the task to recognize 10 isolated letters and used artificial markers on the lips. No visual feature extraction was integrated into their model.

Also of interest are some psychological studies about human speechreading and their approach to describe the human performance. This measurements could also be applied to the performance analysis of automated speechreading systems. Dodd and Campbell [3], and Demorest and Bernstein [2] did some valuable work in this area.

7. CONCLUSION AND FUTURE WORK

We have shown how a state-of-the-art speech recognition system can be improved by considering additional visual information for the recognition process. This is true for optimal recording conditions but even more for non-optimal recording conditions as they usually exist in real world applications. Experiments were performed on the connected letter recognition task, but similar results can be expected for continuous speech recognition as well.

Work is in progress to integrate not only the time independent weight sharing but also position independent weight sharing for the visual TDNN, in order to locate and track the lips. We are also on the way to largely increase our database in order to achieve better recognition rates and to train speaker independently. Investigations of different approaches are still in progress in order to combine visual and acoustic features and to apply different preprocessing to the visual data.

ACKNOWLEDGEMENTS

We appreciate the help from the DEC on campus research center (CEC) for the initial data acquisition. This research is sponsored in part by the Land Baden Württemberg (Landesschwerpunktprogramm Neuroinformatik), and the National Science Foundation.

REFERENCES

[1] Christian Benoit, Tahar Lallouache, Tayeb Mohamadi, and Christian Abry. A Set of French Visemes for Visual Speech Synthesis. *Talking Machines: Theories, Models, and Designs*, 1992.

[2] M.E. Demorest and L.E. Bernstein. Computational Explorations of Speechreading. *In Submission.*

[3] B. Dodd and R. Campbell. Hearing by Eye: The Psychology fo Lipreading. *Lawrence Erlbaum Press*, 1987.

[4] C.G. Fischer. Confusion among visually perceived consonants. *J. Speech Hearing Res.*, 11, 1968.

[5] P. Haffner and A. Waibel. Multi-State Time Delay Neural Networks for Continuous Speech Recognition. In *Neural Information Processing Systems* (NIPS 4). Morgan Kaufmann, April 1992.

[6] H. Hild and A. Waibel. Connected Letter Recognition with a Multi-State Time Delay Neural Network. To appear in *Neural Information Processing Systems* (NIPS 5).

[7] K. Mase and A. Pentland. LIP READING: Automatic Visual Recognition of Spoken Words. *Proc. Image Understanding and Machine Vision, Optical Society of America*, June 1989.

[8] H. Ney. The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1984.

[9] E. Petajan, B. Bischoff, D. Bodoff, and N.M. Brooke. An Improved Automatic Lipreading System to enhance Speech Recognition. In *ACM SIGCHI*, 1988.

[10] D.A. Pomerleau. Neural Network Perception for Mobile Robot Guidance. PhD Thesis, CMU. *CMU-CS-92-115*, February 1992.

[11] P.W. Rander. Facetracking Using a Template Based Approach. *Personal Communication.*

[12] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning Internal Representations by Error Propagation. *Parallel Distributed Processing Vol. 1.* MIT Press, 1986.

[13] David G. Stork, Greg Wolff, and Earl Levine. Neural Network Lipreading System for Improved Speech Recognition. In *IJCNN*, June 1992.

[14] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328-339, March 1989.

[15] B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski. Integration of Acoustic and Visual Speech Signals using Neural Networks. *IEEE Communications Magazine*,

[16] John B. Hampshire II and Alexander H. Waibel. A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Neural Networks*, 1(2), June 1990.

classify /b/ and /p/ based only on visual information would lead to recognition rates not better than guessing, or the net perhaps would get sensitive for features which are uncorelated to the produced speech. This leads to the design of a smaller set of visual distinguishable units in speech, so called “visemes”. We investigate a new set of 42 visemes and a 1-to-n mapping from the viseme set to the phoneme set. The mapping is necessary for the combined layer, in order to calculate the combined acoustic and visual hyphotheses for the DTW layer. For example the hypotheses for /b/ and /p/ are built out of the same viseme /b_or_p/ but the different phonemes /b/ and /p/ respectly.

