TOWARDS UNRESTRICTED LIP READING

Uwe Meier, Rainer Stiefelhagen, Jie Yang, Alex Waibel

{uwem, stiefel, yang+, ahw}@cs.cmu.edu

Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, USA
University of Karlsruhe, Karlsruhe, Germany

ABSTRACT

Lip reading provides useful information in speech perception and language understanding, especially when the auditory speech is degraded. However, many current automatic lip reading systems impose some restrictions on users. In this paper, we present our research efforts, in the Interactive System Laboratory, towards unrestricted lip reading. We first introduce a top-down approach to automatically track and extract lip regions. This technique makes it possible to acquire visual information in real-time without limiting user's freedom of movement. We then discuss normalization algorithms to preprocess images for different lightning conditions (global illumination and side illumination). We also compare different visual preprocessing methods such as raw image, Linear Discriminant Analysis (LDA), and Principle Component Analysis (PCA). We demonstrate the feasibility of the proposed methods by development of a modular system for flexible human-computer interaction via both visual and acoustic speech. The system is based on an extension of an existing state-of-the-art speech recognition system, a modular Multiple State-Time Delayed Neural Network (MS-TDNN) system. We have developed adaptive combination methods at several different levels of the recognition network. The system can automatically track a speaker and extract his/her lip region in real-time. The system has been evaluated under different noisy conditions such as white noise, music, and mechanical noise. The experimental results indicate that the system can achieve up to 55% error reduction using additional visual information.

1. INTRODUCTION

The visual information is complementary to the acoustic information complementary in human speech perception, especially in noisy environments. Humans can determine a confused phoneme using both acoustic and visual information because many of the phonemes, which are close to each other acoustically, might be very different from each other visually. The connection between visual and acoustic information in speech perception was shown by McGurk with the so-called *McGurk Effect* [1]. Visual information from the facial region, such as gestures, expressions, head-position, eyebrows, eyes, ears, mouth, teeth, tongue, cheeks, jaw, neck, and hair, could improve the performance of machine recognition [2]. Much research has been directed to developing systems that combine the acoustic and visual information to improve accuracy of speech recognition. These systems mainly focus on integrating acoustic and visual information from the oral-cavity region of a speaker with acoustic information. Two basic approaches have been used in these systems to combine acoustic and visual information. The first approach uses a comparator to merge the results obtained independently from acoustic and visual sources. The second approach performs recognition using a vector that contains both acoustic and visual information. Most systems reported better performances using both acoustic and visual information than using only one source of information [3, 4, 5, 6, 7, 8, 9, 10].

Most current systems, however, impose certain constraints on users, such as using a head-mounted camera or pacing reflective markers on a user's lips. It is our goal to remove these constraints. In this paper, we present our research efforts towards unrestricted lip reading. Two major reasons cause low quality of visual data in lip-reading: user movement and environment change. We present a top-down approach to automatically track and extract lip regions. We use a real-time face tracker to locate a user while the user moves freely. The lip-finder module locates the lips within the found face and provides the coordinates of the mouth corners to lip/speech-recognition subsystem, which extracts the relevant information from the image. This technique makes it possible to acquire visual information in real-time without limiting user's freedom of movement. We then discuss normalization algorithms to preprocess images for different lightning conditions (global illumination and side illumination) by comparing different visual preprocessing methods such as raw image, LDA, and PCA. We show that an adaptive method can automatically adjust parameters to different noise conditions. We demonstrate the feasibility of the proposed methods by development of a modular system for flexible human-computer interaction via both visual and acoustic speech. The system is based on an extension of an existing state-of-the-art speech recognition system, a modular MS-TDNN system. We have developed adaptive combination methods at several different levels of the recognition network. The system can automatically track a user and extract his/her lip region in real-time. The system has been evaluated under different noisy conditions such as white noise, music, and mechanical noise. The experimental results indicate that the system can achieve up to 55% error reduction using additional visual information.

