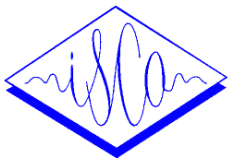


ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

PREPROCESSING OF VISUAL SPEECH UNDER REAL WORLD CONDITIONS

Uwe Meier, Rainer Stiefelhagen, Jie Yang

uwe@ira.uka.de, stiefel@ira.uka.de, yang+@cs.cmu.edu

Interactive Systems Laboratories
University of Karlsruhe, Karlsruhe, Germany
Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

In this paper we present recent work on integration of visual information (automatic lip-reading) with acoustic speech for better overall speech recognition. We have developed a modular system for flexible human-computer interaction via speech. In order to give the speaker reasonable freedom of movement within a room, the speaker's face is automatically acquired and followed by a face tracker subsystem, which delivers constant size, centered images of the face in real time. The image of the lips is automatically extracted from the camera image of the speaker's face by the lip tracker module, which can track the lips in real time. Furthermore, we show how the system deals with problems in real environments such as different illuminations and image sizes, and how the system adapts automatically to different noise conditions.

1. INTRODUCTION

Most approches 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 simltaneously (coctail party effect). Humans deal with these distortions by considering additional sources such as directional, contextual, and visual informations, primarily lip movements. We are interested in emulating some of these capabilities by combining speech recognition with lipreading to improve robustness and flexebiltie by offering complementary information. Publications of other researchers in this area can be found in [1, 2, 3, 4, 5, 6, 7]

2. SYSTEM DESCRIPTION

Our audio-visual speech recognizer [8, 9, 10, 11] has been developed for a German spelling task mainly in speaker-dependent mode, first multi speaker / speaker independent tests show promising results. 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 speaker's face is automatically acquired and followed by the face tracker subsystem, 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 speaker's face by the lip localisation module [12, 13].

The visual data is preprocessed to eliminate problems that raise in real world applications like different illuminations, different face/lip size and different positions in the image. It was shown that the recognitions results decrease if those conditions change within a small range [14]: Experiments with an 100% word accuracy test set have shown that if you shift the image 3 pixels, the word accuracy decreases down to 68%. Changing the illumination by adding a offset of 15 to the grayvalue, the word accuracy is only 93%. Resizing the image with a factor of 1.1 results in 45% word accuracy.

Figure 1 gives an overview on the subsystems of our Lipreading system. We use a Canon VC-C1 Camera with integrated pan-tilt unit. This unit is controled by the Facetracker. The Facetracker sends the postion of the face to the lip localisation module, which stores the positon of the mouth corner for every frame. Tracking of the face and the lip corners is done in realtime during the recording of the data. After that some visual fine-tuning is done to eleminate those online problems in real world applications described above. Then that the data is feeded in a MS-TDNN recognizer. All those submodules are descibed in more detail in the following sections.

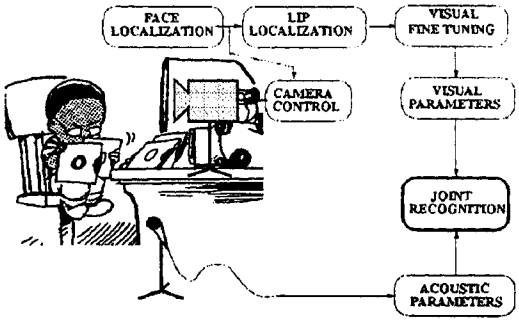


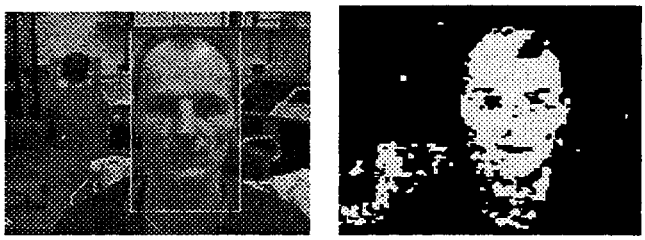
Figure 1. NLIPS - system overview

2.1. Database

We record, in parallel, the acoustic speech and the correspondig series of mouth images of the speaker. The acoustic signal is sampled at 16kHz. The visual evidence is obtained by 'frame-grabbing' the output of a Canon VC-C1 Camera at 20-30 frames/sec with 24-bit RGB resolution. The color images are used for the Face Tracker and Lip Finder, for the Lip Reading modul gray-level images are used.

