ADAPTIVE BIMODAL SENSOR FUSION FOR AUTOMATIC SPEECHREADING

Uwe Meier[1] *Wolfgang Hürst*[1] *Paul Duchnowski*[1,2]

uwe@ira.uka.de, huerst@ira.uka.de, paul_d@ollie.mit.edu

[1]**Interactive Systems Laboratories**
University of Karlsruhe, Karlsruhe, Germany

[2]**Massachusetts Institute of Technology**
Cambridge MA, USA

ABSTRACT

We present recent work on improving the performance of automated speech recognizers by using additional visual information (Lip-/Speechreading), achieving error reduction of up to 50%. This paper focuses on different methods of combining the visual and acoustic data to improve the recognition performance. We show this on an extension of an existing state-of-the-art speech recognition system, a modular MS-TDNN. We have developed adaptive combination methods at several levels of the recognition network. Additional information such as estimated signal-to-noise ratio (SNR) is used in some cases. The results of the different combination methods are shown for clean speech and data with artificial noise (white, music, motor). The new combination methods adapt automatically to varying noise conditions making hand-tuned parameters unnecessary.

1. INTRODUCTION

Automated speech recognition systems still perform poorly in real-world applications. Most approaches are very sensitive to background noise or fail totally when more than one speaker talks simultaneously (cocktail party effect).

It is well known that hearing-impaired listeners and those listening in adverse acoustic environments rely heavily on the visual input to disambiguate among acoustically confusable speech elements. The usefullness of lip movement information stems in large part from its rough complementariness to the acoustic signal [1, 2, 3].

Therefore, it is only natural to try to supplement the acoustic data with lip movement information. Related work on this concept was published by other researchers in [4, 5, 6, 7, 8, 9]. Our own work in this area has been previously reported in [10, 11, 12].

In this paper we focus on combining the acoustic and visual input data to improve recognition performance. The merging of the two information sources is very important for the final results. With only visual input our recognizer obtains recognition rates of up to 55%. Since the pure acoustic recognition accuracy on clean data is over 90% the visual part should presumably be given lower weighting under undisturbed conditions. On the other hand the acoustic-only recognition rate decreases when background noise is present. Here making more use of the visual information seems appropriate. A combination dynamically adapting to the circumstances ought to produce optimal recognition results.

2. SYSTEM DESCRIPTION

In the basic set-up, we record, in parallel, the acoustic speech and the corresponding series of mouth images of the speaker. The speaker and his lips are found and tracked automatically.

We use speaker-dependent continuous spelling of German letter strings (26 letter alphabet) as our task. Words in our database are 8 letters long on average.

noise	signal-to-noise ratio
clean	33 dB
white noise	16 dB and 8 dB
music	20 dB and 16 dB
motor	25 dB and 10 dB

Table 1. Acoustic environments tested (dB SNR).

A modular MS-TDNN [13, 14] is used to perform the recognition. Combining visual and acoustic data is done on the phonetic layer (Fig. 1) or on lower levels (Fig. 3).

As visual input we use Linear Discriminant Analysis coefficients of the gray-scale pictures of the lip region. (top 32 coefficients per image frame). For acoustic preprocessing 16 Melscale coefficients are used.

We have trained the recognizer on 170 sequences of acoustic/visual data from one speaker and tested on 30 sequences of the same person. For each combination method below we have trained the nets on clean acoustic data. We separately trained an acoustic TDNN on the same sequences of clean and corrupted data with white noise at 16 dB SNR. For testing we also added different types of artificial noise to the test-set of clean data (see Tab. 1). As performance measure word accuracy is used (where a spelled letter is considered a word):

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

3. COMBINATION ON PHONETIC LAYER

In the basic system (Fig. 1) an acoustic and a visual TDNN are trained separately. The acoustic net is trained on 63 phonemes, the visual on 42 visemes[1].

[1]viseme = visual phoneme, smallest part of lipmovement that can be distinguished. Several phonemes are usually mapped to each viseme.

