FUSION OF VISUAL AND ACOUSTIC SIGNALS FOR COMMAND-WORD RECOGNITION

Rudolf Kober *Ulrich Harz* *Jutta Schiffers*

Research Institute for Applied Knowledge Processing
PO-Box 2060, D-89010 Ulm, Germany
{kober,harz,schiffer}@faw.uni-ulm.de

ABSTRACT

In this paper, we investigate the question of how the visual information of lip movement contributes to command-word recognition. The fusion of the acoustic and visual signal can be carried out either at the feature level or at the class level. Integration at the feature level means merging of the acoustic and visual features to yield a combined feature vector which is feed into a HMM-system. Fusion at the class level means separate classification of the two sources of information and combination of the classification results. An HMM classifier is used for the acoustic signal and three different classifiers (HMM, DTW and ClaRe) for the visual signal. The classification results are combined using C4.5. The recognition rates of both fusion schemes are comparable. Both yield small improvements at high SNR's using the acoustic/visual system in comparsion to the acoustic system alone. Larger improvements (up to 12%) result at low SNR's.

1. INTRODUCTION

In automatic speechreading, the visual information of lip movement is used in conjunction with the acoustic signal to enhance speech recognition. Speechreading improves the performance of speech recognition, particularly in situations with low signal to noise ratios [1, 2, 3, 4, 5, 6, 7, 8]. Improvements in recognition rate between 0% and 40% have been achieved.

The purpose of this paper is to compare two different fusion architectures for command-word recognition. The first one combines the data at the feature level (Fig. 1), that is, the two feature vectors of the acoustic and visual signal are combined into a joint feature vector, and the joint feature vector is used as input to a classifier. The advantage of this approach is that statistical dependencies between the acoustic and visual signal are taken into account during the classifier training. On the other hand, the fusion process yields higher dimensional feature vectors. The higher the dimensionality of the training vectors, the more complex the classifier becomes and the more samples are necessary to train it. Too few training samples may lead to a reduced ability to generalize, which, in turn, results in poor classification performance on unknown test data.

The second architecture fuses the data at the class level (Fig. 2), that is, the two signals are separately classified and the classification results are combined in a subsequent step. In that case, specific speech and image classifiers can be used for the signals. This approach is preferable if the signals are statistically independent.

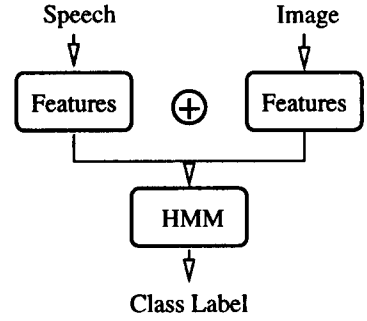

Figure 1. Fusion on feature level.

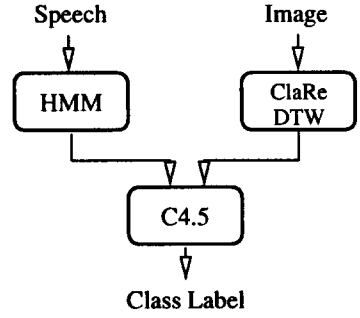

Figure 2. Fusion on class level.

Different levels of fusion in speech recognition have been reported in the literature. Stork [6] used two different neural network architectures to fuse acoustic and visual data on two different levels. Benoît [1] uses HMM's classifiers to investigate fusion on the feature and class levels.

2. FUSION AT THE FEATURE LEVEL

An the feature level, acoustic features and one visual feature are combined into a single feature vector. The combined feature vectors are classified using an HMM approach.

To represent the speech data, the signal is cut into 40 ms frames. Each frame is represented by a 39-dimensional acoustic feature vector (12 mel-cepstral coefficients with 1 energy parameter, and their first and second derivatives). Each of these 40 ms frames corresponds to one lip image. For each image, the height of the mouth is used as the visual feature. This visual feature is extracted from the image sequence using the algorithm described in [9] (Fig. 4).