2. SYSTEM DESCRIPTION

Figure 1 gives an overview on the subsystems and their communication of our Lip-reading system. We use a Canon VC-C1 Camera with integrated pan-tilt unit. This unit is controlled by the Face-Tracker module. The Face-Tracker

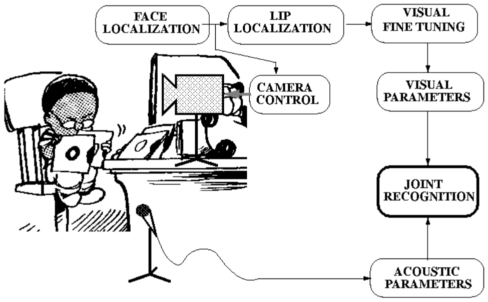

Figure 1. NLIPS - system overview

module sends the position of the face to the Lip-Finder module, which records the acoustic and visual data parallel and stores the position of the mouth-corners for every frame. Tracking of the face and the lip corners is done in real-time during the recording of the data. After that some visual fine-tuning is done to eliminate different illumination conditions from the images. The data is then feeded in a MS-TDNN recognizer [11, 12]. All those submodules are described in more detail in the following sections.

For performance measure 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. The acoustic signal is sampled at 16 kHz. The visual data is grabbed at 20-30 frames/sec with 24-bit RGB resolution. The color images are used for the Face-Tracker and Lip-Finder Modules, for the lip-reading Module gray-level images are used.

3. VISUAL FEATURE EXTRACTION

In our speech reading system we use a top-down approach to automatically track and extract lip regions. This technique makes it possible to acquire visual information in real-time without limiting user's freedom of movement.

To find and track the face, a statistical skin color-model consisting of a two-dimensional Gaussian distribution of normalized skin colors used. The input image is searched for pixels with skin colors and the largest connected region of skin-colored pixels in the camera-image is considered as the region of the face. The color-distribution is initialized so as to find a variety of skin-colors and is gradually adapted to the actual found face [13].

To find the lips a feature based gaze-tracking module is used, which is able to find and track lip-corners in real time. Moreover, the module is able to detect lip localization failures and to automatically recover from failures. Instead of tracking only the lip corners, we also track other facial features such as pupils and nostrils along with them. Tracking all these facial features and using a simple 3D head model, e.g. we know the relative positions of each of the used facial features, outliers in the set of found feature points can be detected and their true positions can be predicted [14].

4. VISUAL PREPROCESSING

In real world applications the conditions like light or size and position of the speaker can change. It was shown [15] that the recognition results decrease drastically if those conditions change within a small range. Therefor the visual data must be preprocessed to eliminate these real-world problems [16, 17].

From the Lip Finder module we get the coordinates of the mouth corners. Using these corners we can cut the lips out of the face image and rescaled to a constant size. Because the Lip Tracking is good enough, no further preprocessing is needed to get constant size and position of the lips in the lip-sequence. In our earlier system we have used frame-correlation with a so-called master-lip to get constant size and position of the lip-images.

For illumination invariance we use an adaptive grayvalue modification. The Normalization of grayvalues is done by using a distribution function of grayvalues, figure 2 shows two possible optimal distributions. Given some images under different lightning conditions, we have to adjust the gray-values in a way, which the distribution matches with an optimal distribution function. Figure 3 gives an example of images in our database, figure 4 shows the distributions before and after the gray-value modification.

In a first approach we used for the adjustment a method (gray-value modification) that is described in [18] in detail: The gray-value distribution is computed, using the accumulated gray-values, it is easy to adjust the gray-values in a way, that the accumulated function is the same as from the target function:

$$f'(p) = T(f(p))$$

where $f(p)$ is the original gray-value, T the modification function and $f'(p)$ the new gray-value.

