Auditory-Visual Speech Processing
(AVSP'99)
Santa Cruz, CA, USA
August 7-10, 1999

AUDIO-VISUAL SENSORFUSION WITH NEURAL ARCHITECTURES

B. Talle A. Wichert

Department of Neural Information Processing

University of Ulm, Germany

ABSTRACT

In this paper we present a new word recognition system for monosyllabic words consisting of two types of neural networks which allows in an easy way the investigation of three different fusion architectures for audio-visual signals. Furthermore, two different kinds of preprocessing are compared: Besides low level data, a linear discriminant analysis is used for the audio and visual signals to reduce the dimensionality. Our cross-validation experiments show a slight advantage for an intermediate fusion model compared with an early fusion model which uses jointly preprocessed audio and visual data.

1. INTRODUCTION

Although automatic speech recognition systems have been highly improved in the recent years, their recognition performance is often impaired in noisy conditions. Therefore it is obvious to use other information sources than the acoustic one alone. It has been shown in many investigations that additional visual information can enhance the speech recognition abilities as well of humans [1][2] as of technical systems [3][4] especially in noisy conditions. But nevertheless it is still an important task to determine the stage of information processing at which the acoustic and visual information should be combined.

In the field of speech recognition by humans this question has been examined for example by Vroomen [5] who used serial-recall as well as coarticulation-compensation experiments for his investigation. A selective adaptation paradigm was used by Roberts and Summerfield [6] for this purpose.

For technical systems this question has been investigated by using different techniques. Adjoudani and Benoit [4] and Rogozan and Deleglise [7] used hidden Markov models. Other groups (see [8],[9]) used neural networks for a comparison as they were also used in our experiments.

In this paper we present the results of investigations of three different fusion architectures: In the early fusion model the feature vectors are combined before they are given to a classifier. In the late fusion model, a separate classification of each channel is realized before both channels are merged. The third model describes the fusion of both channels at an intermediate stage of information processing before the classification takes place (see also [3]). For our investigations we used a word recognition system for monosyllabic words which consists of two different kinds of neural networks and allows the realization all of these fusion models in an easy way.

2. DATABASE

Our database contains the acoustic and visual signals of 1255 spelled letters (so-called letterwords) of the German alphabet belonging to 25 classes. These letterwords are monosyllabic (the polysyllabic letterword 'y' is excluded) and uttered by one speaker.

The acoustic signals are preprocessed by a bank of 16 filters with centerfrequencies equally spaced on a bark scaled frequency axis. The bandwidth of each filter corresponds to the critical bandwidth of the human auditory system. The resulting spectrogram is calculated every 10 msec.

The visual signal consists of 20 x 14 gray value pixel images of the lipregion. To reduce the dimensionality of the visual data we only used the left part of the image. This halfimage of the lipregion is sampled roughly every 33 msec.

3. WORD RECOGNITION SYSTEM

Our word recognition system for the recognition of monosyllabic words was developed on the base of neural networks and contains two modules (Fig. 1):

In the first module a window is shifted over the signal and at the same time each window is attached to one of 30 phonemelike classes. This will be called a segment classification in the following. Accordingly the shifting of the window causes a sequence of such segment classifications.

The task of the second module is the time integration of this sequence of segment classifications so that the entire word is classified.