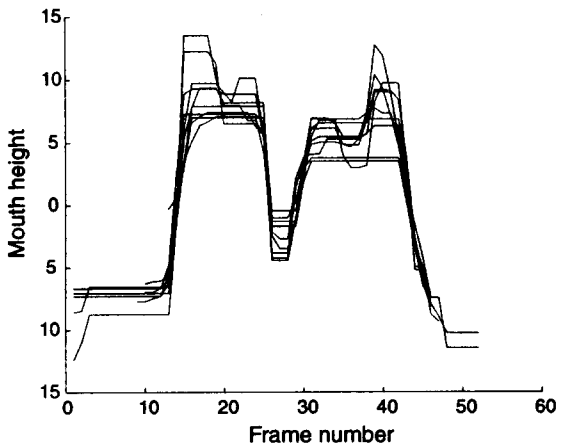

Figure 3. Height of the mouth of 4 persons each speaking three times "cafeteria" after alignment using dynamic time warping (DTW).

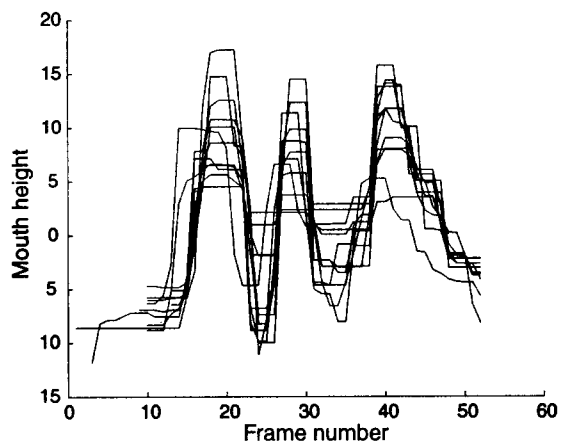

Figure 4. Height of the mouth of 4 persons each speaking three times "minifabrik" after alignment using dynamic time warping (DTW).

The combined 40-dimensional feature vectors are classified with an HMM-approach using the Hidden Markov Model Tool Kit (HTK) [10]. We use 6 word models with 8 to 22 states, depending on the length of the command word; silence is modeled with a 3 state HMM. Emission probabilities are continuous-density Gaussian distributions with diagonal covariance matrices. Normal-distributed noise was added to the acoustic signals in order to evaluate the fusion on the feature level, depending on the SNR.

For the recognition task, the 6 command words "stop", "vorwärts", "rückwärts", "minifabrik", "cafeteria" and "komm her" are used. These command words were spoken by 4 speakers, approximately 10 times each. In total, our data set consists of 262 command word samples. 197 words are used for training and 65 for testing.

The recognition rates for the acoustic signal alone and for the combined acoustic and visual signal are shown in Fig. 5.

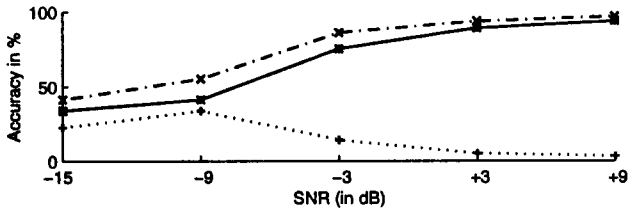

Figure 5. Recognition rates at the feature level for the acoustic signal (indicated by the solid line) and the combined acoustic-visual signal (indicated by the dot dashed line). The improvement is indicated by the dotted line.

3. FUSION AT THE CLASS LEVEL

The acoustic and visual signals are classified separately. The decision tree classifier C4.5 [11] is used to combines the hypotheses of the acoustic and visual classifiers.

To classify the acoustic signal, the HMM approach described above is used again. Three different classification methods are applied to the visual signal: HMM, Template matching with DTW, and ClaRe [12]. The input signals for the HMM and DTW classifier are the median filtered signal of the height of the mouth. For the ClaRe approach, visual data are preprocessed in order to get fixed length input vectors. This preprocessing includes noise/speech segmentation, resampling of the data, and principal component analysis.

The classification rates achieved by the different approaches are: 37% for the HMM, 61% for the DTW and 60% for the ClaRe approach. The results of DTW classification are fused with the HMM results of the acoustic signal.

