"EIGENLIPS" FOR ROBUST SPEECH RECOGNITION

Christoph Bregler [⋆,†] *and Yochai Konig* [⋆]

[⋆]Int. Computer Science Institute
1947 Center Street
Berkeley, CA 94704
U.S.A

[⋆]Computer Science Division
University of California
Berkeley, CA 94720
{bregler,konig}@cs.berkeley.edu

[†]University of Karlsruhe
Institut Prof. Alex Waibel
D-76128 Karlsruhe
Germany

To appear in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia 1994

ABSTRACT

In this study we improve the performance of a hybrid connectionist speech recognition system by incorporating visual information about the corresponding lip movements. Specifically, we investigate the benefits of adding visual features in the presence of additive noise and crosstalk (cocktail party effect). Our study extends our previous experiments [3] by using a new visual front end, and an alternative architecture for combining the visual and acoustic information. Furthermore, we have extended our recognizer to a multi-speaker, connected letters recognizer. Our results show a significant improvement for the combined architecture (acoustic and visual information) over just the acoustic system in the presence of additive noise and crosstalk.

1. INTRODUCTION

Most efforts in robust speech recognition focus on methods that reduce signal distortions. The signal distortions may be caused by background noise (additive noise) and by channel effects (convolutional noise.) We investigate an alternative approach by incorporating additional information from the signal source itself, like positional information about the visible articulators (lipmovements, tongue and teeth positions). In fact it is well known that human speech perception is inherently bi-modal as well [10, 5].

The idea of extending automated speech recognition to the visual modality has already been investigated for a long time. As popular non-connectionist approaches the work of Petajan, Bischoff, Bodoff, and Brooke [11], Mase and Pentland [9] should be mentioned. Just recently Goldschen [6] completed a lip reading system. He trained HMMs to discriminate visual information on a continuous word database. Recent connectionist systems were investigated by Yuhas, Goldstein, and Sejnowski [14], who used static images for vowel discrimination. Wolff, Prasad, Stork, and Hennecke [13] are using a modified TDNN for isolated word segments.

We focus on scenarios where the acoustic modality is degraded in a way that causes state-of-the-art speech recognition systems to achieve poor recognition performance. We simulated such situations by adding car noise and crosstalk of different ratios to clean speech. We have done similar experiments on the same database already described in [3]. In this study, however we use a new visual processing technique, a different acoustic front-end, and an alternative hybrid connectionist recognition architecture (MLP/HMM). Our experiments show that with the incorporation of the additional visual modality, we can significantly reduce the recognition error.

2. VISUAL FRONT END

The crucial part of our lip reading system is extracting the visual features of the speech from the image of the speaker face. Our goal is to reduce the high dimensional sampled image data to low dimensional feature vectors without losing relevant information. As in most vision tasks we have to solve two problems: Spatial segmentation, i.e., finding and tracking the lips, and recognition, i.e., estimating the relevant features about the "lip configuration " for classification. Ideally the spatial segmentation should be driven by the recognition, and the dimension reduction should be invariant against spatial shifting, scaling, rotation, and lighting.

In our earlier system [3] we solve the tracking problem by using a template based pyramid approach. Based on the tracking information the visual input to the TDNN classifier was the gray-level coding of the mouth area. The obvious advantage of direct gray-level coding is that we do not throw away any information. But this increases the number of free parameters that have to be estimated reliably from the data. Furthermore, we leave to the TDNN the nontrivial task of finding the generalization against spatial shifting, scaling, rotation, and lighting.

2.1. ACTIVE CONTOUR MODELS

Our new tracking approach is related to the so called "snakes" [8] and "deformable templates" [15]. Basically a "energy function" which measures the quality of the match between image features and a contour model is minimized. We learn the lip contour model from the training data, in contrast to the snake model (which is a continuous spline) and to the deformable template model (which consists of hand coded linking polynomials). Furthermore, snakes very often align onto undesirable local minima (nose- and eye-boundaries) and deformable templates are difficult to use for representing fine grain lip features. With our learned contour model we get around these limitations.