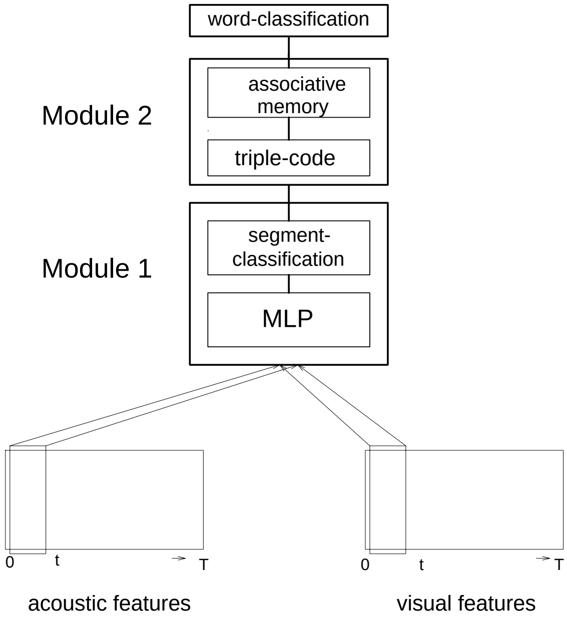

Figure 1: The word recognition system

The first module is realized through a multi-layer perceptron (MLP) which offers the possibility to realize all mentioned fusion architectures. To obtain the input-output pairs required to train the MLP with a standard backpropagation algorithm the first and the last window of each word is labeled as one of 30 phonemelike classes.

The second module is realized by an associative memory. The sequence of segment classifications is triple coded in a binary (a description of the triple code can be found in [10]) and then stored in the associative memory together with the binary vector which indicates the class of the word in a 1-of-n coding by means of a single one-component at a certain position. Because the triple code is independent of the length of the sequence the time integration is performed at this stage.

To classify an unknown word the window is shifted over the signal and classified by the trained MLP. The resulting sequence of segment classifications is triplecoded and fed into the associative memory. The associative memory determines a vector which is most similar to the stored vectors. But sometimes this vector is ambiguous because more than one class can be represented by the one-components. In this case new vectors are formed from the one-components of the answer vector and propagated backward through the associative memory. In this way a comparison can be made with the triplecoded sequence of segment classifications and the different ambiguous classes. The most similar class to the triplecoded sequence is then selected according to the hamming distance. In the case all the classes have an equal similarity to the triple coded sequence of segment classifications, no classification can be determined.

4. PREPROCESSING AND DESIGN OF FUSION ARCHITECTURES

In our experiments we investigated the fusion possibilities mentioned above at different stages of our word recognition system. For a survey of all used architectures see Figure 2.

Fusion in the first module:

Based on the optimized MLPs for each channel, the following fusion architectures were realized (in parenthesis the corresponding fusion model and the abbreviation which will be used in the following. The figure denotes the number of hidden-layers of the architecture):

- Combination of the input-layers (early integration, I1, I2),
- combination of the hidden-layers (intermediate integration, H1, H2),
- combination of the class hypothesis of the trained MLPs for each channel by means of a single layer perceptron (late integration, C).

Architecture C has the same topology as H2 but is trained in two steps. Moreover we investigated architectures with the same topology as I1 respectively I2 but with the same number of weights as architectures H1 respectively H2 (named I1(t) respectively I2(t)). The number of weights were thinned out randomly.

Fusion in the second module:

We combined the classifications of the associative memories of the acoustic and the visual channel. If a clear decision cannot be derived from both classifications and the acoustic classification is unambiguous, the classification of the acoustic channel is preferred. This corresponds to a late fusion model and is named in the following 'L'.

In our experiments we used two kinds of input representations of our data. For the first kind we used a time window of 10 frames of the spectrogram for the acoustic channel. For the visual channel the total time window could not be considered because of the resulting high dimensionality, so only the halfimage of the first frame of the window is used. Altogether this leads to 160 acoustic and 140 visual low level features.