The combined activation (hyp_{AV}) for a given phoneme is expressed as a weighted summation of the phoneme layer activations of this phoneme and the corresponding viseme unit:

$$hyp_{AV} = \lambda_A hyp_A + \lambda_V hyp_V \quad \text{and} \quad \lambda_A + \lambda_V = 1 \qquad (2)$$

The weights λ_A and λ_V for this combination are dependent on the quality of the acoustic data. If the quality is high, i.e. no noise exists, the weight λ_A should be high. In the case of significant acoustic noise, a higher weight λ_V for the visual side has been found to give better results.

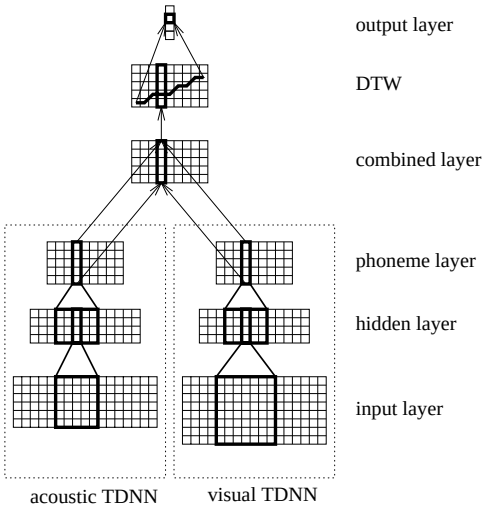

Figure 1. Combination on the phonetic layer.

3.1. Entropy Weights

One way to determine the weights for the combination (2) is to compute the entropy of the phoneme/viseme layer. The 'entropy weights' λ_A for the acoustic and λ_V for the visual side are given by:

$$\lambda_A = b + \frac{S_V - S_A}{\Delta S_{max-over-data}}, \text{ and } \lambda_V = 1 - \lambda_A \qquad (3)$$

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 probability 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 weights to favor one of the modalities. In the results shown here, we have optimized this parameter by setting it by hand, depending on the quality of the actually tested acoustic data.

3.2. SNR Weights

The quality of the speech data is generally well described by the signal-to-noise-ratio (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 the acoustic and visual weights as a function of the SNR (see middle of Fig 2). The SNR itself is estimated automatically every 500 ms from the acoustic signal. Linear interpolation is used to get an SNR value for each frame (i.e. every 10 ms). In several experiments we obtained best results with a maximum and a minimum weight $\lambda_{A\,max} = 0.75$ and $\lambda_{A\,min} = 0.5$ for high (33dB) and low (0dB) SNR respectively and a linear interpolation between them. Fig. 2 shows on an example from the test set, the values of the weights as they vary with the estimated SNR which is shown on top (for more information about this algorithm see [15]).

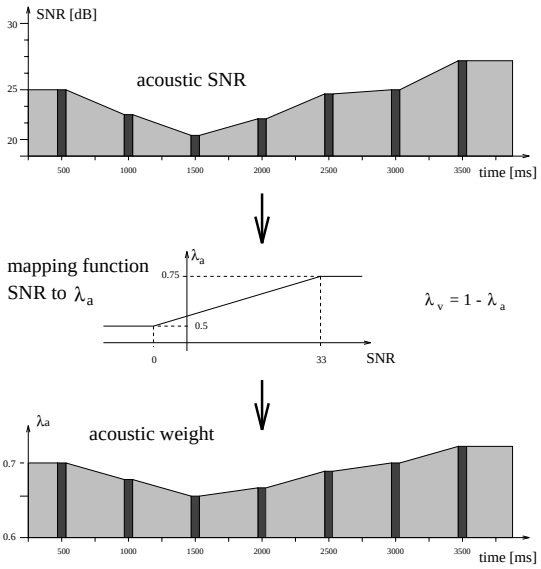

Figure 2. Determining the weights by using the SNR.

3.3. Learning the weights

Another approach is to use a neural network to compute the combination weights at the phoneme level. This method differs form the previous in two ways. First the combination weights are learned from training data and not calculated during the recognition progress. Second, different weights λ_A and λ_V are computed for different features, i.e. for every phoneme/viseme, instead of a weighting common to all phoneme/viseme pairs for a given time-frame as it is in the entropy and SNR-weight cases. The motivation behind this lies in the complementariness of the acoustic and the visual signal: some phonemes which are high confusable even in quiet have corresponding visemes that can be distinguished reliably. So it is only natural to prefer the visual classification for phonemes unclear acoustically and vice versa.

