COMPARISON OF NEURAL ARCHITECTURES FOR SENSORFUSION

Barbara Talle *Gabi Krone* *Günther Palm*

Department of Neural Information Processing,
University of Ulm, D-89069 Ulm, Germany
barbara@neuro.informatik.uni-ulm.de

ABSTRACT

For technical speech recognition systems as well as for humans it has been shown that the combination of acoustic and optic information can enhance speech recognition performance. But it still remains an open question, at which stage of processing the two information channels should be combined.
In this paper we systematically investigate this problem by means of a neural speech recognition system applied to monosyllabic words.
Different fusion architectures of multilayer perceptrons are compared both for noiseless and noisy acoustic data. Furthermore, different modularized neural architectures are examined for the acoustic channel alone. The results corroborate the idea of separate processing of the two channels until the final stage of classification.

1. INTRODUCTION

It has been shown that the combination of acoustic and optic information is useful to enhance speech recognition abilities in machine speech recognition systems [1][2][3][4][5] as well as in humans especially in the presence of acoustic noise [6][7][8]. But still there is the problem to determine at which stage of processing the two streams of information should be combined. Several experiments [9][3][10] suggests, that in technical systems the combination at the stage of class hypotheses should be preferred. Furthermore this seems the more likely model for the human perceptual system [11] but see [12] for a different view.
In this paper we report on a systematic study of neural fusion architectures composed of multilayer perceptrons (MLPs). As in other studies, the number of data is small compared to the number of parameters, which have to be determined, so we have done crossvalidation experiments to get more reliable estimates for the network performances.

2. RECOGNITION SYSTEM

The comparison of architectures is carried out in the framework of the word recognition system shown in Fig.1, which can be considered as a syllabic recognition system and used in this study for a small vocabulary recognition task: the recognition of monosyllabic words. The main features of the system are:

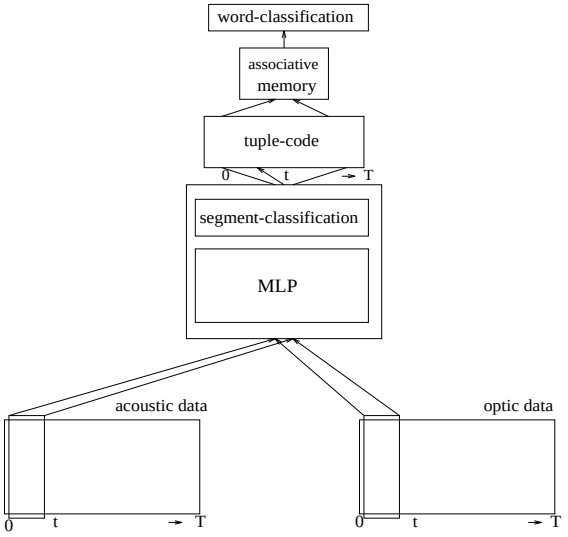

Figure 1. Word recognition system

- The recognition of a syllable is mainly based on its initial and final segment, which cover the important transitions from the consonantal clusters into the center of the syllable [13].

- Timespans longer than the often used 10 - 20 msec should be used, to include longer timedependencies. Therefore we use a timewindow of 100 msec.
  A MLP is trained to classify the initial and final segments. In the testing phase the timewindow is shifted over the signal and the MLP generates a sequence of segment hypotheses.

- Temporal integration of this sequence and the final decision is the task of an associative memory. The sequence of symbols is coded by the symboltriples it contains. The system decides in favour of the class with the best match of the actual and stored vector of symboltriples. This scheme is described in a slightly different framework in [14].

The comparison of fusion architectures, which are discussed below, is done for the stage of segmentclassification.

net (no. of weights)	recognition rate (training)	recognition rate (test)
A (7600)	94.51 (1.11)	85.86 (1.66)
V (3400)	61.76 (3.40)	54.46 (2.80)
AV-O (12800)	98.84 (0.25)	90.27 (1.39)
AV-H (11000)	94.76 (1.22)	86.96 (2.34)
AV-I (19800)	76.77 (8.56)	73.49 (6.88)

Table 1: Average recognition rate [%] of neural fusion architectures

3. DATA AND PREPROCESSING

The database consists of the acoustic and optic signals of the 26 spoken letterwords of the German alphabet. 1255 letterwords were uttered by one speaker. The polysyllabic letterword "Ypsilon" was excluded from the data. This leads to 25 word classes and 30 classes of initial and final segments.
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 optic signal is sampled roughly every 33 msec and represented by an average halfimage of the lip-region. The inputvector of the system consists of a segment of 10 frames of the spectrogram and the halfimage of the first frame (160 acoustic input features, 140 optic input features).