5. SIMULATION RESULTS

Our database consists of 114 and 350 letter sequences spelled by two male speakers. They consist of names and random sequences. The first data set was split into 75 training and 39 test sequences (speaker msm). The second data set was split into 200 training and 150 test sequences (speaker mcb).

Best results were achieved with 15 hidden units in the acoustic subnet and 7 hidden units in the visual subnet. Obviously visual speech data contains less information than acoustic data. Therefore better generalization was achieved with as little as 7 hidden units.

Backpropagation was applied with a learning rate of 0.05 and momentum of 0.5. We applied different error functions to compute the error derivatives. For bootstrapping the McClelland error measure was applied, and for the global training on letter targets the Classification Figure of Merit [16] was applied.

	Acoustic	Visual	Combined
msm/clean	88.8%	31.6%	93.2%
msm/noisy	47.2%	31.6%	75.6%
mcb/clean	97.0%	46.9%	97.2%
mcb/noisy	59.0%	46.9%	69.6%

Table 1: Results in word accuracy (words correct minus insertion and deletion errors)

Table 1 summarizes the recognition performance results on the sentence level. Errors are misclassified words, insertion, and deletion errors. For speaker “msm”, we get an error reduction on clean data from 11.2% (acoustic only) down to 6.8% with additional visual data. With noise added to the acoustic data, the error rate was 52.8%, and could be reduced down to 24.4% with lipreading, which means an error reduction to less than half of the pure acoustic recognition. For speaker “mcb”, we could not get the same error reduction. Obviously the pronunciation of speaker “mcb” was better, but doing that, he was not moving his lips so much.

It also should be noted that in the pure visual recognition a lot of the errors are caused by insertion and deletion errors. When we presented the letters with known boundaries, we came to visual recognition rates of up to 50.2%. The results of table 1 were achieved with histogram-normalized grey-value images. Experiments with 2D-FFT images are still in progress. In our initial 2D-FFT simulations we come to visual recognition errors, which are on average about 8% higher than the grey-level coding recognition errors.

We also took a closer look to the dynamic behavior of the entropy-weights. Figure 3 shows the weights from the acoustic and visual TDNN to the combined layer over time during the letter sequence M-I-E was spoken. The upper dots represent the acoustic weight A and the lower dots the visual weight V, where

$$A=0.5+ (entropy(Visual\text{-}TDNN)-entropy(Acoustic\text{-}TDNN))/2K$$

and

$$V=1.0-A.$$

Big white dots represent weights close to 1.0 and big black dots weights close to 0.0. K is the maximum entropy difference in the training set. At the end of the /m/-phoneme when the lips are closed, V is higher than A. Obviously there the visual hypotheses are more certain than the acoustic ones. During the /ie/-phoneme the acoustic hypotheses are more certain than the visual ones, which also makes sense.

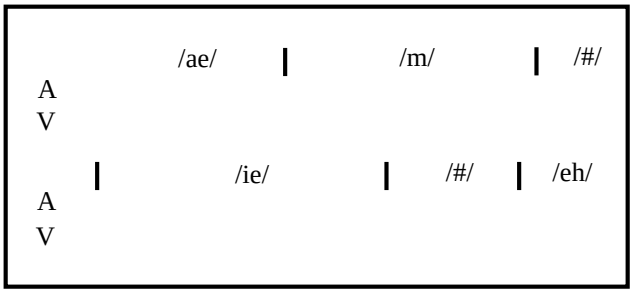

Figure 3: Entropy-Weights

6. RELATED WORK

The interest in automated speechreading (or lipreading) is growing recently. As a non-connectionistic approach the work of Petajan et al. [9] should be mentioned. Yuhas et al. [15] did use a neural network for vowel recognition, working on static images. Stork et al. [13] used a conventional TDNN (without DTW) for speechreading. They limited

detection, some “hard decisions” are made, which may hide useful information for the later learning scheme. In fact it has been reported [10] that such edge detectors are learned automatically in cases were it is necessary.