2.2. Face Tracking Module

To find and track the face, a statistical skin color-model consisting of a two-dimensional Gaussian distribution of normalized skin colors is used [15]. The input image is searched for pixels with skin colors and the largest connected region of skin-colored pixels in the camera-image is considered as the region of the face. The color-distribution is initialized so as to find a variety of skin-colors and is gradually adapted to the actual found face.



input image (color!) face-colored regions

Figure 2. Application of the color model to a sample input image. The face is marked in the input image

2.3. Lip Tracking Module

In our lipreading system as described in [9], we used a neural net based lip-localization module to find the lip-corners. One disadvantage of this module was, that it did not perform in real time – the facial region had to be stored and lip-localization and extraction had to be done off-line which slowed down the system significantly.

In our current system we are using a new feature based lip-tracking module which is able to find and track lip-corners in real time. Moreover, the module is able to detect lip localization failures and to automatically recover from failures. In the new lip localization approach, instead of tracking only the lip corners, we also track other facial features such as pupils and nostrils along with them. Tracking all these facial features and using a simple 3D head model, e.g. we know the relative positions of each of the used facial features, outliers in the set of found feature points can be detected and their true positions can be predicted. The new lip-localization module is described in detail in [12].

2.4. Preprocessing

Size and Translation Invariance From the Lip Finder modul we get the coordinates of the mouth corners. Using these corners we can cut the lips out of the face image and rescale to a constant size. Because the Lip Tracking is good enough, no further preprocessing is needed to get constant size and position of the lips in the lip-sequence.

Illumination Invariance Figure 3 shows the distribution of the grayvalues in images with 'normal' lightning. If you adjust the grayvalues of your given images in that way, that they have the grayvalue distribution of the target images you can eliminate most illuminations differences. Figure 4 gives an example of images in our database, figure 5 shows the distributions before and after the grayvalue modification.

We used for the adjustement a method (grayvalue modification) that is described in [16] in detail: The grayvalue distribution is computed, using the accumulated grayvalues, it is easy to adjust the grayvalues in a way, that the accumulated function is the same as from the target function:

$$f'(p) = T(f(p))$$

where $f(p)$ is the original grayvalue, T the modification function and $f'(p)$ the new grayvalue.

One problem with this method is, that it has no effect on side illumination. We solved this problem by developing a adaptive grayvalue modification: The image is divided in 4 parts Q_k (figure 6). Now we can compute the grayvalue modification T_1, T_2, T_3 and T_4 for each part seperate. The adaptive grayvalue modification is a linear combination of these grayvalue modifictions:

$$T(f(p)) = \sum_{i+1}^{4} w_i T_i(f(p))$$

To compute the w_i each of the 4 parts is seperated again in 4 parts (q_{ij}). There are 3 kinds of neighbourhood (Region A, B and C in figure 6): q_{ij} has no, one or three Q_k neighbours. On the example of the points P_1, P_2 and P_3 in figure 6 we show how to compute the transformation:

$$\begin{aligned}
T(P_1) &= \frac{y_u}{y_o + y_u}\left(\frac{x_r}{x_l + x_r}T_1(P_1) + \frac{x_l}{x_l + x_r}T_2(P_1)\right) \\
&+ \frac{y_o}{y_o + y_u}\left(\frac{x_r}{x_l + x_r}T_3(P_1) + \frac{x_l}{x_l + x_r}T_4(P_1)\right) \\
T(P_2) &= \frac{y_u}{y_o + y_u}T_1(P_2) + \frac{y_o}{y_o + y_u}T_3(P_2) \\
T(P_3) &= T_3(P_3)
\end{aligned}$$

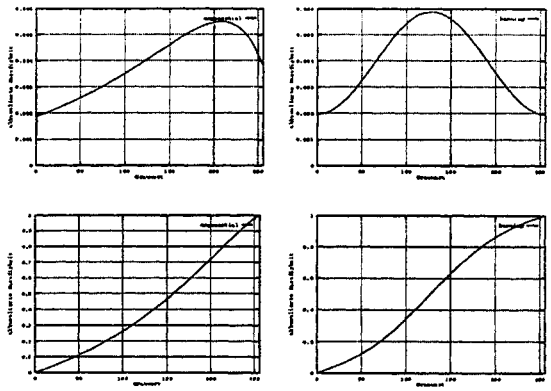


Figure 3. grayvalue modification: target distributions

2.5. The Recognizer

A modular MS-TDNN is used to perform the recognition. Combining visual and acoustic data on the phonetic level has shown to give the best results. As visual input we use gray-level images (24x16 pixel = 384 Parameter) or Linear Discriminant Analysis (16 Parameters). For the acoustic preprocessing 16 Melscale coefficients are used. The architecure of the MS-TDNN with combining on the phonetic layer is shown in figure 8. The acoustic and visual TDNN are trained seperately. For the visual TDNN a visem set, that is a visual subset of our phoneset, was developed. The combining of the phoneme- and visems- activations (hyp_A and hyp_V) is done by a weighted sum:

$$hyp_{AV} = \lambda_V hyp_V + \lambda_A hyp_A$$

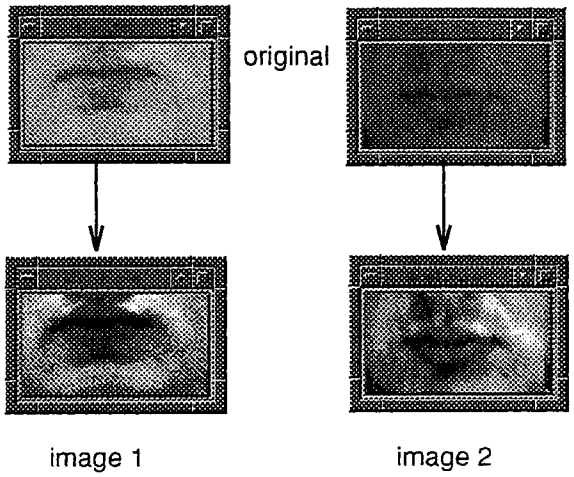


Figure 4. grayvalue modification: example

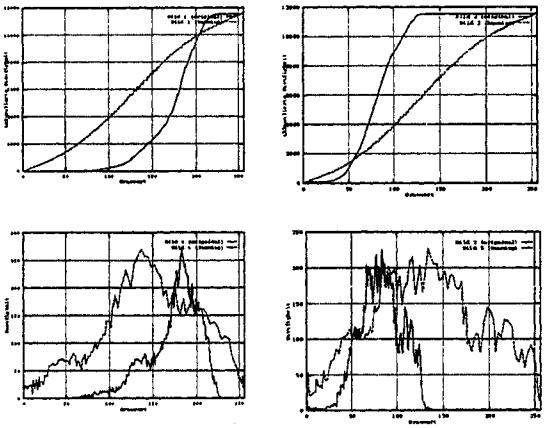


Figure 5. grayvalue modification: example histograms

$$\lambda_A = b + \frac{S_V - S_A}{\Delta S_{max-over-data}}, \text{ and } \lambda_V = 1 - \lambda_A$$

The entropy quantities S_A and S_V are computed for the acoustic and visual activations by normalizing these to sum to one (over all phonemes or visemes, respectively) and treating them as probabilities mass functions. High entropy is found when activations are evenly spread over the units which indicates high ambiguity of the decision from that particular modality. The bias b pre-skews the weight to favor one of the modalities. This bias is set depending on the signal-to-noise-ration (SNR). The quality of the speech data is generally well described by the SNR. Higher SNR means higher quality of the acoustic data and therefore the consideration of the acoustic side should increase for higher and decrease for smaller SNR-values. We used a piecewise-linear mapping to adjust bias b as a function of the SNR.

3. PERFORMANCE

We have trained a speaker dependend recognizer on 170 sequences of acoustic/visual data from one speaker and tested on 30 sequences of the same person. For testing we also added white noise to the test-set. The results are shown in table 1, as performance measure word accuracy is used (where a spelled letter is considered a word):

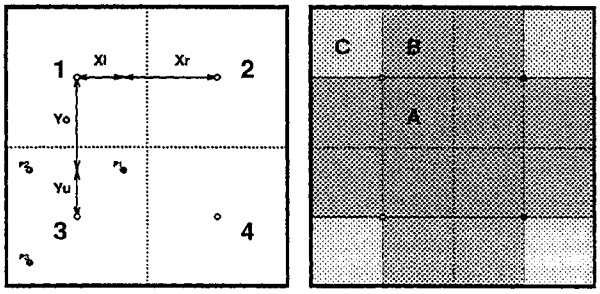


Figure 6. adaptive grayvalue modification

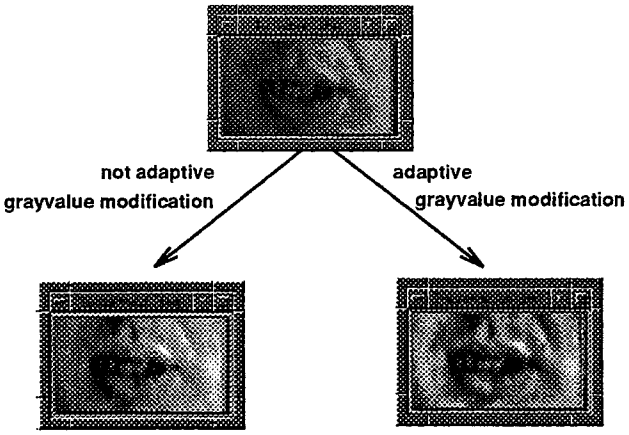


Figure 7. adaptive grayvalue modification, example

$$WA + 100\%(1 - \frac{\#SubError + \#InsError + \#DelError}{\#Letter}$$

We have also started to train a speaker independent system. The database used for this experiments has only 7 speakers (3 female, 4 male, 100 utterances per speaker). Therfore we have used the round robbin method: we have trained on 6 speakers and tested on 1 new speaker, this was done for all permutations of the 7 speakers, table 3 shows the visual word accuracy of these experements. The speaker dependend results for these speakers are shown in table 2.