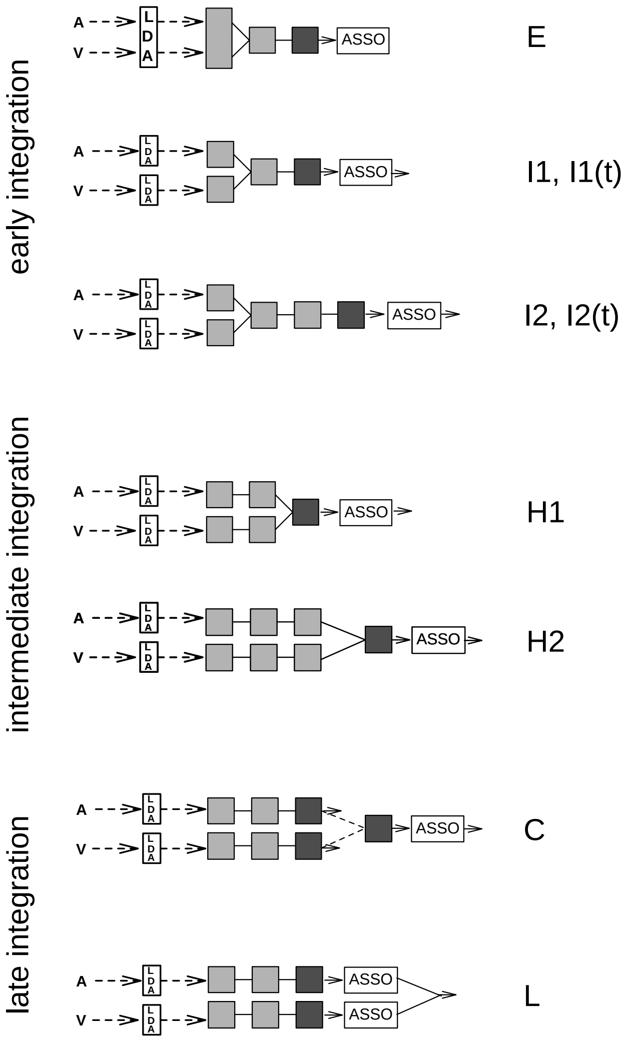

Figure 2: Survey of the fusion architectures. For low level data, the linear discriminant analysis is not performed which is indicated by the dotted arrows. The filled squares denote the different layers of the MLP, the dark squares are output-layers. The boxes concerning the segment-classification, the tuple code and the word classification are not drawn in for easy visualization reasons.

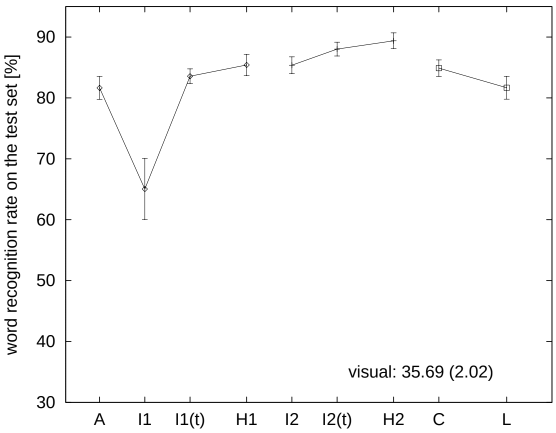

Figure 3: Word recognition rates on the test set for the low level data in percent. The error bars indicate the 95 percent confidence intervals.

To reduce the dimensionality of the data in the second kind of representation a linear discriminant analysis (LDA) is used to preprocess the low level features. Thereby two possibilities are investigated. First the LDA is calculated for each channel separately, second for the concatenation of the acoustic and the visual feature vectors, which corresponds to an early fusion model (see architecture 'E' in Figure 2). Besides the preprocessing of the data with LDA allows us to consider not only the first frame but also the total time window of the visual channel. Furthermore, data with a noisy acoustic channel were used (SNR=-3dB). Preliminary experiments led to the use of 20 LDA-coefficients for the acoustic and 9 LDA-coefficients for the visual data.

Due to the number of data 5-fold cross-validation experiments are performed on 5 different initializations of the networks, so altogether 25 experiments per architecture are executed. The word classes are not equally distributed in the data set, so we used a special measure in the following called the "class specific middle". For that purpose the word recognition rate is calculated for each class separatedly. Then the mean is calculated for all classes. In the following figures the mean of the class specific middle of the 25 experiments and as errorbars the 95 percent confidence intervals are shown.