In this method only global histogram of the images is adapted. The method gives not the desired result if side-illumination occurs in the image. We solved this problem by developing an adaptive gray-value modification: The image is divided in 4 parts Q_k (figure 5). Now we can compute the gray-value modification T_1, T_2, T_3 and T_4 for each part separate. The adaptive gray-value modification is a linear combination of these gray-value modifications:

$$T(f(p)) = \sum_{i+1}^{4} w_i T_i(f(p))$$

To compute the w_i each of the 4 parts is separated again in 4 parts (q_{ij}). There are 3 kinds of neighborhood (Region A, B and C in figure 5): q_{ij} has no, one or three Q_k neighbors. On the example of the points P_1, P_2 and P_3 in figure 5 we show how to compute the transformation:

$$\begin{aligned}
T(P_1) &= \frac{y_u}{y_o + y_u}\left(\frac{x_r}{x_l + x_r} T_1(P_1) + \frac{x_l}{x_l + x_r} T_2(P_1)\right) \\
&+ \frac{y_o}{y_o + y_u}\left(\frac{x_r}{x_l + x_r} T_3(P_1) + \frac{x_l}{x_l + x_r} T_4(P_1)\right) \\
T(P_2) &= \frac{y_u}{y_o + y_u} T_1(P_2) + \frac{y_o}{y_o + y_u} T_3(P_2) \\
T(P_3) &= T_3(P_3)
\end{aligned}$$

5. VISUAL DATA REPRESENTATION

The dimensionality of the normalized pixel vector is quite high (24x18pixel = 384), especially when compared with the acoustic input vector. Unlike for acoustic speech data, there are no generally agreed-upon parameterization strategies for the visual lip image. Since we are using a connectionist algorithm for recognition we have followed the philosophy of avoiding explicit feature extraction and segmentation of the image. Instead, we rely on the network to develop appropriate internal representations of higher level features. We

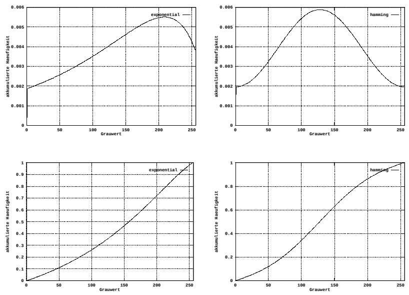

Figure 2. gray-value modification: target distributions

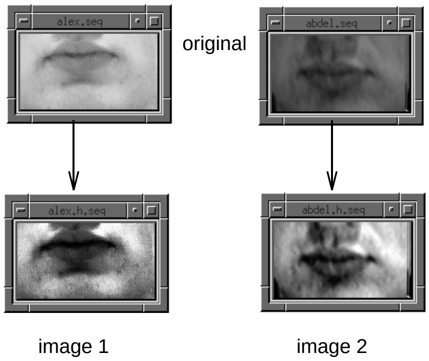

Figure 3. gray-value modification: example

have been investigating several alternative visual data representations consistent with this strategy. There is clearly much that changes little from image to image, for instance the appearance of the cheek area around the lips. While it is possible that the network will learn to ignore such redundant information, it was hypothesized that reducing the dimensionality of the input would be advantageous to generalization performance, especially under limited training data [19]. Besides the gray-level images we made experiments with PDA and LDA to reduce the input vector to 16 Coefficients.

6. COMBINATION ALTERNATIVES

A modular MS-TDNN [11, 20] is used to perform the recognition. Figure 7 schematically shows the architecture of the MS-TDNN used for acoustic recognition. Combining visual and acoustic data is done on the phonetic layer (Fig. 8) or on lower levels (Fig. 9) [21, 22].

