatically controls the position and focus of the camera to maintain the view of a speaker's face. The lip-finder module locates the lips within the face image and provides the coordinates of the mouth corners to the lip-reading/speech-recognition subsystem which extracts the relevant information from the image and combines it with the acoustic input to recognize the utterance. In this paper we present an overview of each of these three components, their combination within the overall scheme, and the performance of the integrated system.

2. FACE TRACKING

The task of the face tracking system, described in detail in [5], is to support the lip-locating/reading system with a stable image of the speaker's face. The face-tracker can locate faces in arbitrary environments. 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. The system's output also includes the position and size values of the observed face, so that the lip-reading system independently can grab the same camera image in higher resolution and faster frame rate. The face coordinates aid the the lip-locator in isolating the relevant part of the image.

2.1. System Structure

The system has two main modes: locating an arbitrary face and tracking the located face. A conventional camcorder, mounted on a computer-steerable pan/tilt unit (PTU), supplies roughly 10 images per second. Color information is extracted by the Face Color Classifier (FCC) and movement is computed from successive frames. These data are merged and 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.

2.2. Features for Classification

Though extremely helpful, using color for locating human faces presents several problems. Color values of an object vary with the camera, framegrabber, and illumination. The color composition of human skin differs surprisingly little across individuals, but total intensity of the reflection varies over a wide range. The color dependencies can be largely resolved by the FCC, which groups different hues as skin- or non-skin-color. Brightness dependencies are eliminated by dividing each of the three color values by their sum. This maps each pixel into a two-dimensional normalized color space which the FCC divides 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 which is smoothed to be representive of arbitrary faces. During the locating phase the system finds the image regions whose color corresponds to the skin color distribution. An example of the resulting pixel assignment is shown in Figure 1.

Two other features are used to confirm the face location. Motion, computed as the difference between two successive frames, is helpful in avoiding such distractions as faces in pictures hanging on a wall. The shape of moving objects containing skin-like color is eventually used to classify an object as a face, to eliminate arms, hands, etc.

Figure 1. Example of the output of the Face Color Classifier. Dark areas are classified as skin-colored.

2.3. Tracking

Tracking a known face amounts to locating a face while taking advantage of what is already known about the subject. The virtual camera image is searched first, since that was the last location of the relevant face. The FCC is adjusted to the color of the face being tracked. By repeating this adjustment the system can automatically adapt to changing ambient and recording conditions. If detected, motion is used to obtain additional clues about the object's shape. The shape of objects having the same color distribution as the located face is considered by a neural network to determine the current position and size of the face, so that the camera and zoom lens can be adjusted.

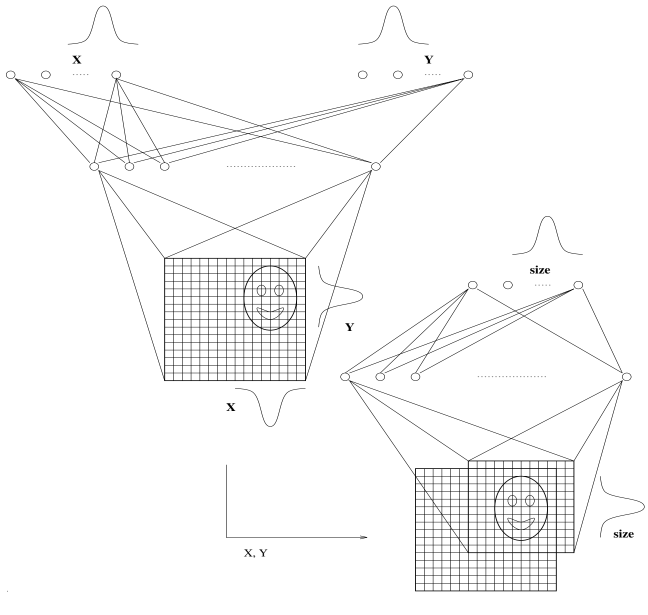

Figure 2. Neural network arrangement for face centering and size estimation.