We apply two alternative preprocessing techniques: Histogram normalized grey-value coding, or 2 dimensional Fourier transformation. In both cases we just consider an area of interest (AOI) centered around the lips, and low pass filter these AOIs. The AOIs were initially segmented by hand, but an automatic procedure is now also available [11].

Grey-Value coding: We found that a 24x16 pixel resolution is enough to recognize lip shapes and movements (Figure 1). Each of these AOI pixels is the average grey-value of a small square in the original image (low pass filter). The grey-levels are rescaled in such a way that the darkest/brightest 5% in the histogram are coded with -1.0/ 1.0. The remaining 90% are scaled linear between -1.0 and 1.0.

2D-FFT: The AOI is rescaled to a 64x64 pixel image so that the 2 dimensional FFT results also with 64x64 coefficients. We just consider the log magnitudes of the first 13x13 FFT coefficients and rescale them to [-1.0, 1.0]. (After multiplying the complex FFT space with a 13x13 window and applying the inverse FFT, we could still recognize in the resulting low passed original image the distinct lip shapes and movements.) The motivation for considering the FFT is, that this coding is spatial shift invariant. It makes the recognition more stable against inaccurate AOI positioning.

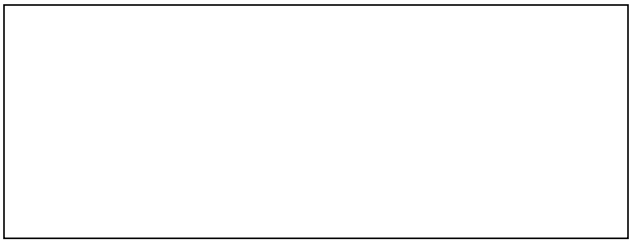

Figure 1: Typical AOIs

3. SYSTEM ARCHITECTURE

As recognition system we use a modular MS-TDNN [6]. Figure 2 shows the architecture. The preprocessed acoustic and visual data are fed into two front-end TDNNs [14], respectively. Each TDNN consists of an input layer, one hidden layer and the phone-state layer. Backpropagation was applied to train the networks in a bootstrapping phase, to fit phoneme targets.

Above the two phone-state layers, the Dynamic Time Warping algorithm [8] is applied (in the DTW layer) to find the optimal path of phone-hypotheses for the word models (German alphabet). In the letter layer the activations of the phone-state units along the optimal paths are accumulated. The highest score of the letter units represents the recognized letter. In a second phase the networks are trained to fit letter targets. The error derivatives are backpropagated from the letter units through the best path in the DTW layer down to the front-end TDNNs, ensuring that the network is optimized for the actual evaluation task, which is letter and not phoneme recognition. As before, the acoustic and visual subnets are trained individually.

In the final “combined mode” of the recognizer, a combined phone-state layer is included between the front-end TDNNs and the DTW layer. The activation of each combined phone-state unit is the weighted sum of the regarding acoustic phone-state unit and visual phone-state unit. We call these weights “entropy-weights”, because their values are proportional to the relative entropy between all acoustic phone-state activations and all visual phone-state activations. Hypotheses with higher uncertainty (higher entropy) are weighted lower than hypotheses with lower uncertainty.

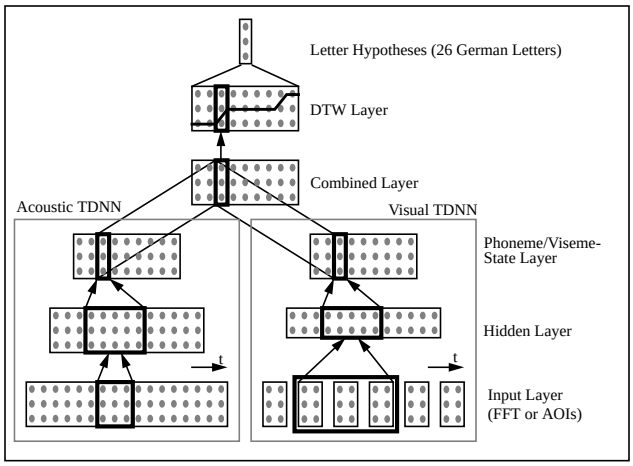

Figure 2: Neural Network Architecture

4. PHONEMES & VISEMES

For the acoustic classification we use a set of 65 phoneme-states (phoneme-to-phoneme transition states included). They represent a reasonable choice of smallest acoustic distinguishable units in German speech, and the TDNN architecture is very well suited to be trained as a classifier for them.