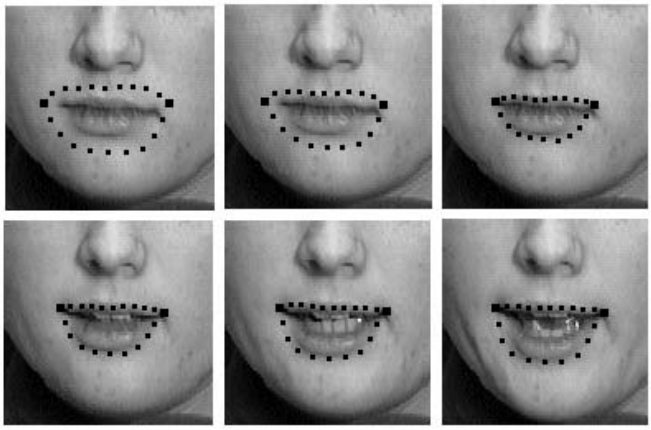

Figure 1: A typical relaxation and tracking sequence of our lip tracker

The training set consists of 4500 images of 6 speakers uttering letter sequences. The images are initially "labeled" with the conventional snake algorithm, and misaligned snakes are removed from the database by hand. Each lip boundary contour is coded as a 80 dimensional vector (40 two dimensional points evenly spaced along the contour and normalized). All legal lip contour shapes lie only in a small lowdimensional subspace. We represent this constraint "lip-space" as a 5 dimensional surface embedded in the 80 dimensional vector space. If we constrain the surface to be linear, it is similar to the space spanned by the first 5 principal components of the 80 dimensional contour database. In the nonlinear case we currently are using a learnable "mixture of local patches" representation which is described in [4]. (But the experiments reported here are only based on linear subspaces.)

For finding and tracking the lips in new images we compute an energy term, which is the negative sum of all gray-level gradient estimates along the contour. A local energy minima represents a match of the contour model with the real boundary of the lips. We minimize this term with gradient descent, but constrain the shapes to lie in the learned legal shape space (project the contour onto the principal components). This constraint represents the desired top down flow, because the local search for maximum gray-level gradients is guided by the global learned space of legal lip shapes.

Figure 1 shows some typical gradient descent. The first three images show iterations on an initial lip position. The next three images show the different energy minima for consecutive lip positions.

2.2. EIGENLIPS

Once we have estimated position and scale of the lip contours, we feed either the first n principal components of the contours, or the first n principal components of a 24x16 gray-level matrix centered and scaled around the lips into the recognizer. The outer boundary of the lips is a fairly robust feature for tracking but probably too coarse grained for classifying visemes (visual phonemes). For that reason

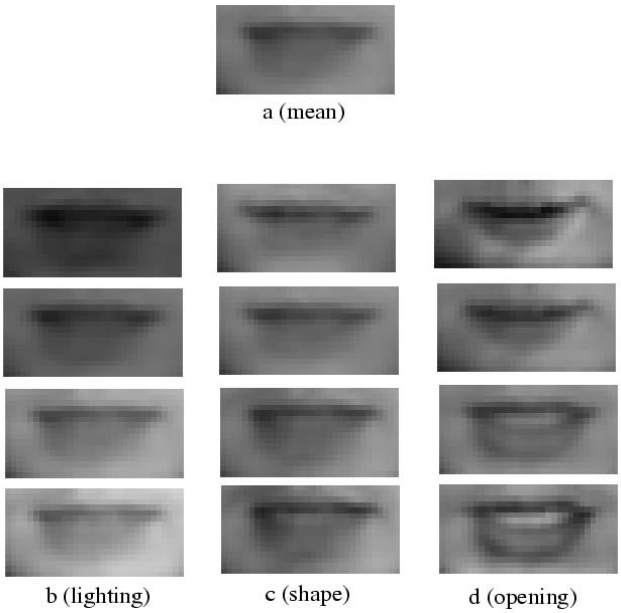

Figure 2: a: Mean vector, b: Variations along 1st principal axis, c: Variations along the 2nd principal axis, d: Variations along the 3rd principal axis

we investigate the gray-level matrix coding as well. We call this approach "Eigenlips" in regards to a similar approach from Turk and Pentland [12] for face recognition, called "Eigenfaces".

Figure 2 shows the mean and some variations along the principal axis (or eigenvectors) of the lips. The first 10 principal components are enough to distinguish the full gray-level lip shapes.

The contour coding is invariant against spatial shifting, rotation, scaling, and lighting. Also the gray-level matrix coding is invariant against shifting, rotation, and scaling, because we dynamically place and scale the matrix around the tracked contour. Unfortunately it is not lighting invariant. (The first principal gray-level axis represents variations in lighting.) However, in our experiments we achieved the best results with the gray-level coding. The missing lighting invariance might be eliminated by the MLP classifier, because only 1 of 10 features (first principal component) was highly dependent on the absolute brightness.