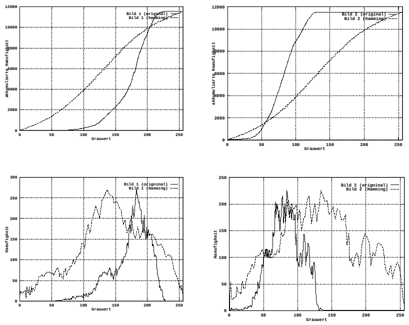

Figure 4. gray-value modification: example histograms

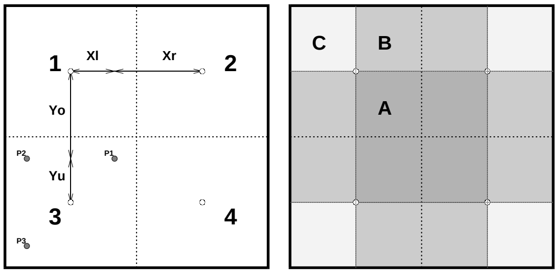

Figure 5. adaptive gray-value modification

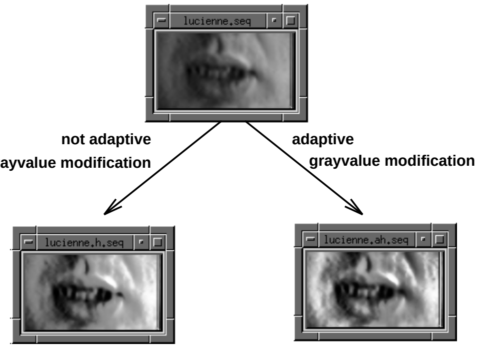

Figure 6. adaptive gray-value modification, example

6.1. Phonetic Layer Combination

In the basic system (Fig. 8) an acoustic and a visual TDNN are trained separately. The acoustic net is trained on 63 phonemes, the visual on 42 visemes[1].

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

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

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.

6.1.1. Entropy Weights

One way to determine the weights for the combination (1) 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:

$$\lambda_A = b + \frac{S_V - S_A}{\Delta S_{max-over-data}}, \quad \text{and} \quad \lambda_V = 1 - \lambda_A \qquad (2)$$

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

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

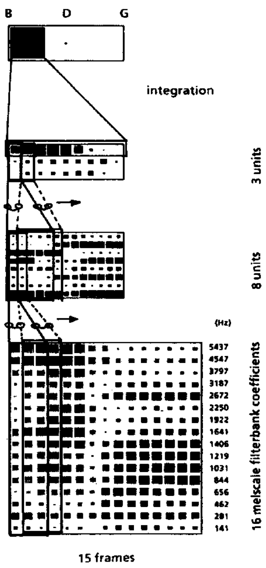

Figure 7. MS-TDNN Architecture

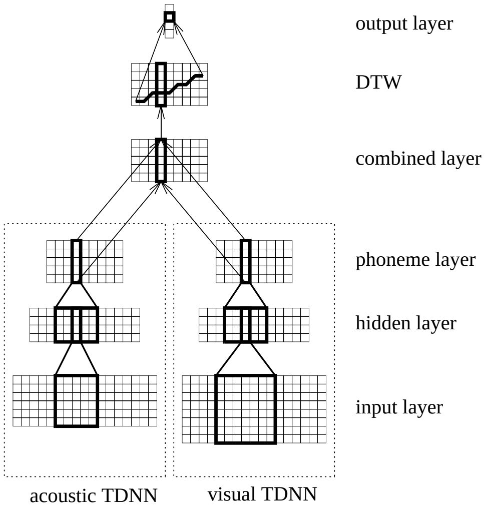

Figure 8. Combination on the phonetic layer.

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.

6.1.2. 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. 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 $\lambda_{Amax} = 0.75$ and $\lambda_{Amin} = 0.5$ for high (33dB) and low (0dB) SNR respectively and a linear interpolation between them. For more information about the SNR algorithm sees [23, 24].

6.1.3. Learning 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 back-prop net with two input layers (phonemes and visemes), one output layer (phonemes), and no hidden layer. Each unit of the combination layer is fully connected with the corresponding acoustic and visual frame.

6.2. 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 except 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.