For visual features this will be different. Distinct sounds are generated by distinct vocal tract positions, and voiced/unvoiced excitations. External features of the vocal tract like the lips, part of the tongue and teeth, contribute only in part to the sound generation. I.e. /b/ and /p/ are generated by similar lip-movements, and cannot be distinguished with pure visual information. Training a TDNN to

IMPROVING CONNECTED LETTER RECOGNITION BY LIPREADING

Christoph Bregler, Hermann Hild, Stefan Manke, and Alex Waibel*

University of Karlsruhe
Department of Computer Science
Am Fasanengarten 5
7500 Karlsruhe 1
Germany
bregler@ira.uka.de, manke@ira.uka.de

Carnegie Mellon University
School of Computer Science
Pittsburgh
Pennsylvania 15213
U.S.A.
hhild@cs.cmu.edu, ahw@cs.cmu.edu

ABSTRACT

In this paper we show how recognition performance in automated speech perception can be significantly improved by additional Lipreading, so called "Speechreading". We show this on an extension of an existing state-of-the-art speech recognition system, a modular MS-TDNN. The acoustic and visual speech data is preclassified in two separate front-end phoneme TDNNs and combined to acoustic-visual hypotheses for the Dynamic Time Warping algorithm. This is shown on a connected word recognition problem, the notoriously difficult letter spelling task. With speechreading we could reduce the error rate up to half of the error rate of the pure acoustic recognition.

1. INTRODUCTION

Automated speech perception systems still perform poorly, when it comes to real world applications. Most approaches are very sensitive to background noise or fail totally when more than one speaker talks simultaneously (cocktail party effect), as it often happens in offices, cockpits, outdoors and other real world environments.

Humans deal with this distortions in considering additional sources. Very often misclassified acoustic signals can be corrected with the use of higher level context information. In recognition systems this is partly covered by language models or grammars. Psychological studies have shown [3], that on the lower level additional information contributes to human hearing as well. Besides the acoustic signal from both ears, visual information, mostly lipmovements, are subconsciously involved in the recognition process. This source is even more important for hearing impaired people, but also contributes significantly for normal hearing recognition.

We investigate this phenomena on the letter spelling task. No grammars or other higher level information are employed. If visual information is missing as well, even humans perform poorly. Just remember how hard it is to recognize spelled names at the telephone.

The spelling task is seen as a connected word recognition problem. As words we take the highly ambiguous 26 German letters. A test person in front of a microphone and video camera is spelling names and random letter sequences in German. We did not care about high quality recordings, we even degraded the acoustic signal with artificial noise to simulate some real world conditions.

As speech recognition system we present an extension of an existing Multi-State Time Delay Neural Network architecture (MS-TDNN) [6] for handling both modalities, acoustic and visual sensor input. It is shown how recognition performance with integrated acoustic and visual information achieves significant improvements over acoustic input only.

2. BIMODAL ACQUISITION AND PREPROCESSING

Our recording setup consists of a conventional NTSC camera and microphone. The video images are grabbed in real-time (30 fullframes/sec) into our workstation and are saved as 256x256 pixel images with 8bit grey-level information per pixel. This squared region covers the full face of the speaker. In parallel the acoustic data is sampled at a 16KHz rate and 12bit resolution. Also timestamps were saved, because the correct synchronization between audio and video signals is critical for later processing.

For acoustic preprocessing we follow the established approach to apply FFT on the Hamming windowed speech data in order to get 16 Melscale Fourier coefficients at a 10 msec frame rate. For visual preprocessing there is still the active discussion, how much preprocessing heuristics is appropriate before some connectionist classification schemes are applied to the data. We follow the idea to allow only transformations with fairly low information reduction. In other preprocessing algorithms like edge

* The author is now with International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704