5. EXPERIMENTS WITH LOW LEVEL DATA

Figure 3 shows the mean of the word recognition rates for the low level data on the test set. The first group presents the results for MLPs with one hidden layer (on the left side the performance of the acoustic channel alone indicated by an 'A'), the second group presents the results for MLPs with two hidden layers and the third group the results for the late fusion architectures. The architectures with two hidden-layers as well as architecture H1 and C show a significantly better recognition performance than the performance for the acoustic channel alone. The performance of architecture 'L' is equal. The failure of I1 is traced back to the number of weights and not to the architecture because I1(t) performs nearly as well as H1. The architectures with two hidden-layers and less weights (H2, I2(t)) perform best. Among the late integration architectures, architecture C is significantly better than architecture L.

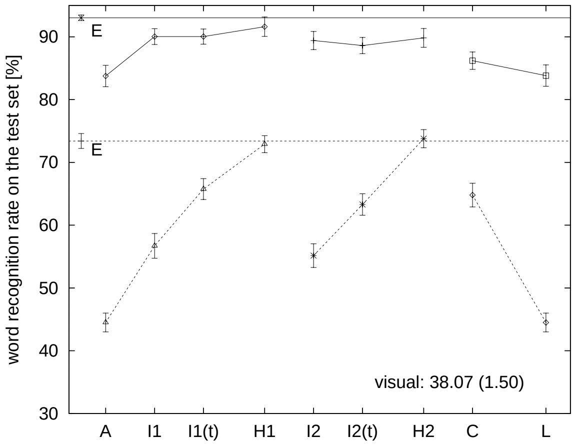

Figure 4: Word recognition rates on the test set for LDA preprocessed data in percent, visual features: one frame. The errorbars indicate the 95 percent confidence intervals.

6. EXPERIMENTS WITH LDA FEATURES

To reduce the dimensionality of the data and thus the size of the MLPs we used an LDA in the second experiment. The recognition performances for all fusion architectures can be taken from Figure 4. The visual features consist only of the first frame of the time window. The arrangement is the same as in Figure 3. Additionally the performance of the best architecture using the concatenated feature vectors (early fusion model, indicated by 'E') is outlined through a straight line. The dotted graphes show the recognition performances for architectures which were trained and tested with the noisy data, which is a more difficult task.

As in the case of the low level data, the additional visual information leads to a better performance especially in the case of the noisy acoustic channel except the architecture L. Furthermore the late fusion architectures perform significantly worse than the early and intermediate fusion architectures in the noiseless case. For noisy acoustic data architecture C performs as good as architectures I1(t) and I2(t), architecture L performs worst. Now there is only a slight difference between the architectures with one and two hidden layers: H1 (and H2 in the noisy case) yields the best results but architecture E is equally good.

Figure 5 shows the results of experiments where the whole time window for the visual features is used. All architectures show the same tendencies as in the preceding case, but now the architectures H1 (and H2 in the noisy case) are significantly better than of architecture E. It is remarkable that the performance most architectures is worse than in the case of Experiment 2, although the performance for the visual channel alone is significantly better. Tendentious, the earlier the fusion takes place the larger the difference between the performances is.

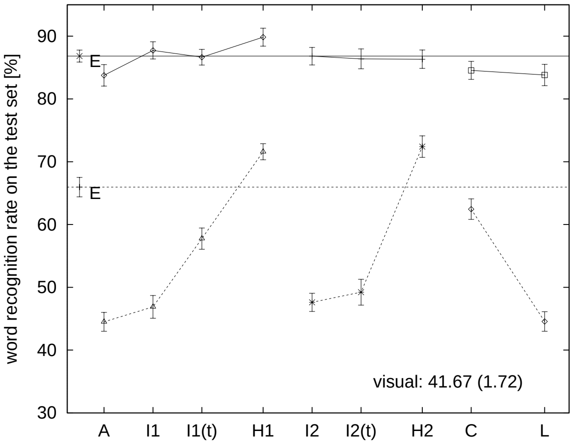

Figure 5: Word recognition rates on the test set for LDA preprocessed data in percent, visual feature: whole window. The error bars indicate the 95 percent confidence intervals.