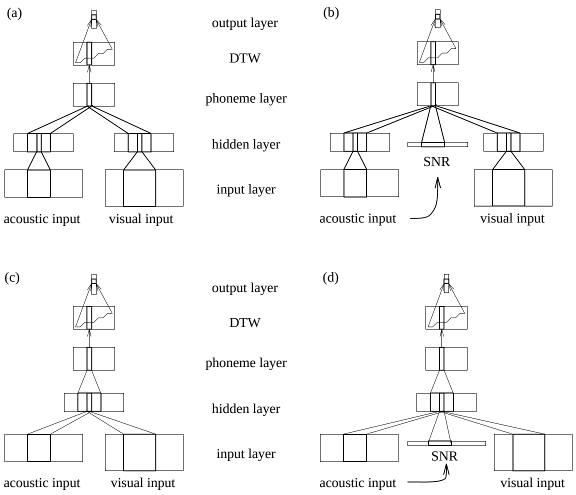

Figure 9. Lower level combination: (a) hidden layer (b) hidden layer and SNR (c) input layer (d) input layer and SNR.

Two alternatives are to combine visual and acoustic information on the input or on the hidden layer (see Fig 9 (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 9 (b) and (d)).

6.3. Performance

We have tested the recognizer on datasets 200 letters sequences (continuous spelling) from one single speaker. As performance measure we used the Word Accuracy (where each letter is seen as a word):

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

Visual Data Representation Figure 10 shows the combined word accuracy on test-set 1 with different preprocessing methods. We show the scores for clean acoustic data and for two cases where increasing amounts of white noise were artificially added to degrade the acoustic-only recognition rates. In general, best performance was achieved with gray level and LDA input. The results indicate that of the tested visual input representations, the gray-level and LDA gave very similar performance under most conditions. Thus with a proper choice of transformation we can significantly (factor 12) reduce the dimensionality of the input without sacrifing performance. Note that the reduction is done without any heuristic feature extraction. One disadvantage of PCA and LDA preprocessing is, that they are more sensible against online conditions (size, illumination) than the raw gray-level image.

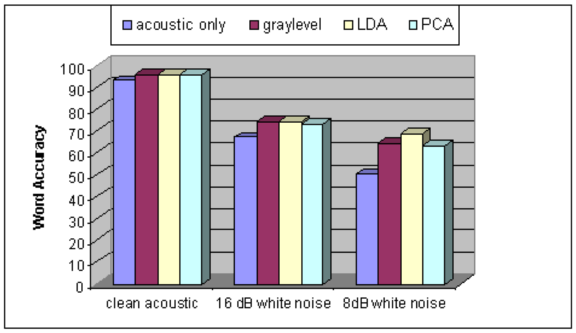

Figure 10. visual audio combined recognition rates for different data representations

Combination Alternatives 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: white noise (16 and 8 dB), music (20 and 8 dB), and mechanical noise (25 and 10 dB).

Figure 11 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. 9 (b) and (d) were not trained with the clean dataset because the additional information (SNR) does not appear in this training set (e.g. 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 pre-skew 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 that 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 except 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.

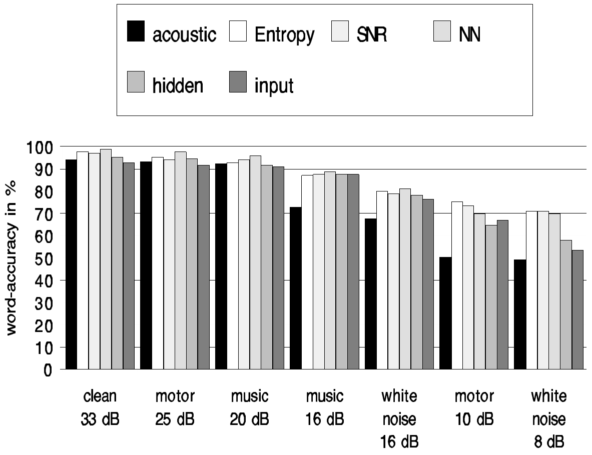

Figure 11. Combination on input, hidden, and phone layer; trained with clean data.