3. ACOUSTIC FRONT END

Our goal is to build a system, which is robust against various signal distortions. With the visual information we try to substitute missing acoustic information of the speech. Another error source which breaks most state-of-the-art systems, is caused by non-speech factors like distortion of the channel itself. To be invariant against this, we use the RASTA-PLP method [7]. RASTA-PLP replaces the conventional short-term absolute spectrum by a band-pass filtered spectral estimate. It is invariant to steady-state spectral distortions and has been shown to improve the recog-

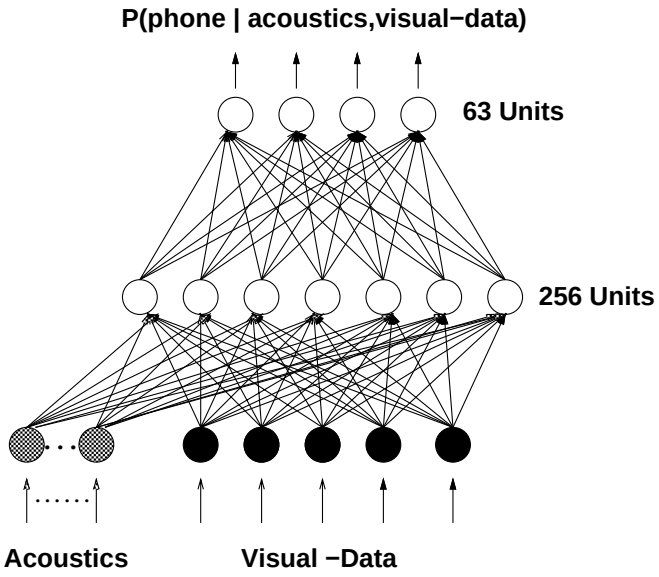

Figure 3: Connectionist architecture.

nition performance by an order of magnitude with realistic channel distortions.

4. ARCHITECTURE AND BIMODAL SENSOR FUSION

When combining the visual information into our hybrid connectionist MLP/HMM speech recognition system [2] we have to decide on the level of integration of the two modalities. In our former system [3] we trained a separate TDNN for each modality and combined the viseme and phoneme hypotheses. This was necessary, because the visual features were with a higher dimension than the acoustic features, but it also had the flexibility of dynamically weighting the modalities differently, according to the cross-modal entropy. With our new visual preprocessing technique the dimension of the visual features is much lower, which allows us to combine the two modalities already at the input level of our connectionist system.

On the one hand the input level approach does not require strong assumptions, e.g., independence, about the nature of interaction between the two modalities. On the other hand when the correct assumptions are taken it simplifies the architecture and leads to an improved performance. In this study we decided to combine the two modalities by concatenating the visual and acoustic features to one bimodal vector and "feed" our recognizer with the extended vector every time frame (10 msec). The visual features, which are generated with a rate of 33 msec by the visual front-end have to be interpolated to our 10 msec rate. Currently we are using linear interpolation, but in the future we will perform nonlinear interpolation with the help of our learned "lip-surface". Given the bimodal feature vectors we train the MLP to estimate the following posterior probability $P(phone|acoustic-data, visual-data)$. Then we divide the posterior probabilities by the priors of the phone classes to get the likelihoods $P(acoustic-data, visual-data|phone)$, according to Bayes law. These Likelihoods are used as the emission probabilities of the Hidden Markov Models.

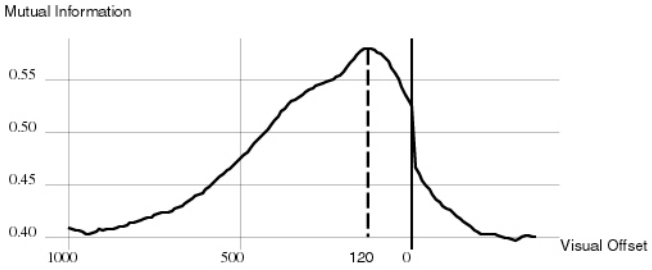

Figure 4: Cross-modal mutual information measurements. The X-axis shows the the visual to acoustic offset and the Y-axis shows the cross-modal mutual information

All the nets used in this experiment are fully connected MLP's with 256 hidden units, 63 output units (the size of our phoneme set), and we use temporal window of 19 (9 past frames, and 9 future frames) as shown in figure 3. The large window is necessary, because some lip movements start much earlier than the corresponding acoustic cases. To confirm this, we looked at cross-modal mutual information measurements. Figure 4 shows the mutual information between the acoustic feature vectors and the visual feature vectors with various temporal offsets. The X-axis describes the cross-modal offset in msec and the Y-axis the mutual information. As we see, at an offset with -120 msec we get maximum mutual information, that means on average the acoustic features are maximal correlated with visual features of 120 msec in the past. In part this offset is caused by different channel delays, but this "forward-articulation" is also confirmed by psychological experiments [1]. As a result we experimented with changing the temporal window from a symmetric window to an asymmetric window, i.e., the 19 frames are combined from 15 frames to the past and 3 future frames. However our recognition results were inferior to the results obtained with the symmetric window reported below.