4. EXPERIMENTS

In the context of MLPs the combination of the acoustic and optic information can be done at three layers of the individual MLPs:

- combination of the output layers (AV-O)
- combination of the hidden layers (AV-H)
- combination of the input layers (AV-I).

For these different architectures 5-fold crossvalidation experiments have been performed. For each crossvalidation dataset we use 5 different initializations for the MLPs. This leads to 25 experiments for each architecture. All nets are trained for 1000 epochs with Backpropagation. In the tables of the next sections the average performance of these nets and in parenthesis the standard deviation are given.

net (no. of weights)	recognition rate (training)	recognition rate (test)
totaly conn. (5700)	94.62 (1.58)	85.67 (1.57)
modular frame (1380)	96.84 (0.43)	86.92 (1.47)
modular frequency (1280)	96.25 (0.42)	86.46 (1.23)
random (1380)	96.86 (0.41)	86.62 (0.99)

Table 2: Average recognition rate [%] of modular acoustic nets

Primary experiments with different topologies for the individual channels show small differences of performance over a wide range of the number of hidden units and reveal the best performance in case of about 40 hidden units for the acoustic channel and 20 hidden units for the optic channel. Thus in our experiments the single channel networks have 160 input, 40 hidden, 30 output units and 140 input, 20 hidden, 30 output units, respectively. The combined layers contain 300 input units and 60 hidden units in the corresponding cases. In the AV-O architecture the output layers of the two channels are combined into a final output layer of again 30 units.
Three different series of experiments have been performed:
1. fusion architectures in the case of a noiseless acoustic channel
2. modularization of the acoustic channel
3. fusion architectures in the case of a noisy acoustic channel.

5. RESULTS

Table 1 gives the results of the first experiments. Even in the noiseless case the optic information improves recognition performance for most architectures compared to the recognition performance of the acoustic channel alone.
The only exception is architecture AV-I. The reason for the low recognition performance can be twofold: (1) the net has to be trained further or (2) the number of weights is large compared to the available data. Further training however shows only slight improvements of performance in the range of 2%. To study the second possibility, we randomly eliminate connections between input and hidden layer so that the number of connections is equal to the number of connections in AV-H.
Thinning out connections yields recognition rates of 94.60% (1.70%) for the training and 85.41% (2.03%) for the test set, i.e. a considerablly better performance which is similar to that of AV-H.

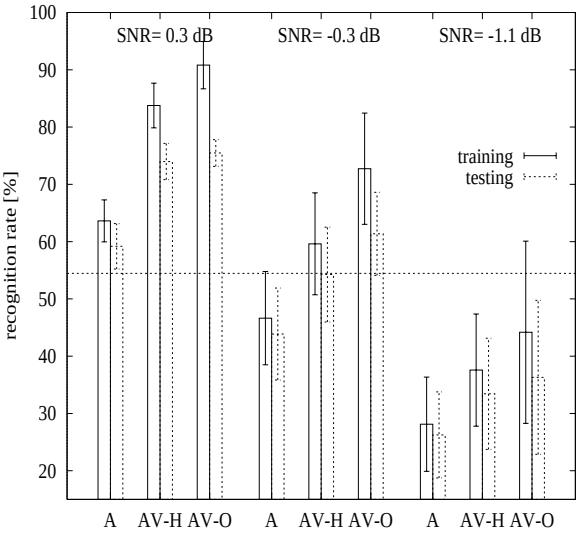

Figure 2. Average performance of neural fusion architectures in the presence of noise

In an accompanying experiment with the acoustic net alone we find the same effect, if we compare different modularization techniques (modularization with regard to frames and frequency channels) and a net, we randomly thinned out. Table 2 indicates that regardless of the kind of weight elimination all nets with less weights than the fully connected acoustic net (1) perform about equally well and (2) show slight improvements of performance compared to the fully connected network.

The results of both experiments suggest that the fully connected nets performs less because of the large number of connections.