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. Here we also trained the architectures from Fig. 9 (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 reduction in this case.

Under both, clean and noisy, conditions we get the best performance with combining on the phonetic level. The advantage in doing this is, that the acoustic and visual net are trained separately. This means that the parameters for training can be optimized separately, i.e. the epochs for training the visual nets are three times higher than for the acoustic net.

7. CONCLUSION

We have summarized our efforts towards unrestricted lipreading in this paper. Using a top-down approach, a robust real-time tracking system can extract a user lip region while the user moves freely. The illumination changes can be handled effectively by the adaptation method. We have demonstrated the proposed methods by the continuously spelling task of German letters. The system can achieve up to 55% error reduction using additional visual information under noisy conditions. We are currently developing a user independent system for language training applications.

8. ACKNOWLEDGEMENTS

Supports for this work has come from the state of Baden Würtemberg, Germany (Landesforschungsschwerpunkt Neuroinformatik), the NSF (USA), under contract CDA-9726363, and the DARPA (USA), under contracts N00014-93-1-0806 and N6601-97-C8553.

REFERENCES

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

[2] Catherine Pelachaud, Norman Badler, and Marie-Luce Viaud. Final report to nsf of the standards for facial animation workshop. *Technical report, University of Pennsylvania*, 1994.

[3] M.E. Hennecke, K.V. Prasad, and D.G. Stork. Using deformable templates to infer visual speech dynamics. *28th Annual Asimolar conference on Signal speech and Computers.*

[4] A.J. Goldschen, O.N. Garcia, and E. Petajan. Continuous optical automatic speech recognition by lipreading. *28th Annual Asimolar conference on Signal speech and Computers.*

[5] P.L. Silsbee. Sensory integration in audiovisual automatic speech recognition. *28th Annual Asimolar conference on Signal speech and Computers*, 1994.

[6] K.V. Prasad, D.G. Stork, and G.J. Wolff. Preprocessing video images for neural learning of lipreading. *Ricoh California Research Center, Technical Report CRC-TR-93-25.*

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

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

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

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

[11] 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.

[12] H. Hild and A.Waibel. Multi-speaker / speaker-independent architectures for the multi-state time delay neural network. *Proc. Intern. Conference on Acoustics, Speech and Signal Processing ,IEEE*, 1993.

[13] J. Yang and A. Waibel. a real-time face tracker. *WACV 96.*

[14] R. Stiefelhagen, Jie Yang, and Uwe Meier. Real time lip tracking for lipreading. *Eurospeech 97.*

[15] U. Meier. Lippenlesen: verschiedene Methoden der visuellen Vorverarbeitung und Merkmalsextraktion. Studienarbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universit"at Karlsruhe (TH), Germany, 1994.

[16] U. Meier. Robuste Systemarchitekturen für automatisches Lippenlesen. Diplom-Arbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[17] Uwe Meier, Rainer Stiefelhagen, and Jie Yang. Preprocessing of visual speech under real world conditions. *Proceedings of European Tutorial & Research Workshop on Audio-Visual Speech Processing (AVSP 97).*

[18] W.K. Pratt. *Digital Image Processing.* A Wiley-Interscience Publication.

[19] 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.

[20] 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.

[21] U. Meier, W. Hürst, and P. Duchnowski. Adaptive bimodal sensor fusion forautomatic speechreading. *Proc. ICASSP*, 2:833–837, 1996.

[22] W. Hürst. Adaptive bimodale Sensorfusion für automatische Spracherkennung und Lippenlesen. Studienarbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[23] 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.*

[24] M. Schoch. Sch"atzung des Signal-Rausch-Abstandes. Studienarbeit, Institut f"ur Logik, Komplexit"at und Deduktionssysteme, Universit"at Karlsruhe (TH), Germany, 1995.