We have used a simple backprop net with two input layers (phonemes and visems), one output layer (phonems), and no hidden layer. Each unit of the combination layer is fully connected with the corresponding acoustic and visual frame.

4. LOWER LEVEL COMBINATION

The combination of acoustic and visual information on the phoneme/viseme layer offers several advantages. There is independent control of two modality networks, allowing for separate training rates and number of training epochs. It

is also easy to test uni-modal performance simply by setting λ_A and λ_V to zero or one. On the other hand, this method forces us to develop a viseme alphabet for the visual signal, as well as a one-to-many correspondence between the visemes and phonemes. Unlike phonemes, visemes have proven much more difficult to define consistently exept for a few fairly constant sets. Combination of phonemes and visemes further prevents the recognizer from taking advantage of lower level correlation between acoustic and visual events such as inter-modal timing relationships.

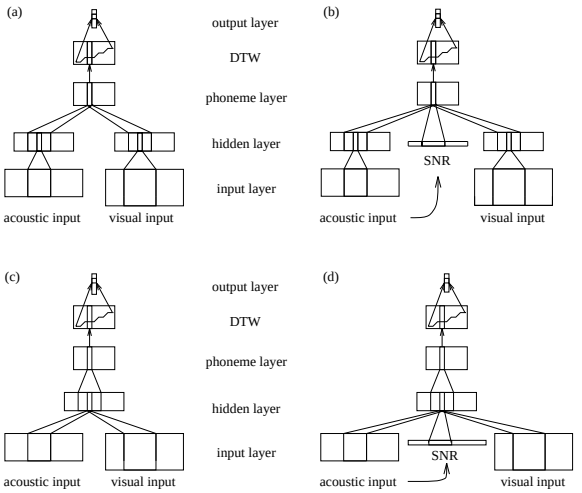

Figure 3. Lower level combination: (a) hidden layer (b) hidden layer and SNR (c) input layer (d) input layer and SNR.

Two alternatives are to combine visual and acoustic information on the input or on the hidden layer (see Fig 3 (a) and (c)). In another approach, we have used the estimated SNR of the acoustic data as an additional input to both networks (see Fig 3 (b) and (d)).

5. RESULTS

Figure 4 shows the results for the three combination methods on the phonetic layer and on the input and hidden layer in comparison to the acoustic recognition rate in different noise environments. All the nets were trained on clean acoustic data. The recognition rate on the visual data (without acoustic information) was 55%. The architectures in Fig. 3 (b) and (d) were not trained with the clean dataset because the additional information (SNR) does not appear in this training set (i.e. the SNR is approximately constant for all the words in this database). So recognition improvements from this kind of architecture could not be expected in this case of training data.

With all combination methods we get an improvement compared to the single acoustic recognition, especially in the case of high background noise. We obtain the best results using the combination on the phonetic layer. Using the entropy weights yields good recognition results but has a great disadvantage: a bias b which is necessary to preskew the weights is needed and has to be optimized by hand. In contrast, the SNR weights were determined automatically. They result in roughly the same performance without having to 'hand-optimize' any parameters during the recognition progress. We have also tested a combination of this two methods, i.e. computing the bias b of the entropy weight from the SNR instead of setting it by hand. The results were approximately the same as with hand-optimized entropy weights.

Both combination methods have the disadvantage that they do not take into consideration the inherent confusability of some phonemes and visemes, but use a single weight in each acoustic/visual time frame depending only on the quality of the acoustic data. The approach which uses a neural network for combination relies on the fact that some phonemes are easier to recognize acoustically while some can be more reliably distinguished from the visual input, by using different weights for each phoneme/viseme pair. As expected, this method delivers the best results exept in the case of high background noise (i.e. motor 10 dB and white noise 8 dB).

Similarly, the hidden- and input-combination recognition performance suffers more in these cases. However, when evaluating the different approaches one has to remember that the neural net combination, just as the hidden- and input-combination, has no explicit information about the quality of the acoustic input data which can be used during the recognition progress as it is done by the combination at the phonetic level with the entropy- and the SNR-weights.