Two neural networks are used for finding the face coordinates. Their structure is similar to that used in [8] and is shown schematically in Figure 2. The input retina receives data from the motion and color analysis and determines the position of the face with the first network. The second network uses a centered area about the located face as input to estimate the size of the face. The seperation into two networks allows the size estimation to be reduced to centered faces. The nets 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 in front of different backgrounds. All subjects were asked to perform arbitrary movements in front of the camera. 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.

3. LIP LOCALIZATION

The task of the lip-locator is finding the corners of the mouth in the image grabbed from the camera as centered and zoomed by the face tracker. Additional requirements include operation without special lighting and with possibly cluttered background. The detection should also function independently of the mouth shape. Images of 256×256 pixels, at 30 frames/sec and at 8 bit grayscale resolution were used as raw input.

Initial experiments indicated that manual design of a robust lip feature detector was not feasible. Also, using other features such as the outline of the face and the relative location of the eyes was considered advantageous in pinpointing the lips. Accordingly, a system consisting of two neural networks was designed [2]. The first network gives a coarse estimate of the position of the mouth. The second locates the two corners of the mouth within a window around the position that was estimated by the first network. Constraining the exact search to a more confined area speeds up the total localization process.

The network used for initial position estimation is shown in Figure 3. The size (and thus the resolution) of the input image used here is first reduced to 32×32 pixels and edges are found using the Sobel operator. Two directional edgemaps are extracted and normalized. They constitute the input into the three-layer locator network.

In order to reduce the amount of computation the hidden units have restricted "receptive fields". The input grid is divided into 16 8×8 nonoverlapping fields. Each field is connected with the same number of distinct hidden units (currently 8). The output units form a 9×11 grid with activation 1.0 indicating the center of the mouth at that location (zero otherwise). During training 30 pictures of 10 different faces were presented to the network in several positions and in 8 different sizes (face always fully visible, scaling and shifting done artificially).

The architecture of the second network for the detection of the corners of the mouth is similar. A window centered on the position estimated by the first network is extracted from the image and reduced to 60×36 pixels. Only the horizontal edge map is computed. Again, receptive fields project onto the hidden layer which in turn connects to 874 output units arranged in a rectangular grid covering every second row and column. Training was performed with translated and scaled versions of a small number of examples (50 images from 10 subjects for the network evaluated below).

For testing, 212 pictures of 10 persons whose pictures where not used during training were processed by the two networks. The position of the center of the mouth that was estimated by the first

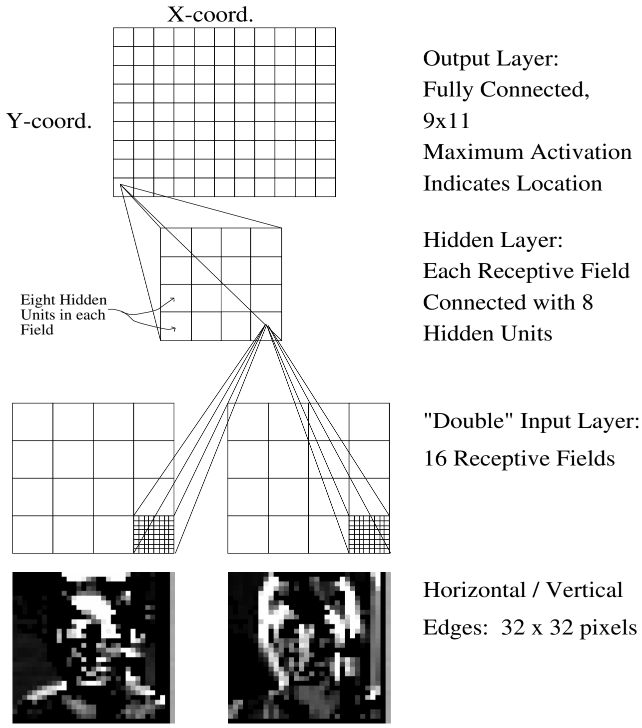

Figure 3. Neural network architecture for approximate lip position estimation.