The input to the C4.5 decision tree are the indices of the 3 best classes of the acoustic HMM classifier and the scores for all 6 classes of the visual DTW classifier. Thus a 9-dimensional vector with 3 symbolic and 6 numeric values is fed into C4.5. The ability to combine symbolic and numeric data is one of the advantages of C4.5. We do not use the total log probabilities of the HMM classifier because in our experiments they did not represent the reliability of the N-best classes. The recognition rate for the acoustic signal alone and the combined acoustic and visual signal is shown in Fig. 6.

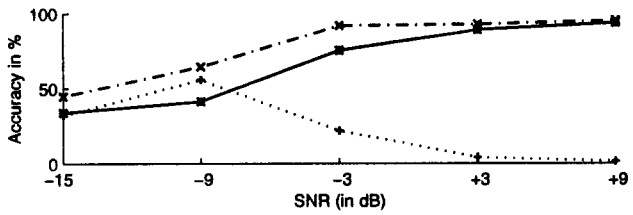

Figure 6. Recognition rates at the class level for the acoustic signal (indicated by the solid line) and the combined acoustic-visual signal (indicated by the dot dashed line). The improvement is indicated by the dotted line.

4. CONCLUSIONS

Our experiments show an improved recognition rate for the fusion at the class level as compared with the fusion at the feature level. In particular, at an SNR of -9 dB the class-fusion system outperforms the feature-fusion system. These results can be attributed to the independence of the acoustic and visual signal, which is captured by the specific adaptation of the acoustic and visual classifier. Although, similar results have been reported in the literature [6] and [1], our approach is novel in that it uses decision tree classification for fusion of the first level classification results. This enables us to combine both numerical and symbolic data and to produce explainable results.
[1]

REFERENCES

[1] A. Adjoudani and C. Benoît, *Audio-Visual Speech Recognition Compared Across Two Architectures*, ESCA. EUROSPEECH'95. 4th. European Conference on Speech Communication and Technology. Madrid, September 1995, pp.1563-1566

[2] C. Bregler, S. Manke, H. Hild and A. Waibel, *Improving Connected Letter Recognition by Lipreading*, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Minneapolis, 1993

[3] Brooke, N.M., Tomlinson, M.J. (to appear) *Processing facial images to enhance speech communication*, in M.M. Taylor, F. Neel and D.G. Bouwhuis (eds.) Proceedings of the Second Venaco Conference on the Structure of Multimodal Dialogue.

[4] Pierre Jourlin Marc El-Béze Henri Mèloni, *Automatic bimodal speech recognition* , Proc. of International Workshop on Automatic Face- and Gesture-Recognition, Zurich, 1995

[5] Eric D. Petajan, *Automatic Lipreading to Enhance Speech Recognition*, Proc. Computer Vision and Pattern Recognition. IEEE Computer Soc. Conf., San Francisco, USA, June 19-23, 1985, pp. 40-47

[1]This work was supported by the state of Baden-Württemberg, Germany, Landesschwerpunktprogramm Neuroinformatik.

[6] David G. Stork, Greg Wolff and Earl Levine, *Neural network lipreading system for improved speech recognition*, IJCNN-92, Baltimore MD, 1992

[7] Jian-Tong Wu and Shinichi Tamura et. al., *Neural Network Vowel-Recognition Jointly Using Voice Features and Mouth Shape Image*, Pattern Recognition. Vol 24. No 10. pp. 921-927, 1991

[8] Ben P. Yuhas, Moise H. Goldstein, Jr., Terrence J. Sejnowski, *Integration of Acoustic and Visual Speech Signals Using Neural Networks*, IEEE Communications Magazine, pp. 65-71, November 1989

[9] R. Kober, J. Schiffers, K. Schmidt, *Model-Based versus Knowledge-Guided Representation of non-rigid Objects: A Case Study*, Proceedings ICIP 94, IEEE International Conference on Image Processing, Vol. I of III, 1994, pp. 973-977

[10] S. Young, J. Jansen, J. Odell, D.Ollason, P. Woodland, *The HTK Book*, Cambridge University Engineering Department Speech Group, 1995

[11] J. R. Quinlan, *C4.5: Program for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993

[12] J. Schiffers and R. Kober, *A Self-Organizing Tree-Structured Classifier for High-Dimensional Data Sets*, Proc. of IDA-95, Int. Institute for Advanced Studies in Systems Research and Cybernetics, 1995