Table 1 shows, that AV-O outperforms AV-H by roughly 4%, despite the fact that the number of connections is larger in the first case. A more significant enhancement of 8% can be observed, if one compares the class-specific recognition rates. This is due to the fact, that this measure equally considers all classes, which are not evenly distributed in our dataset. Compared to the acoustic channel, the improvement is about 4% for segment classification. The result for the final decision of the associative memory is 93.23% word recognition performance compared to 90.04% word recognition performance of only the acoustic channel.

This suggests, that the separate processing of the two channels up to the stage of class hypotheses seems more favourable. To decide this definitely we will carry out experiments with more data and with MLPs which have comparable numbers of weights between the different layers.

In a third series of experiments we investigated the influence of acoustic noise for the architectures AV-H and AV-O. We compare the performance to the acoustic channel alone. All MLPs were initialized by the weights that were learned in the noiseless case. We modularize the acoustic channel with regard to frames. The results are the following: 1. The trainingcurves (not shown here) indicate that further training the nets with noisy data considerably improves the performance. 2. For all signal-to-noise-ratios (SNR) the combination of the outputlayers shows a better improvement than the combination of the hidden layer compared to the acoustic channel alone. 3. For very low acoustic SNR the fusion architectures perform less than the visual channel alone (horizontal line in Fig. 2).

In summary, our findings corrobarate the idea that in audio-visual speech recognition the optic and the acoustic channel should be processed separately until the final stage of classification. This corresponds to the supposition that the stochastic fluctuations disturbing the two channels are independent. However the combination should not necessarily be a single weighted average of the two classifications, because the confusion matrices of the two channels need to be taken into account.

Acknowledgements: This work as well as the recording of data by the group of Prof. Waibel has been supported by the Federal Government of Baden-Württemberg (Forschungsverbund Neuroinformatik). We thank A. Wichert for his work on preprocessing the data.

6. REFERENCES

[1] C. Bregler, and Y. Konig, " 'Eigenlips' for Robust Speech Recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Adelaide, 1994

[2] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, pp. 547 - 550, 1994

[3] M. Hennecke, K. Prasad, and D. Stork, "Automatic speech recognition system, using acoustic and visual signals," Tech. Rep. Ricoh California Research Center, 1995

[4] E. Petajan, B. Bischoff, and D. Bodoff, "An improved automatic lipreading system to enhance speech recognition," ACM SIGCHI-88, pp. 19 - 25, 1988

[5] B. Yuhas, M. Goldstein, T. Sejnowski, and R. Jenkins, "Neural network models of sensory integration for improved vowel recognition," Proceedings of the IEEE, vol. 78, no. 10, pp. 1658 - 1668, 1990

[6] C. Benoit, T. Mohammadi, and S. Kandel, "Audio-visual intelligibility of French speech in noise," Journal of Speech & Hearing Research, vol. 37, pp. 1195 - 1203, 1994

[7] N. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," Journal of Speech & Hearing Research, vol. 12, pp. 423 - 425, 1969

[8] A. MacLeod, and Q. Summerfield, "A procedure for measuring auditory and audio-visual speech-reception measuring thresholds for sentences in noise: rational, evaluation and recommendations for use," British Journal of Audiology, vol. 24, pp. 29 - 43, 1990

[9] A. Adjoudani, and C. Benoit, "Audio-visual speech recognition compared across two architectures," Proceedings of the 4th European Conference on Speech Communication and Technology, pp. 1563 - 1566, 1995

[10] J. Movellan, "Visual speech recognition with stochastic networks," In G. Tesauro, D. Touretzky, T. Leen, editors: Advances in Neural Information Processing Systems, MIT Press, vol. 7, pp. 851 - 858, 1995

[11] D. Massaro, M. Cohen, and P. Smeele, "Cross-lingustic comparisons in the integration of visual and auditory speech," Memory & Cognition, vol. 23, pp. 113 - 131, 1995

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

[13] S. Furui, "On the role of spectral transitions for speech perception," J. Acoust. Soc. Am., vol. 80, no. 4, pp. 1016 - 1025, 1968

[14] G. Palm, F. Schwenker, and F. Sommer, "Associative memory networks and sparse similarity preserving codes," In: Charkassky, V. (Ed.): From statistics to neural networks: Theory and pattern recognition applications. Nato ASI Series F, Springer, 1994