5. EXPERIMENTS

Our experiments were based on a German multi-speaker spelling task database. The training set (2 female, 4 male speakers) consists of 2955 connected letters. For cross-validation we have used additional 364 letters. Our independent test set was combined from 346 spelled letters across all speakers. Each utterance was a sequence of 3-8 spelled letters. We trained 3 versions of the networks: one pure acoustic network based on the 8 RASTA-PLP cepstral features and the acoustic energy, and two bimodal networks: The "Eigenlip"-net, based on the acoustic features and an additional 10 eigenlip coordinates, and the "Delta-Eigenlip"-net, which has the 10 eigenlip coordinates and an additional 10 "Delta-features". The real information in lipreading lies in the temporal change of lip positions, rather than the absolute lip shape. All nets were trained on 8KHz sampled clean speech.

Task	Acoustic	Eigenlips	Delta-Lips
clean	11.0 %	10.1 %	11.3 %
20db SNR	33.5 %	28.9 %	26.0 %
10db SNR	56.1 %	51.7 %	48.0 %
15db SNR crosstalk	67.3 %	51.7 %	46.0 %

Table 1: Results in word error (wrong words plus insertion and deletion errors)

We generated several test sets covering the 346 letters: one set with clean speech, two with 10db and 20db SNR additive noise (recorded inside a moving car), and one set with 15db SNR crosstalking.

Table 1 summarizes all our simulation results. In clean speech we could not get any significant improvements. In noise degraded speech the improvement was significant at the 0.05 level, as well as in the crosstalk experiment, which showed the largest improvement.

6. SUMMARY

We demonstrated a new visual processing technique for tracking and recognizing human lips. Based on the combination of this new visual front-end and the so called RASTA-PLP acoustic front-end we performed bimodal speech perception using a hybrid connectionist architecture for continuous word recognition (MLP/HMM). We have shown significant recognition performance improvements in noisy environments in considering both speech modalities instead of just the single acoustic modality.

7. ACKNOWLEDGEMENTS

We would like to thank Jerry Feldman, Hermann Hild, Joachim Koehler, Philip Kohn, Nikki Mirghafori, Nelson Morgan, Steve Omohundro, Gary Tajchman, and Alex Waibel for their support and helpful discussions, and Uwe Maier and Peter Sheytt for helping us with recording the database. This research was funded by the Advanced Research Project Agency, under contract #N0000 1493 C0249. The database was collected with funds from Land Baden Wuerttemberg (Landesschwerpunkt Neuroinformatik) in Alex Waibel's research group.

8. REFERENCES

[1] C. Benoit, *The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking Faces* in "HiradaTechnika" (Journal of the Hungarian Telecommunication Association), in 1992.

[2] H.A. Bourlard and N. Morgan.*Connectionist Speech Recognition, A Hybrid Approach.* Kluwer Academic Publishers, 1993.

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

[4] C. Bregler and S. Omohundro, *Surface Learning with Applications to Lip-Reading*, in Cowan, J.D., Tesauro, G., and Alspector, J. (eds.), Advances in Neural Information Processing Systems 6. San Francisco, CA: Morgan Kaufmann Publishers, 1994.

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

[6] A.J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* Ph.D. Dissertation, School of Engineering and Applied Science of the George Washington University, Sep 10, 1993.

[7] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, *RASTA-PLP speech Analysis Technique*, in Proc. Int. Conference on Acoustics, Speech, and Signal Processing, San Francisco 1992.

[8] M. Kass, A. Witkin, and D. Terzopoulos, *SNAKES: Active Contour Models*, in Proc. of the First Int. Conf. on Computer Vision, London 1987.

[9] 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.

[10] D.W. Massaro and M.M. Cohen, *Evaluation and Integration of Visual and Auditory information in Speech Perception.* Journal of Experimental Psychology: Human Perception and Performance, 9, 1983.

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

[12] M. Turk and A. Pentland *Eigenfaces for Recognition* Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.

[13] G.J. Wolff, K.V. Prasad, D.G. Stork, and M.Hennecke *Lipreading by Neural Networks: Visual Preprocessing, Learning and Sensory Integration.* in Cowan, J.D., Tesauro, G., and Alspector, J. (eds.), Advances in Neural Information Processing Systems 6. San Francisco, CA: Morgan Kaufmann Publishers, 1994.

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

[15] A. Yuille, *Deformable Templates for Face Recognition*, Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.