Testset	visual only	acoustic only	combined
clean	55%	98.4%	99.5%
16 dB SNR	55%	56.9%	73.4%
8 dB SNR	55%	36.2%	66.5%

Table 1. speaker dependent results

4. CONCLUSION

We have presented the componentes 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. The System adapts automatically to different noise and lightning conditions.

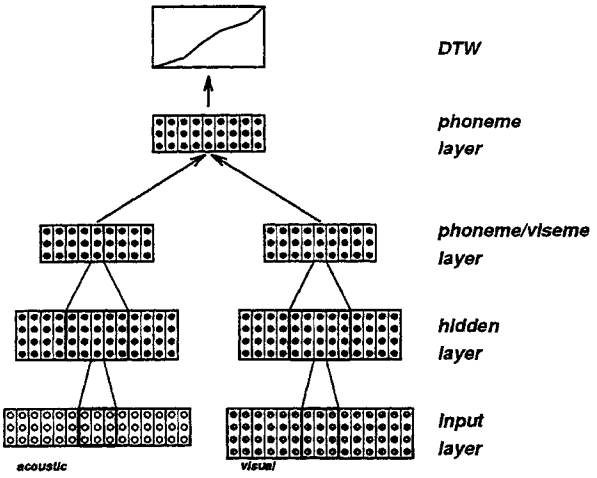


Figure 8. audio visual MS-TDNN

Training/Test	visual
speaker 1	11.7%
speaker 2	16.7%
speaker 3	11.1%
speaker 4	7.6%
speaker 5	12.9%
speaker 6	16.1%
speaker 7	25.3%
average	16.3%

Table 2. speaker dependent results

Combining the acoustic and visual signal, we get an error reduction up to 68%.

5. ACKNOLEDGEMENTS

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

REFERENCES

[1] A.J. Goldschen, O.N. Garcia, and E. Petajan. Continuous optical automatic speech recognition by lipreading. *28th Annual Asimolar conference on Signal speech and Computers.*

[2] P.L. Silsbee. Sensory integration in audiovisual automatic speech recognition. *28th Annual Asimolar conference on Signal speech and Computers*, 1994.

Test	visual
speaker 1	7.1%
speaker 2	45.9%
speaker 3	47.3%
speaker 4	37.9%
speaker 5	10.5%
speaker 6	35.4%
speaker 7	9.2%
average	27.6%

Table 3. speaker independent results (round robbin training/testing, 1 speaker testset 6 speaker trainingsset

[3] M.M. Cohen and D.W. Massaro. What can visual speech synthesis tell visual speech recognition. *28th Annual Asimolar conference on Signal speech and Computers.*

[4] J. R. Movellan. Visual speech recognition with stochastic networks. *NIPS 94*, 1994.

[5] K. Mase and A. Pentland. Automantic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76, 1991.

[6] B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, pages 65–71, November 1989.

[7] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. *IJCNN*, June 1992.

[8] P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: Integrating automatic speech recognition and lipreading. *International Conference on Spoken Language Processing, ICSLP*, pages 547–550, 1994.

[9] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel. Toward movement-invariant automatic lipreading and speech recognition. *Proc. ICASSP*, 1995.

[10] U. Meier, W. Hürst, and P. Duchnowski. Adaptive bimodal sensor fusion forautomatic speechreading.

[11] M.T. Vo, R. Houghton, J. Yang, U. Bub, U. Meier, A. Waibel, and P. Duchnowski. Multimodal learning interfaces. *ARPA Spoken Language Technology Workshop*, 1995.

[12] R. Stiefelhagen, Jie Yang, and Uwe Meier. Real time lip tracking for lipreading. *Eurospeech 97.*

[13] R. Stiefelhagen and Jie Yang. Gaze tracking for multimodal human-computer interaction. *ICASSP 97.*

[14] U. Meier. Robuste Systemarchitekturen für automatisches Lippenlesen. Diplom-Arbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[15] J. Yang and A. Waibel. a real-time face tracker. *WACV 96.*

[16] W.K. Pratt. *Digital Image Processing.* A Wiley-Interscience Publication.