Error (pixels)	Percent of Images direction from corner		
	horizontal	vertical	total
0	14.9	11.4	2.5
1-2	43.3	40.8	29.3
3-4	19.9	23.4	29.8
5-10	20.5	22.0	33.9
> 10	1.4	2.4	4.5

Table 1. Error distribution for localization of the left corner of the mouth. Average mouth width was 42 pixels.

network differed from the observed real position with a mean distance of 17.7 pixels (mean horizontal: 6.7, mean vertical: 14.7). The typical lip width in the images was 42 pixels.

The mean error distance for the second network was 6.4 pixels for both corners of the mouth. 11 of the 212 pictures were then identified as outliers (no valid estimation could be made). The remaining 201 images showed mean errors of 4.8 and 5.5 pixels for left and right corner, respectively. Table 1 gives a more detailed breakdown of error distribution for the left corner.

The lip-locator passes the grabbed image along with the mouth position estimate to the automatic lip-reading/speech recognition subsystem.

4. 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.

4.1. System Description

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 speech waveform produces 16 Melscale Fourier coefficients at a 10 ms frame rate as the acoustic input to the recognizer.

The visual evidence is obtained from the image already "grabbed" by the lip-finder. We fine-tune the estimate of lip position delivered by the lip locator by finding the maximum normalized cross-correlation between images in temporally adjacent frames[1]. This ensures the stability of the lip image *sequence*. Similarly, the very first frame of a sequence, where the speaker's lips are assumed closed, is correlated with a stored mouth template to compute a uniform scaling factor. Given the final scale and coordinates the mouth image is centered within a 144×80 pixel frame such that the width of closed lips occupies roughly 2/3 of the frame width. To remove dependence on varying lighting, including illumination gradients, we use adaptive histogram modification. Together, these pre-processing steps normalize the lip images for location, size, and brightness.

A data vector that is used by the recognition algorithm is then extracted. We have investigated several representations of the visual data: 1) Direct gray-level values of low-pass filtered and downsampled (to 24×16 pixels) images, 2) Band-pass Fourier magnitude coefficients (averaged in rings in the frequency domain), 3) Principal Components of the downsampled image, 4) Linear Discriminant Analysis coefficients of the downsampled image. Representations 2–4 were chosen with the goal of preserving the relevant information in the lip image while substantially decreasing the parameter count.

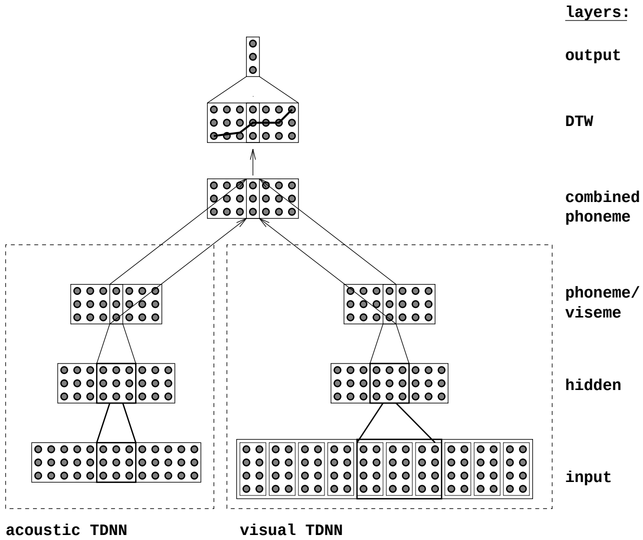

Figure 4. Basic recognition network architecture (integration at the phoneme/viseme level).

In the basic system a modular Multi-State Time Delay Neural Network (MS-TDNN) [4] performs the recognition. Figure 4

[1]Since the lip locator is accurate to within a few pixels, the cross-correlation has to be computed for only a small number of shifts of the second image about the position indicated by the lip finder.