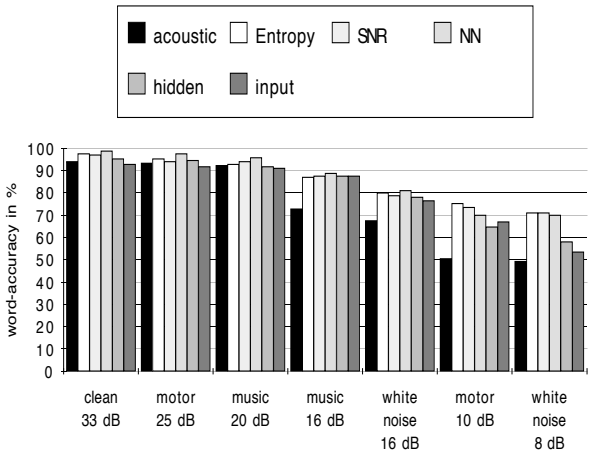

Figure 4. Combination on input, hidden, and phone layer; trained with clean data.

Motivated by this we have trained the net on a set of clean and noisy data, i.e. the 170 sequences used before and with the same sequences with 16 dB white noise. The results are presented in Fig. 5. Here we also trained the architectures from Fig. 3 (b) and (d), i.e. hidden and input combination with additional input of the SNR. In some cases we get small improvements with that kind of combination.

On the slightly noisy data we get improvements in comparison to the results achieved with the clean training data set. The improvements in the case of white noise are predictable since the training data contains utterances contaminated with 16 dB SNR white noise. The improvements obtained with the motor 10 dB SNR test set are most remarkable. Here an error reduction of about 50% was found in the case of phonetic combination with entropy- and SNR-weights compared to the results obtained with the exclusively clean training data set. Unfortunately the combination with a neural network did not lead to such a good error

Figure 5. Combination on input, hidden, and phone layer; trained with clean data and artificial noise.

6. CONCLUSION

In this paper we have presented different types of sensor fusion for automatic speech recognition and Lip-/Speechreading. We get an error reduction of up to 50% in comparison to the acoustic-only recognition results. The adaption to different noise environments is done automatically. The investigated methods differ in the combination level (high or lower layer of the TDNN) at which they are invoked and in the method of computing the combination weights (frame and feature dependent). Another difference is the fact that some combination methods (entropy- and SNR-weights on phonetic-level-combination) make use of automatically extracted information about the quality of the acoustic data during the recognition process.

Good results were obtained with the combination via neural network on the phoneme level. This kind of high level combination with different weights for different features (i.e. phonemes/visemes) yields good results although it does not use information about the quality of the acoustic data during the recognition process.

ACKNOWLEDGEMENTS

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] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 1976.

[2] M. McGrath and Q. Summerfield. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, 77(2):678–685, February 1985.

[3] A.A. Montgomery, B. Walden, and R. Prosek. Effects of consonantal context on vowel lipreading. *Journal of Speech and Hearing Research*, 30:50–59, 1987.

[4] A.J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* Dissertation, The School of Engineering and Applied Science of The George Washington University, September 1993.

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

[6] P.L. Silsbee and A.C. Bovic. Audio-visual speech recognition for a vowel discrimination task. *SPIE*, 2049:84–95.

[7] E.D. Petajan. Automatic lipreading to enhance speech recognition. *Proc. IEEE Communications Society Global Telecommunications Conference*, 1984.

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

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

[10] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. *Proc. ICASSP*, 1993. Minneapolis.

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

[12] 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.

[13] A. Waibel, T. Hanazawa, G. Hinton, and K. Shikano. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):328–339, 1989.

[14] Hermann Hild and Alex Waibel. Speaker-Independent Connected Letter Recognition With a Multi-State Time Delay Neural Network. In *3rd European Conference on Speech, Communication and Technology (EUROSPEECH) 93*, September 1993.

[15] H. Günther Hirsch. Estimation of Noise Spectrum and its Application to SNR-Estimation and Speech Enhancement. *Technical Report, International Computer Science Institute, Berkeley, California, USA.*