The use of the LDA leads to at least equally good results as the use of the low level features.

7. CONCLUSION

We presented a new neural word recognition system for monosyllabic words which allows in an easy way the realization of different audio-visual fusion models and needs only a small part of the signal as training data. The use of the triple code together with an associative memory offers a new possibility to handle signals with different lengths. To investigate three different fusion models cross-validation experiments were performed to get more reliable results. Two kinds of data were investigated: Besides low level data an LDA was used to reduce the dimensionality. This leads to at least equally good results but a simpler treatment of the data. Best results are obtained for an early and an intermediate fusion model with a slight advantage for the intermediate model whereas the early fusion model performs only well when both channels are preprocessed jointly.

The advantage of the intermediate fusion model is in part in agreement with findings of Duchnowski et al. [8] who used Time Delay Neural Networks (TDNNs) for the implementation of different fusion architectures which were tested with different SNR conditions of the acoustic data. Their results were equivocal: A comparison between the early and the intermediate model leads to marginally better results for the intermediate model except in the case of the highest noise level of the acoustic data. On the other hand Stork et al. [9] found that an intermediate model did not show a better performance than a late fusion model. Both of them were also constructed on the base of TDNNs. Nevertheless, the failure of the intermediate fusion model was possibly due to insufficient training data. Furthermore Robert-Ribes et al. [11] investigated early, late and intermediate fusion models. They preferred an intermediate integration where the fusion takes place in a common representation space.

Acknowledgements: We like to thank the group of Prof. Waibel, Karlsruhe, for providing the data.

8. REFERENCES

1. Erber, N, *Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli*, Journal of Speech & Hearing Research, vol. 12, pp. 423-425, 1969

2. Sumby, W. H., Pollack, I., *Visual Contribution to Speech Intelligibility in Noise*, J. Acoust. Soc. Am., 26, pp. 212-215, 1954

3. Hennecke, M., Stork, D., Prasad, V., *Visionary Speech: Looking Ahead to Practical Speechreading Systems,* In: Stork, D., Hennecke, M. (Eds.), Speechreading by Humans and Machines, NATO ASI Series, Series F: Computer and Systems Science, Vol. 150, pp. 331-351, 1996

4. Adjoudani, A., Benoit, C., *Audio-visual Speech Recognition Compared Across Two Architectures,* Proc. of the 4th European Conference on Speech Communication and Technology, Madrid, pp. 1563-1566, 1995

5. Vroomen, J. H. M., *Hearing Voices and Seeing Lips: Investigations in the Psychology of Lipreading*, Doctoral dissertation, Katolieke Univ. Brabant, 1992

6. Roberts, M., Summerfield, Q., *Audiovisual Presentation Demonstates that Selective Adaptation in Speech Perception is Purely Auditory*, Perception and Psychophysics, 30, pp. 309-314, 1981

7. Rogozan, A., Deleglise, P., *Adaptive Fusion of Acoustic and Visual Sources for Automatic Speech Recognition*, Speech Communication 26, pp. 149-161, 1998

8. Duchnowski, P., Meier, U., Waibel, A., *See Me, Hear Me: Integrating Automatic Speech Recognition and Lipreading,* Proc. ICSLP 94, Yokohama, Japan, pp. 547-550, 1994

9. Stork, D. G., Wolff, G. J., Levine, E. P., *Neural Network Lipreading System for Improved Speech Recognition,* Proceedings of the IJCNN-92, Baltimore MD, vol. 2, pp. 289-295, 1992

10. Wickelgren, W.A., *Context-Sensitive Coding, Associative Memory, and Serial Order in (Speech) Behavior* , Psychological Review, 76, pp. 1-15, 1969

11. Robert-Ribes, J., Schwartz, J.-L., and Escudier, P., *A Comparison of Models for Fusion of the Auditory and Visual Sensors in Speech Perception,* Artificial Intelligence Review, vol. 9, pp. 323 - 346, 1995