Acoustic Environment (dB SNR)	Word Accuracy (%)		
	Acoustic alone	Acoustic + Visual parameters	
		Gray Levels (30% alone)	LDA (53% alone)
Quiet	97.0	97.6	97.6
White Noise (16)	67.3	75.2	78.2
White Noise (8)	42.4	49.1	59.4
Motor Noise (25)	95.8	95.8	97.0
Motor Noise (16)	49.1	52.1	61.8
Radio (16)	87.3	90.3	93.9
Radio (10)	58.2	72.1	80.0

Table 2. Speaker-dependent recognition rates for purely acoustic and combined input shown for several different sources of background disturbance and visual representations.

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[2] 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 (so called "entropy weights", see [1]) 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 can be found in [3]. This approach is possibly more reflective of the way humans integrate audio-visual input [10] but complicates somewhat the training process. We are also currently experimenting with guiding the combination by an explicit estimate of the acoustic signal-to-noise ratio (SNR).

4.2. Results

In experiments with an earlier version of the automatic lip-reader [3] we established that the gray-level and LDA image representations deliver generally best results. Therefore, we have concentrated on these parameters for the movement-invariant system.

We have trained the recognizer on visual/acoustic data from 200/1500 letter sequences from a single speaker and tested on 30 sequences. Table 2 gives results in terms of word accuracies (reflecting substitutions, deletions, and insertions) for the two tested visual representations. As indicated at the top, the accuracy when *visual input alone* was used was 30 and 53 percent for gray levels and LDA respectively. Combined recognitions scores are further shown for a variety of audio conditions: in quiet and for different noise sources and intensities.

The results indicate a substantial improvement in recognition when acoustic input is augmented with automatic lip-reading. The LDA representation of the visual information is again seen as uniformly superior to the direct gray level values. In addition, LDA allows a factor of 12 reduction in data rate, requiring only 32 parameters per frame vs. 384 gray levels. For LDA the error rate reduction over audio-alone scores ranges from 20 to over 50 percent, all achieved with completely automatic acquisition of the visual data. This magnitude of improvement is comparable to that achieved by the earlier system which required a highly cooperative speaker.

[2] A viseme, the rough visual correlate of a phoneme, is the smallest visually distinguishable unit of speech.

5. CONCLUSION

We have presented the components of a lip-reading/speech recognition system that *non-invasively* and automatically captures the required visual information. The system which comprises them performs automatic lip-reading in realistic situations where lip motion information enhances speech recognition under both favorable and acoustically noisy conditions. Simultaneously, the speaker is allowed a reasonable freedom of movement within a room, with no need to position himself in any particular location. We are proceeding to investigate the system for speaker-independent tasks and plan to adapt it to large-vocabulary continuous speech recognition tasks. Eventually it will also be integrated with other components of a complete multimodal interface.

ACKNOWLEDGEMENTS

This work is sponsored by the state of Baden-Württemberg, Germany (Landesschwerpunkt Neuroinformatik) and by the Advanced Research Projects Agency (USA). The views and conclusions stated in this paper are those of the authors.

REFERENCES

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

[2] D. Büsching. Automatische Lokalisierung der Lippenregion in Videobildern von Gesichtern. Masters Thesis, Fakultät für Informatik, Universität Karlsruhe, 1994.

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

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

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

[6] K. Mase and A. Pentland. Automatic Lipreading by Optical-Flow Analysis. *Systems and Computers in Japan*, 22(6), 1991, pp. 67–76.

[7] E.D. Petajan. Automatic lipreading to enhance speech recognition. in *Proc. IEEE Communications Society Global Telecom. Conf.*, Atlanta GA, Nov. 1984.

[8] D.A. Pomerleau. Neural Network Perception for Mobile Robot Guidance. Technical Report CMU–CS–92–115, Carnegie Mellon Univ., 1992.

[9] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. in *Proc. IJCNN'92.*

[10] Q. Summerfield. Audio-visual Speech Perception, Lipreading, and Artificial Stimulation. in *Hearing Science and Hearing Disorders*, M.E. Lutman and M.P. Haggard eds., New York: Academic Press, 1983.

[11] A. Waibel, M.T. Vo, P. Duchnowski, and S. Manke. Multimodal Interfaces. to appear in *Artificial Intelligence Review Journal*, special issue, 1994.